IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 June 2024, accepted 5 August 2024, date of publication 14 August 2024, date of current version 29 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3443520

SURVEY

Managing Distributed Machine Learning Lifecycle for Healthcare Data in the Cloud

ENGIN ZEYDAN[1], **(Senior Member, IEEE), SUAYB S. ARSLAN**[2,3], **(Senior Member, IEEE), AND MADHUSANKA LIYANAGE**[4], **(Senior Member, IEEE)**

[1]Centre Tecnològic de Telecomunicacions de Catalunya, Castelldefels, 08860 Barcelona, Spain
[2]Department of Computer Engineering, Boğaziçi University, 34342 İstanbul, Türkiye
[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]School of Computer Science, University College Dublin, Dublin, D04 V1W8 Ireland

Corresponding author: Madhusanka Liyanage (madhusanka@ucd.ie)

This work was supported in part by the COST Action CA22104 Beingwise, Cybersecurity Research Program Flanders–Second Cycle (VOEWICS02) Project, Science Foundation Ireland under CONNECT Phase 2 Project under Grant 13/RC/2077_P2; in part by the Spanish Ministry of Economy and Competitiveness (MINECO)–Program UNICO I+D under Grant TSI-063000-2021-54 and Grant TSI-063000-2021-55; in part by the ERDF a way of making Europe Project funded by MCIN/AEI/10.13039/501100011033 under Grant PID2021-126431OB-I00; and in part by Generalitat de Catalunya, in 2021, under Grant SGR 00770.

ABSTRACT The main objective of this paper is to highlight the research directions and explain the main roles of current Artificial Intelligence (AI)/Machine Learning (ML) frameworks and available cloud infrastructures in building end-to-end ML lifecycle management for healthcare systems and sensitive biomedical data. We identify and explore the versatility of many genuine techniques from distributed computing and current state-of-the-art ML research, such as building *cognition-inspired learning pipelines* and *federated learning (FL)* ecosystem. Additionally, we outline the advantages and highlight the main obstacles of our methodology utilizing contemporary distributed secure ML techniques, such as FL, and tools designed for managing data throughout its lifecycle. For a robust system design, we present key architectural decisions essential for optimal healthcare data management, focusing on security, privacy and interoperability. Finally, we discuss ongoing efforts and future research directions to overcome existing challenges and improve the effectiveness of AI/ML applications in the healthcare domain.

INDEX TERMS Machine learning, healthcare, biomedicine, data management, federated learning, cloud, coded computation, distributed systems.

I. INTRODUCTION

Traditional healthcare, as a mission-critical domain, is in flux for various reasons: Cost and capacity concerns, ageing populations, personalized treatments, and technological advancements that enable speedy pattern analysis. Data analytics and Artificial Intelligence (AI)/Machine Learning (ML) techniques are paving the way for many use cases in medicine that are quite critical today. At the same time, healthcare and biomedical applications are constantly generating increasingly large and diverse datasets. Information can include a variety of data types, spanning from genetic information to Electronic Health Records (EHRs), which can cover details like patient medical records, important physiological measurements, results from laboratory tests, inherited disorders, behavioural tendencies, and more. This information can also be gathered from wearable gadgets utilizing headgear (AR and VR technologies) or networks of body sensors, largely constituting time-series biomedical data. The range of modern biomedical applications is already huge and growing each day. AI in healthcare is shown to help transform medical techniques into better disease treatment and prevention, therapy recommendations, and health management that considers various factors such as a patient's genetic, environmental, and usual lifecycle behaviors. In the present era, AI/ML methods are employed for diverse purposes: Foreseeing illnesses in high-risk patients requiring additional support [1], aiding individuals

The associate editor coordinating the review of this manuscript and approving it for publication was Chin-Feng Lai.

with motor impairments [2], uncovering potential drug combinations for novel medications [3], and identifying cancer through X-Ray analysis [4].

Regarding intelligent decision-making in healthcare, various stakeholders possess distinct objectives. For instance, pharmaceutical firms seek enhanced market comprehension and product distinction, while healthcare providers aim to optimize treatment safety and effectiveness, reducing costs. Additionally, patients necessitate disease-specific medications/therapies with minimal side effects, improving their quality of life affordably. Simultaneously, public or social institutions, often funding healthcare expenses of citizens, strive for humane clinical outcomes while ensuring the long-term safety and efficiency of the healthcare system. Recent advances in data engineering and data science have enabled the easy integration of a variety of datasets from each patient, rapid processing, and powerful analytical results. By bringing together diverse health datasets per patient and leveraging advancements in data engineering/science, it becomes feasible to advance, administer, and bolster individualized therapies and treatments, often termed as precision medicine [5]. This approach facilitates accurate diagnosis, efficient patient monitoring, and informed decision-making regarding the health status of patient. On the other hand, data analysis has recently been revolutionized by decentralizing computational tasks across multiple nodes or devices, enhancing efficiency and scalability for handling large and complex healthcare datasets. The latter is usually referred to as distributed ML. With the increasing adoption of distributed ML techniques, several economic impacts emerge.

Distributed ML allows healthcare providers to leverage large and diverse datasets more efficiently by merely distributing the overall workload, reducing the need for expensive localized hardware infrastructure. This cost-saving benefit enables healthcare organizations to allocate resources more effectively, potentially reducing operational costs and improving financial sustainability. Additionally, pharmaceutical companies can utilize distributed ML to gain deeper insights into market trends, patient demographics, and treatment efficacy, analyzing large volumes of healthcare data from various sources such as EHRs, genetic information, and wearable device data. This enhanced market understanding can lead to more effective and profitable medications. Moreover, distributed ML enables healthcare providers to deliver more personalized treatments (precision medicine) by analyzing patient-specific data, including genetic information, medical history, and lifestyle factors, tailoring treatments to individual patients. This personalized approach to healthcare can result in better patient outcomes, reduced hospital readmission, and lower healthcare costs dramatically. Furthermore, by streamlining data processing and analysis through distributed means, healthcare organizations can optimize resource allocation and workflow efficiency, improving patient scheduling and inventory management. This leads to reduced wait times, enhanced patient satisfaction, and overall operational financial gains, with the ability to proactively allocate resources to areas with the greatest need, such as disease prevention and population health management.

In this study, we consider the end-to-end management of the health/biomedical data lifecycle in the cloud from both data science and data engineering perspectives in an Internet of Things (IoT) context. The main goal of this work is to outline the research directions and explain the role of current AI/ML frameworks in building such end-to-end lifecycle management systems for healthcare, provide an overview of the current state-of-the-art ML methods and tools for lifecycle management along with their benefits, challenges, and future directions. We also explain how Federated Learning (FL) approaches can be used for biomedicine, their benefits, associated challenges and recent advances. At the end of the paper, we highlight the potential issues and future directions toward the convergence of biomedical data engineering approaches in the healthcare sector.

Fig. 1 summarizes the outline of the our paper. The rest of the paper is organized as follows: Section II provides the related work. In Section III, we provide background information and relevant use cases in healthcare and explain the main motivation for this paper. In Section IV, we provide a detailed insight into the structures of an end-to-end machine learning pipeline using known data engineering frameworks in healthcare. Section V discusses the FL approach, including healthcare datasets, its mechanisms, challenges, and recent advances. In Section VI, we present some potential issues and further discussions in healthcare and the integration of data engineering frameworks. Finally, in Section VIII, we draw conclusions from the paper.

II. RELATED WORK AND MAIN CONTRIBUTIONS

Many recent technological advances such as FL, blockchain, explainable-AI [6], IoT and data engineering can be used to find solutions to some of the healthcare domain problems. These advancements can be used individually or in combination to address a specific set of problems [7], [8]. For instance, one specific recent health-related issue was the COVID-19 pandemic, which led to a burst of studies to cure and predict its effects. An overview of different research, platforms, services, and products where IoT technology is used to combat the COVID-19 pandemic is studied in [9]. The heavy use of AI technologies in the healthcare sector has become quite common both for research and industry to improve the performance and efficiency of handling mechanisms of biomedical data. A review of cutting-edge network architectures, applications, and industrial trends using a deep learning approach applied to healthcare systems can be found in [10]. The authors in [11] use an active learning approach for heart disease prediction. The study in [12] uses an autoencoder-based semi-supervised deep learning approach to identify rare thyroid nodules. The authors in [13] use AI to improve the interpretation of the

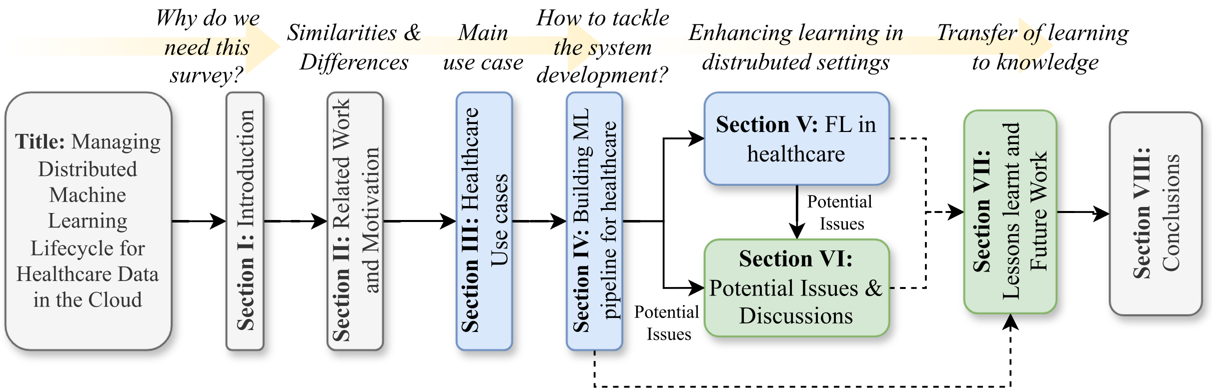
Why do we need this survey?
Similarities & Differences
Main use case
How to tackle the system development?
Enhancing learning in distrubuted settings
Transfer of learning to knowledge
Title: Managing Distributed Machine Learning Lifecycle for Healthcare Data in the Cloud
Section I: Introduction
Section II: Related Work and Motivation
Section III: Healthcare Use cases
Section IV: Building ML pipeline for healthcare
Section V: FL in healthcare
Potential Issues
Section VI: Potential Issues & Discussions
Potential Issues
Section VII: Lessons learnt and Future Work
Section VIII: Conclusions

FIGURE 1. The outline of the paper.

role of genetic risk elements in Neurodevelopmental and Neurodegenerative Disorders (NNDs) pathogenesis. Long-Short Term Memory (LSTM) based Deep Learning (DL) models are used for heartbeat classification in [14]. ML-based approaches are used to identify chronic urticaria subtypes in [15]. Yet another technique called Bayesian Deep Learning (BDL) is used to identify the emotions from the patterns of the heartbeats [16]. A comprehensive survey about this novel technique specifically applied to healthcare is surveyed in [17]. Additionally, using ML, the authors in [18] classify schizophrenia and healthy control cohorts. Patient time series data is predicted in [19]. Another prediction solution for Hepatitis B surface antigen levels in inactive carrier patients is performed in [20] using Deep Neural Networks (DNNs) and acute coronary syndrome and death from non-steroidal anti-inflammatory drugs is predicted using ML models in [21].

The use of deep networks has been extended to electro-physiological brain signals. For example, the authors of [22] have used Deformable Convolutional Networks (DCN) to detect ocular artefacts from the input Electroencephalography (EEG) signal. Recently, the Compact Multi-Branch One-dimensional Convolutional Neural Network (CMO-CNN) has been used to decode Motor Imagery EEG signals by utilising the original signals and enhanced ERD/ERS pattern recognition as an example for streamlined preprocessing-free decoding techniques [23]. The study in [24] proposes an improved graph convolution model with dynamic channel selection to address the challenge of variable EEG channel data, which is critical for accurate emotion classification and important applications in healthcare, including autism research. Equally important is how health data is stored and managed throughout its lifetime. In [25], for example, the researchers investigated a machine learning model that suggests allocating health data blocks to specific storage media and determining their retention period. The storage and transmission of such biomedical data have also led to associated privacy apprehensions. To this end, FL-based studies have recently appeared in healthcare with promising results. A privacy infrastructure based on FL and blockchain technology is proposed to disseminate COVID-19 information in [26]. One of the problems of modern FL is the availability of heterogeneous devices and the statistical imbalance of data between the participating nodes. For example, the collected Magnetoencephalography (MEG) signals are dependent on the biophysical properties of the skull, and therefore, the signals yield different features and statistics [27]. Studies such as [28] address the problem of availability with self-adaptive techniques. On the other hand, the distribution of computations to different clients is also considered in [29], where the statistical imbalance is addressed by an adaptive workload distribution strategy. A comprehensive state-of-the-art survey on the use of FL in smart healthcare can be found in [30].

Building data engineering pipelines has been the focus of many previous works in different domains (see, e.g. [31] for application of data engineering concepts to network management and orchestration). The authors in [32] study technological advances in platforms, tools, methodologies and various challenges when applying big data analytics in healthcare. A review of recent advances in big data generated from biological research (cancer, infectious diseases, etc.) and major computational techniques, algorithms, and their outcomes have been given in [33]. From existing survey papers in the literature, the authors in [10] provide a taxonomy and overview of deep learning approaches in healthcare, identifying challenges and open issues, but lack focus on lifecycle management and IoT integration. The paper in [17] reviews Bayesian deep learning applications in healthcare, discussing challenges but with limited emphasis on IoT integration and data lifecycle. The authors in [30] surveys FL in smart healthcare, highlighting its applications and challenges but lacking detailed lifecycle management and data engineering analysis. In [26], a COVID-19 healthcare system using FL and blockchain for privacy is discussed but does not cover end-to-end data lifecycle management. The paper in [33] explores big data advances and trends in biomedical research without addressing comprehensive lifecycle management. The paper in [34] offers a thorough review of big data analytics platforms and tools in healthcare, with a limited focus on FL applications. Authors in [35] survey AI and big data analytics applications in m-health, lacking a detailed focus on lifecycle management and data engineering. The paper in [36] reviews IoT and cloud computing in healthcare without comprehensive data lifecycle analysis. Finally, [37] surveys privacy-preserving data mining and ML in healthcare, limited in focus on FL and lifecycle management. However, all these analysis on big data platforms is limited in the sense that they do not include an end-to-end management of biomedical health data lifecycle, which includes stages such as *data connection, data ingestion, data processing* and *analysis, data storage, data visualization and data management* and *orchestration* stages. In this survey, We present a comprehensive analysis of lifecycle management for healthcare, focusing on data engineering aspects, FL efforts, and future research directions.

Despite the progress made in these studies, several deficiencies and gaps remain in the literature. One notable gap is the lack of an end-to-end management framework for the biomedical health data lifecycle, encompassing data connection, data ingestion, data processing and analysis, data storage, data visualization, and data management and orchestration. Additionally, many studies focus on specific aspects of AI/ML applications in healthcare without addressing the integration of data engineering principles necessary for robust and scalable solutions. There is also a limited exploration of the potential synergies between advanced data engineering techniques and traditional healthcare systems, which is critical for practically deploying AI/ML solutions in real-world settings. Finally, to highlight the existing gaps in the literature, we summarize relevant survey papers and their corresponding descriptions and limitations in Table 1.

This paper aims to provide an overview of data engineering/science frameworks suitable for solving healthcare problems. We explore the necessary link between recent advances in data engineering and traditional healthcare ecosystems in the context of end-to-end machine learning lifecycle management, which, to our knowledge, has not been explicitly addressed in any previous work. This paper particularly emphasizes how these data engineering principles can be integrated and enhanced in IoT-based biomedical systems to ensure efficient and secure data flow from IoT devices through the entire ML pipeline. The main contributions of this paper can be summarized as follows:

- We begin by providing an explicit overview of building a ML pipeline and a carefully designed platform for the healthcare domain.
- The general proposed platform is categorized into three modules: data sources, data engineering layer and target systems. This modular architectural design allows simplicity, easy management and support for building robust AI/ML pipelines for healthcare applications.
- One of the main objectives is to link the capabilities of the data engineering ecosystem with a possible link to future healthcare systems. Unlike previous work on data engineering, this paper also explores the necessary link that needs to be established between recent advances in data engineering and traditional healthcare ecosystems.
- We focus on the mechanics, challenges, and recent advances of FL and how the FL approach can be applied in the healthcare sector. Since its inception, FL has become an imperative component of AI-powered automation technologies.
- Finally, we highlight some potential issues and provide further discussions on achieving end-to-end ML lifecycle management and data engineering pipelines in real-world healthcare scenarios. We conclude by identifying the research directions that may naturally arise in the future.

III. HEALTHCARE USE CASES

The data landscape is constantly evolving over time. As a result, many new technologies, frameworks, and tools are introduced over time. However, healthcare organizations must combine these evolving data landscape technologies in a meaningful way to meet their needs, i.e., to gain intelligent insights from data. The healthcare sector is expected to change rapidly with advances in AI/ML technologies. For example, some of the time-critical decisions can be transferred to the intelligent systems of AI/ML [39]. Thus, depending on the symptoms of the patient symptoms, better conclusions can be drawn about their condition (both current and future) [40]. Some of the main reasons for the growing interest in AI/ML in healthcare are: (i) large datasets, (ii) diversity of digital health data with different characteristics (imaging, lab tests, devices, genomics, sensor data, etc.), (iii) breakthroughs in AI/ML algorithms (semi-supervised, self-supervised and unsupervised learning) and availability of open-source software tools and libraries (Tensorflow, PyTorch, scikit-learn, etc.), and (iv) industry investments.

Deep learning has been used extensively in imaging technology in the last decade. Efficient and accurate results make it a breakthrough technology in medical imaging [41], [42], [43]. Deep learning has enabled the analysis of medical images for oncology and radiology to be better and faster than expert human accuracy with computer-aided signal detection (e.g., detecting regions with suspicious image characteristics). Applications range from chest X-rays, brain scans, Alzheimer's dementia, or cancer (breast, prostate, etc.) in patients. The authors in [43] have given many examples, use cases, and a new proposal for the use of computer vision techniques (especially deep learning based on Convolutional Neural Networks (CNN) and Vision Transformers (ViT) [44]) in cardiology, pathology, dermatology, ophthalmology, haematology, serology, and gastroenterology. Some of the important use cases relevant to data analytics with data from, e.g., patients, caregivers, healthcare organizations and society, community service providers, etc., can be enumerated as follows:

(i) *Diagnosis using medical imaging:* Leverage clinical and other relevant data to enable early diagnosis and

TABLE 1. Summary of Important Surveys on Managing Healthcare Data in the Cloud.

	Analysis of Lifecycle Management for Healthcare	Data engineering analysis in Healthcare	Analysis of Machine Learning for Healthcare	Analysis on architectural design for healthcare	Concrete application of FL in healthcare	Analysis on future research directions	Remarks
[10]	L	L	H	M	L	M	Focuses on deep learning approaches in healthcare but lacks comprehensive lifecycle management and integration with IoT.
[17]	L	L	M	L	L	M	Provides an overview of Bayesian deep learning in healthcare, limited focus on integration with IoT-based systems.
[30]	M	L	M	L	H	M	Discusses FL in smart healthcare, lacks detailed data lifecycle management and engineering principles.
[26]	L	L	M	M	H	L	Focuses on privacy infrastructure using FL and blockchain, lacks end-to-end data lifecycle management.
[33], [38]	M	L	M	L	L	M	Reviews big data analytics in biological research, does not cover end-to-end management of biomedical health data lifecycle.
[34]	H	M	H	H	L	H	Comprehensive survey on big data analytics platforms and tools in healthcare but limited in FL applications.
[35]	M	M	H	M	M	H	Provides a survey on AI and big data analytics in healthcare but lacks detailed focus on lifecycle management and data engineering.
[36]	M	L	H	M	M	M	Surveys IoT and cloud computing for healthcare but lacks comprehensive analysis of the data lifecycle.
[37]	M	M	M	L	M	M	Surveys privacy-preserving data mining in healthcare, limited focus on FL and lifecycle management.
This survey	H	H	H	H	H	H	This survey presents a comprehensive analysis on lifecycle management for the healthcare with analysis on data engineering aspects, spotlights on FL efforts in healthcare, and the future research directions.

H	Explores the field in detail.	M	Provides some information about the field.	L	No information or explores the area only briefly.

treatment of diseases. Analyse and transform images using advanced medical imaging technology to gain medical insights (e.g., to help pathologists distinguish cancer types) and model potential diseases. For instance, the use of AI in mammography has significantly improved the early detection rates of breast cancer, reducing false positives and enhancing diagnostic accuracy [45].

(ii) *Healthcare management:* Determine the optimal price for healthcare services provided, gain competitive intelligence between hospitals, use AI-powered automation technologies to automate routine operations, reports, etc. and reduce CAPEX and OPEX costs. For example, applying predictive analytics in hospital resource management has led to better allocation of staff and equipment, reduced waiting times and improved patient care outcomes [46].

(iii) *Patient care based on patient data:* Use prescriptive analytics on patient data to enable real-time triage and prioritization of cases by automatically identifying and predicting high-risk and adverse health conditions (e.g., frail older adults, patients with chronic conditions). Find the best patient care while reducing costs and increasing efficiency thanks to advances in data analytics and related techniques. Predictive modelling could identify patients at risk of sepsis, enabling timely intervention and significantly lowering mortality rates [47].

(iv) *Remote monitoring:* The benefits of remote patient monitoring systems include increasing patient awareness and accountability, saving overall healthcare costs by reducing hospitalization and admission costs, and providing real-time recommendations based on patient health status. For example, using remote monitoring systems to patients with chronic heart failure showed a marked decrease in emergency room visits and hospital admissions, enhancing patient quality of life and reducing healthcare costs [48].

Finally, Fig. 2 shows various use cases in healthcare (diagnosis with medical imaging, health management, patient care, remote monitoring) enabled by AI/ML technologies and illustrates their benefits and the diversity of digital health data sources.

IV. BUILDING A MACHINE LEARNING PIPELINE FOR HEALTHCARE

With the advent of open-source software, various tools and frameworks have emerged for advanced data engineering. As the ecosystem for data engineering and AI/ML grows and matures, so do the opportunities for building a scalable production-level data science pipeline. Several sectors, including the medical domain and healthcare services, can benefit from this. For example, better design for a data-sharing platform for health can provide access to big data sources that can help develop better solutions (e.g., disease detection and treatment). In addition, mobile applications

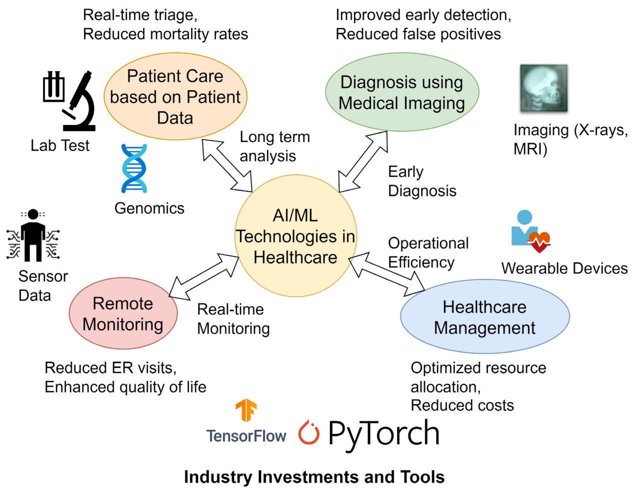

FIGURE 2. **AI/ML applications in healthcare: Use cases and benefits.**

connected to AI/ML platforms can enable patients to access health services in their own personal environment. Assembling large and complex datasets that can fulfil some of the functional and non-functional requirements of healthcare requires data pipelines. Data pipelines help develop and deploy the AI/ML models at scale. The analytics tools and models used in the data pipeline provide actionable insights into key performance metrics (e.g., operational efficiency and patient satisfaction) and a range of healthcare problems, from predictive to prescriptive.

Fig. 3 shows an overview of the three layers of the proposed platform adapted for healthcare. In this figure, the first layer is the data source layer, the second layer in the middle is the data engineering layer, and the top layer is the target systems. The layered architecture helps healthcare administrators to implement the desired functionalities at any layer without affecting the elements of the other layers.

A. DATA SOURCE LAYER

This layer contains the data sources that are the lifeblood of the health system and represent the most valuable asset. Multiple data sources can provide a complete picture of the user's health status, including physical, cognitive, and social conditions. Data sources can include databases, sensor data, EHRs, Comma Separated Value (CSV) files, remote services, distributed file systems such as the Hadoop Distributed File System (HDFS) [49], and so on. Some examples of real-world healthcare data sources include pharmacy data, wearables, surveys, hospital data, lab/biomarker data, electronic medical and health records, social media data, disease registries, claims data, experimental (cell lines, clinical trials, histology, etc.), electrophysiology data (EEG, MEG, ECoG, etc.) biological (genome, epigenome, etc.) and clinical data (family history, medications, disease history, laboratory tests, etc.). With the possible invasive and non-invasive techniques, vital signs such as blood pressure, heart rate, EEG, Electrocardiography (ECG), etc., of patients inside or outside the hospital can be continuously monitored via IoT systems. In addition, IoT devices can also capture other context variables (such as room temperature, pressure, or oxygen levels) to extend sensing data collection through both environmental activities and wearable devices.

Sensors can be either implanted or carried through external means. Implanted sensors are located in the patient's body, while external sensors are either placed on the patient's skin or at a distance of 2-5 cm distance from the patient [50]. Some other devices such as a Brain–Machine Interface (BMI), can help translate neural information into computer commands to control external software or hardware. Non-invasive techniques such as Functional Near-Infrared Spectroscopy (fNIRS) can record brain signals. Medical images can be acquired using electron radiation (ECG, Magnetic Resonance Imaging (MRI), Near-infrared spectroscopy (NIRS)), sound (ultrasound, echocardiogram), or ionizing radiation (X-ray, PET), all of which serve as important data sources for further processing and reliable inference.

Data may also be collected in observational, noncontrolled, and nonexperimental studies. Data sources can be either real-time data or data at rest. However, collecting real-world data on which the decision-making process depends is difficult. Real-time data collection includes time series data for blood pressure, pulse oximetry, respiration, skin temperature, and activities. In addition to proprietary datasets from individual healthcare institutions, some of the other public and private health related data can be collected from portals and competitions (Kaggle,[*] DrivenData,[†] AiCrowd,[‡] Codalab,[§] Bitgrid,[¶] IEEE Dataport,[‖] Medical Segmentation Decathlon,[**] etc.) and open health dataset providers for researchers (CheXpert (a large dataset of chest x-rays) [51], MIDRC (Medical Imaging Data Resource Center, an open radiology database),[††] NIH Clinical Center Dataset [52], Image Processing Portal of CBICA,[‡‡] Genomic Data Commons Data Portal,[§§] and Mendeley Data[¶¶]).

B. DATA ENGINEERING LAYER

This layer (shown in the middle of Fig. 3) consists of several interconnected modules. Note that the interconnection between each module is only roughly shown and multiple interfaces may connect these modules, depending on the use case. Therefore, multiple pipelines may be based on Service Level Agreements (SLAs) or non-functional requirements. One pipeline may be appropriate for real-time notifications,

[*] Online: https://www.kaggle.com/, [accessed Dec-2023]
[†] Online: https://www.drivendata.org/, [accessed Dec-2023]
[‡] Online: https://www.aicrowd.com/, [accessed Dec-2023]
[§] Online: https://competitions.codalab.org/, [accessed Dec-2023]
[¶] Online: https://bitgrit.net/, [accessed Dec-2023]
[‖] Online: https://ieee-dataport.org/, [accessed Dec-2023]
[**] Online: http://medicaldecathlon.com/, [accessed Dec-2023]
[††] Online: https://www.midrc.org/, [accessed Dec-2023]
[‡‡] Online: https://ipp.cbica.upenn.edu/, [accessed Dec-2023]
[§§] Online: https://portal.gdc.cancer.gov/, [accessed Dec-2023]
[¶¶] Online: https://data.mendeley.com/, [accessed Dec-2023]

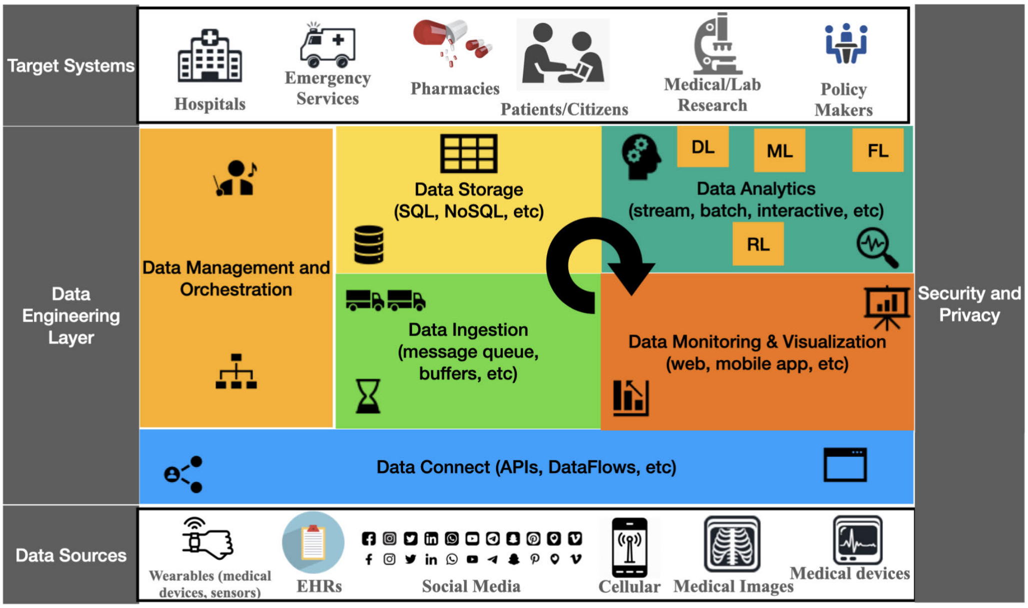
Target Systems
Hospitals
Emergency Services
Pharmacies
Patients/Citizens
Medical/Lab Research
Policy Makers
Data Engineering Layer
Data Management and Orchestration
Data Storage (SQL, NoSQL, etc)
DL
ML
FL
Data Analytics (stream, batch, interactive, etc)
RL
Data Ingestion (message queue, buffers, etc)
Data Monitoring & Visualization (web, mobile app, etc)
Security and Privacy
Data Connect (APIs, DataFlows, etc)
Data Sources
Wearables (medical devices, sensors)
EHRs
Social Media
Cellular
Medical Images
Medical devices

FIGURE 3. A high level illustration of a general platform to support AI/ML in healthcare domain.DL: Deep Learning, ML: Machine Learning, FL: Federated Learning, RL: Reinforcement Learning.

while another may be more appropriate for non-real-time requests. In addition, in some scenarios, such as IoT-based healthcare network services, raw data collected via the Data Connect module (e.g., blood pressure, heart rate) can be integrated with the data visualization component either via mobile applications or web user interfaces for direct visualization in a dashboard through thin-layer interface.

In other scenarios, further data analytics and processing may be required between data connection and data visualization modules, leveraging recent advances in AI/ML algorithms (e.g., risk factor prediction or statistical analysis of overall health). In other scenarios that require data reliability and swift response (e.g., predicting neurological or cardiovascular disorders in health of a patient), the data ingestion and data analytics modules (for real-time ultra-low latency event processing) must be embedded in the data engineering pipeline. The key steps in building a healthcare-focused data engineering pipeline can be summarized as follows.

1) DATA CONNECTION MODULE

The first step is to collect health data needs from multiple sources using existing data connectivity tools to connect to each source separately through a reliable channel. This step is critical because to represent the performance of a model in real-world applications accurately. The data connection process must retrieve the correct data that meets the needs of the use case. During the data collection procedure, patient data, e.g., from wearable sensors (e.g., accelerometers and electrodermal activity (EDA)), can be transmitted either to the cloud or to the on-premise data ingestion layer, mainly in four modes: (i) continuous transmission for all data, (ii) continuous transmission at specific times, (iii) event-driven transmission, and (iv) on-demand transmission.

2) DATA INGESTION MODULE

After the data connection between the data sources and the data engineering layers is established, data from various data sources must be ingested into a buffer or temporary storage systems in a reliable and timely manner. Time-sensitive events are triggered when healthcare users perform certain actions, such as real-time interactions in response to critical events for data sharing between physicians, sending real-time updates on health status of a patient to healthcare professionals, real-time monitoring of heart disease, etc. For this reason, hundreds of terabytes of data must be retrieved daily from heterogeneous data sources and ingested for further low-latency processing. At the data ingestion stage, data can also be optimally processed, cleaned, corrected, or enriched so that downstream pipeline modules (e.g., data storage and data analytics modules) can consume it in the format they understand. Healthcare data can be unstructured (text, images, audio, sensor readings, binary data, etc.) [53] and must be converted to numeric features for downstream AI/ML applications. Thanks to wearable sensors, remote monitoring platforms can provide 24/7 care at home at little or no cost. These sensors measure patients' vital signs and transmit the data continuously in real-time. In real-time applications, three important requirements must be met: *(i)* Each dataset must be processed as fast as possible (low latency). *(ii)* Data must be ingested reliably without losing data in case of failures (resiliency). *(iii)* The growing volume of data should be processed without any problems (scalability).

Even streaming platforms such as Apache Kafka [54], Apache Flume [55] and Amazon Kinesis [56] are some examples of distributed data ingestion platforms that can collect, aggregate, and transfer large amounts of data seamlessly and reliably for real-time applications. These platforms enable applications to publish and consume messages stored as records in a "topic" and react to real-time data streams.

3) DATA ANALYSIS AND PROCESSING MODULE

After data ingestion, this module ensures the creation of insights from data and translates them into intelligent decisions. The data can be aggregated and transmitted to the cloud or a local site for analysis and processing with analytics tools. Local sites can perform data analysis (e.g., heart illness prediction using ML, DL, Reinforcement Learning (RL) or FL prediction module) in a real-time scenario for internal patients. In contrast, the cloud layer can perform the same tasks for remote patients. In this module, specific features relevant to the patient status problem under consideration are identified, correlations are explored and AI/ML techniques (model training and evaluation, cross-validation, etc.) are applied. Finally, the AI/ML model that is evaluated and performs best according to the requirements of the problem domain and uses case (e.g., selection of the best model based on accuracy, precision, recall, etc.) is deployed in the production environment.

The data analyzed in this module can help physicians identify important patient condition changes and make real-time or non-real-time decisions or recommendations about the current or future condition of patient. For example, brain signals recorded with fNIRS, EEG and Magnetoencephalography (MEG) techniques can be used to interact with the environment using deep reinforcement learning algorithms [57] within this module. Fig. 4 shows a general integrated AI/ML system for end-to-end lifecycle management that consists of the following ten major components:

(1) *Data collection*: This component is important to train the model and make inferences. Healthcare-related data can be collected from various internal and external sources, as described in the data source layer in Fig. 3. Due to the sensitivity of the healthcare data, most institutional review boards (IRB) must provide appropriate consent before executing the experiment, which may cause various delays and biases in the data collection process.

(2) *Data clean-up and preparation*: Healthcare data may be messy and contain missing values or erroneous features (data may need to be extracted from electronic health records or aggregated from multiple sources), which can be a challenge for training correct ML models. In addition, due to the variety of error root causes/types, the generalizability of the ML models is pretty limited. This component includes data cleansing, filtering, and transformation and is used to prepare the data for ML tasks.

(3) *Feature engineering*: This module performs feature wrangling and ensures that raw input data is transformed into feature vectors (i.e., integer mappings). The data analyst also applies his or her expertise to feature selection.

(4) *Data analysis and visualization*: This component is used to validate the input data fed into the system. It ensures that the data is error-free. It can compute and display descriptive statistics of the data, infer data schemas, and detect anomalies - all using exploratory data analysis. Note that it can also be used to identify patterns and trends in patient data [58].

(5) *Model training*: This component is responsible for training the model. For example, a model could be trained to predict the likelihood of a patient developing a particular disease based on his or her medical history [59], or to determine/recommend the most effective treatment for a patient based on the characteristics of his or her disease [60]. For imbalanced data, a data augmentation must be accompanied by the training data set for better accuracy. If the data does not fit in memory for training, the model training components should be able to parallelize data and model and process out-of-memory data.

(6) *Model evaluation*: In this module, a thorough analysis of the training results is performed to ensure that the exported models perform adequately in the production environment. In this phase, the best models are tested and selected. This phase is essential in healthcare to ensure the model is reliable and makes accurate predictions.

(7) *Model deployment*: This component ensures that the models are ready for the production system (such as a hospital or clinic) and can accept user queries to the model to make appropriate decisions. Deployment can be done on servers or mobile devices themselves.

(8) *Model serving*: This component is responsible for responding to user requests by minimizing response time and maximizing throughput (e.g., the number of served requests). When changes are made to the model or data is updated, this component can easily update newer versions of the trained models. In this phase, the model is hosted on a server (either in a hospital or in the cloud) and an API is created so that other healthcare applications (e.g. mobile applications) can access the predictions of the model.

(9) *Model monitoring*: This component is responsible for automatically monitoring and logging all steps from model training to production. Suppose the performance of the ML model degrades over time (e.g., predicting patient outcomes or identifying potential diseases). In that case, this component ensures to send notifications, perform rollbacks for deviating values or possibly invoke a new ML process iteration.

(10) *Model maintenance*: This component is used to troubleshoot and decide when and how to update models in production. The feedback collected from healthcare providers is used to retrain the model on new data or to fine-tune the hyper-parameters of the model.

Project management tools are also required for the project setup step. A good and detailed explanation of most

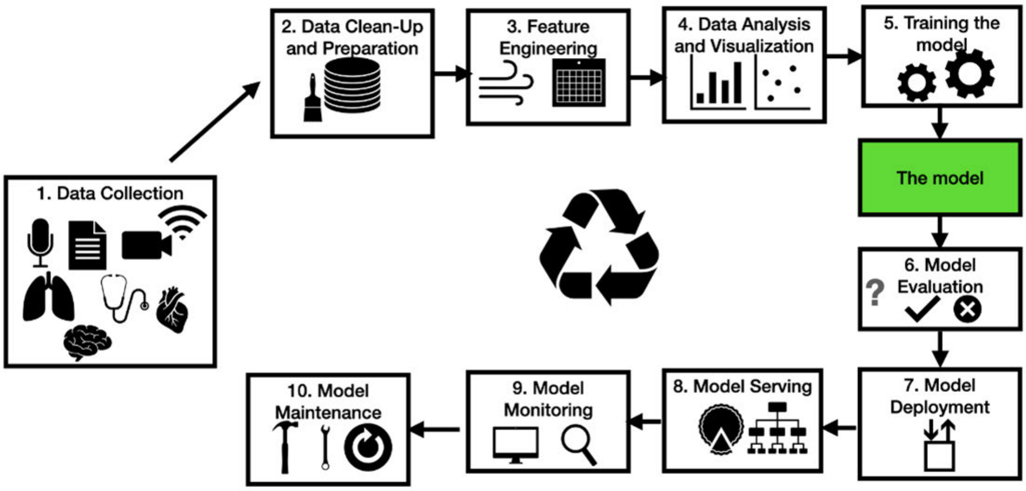

FIGURE 4. Creation and maintenance of AI/ML models over the end-to-end project lifecycle.

components of the ML lifecycle can be found in [61]. Data processing can be done either in real-time or in batch mode:

- *Real-time processing*: Some of the vital signs of patient (SpO2, ECG, blood pressure, blood glucose levels) need to be processed in real-time to enable continuous tracking and risk mitigation. For this reason, critical health applications such as vital sign monitoring and health status assessment must rely on a streaming pipeline of data ingestion frameworks for real-time interactions in response to critical events. Application requirements can vary depending on real-time (in less than 10 ms) or near-real-time (less than a few minutes). For example, a ECG healthcare service requires a delay of about one second at a data rate of 1.0–20.0 kbps and with zero packet loss, while remote teleoperation requires less than 300.0 milliseconds of raindrop time [62]. Various modifications can be made in the data ingestion frameworks to meet these different application requirements (e.g., using RabbitMQ for latency-sensitive or Kafka for throughput-sensitive applications [63]).
- *Batch processing:* Some healthcare applications may require high throughput (processing data on the order of PB/day in a single cluster). For example, ML can be used to analyze, large volumes of patient medical records to predict readmissions [64], mortality [65], [66], intensive care units complications [67], [68], sepsis [69], [70], diabetes [71], etc. among them. In addition, body sensor network data can also be analyzed in batch mode to understand patient behavior and provide tailored services to each patient [72], [73]. Some leading open-source frameworks and tools in this module are Tensorflow [74], PyTorch [75] and Keras [76] frameworks. Some examples of cloud tools are BigQuery with SQL and BigQueryML to analyze data stored in BigQuery [77].

Overall, AI/ML model in a healthcare setting requires careful planning and attention considering the unique challenges of the healthcare domain. Note that many companies want to develop AI models on healthcare data, but the data is often not readily available. Additionally, some companies are looking to sell back data analytics [78] by having hospitals upload healthcare data to the cloud. Still, hospitals may prefer vendors to deploy their tools/models locally to keep the data within the domain of the hospital. Therefore, effectively utilising healthcare data is essential to the success of AI/ML in healthcare. Furthermore, understanding the challenges of, and potential solutions for, infrastructure related to healthcare AI/ML is vital to ensure the efficient and effective utilization of healthcare data. We have discussed some potential challenges and solutions in Section VI.

4) DATA STORAGE MODULE

All data can be stored in the cloud or on the hospitals' local premises. Stored data can be kept in a data lake, data marts, or data warehouse for real-time reporting and analysis or long-term archival purposes. Stored data should be securely accessible from multiple sources, including laboratories, pharmacies, ambulances, clinics, etc. The backend storage systems in the cloud can collect patient data either through direct communication with the data connection module (e.g., HTTP, REST API calls) or through the data ingestion module (e.g., through the streaming integration tools Kafka, Pulsar [79], Spark Streaming [80]). Canonical examples of object storage in cloud environments include Amazon S3 and Google Cloud Storage (GCS). The GCS* platform can provide four types of storage options:

(i) *Standard storage:* This option allows frequent access to data, i.e., it is well suited for hot data such as hospital websites, mobile applications, streaming videos, etc.

*Online: https://cloud.google.com/storage, [accessed June-2022] *(ii)* *Nearline storage:* This option provides at least 30 days of storage for backup and long-tail multimedia content.

(iii) *Coldline storage:* For disaster recovery scenarios, this option enables storage for at least 90 days.

(iv) *Archive storage:* For data to be stored for a period of at least one year. Existing optimized data formats, such as AVRO, Parquet, Pbuffer, etc., can be exploited to store health data for further analysis. Cloud tools such as BigQuery (a serverless data warehouse) can be used to store data. On the other hand, tools like Elasticsearch can be used as a search engine to make medical data (images, text, etc.) searchable via metadata or tags.

5) DATA VISUALIZATION AND MONITORING MODULE

This module allows users to visually abstract or aggregate data to understand the content of the data and monitor progress. AI/ML-based data monitoring is important in production systems to decide which model from one or more available base models to adopt in the final application. This can be accomplished by tracking and visualizing the performance of production data. It may happen that the performance of a model suddenly drops. Then it is important to find the root cause of the failure in a data engineering pipeline. Some options for monitoring are Prometheus [81] and Grafana* as a data engineering monitoring stack or using popular techniques like SHapley Additive exPlanations (SHAP) for model interpretability and analysis, etc. Tools such as Kibana as a visualization tool can enable interactive exploration of e.g. Digital Imaging and Communications in Medicine (DICOM) data already indexed by Elasticsearch.

6) DATA MANAGEMENT AND ORCHESTRATION

Data management frameworks ensure that all phases of data lifecycle management (connection, ingestion, retention, compliance, and export) are covered for industries with large datasets. Features such as auto-scaling, fault tolerance, and managing data model evolution make data management frameworks the first choice across industries. For medical field analysis, tools such as AirFlow can be used for workflow management systems to enable complex and flexible data processing workflows. Using data orchestrators, models can be updated regularly, such as every week. Schedulers can also easily be part of Continuous Integration (CI)/Continuous Development (CD) pipeline via Application Programming Interfaces (APIs). Together with the monitoring capabilities of the monitoring components, the orchestrator/scheduler can evaluate the performance of the models. If a model is deemed good enough, it can be made available via APIs to client software (e.g., ML-driven devices used by hospital staff on the front end).

In the application scenario for predicting a disease using medical images and advanced AI/ML techniques, for example, different models can be compared with the actual diagnosed disease after important medical tests have been performed. When a new and better version of the model is available, requests for predictions can be forwarded to the API of the new model via the front end. This process can be performed in a controlled environment for each new incoming user request while monitoring performance (on the dashboard of the system owner) and verifying whether the new model behaves as expected. Cloud orchestration tools such as Kubernetes (along with Docker instances) can manage multiple health applications at scale as micro-applications.

On the other hand, cloud systems offer several benefits in healthcare, such as reducing costs and improving efficiency in healthcare system development. Since every healthcare application generates data whose volume is constantly growing, data engineering pipelines must be scalable to adapt to dynamic changes in data growth. For this reason, establishing pipelines in the cloud makes sense because the cloud offers scalability and flexibility on demand. The cloud portion acts as a bridge between patients and the hospital in a cloud-enabled system. Cloud computing can help unify EHRs between hospitals and improve the interoperability, maintainability, and scalability of the healthcare system. A cloud-based system can easily handle large amounts of patient data from various sources, such as remote monitoring systems, databases, file servers, etc.

C. TARGET SYSTEMS, SECURITY AND PRIVACY

The target users and institutions are at the top level of Fig. 3. At this level, the end user may be a patient in a hospital or an elderly person being remotely monitored in a smart home facility. Other healthcare providers (a mobile ambulance, a drugstore, a research facility) can also benefit from the platform as end users. In building a ML pipeline for healthcare, sensitive personal and medical information must be handled appropriately. For this reason, prioritizing the security and privacy of such data is paramount in gaining the trust of patients and other entities involved during the build process. Some of the key considerations that must be tackled are:

1) **Data protection:** Employing encryption, access control and backups can enhance data protection from unauthorized access, deletion or modification. These measures ensure that data integrity and confidentiality are maintained throughout the data lifecycle.
2) **Data privacy:** Compliance with general personal and medical data privacy regulations such as the General Data Protection Regulation (GDPR) in the European Union (EU) or the Health Insurance Portability and Accountability Act (HIPAA) in the United States (US) is essential. These regulations ensure that the healthcare data is protected and handled in a way that preserves patient privacy and meets legal requirements.
3) **Data governance:** While building ML pipeline, data should be managed according to clear procedures and policies to comply with relevant rules and regulations [82]. This includes establishing data ownership,

[*] Online: https://grafana.com/, [accessed June-2022]

accountability, and stewardship practices to ensure data is used responsibly and ethically.

4) **Data security:** Stored or processed data can be secured via secure communication protocols, firewalls, or frequent security updates. Implementing multi-factor authentication and regular security audits contributes to a robust security posture.
5) **Data minimization:** Data protection can be further enhanced by collecting and processing only a specific data set needed for a specific healthcare task. For instance, when developing an ML model for image analysis (e.g., image dataset ImageNet), only the relevant image data should be collected and used, minimizing the exposure of unnecessary personal information.
6) **Data de-identification:** Techniques such as differential privacy [83], Multi-party Computation (MPC), Privacy-Preserving ML (PPML) or blockchain and cryptography encryption [84] can help anonymize personal identifiers in personal and medical data. These techniques help ensure that individuals cannot be easily identified from the data, thus protecting their privacy.

V. FEDERATED LEARNING IN HEALTHCARE

There are several distributed learning strategies to increase the security and durability of sensitive data, such as gossip learning [85]. However, the most prominent and successful approach has been FL [86] due to its privacy-preserving and half-centralized (aggregation) nature. In addition, relevant alternatives seem to share much in common regarding system components, architecture and performance. Therefore, this subsection will heavily focus on FL. We would like to begin by observing that healthcare has experienced an unprecedented boom in the last decade, thanks to analytics and big data generated by the widespread use of IoT devices to collect sensitive health data. This data typically comes from clinical facilities, patients wearing small gadgets (such as headsets), the pharmaceutical industry, and various insurance companies. The high quality of healthcare can only be maintained if a high computational burden due to the processing of large data sets is viable with data security and privacy guarantees. Healthcare data is usually disease- and patient-oriented; finding a common pattern in such a large dataset can be extremely difficult. Additionally, one of the challenges for the modern ML ecosystem for healthcare is the unavailability of large amounts of training data to provide reliable insights into inherently fragmented data. Similarly, adopting data science and engineering principles and potential data-sharing processes between organizations is slow due to the sensitivity of patient data and format variability.

Data privacy issues play an important role in developing healthcare applications and services. Most patients using systems such as remote patient monitoring are concerned about the security and privacy of their data. The access, analysis, modelling, and use of such data are extremely protected by various regulatory acts such as the Health Insurance Portability and Accountability Act [87]. Therefore, patient data must be processed confidentially and securely while complying with required governance and privacy regulations. Data privacy/security technologies combined with data engineering frameworks can help address these issues. Techniques for privacy-preserving AI/ML such as FL, Differential Privacy (DP) and Homomorphic Encryption (HE) are increasingly becoming promising approaches to data privacy preservation. For example, recent frameworks such as *CrypTFlow*[*] can securely run trained models over Secure MPC protocols without sharing sensitive data of patients with privacy-preserving ML in healthcare.

The concept of FL was originally proposed by Google researchers back in 2017 [88] and has since been of interest to the broader community, particularly in the area of healthcare [89]. Thanks to its privacy-centric approach, it is evident that approaches such as FL are critical to healthcare data processing. Due to policies and regulations governing the sharing of sensitive patient data, AI/ML is perfect for developing life-saving tools without moving sensitive health data from its original location and exposing it to privacy breaches. FL is about training privacy-preserving machine learning models where computation is moved to where the data is originated, stored and preserved. The main goal is collectively training a global model by exchanging model parameters instead of sensitive data between distributed computing nodes. There is also a close relationship between FL and consensus algorithms, as local models achieve consensus using their local data by exchanging their model parameters. Once the convergence is established, we hopefully have a globally trained model that works best with the entire dataset (good generalization properties). Finally, we must distinguish between a collaborating single organization and multiple organizations to train a model, as the latter involves considering reliable entities. In other words, we must ensure that the participating entities do not tamper with the model parameters during the data exchange. Accordingly, we can refer to FL as "cross-device" and "cross-silo" depending on the scope of the learning. A good example of cross-device FL is Apple's approach [90] in iOS 13. As for the healthcare sector, doc.ia [91] develops FL solutions for medical research. "Cross-silo" has also attracted attention and is used in healthcare sector [92].

A. FEDERATED LEARNING FROM DATASET-PERSPECTIVE IN HEALTHCARE SECTORS

Datasets collected or accumulated over time in healthcare may not have the same feature space for all medical devices. For example, a standardized Magnetic Resonance Imaging (MRI) device or Computed Tomography (CT) scan generates patient information in the same data format, whereas blood-test devices from different manufacturers

[*] Available online: https://github.com/mpc-msri/EzPC, [accessed June-2022]

output different formats. Similarly, DNA sequencers used in drug development and pharmaceuticals also have a few common data formats, such as FASTQ [93]. Standardization can be expected only after their commercialization in the market. In the latter case, datasets with different feature spaces may have to be used to train the global model.

Indeed, the homogeneity assumption across devices (regarding hardware and statistical data distribution) would oversimplify the large and diverse nature of the healthcare sector and patient portfolio. Datasets with the same feature space for all devices have been referred to in the literature as "Horizontal FL". In contrast, datasets with different feature spaces are referred to as "Vertical FL". Vertical FL is more appropriate for domain-specific training. However, the common practice is to use it with a pre-trained model that uses a similar feature dataset and then specialize its parameters for a particular topic. This technique is known as *federated transfer learning*, which is based on vertical FL and is the foundation for personalized patient-centred recommendation systems in healthcare.

B. FEDERATED LEARNING FUNDAMENTALS

Suppose there are K (patient) clients, where a client could be a bioinformatics tool, a cell phone, a sensory wearable device, or a clinical warehouse. Each client can provide n_k data samples. Given a total number of $n = \sum_{k=1}^{K} n_k$ data samples, we minimize empirical risk as follows

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} F(\boldsymbol{w}) = \min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{n_k}{n} F_k(\boldsymbol{w}), \tag{1}$$

where

$$F_k(\boldsymbol{w}) = \frac{1}{n_k} \sum_{x_i \sim \mathcal{D}_k} f_i(\boldsymbol{w}, \boldsymbol{x}_i), \tag{2}$$

and $\boldsymbol{w} \in \mathbb{R}^d$ is the model parameter ($\boldsymbol{w}^*$ being the optimal that minimizes Eqn. (1)), $\mathcal{D}_k$ is the data distribution for the k-th client and $\boldsymbol{x}_i \sim \mathcal{D}_k$ means that $\boldsymbol{x}_i$ is distributed according to $\mathcal{D}_k$. Here $f_i(\boldsymbol{w}, \boldsymbol{x}_i)$ is a loss function computed based on the input-label pairs $(\boldsymbol{x}_i, y_i)$. For example, in the case of logistic regression (binary classification), we have $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$ and $f_i(\boldsymbol{w}, \boldsymbol{x}_i) = -\log\left(1 + e^{-y_i \boldsymbol{x}_i^T \boldsymbol{w}}\right)$ in Eqn. (2). Accordingly, each local institution performs its own training to find the model parameters specific to its training data as a result of the minimization given in Eqn. (1). Then, they submit these local parameters ($\boldsymbol{w}^*$) as updates to a central server and contribute to the global model learning process. The central server collects all the local updates. It combines them (based on a combining algorithm such as averaging) to generate the updated global model parameters before sharing them with the local institutions.

C. COMMUNICATION COST OF FL

During the execution of FL, model sharing takes place both from client-to-server (uplink) and from server-to-client (downlink) in both directions. To minimize the communication between the client and the server, lossless model compression is usually used [94]. Ideally, both the model and its updates can be compressed to its entropy $H(\boldsymbol{\Delta w})$ where $\boldsymbol{\Delta w}$ represents the updates to the model parameters. Thus, the total number of bits transmitted in either the uplink or downlink can be expressed as $O(U \times |\boldsymbol{w}| \times H(\boldsymbol{\Delta w}))$ where $|\boldsymbol{w}|$ is the size of the model and U is the number of iterations performed during the update phases of the FL process [95]. As can be seen, there are three main ways to reduce the traffic through FL: (1) enabling a smaller number of clients K, (2) smaller update size $|\boldsymbol{w}|$ and (3) smaller number of update iterations U.

One of the best-known methods for reducing traffic is model compression. Models can either be compressed losslessly using standard compression techniques (data compression) or created by observing structures whose parameters are learned from only a few variables [86]. On the other hand, there are also lossy model compression techniques that come at the cost of reduced learning accuracy or an increased number of iterations U. These techniques include pruning the least useful connections of a deep network, weight quantization [96], and model distillation [97]. Lossy compression can also be applied to full model updating, which goes through standard encoding stages such as transformations, subsampling, quantization, and rotations. On the server side, decoding is performed before aggregating the updates. Another idea is the so-called "federated drop-out", where clients generate updates for a subset of the global model (a sub-model) rather than for the global model itself [98]. These updates, which affect a reduced set of parameters, are typically lower in size and can, therefore, significantly reduce the computational requirements of individual clients. The computed updates for the subsets of the global model can be interpreted and modified to aid in updating the global model. The same study shows that the drop-out idea reduces the uplink and downlink communication costs incurred by processing/generating an update for a subset model.

Client selection is another simpler method to reduce communication within FL. Depending on the client data and the trained model, some model updates may be more informative for the global model than the rest. For example, in the case of stochastic gradient descent, the parameters can be selected based on the distance of the current value to the local optima, i.e., those with larger gradients [99]. Let us consider the practical healthcare sector and the heterogeneous devices used in our healthcare system. It is advisable to have a reliable server that manages the resources of the other nodes based on their channel quality, the ability of each client to access them and the type of data stored in that client. Usually, there are trade-offs between these resources, and if one aims to minimize bandwidth cost, the server must make appropriate decisions. Note that the average energy consumption or training latency may not be optimal and may need to be optimized in such grounds [100].

In addition to client selection, updates can also be reduced through delayed messages and averaging models. This way, communication costs can be reduced by an order of magnitude [101]. In some special cases of deep networks, it has been shown that passing local data multiple times without sending parameter updates to a central server improves the communication cost of FL at the expense of a small change in learning time [102]. Recently, two techniques have been used to reduce the total number of iterations to one: (1) one-shot FL and (2) data distillation [103]. In the latter work, each client applies distillation (a purification of data) to its private data and sends the synthetic data (e.g., images or sentences instead of parameter updates) to a central server. The distilled data usually looks like noise and is treated as unusable after model fitting.

D. RECENT ADVANCES IN FL

In this section, we would like to present two important advances related to FL that can be directly applied to the health sector: (1) The decentralization of FL and (2) The quantum computing counterpart of FL, which we discuss in more detail below.

1) DECENTRALIZED FL

First, we note that all previous works mentioned so far usually rely on a trusted central server, for example, a primary hospital which manages the interoperability of data collected at the edges. However, the total communication cost may be prohibitive with many clients involved in global model training. In addition, servers are subject to unpredictable failures that can lead to training interruptions or total loss of the global model. To address this problem, studies such as [104] investigate the decentralized environment with no trusted central server. Peers (local clients) exchange model update information with each other until they agree on the global model. However, the fully decentralized approach presents many challenges, such as lacking training capacity at all federated nodes, lacking high-quality training datasets, and the authentication requirements for each participating client. To overcome these challenges, blockchain-based FL solutions have been presented in the literature related to the Internet of Health Things (IoHT) [105]. The main purpose of such studies is to trust the decentralized network of healthcare with carefully designed consensus protocols that would also be useful for FL, such as measuring the training quality for each peer and validating their work quality through such consensus protocols (proof of training quality) [106]. These studies still need some maturity as security attacks may reveal sensitive information to undesired third parties.

2) QUANTUM COMPUTING FOR PRIVATE DISTRIBUTED LEARNING

The use of quantum computers to accelerate ML is a well-researched topic in the literature [107], [108]. This work focuses on exploiting potential quantum mechanical advantages over classical systems to improve the execution speed of ML algorithms on a large scale. However, the special nature of quantum mechanics can also be exploited to solve the privacy problems we face everyday such as key distribution [109]. In the context of FL for the healthcare sector, the goal is to learn a global model while protecting the privacy of patient data. Therefore, quantum techniques can be applied to private distributed learning to solve the interplay between individual patient privacy and machine learning techniques. Most conventional approaches to extending FL in quantum computing involve retaining the protocol but replacing the client's classical computations with quantum computations. For example, in [110], variational quantum classifiers (QVCs) are run on the clients, while the server aggregates the local parameters of the QVCs and computes global parameters that are shared. They use transfer learning to retrain the local QVCs while keeping the local data private.

An alternative approach is based on blind quantum computing. The idea behind the blind quantum computing protocol is based on a quantum server that can perform quantum computations for a client without explicitly knowing the client's data and computations (because that would require measurement and hence the loss of data). In such an environment, the client has no quantum computation or memory capabilities [111]. This seems to be in contrast to the original definition of FL, where clients are an active part of the overall computation, computing gradients in the context of ML, for example. However, given the potential capabilities of quantum computers (inherent parallelism), limited mobility, and large number of clients, it makes more sense to move the computation to the server side. On the other hand, we note that such a large-scale environment must enable unconditionally secure private distributed learning. Thus, there are two clear advantages: (1) offloading local model training to a centralized, untrusted quantum computer while preserving the privacy of patient data, and (2) leveraging the potential quantum advantages in accelerating various ML algorithms in healthcare [112].

3) SUITABILITY OF FL FOR HEALTH DATA

Perhaps the most cited accumulation of data in the healthcare sector is due to EHRs [113] which typically provide biased patient-related data for analysis. The basic randomness or inherent bias in the data does not help machine learning algorithms with their generalization capabilities. To avoid wrong inference on local data, it is beneficial to use EHRs, develop models for global population and health trends. However, mainly for privacy reasons, collected data are rarely merged into a single database for combined processing, even within the same medical institution. Moreover, some of the data have many common records, leading to replication-discovery methods [114] as found in all deduplication systems [115]. Based on all these observations, FL seems to be a good candidate to address the problems of data agglomeration and global model fitting simultaneously. With recent advances in the FL literature, new perspectives on data sharing, high levels of privacy, client verification, on-device ML, incentive-based training protocols, and rapid access to intelligent diagnostic decisions can be achieved without leaking sensitive patient data.

E. FEDERATED LEARNING APPLICATIONS IN HEALTHCARE

FL has emerged as a transformative approach in healthcare, significantly enhancing applications such as Remote Patient Monitoring (RPM), Medical IoT, wearable devices, Home Care, and telemedicine. By training machine learning models directly on devices at the edge nodes, FL fundamentally shifts away from the traditional centralized approach of managing sensitive patient data, placing privacy and compliance at the forefront in alignment with regulations like HIPAA [116]. In the realm of RPM, FL enables real-time refinements for personalized care models, ensuring continuous and secure monitoring of a patient's health status with updated models through the latest data without compromising individual privacy. This allows for more responsive and adaptive healthcare, as the most current data can be used to tailor interventions and treatments. For medical IoT and wearable health devices, FL leverages the vast amounts of data generated by these devices to gain deeper insights into patient health trends and disease patterns. The decentralized nature of FL means that the raw data does not need to be transferred from the device, significantly mitigating the risks of data breaches and reducing the latency and bandwidth requirements associated with large data transfers. The result is faster, more scalable and more efficient solutions for the healthcare sector.

Home care and telemedicine also benefit greatly from FL. In-home care, FL enables continuous learning and improvement of models based on data from the home of patient without violating their privacy. This is particularly beneficial for elderly or chronically ill patients who require constant monitoring. Telemedicine platforms can use FL to integrate data from multiple sources, such as patient self-reports, home monitoring devices and remote consultations, to enable comprehensive care and timely interventions while ensuring privacy and security. The advantages of FL in these contexts include:

- **Enhanced Privacy:** By keeping data localized on the devices, FL minimizes the risk of data breaches and ensures compliance with stringent privacy regulations.
- **Personalized Healthcare:** Continuous model updates using real-time data allow for highly personalized treatment plans and interventions.
- **Scalability and Efficiency:** Reducing the need for central data aggregation decreases latency and bandwidth use, facilitating quicker and more scalable solutions.
- **Collaborative Ecosystem:** FL promotes a collaborative approach where multiple healthcare institutions can contribute to and benefit from shared AI models without exposing proprietary or sensitive information, fostering innovation and improving overall healthcare quality.

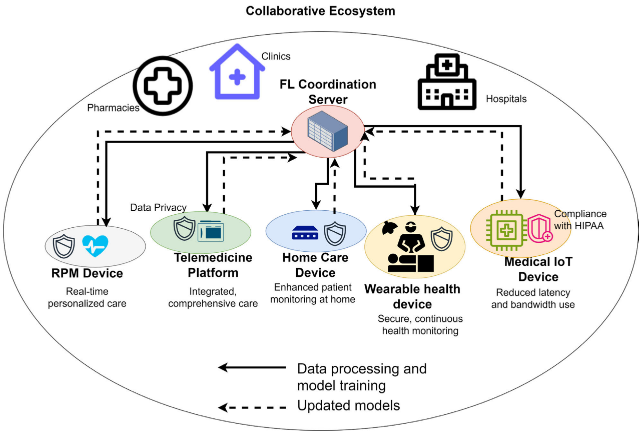

FIGURE 5. FL applications in healthcare in a collaborative ecosystem.

Consequently, FL envisions a secure, patient-centric collaborative ecosystem that drives innovation in healthcare by enabling diverse healthcare institutions to share and utilize AI models safely. This approach not only preserves patient privacy but also accelerates the development of advanced, data-driven healthcare solutions that can adapt to the dynamic needs of patients and healthcare providers. Fig. 5 shows the collaborative ecosystem and FL components for healthcare applications.

Finally, Table 2 provides an extended overview of the practical applications and case studies of Federated Learning in healthcare, highlighting the benefits of each application.

F. PERFORMANCE ADVANTAGES AND DISADVANTAGES OF FEDERATED LEARNING IN HEALTHCARE

FL offers a promising approach to improving healthcare through decentralized machine learning, enabling better data privacy, real-time personalization, and collaborative innovation. However, like any technology, it has its own challenges and limitations. Understanding the pros and cons of FL is crucial for effective implementation in healthcare. One of the key benefits of FL is its ability to protect privacy by localizing data. This minimizes the risk of data breaches as sensitive patient information is not transmitted to a central server. This aspect is crucial in the healthcare sector, where patient confidentiality is of paramount importance. However, ensuring complete security against sophisticated cyber-attacks remains a challenge. Even if the data is not centralized, the model parameters exchanged during the training process can still be vulnerable to inference attacks. Advanced encryption techniques and robust security protocols are essential to minimize these risks. FL enables real-time model updates and thus highly personalized treatment plans and interventions tailored to the individual patient. This is particularly beneficial in scenarios that require immediate responses, such as chronic disease monitoring or critical care management. However, the heterogeneity of data from different sources can complicate the modeling process. Variations in data quality, format and

TABLE 2. Practical Applications and Case Studies of Federated Learning in Healthcare.

Application	Case Study	Benefit
Remote Patient Monitoring	Implement an FL-based RPM system to monitor heart disease patients [117]. The system can allow continuous updates to the predictive models using real-time data from patients' homes, leading to timely interventions and reduced hospital readmissions.	Real-time personalized care with enhanced privacy, reducing the risk of data breaches.
Medical IoT and Wearables	Use FL to analyze data from diabetic patients in large scale deployment of wearables [30], [118]. The insights gained can help in early detection of complications, significantly improving patient outcomes.	Secure, continuous health monitoring with reduced latency and bandwidth use.
Home Care	Employ FL to monitor residents' vital signs [119]. The decentralized approach enabled continuous model improvements without compromising residents' privacy.	Enhanced patient monitoring at home with minimal privacy concerns.
Telemedicine	Integrate FL to manage data from remote consultations and home monitoring devices [120]. This approach ensured comprehensive care while maintaining data security.	Integrated, comprehensive care with secure data management.
Hospital Readmission Prediction	Use FL to predict patient readmissions by training models on decentralized patient data from multiple institutions [121]. This approach can help identify high-risk patients and reduce overall readmission rates.	Improved predictive accuracy and proactive patient management while preserving patient data privacy.
Cancer Detection	Apply FL to develop models for early cancer detection using imaging data [122], [123], [124]. Enable the pooling of knowledge and data without sharing sensitive patient information.	Enhanced early detection capabilities with collaborative model development, maintaining patient confidentiality.
Chronic Disease Management	FL to manage chronic diseases like diabetes and hypertension [118]. The system can aggregate data from various local health centres to provide better disease management insights.	Better disease management through aggregated data insights while ensuring patient data security.
Pharmaceutical Advancements	Introducing FL to collaborate on drug research [125]. By sharing model updates instead of raw data, the process could be accelerated while maintaining the security of proprietary data.	Accelerated drug discovery process with secure collaboration and protection of proprietary data.

distribution require sophisticated aggregation and normalization techniques to ensure that personalized models remain accurate and effective.

FL reduces latency and bandwidth utilization by processing data locally, making it a scalable solution suitable for large healthcare systems. This decentralized approach enables faster updates and scalability. However, frequent updates can lead to high communication costs and latency issues, especially in resource-constrained environments. Optimizing communication protocols and using model compression techniques can help alleviate these issues and ensure efficient and scalable FL deployment. FL fosters a collaborative environment where multiple healthcare organizations can share AI models without disclosing sensitive information. This collaborative innovation can accelerate progress in healthcare by utilizing different data sets from different institutions. However, standardizing data across different devices and institutions remains complex and time-consuming. The development of interoperable standards and frameworks can streamline this process and enable smoother collaboration and integration. By using edge devices for processing, FL reduces the need for centralized data aggregation, leading to resource efficiency. This can reduce infrastructure costs and improve the responsiveness of healthcare applications. However, the limited computing resources on edge devices can affect the efficiency of model training. Balancing the computational load between edge devices and central servers and using lightweight model architectures can improve resource efficiency without compromising performance. Finally, FL aligns well with privacy regulations such as HIPAA to ensure compliance and protection of patient data. This alignment with regulatory frameworks is critical for trust and adoption in the healthcare industry. However, compliance requires constant vigilance and the use of advanced techniques to secure data throughout its lifecycle. Continuous monitoring, regular audits and updates to security protocols are necessary to ensure compliance and protection against new threats.

Table 3 provides an overview of the performance advantages and disadvantages of FL in healthcare, highlighting the key aspects of this technology.

VI. POTENTIAL ISSUES AND DISCUSSIONS

A. CHALLENGES OF FEDERATED LEARNING AND POTENTIAL SOLUTIONS

There are multiple challenges related to FL, which are addressed by a few recent studies, as will be explored later. These challenges influence how FL operates regarding convergence, consensus, accuracy, and latency to obtain reliable results. These challenges are significant in healthcare applications where reliable results are critical.

1) STATISTICAL CHALLENGE

This refers to the fact that health data originating from multiple resources and geographic locations may have dramatic differences, i.e., the nature of the data distribution may be quite different for each patient. The presence of signal/data differences on a large scale can cause difficulties for the learning algorithms to converge within a reasonable time. For example, patient data from different regions may have different disease prevalences and health indicators, making it difficult to train a unified model. Personalized FL techniques can be used to tailor the models to individual clients or groups of clients with similar data distributions. In addition, transfer

TABLE 3. Performance Advantages and Disadvantages of Federated Learning in Healthcare.

Aspect	Advantages	Disadvantages
Privacy Preservation	Minimising risk of data breaches by keeping data localized	Ensuring complete security against cyberattacks remains challenging
Personalization	Real-time model updates allow for highly personalized treatment plans and interventions	Data heterogeneity can complicate the model training process
Scalability	Reduces latency and bandwidth use, facilitating quicker and more scalable solutions	Frequent updates can lead to high communication costs and latency issues
Collaborative Innovation	Promotes a collaborative approach where multiple healthcare institutions can share and utilize AI models without exposing sensitive information	Standardizing data across different devices and institutions can be complex and time-consuming
Resource Efficiency	Utilizes edge devices, reducing the need for central data aggregation	Limited computational resources on edge devices can hinder the efficiency of model training
Regulatory Compliance	Aligns with privacy regulations like HIPAA, ensuring compliance and protecting patient information	Constant vigilance and advanced techniques are required to maintain compliance and security

learning and domain adaptation methods can help align the models so that they can handle different data sources better. The data engineering layer in Section IV with its data analysis and processing modules can use personalized federated learning techniques to process different data sources and ensure that the models are accurate despite different data distributions.

2) NETWORK RESOURCE DYNAMICS

The dynamic capacity of network resources poses another challenge. Since the number of patients could be quite large, typically larger than the average number of samples per patient, FL may suffer from the high-load communication requirements since transmitting the patient model over the network would cause too much data traffic, making the data communication to be the main bottleneck of the system. Moreover, this traffic can vary due to leaving and arriving participants in dynamic network environments [28]. Implementing efficient communication protocols and optimizing the frequency of model updates can reduce communication overhead. Techniques such as asynchronous FL, where clients update the central model at different times, can also help to manage the network load more effectively. Efficient communication protocols and asynchronous FL can be integrated into the data connection and ingestion modules discussed in Section IV to manage network load and ensure timely data processing without bottlenecks.

3) TRUST AND SECURITY

among participating entities: Ensuring trust among participating entities is crucial. For instance, some of the clients could be malicious. Therefore, robust authentication mechanisms are required to ensure participants' reliability and prevent untrusted parties from disrupting the training phase of FL. It would harm the learning convergence if untrusted parties participate in the training phase of FL. As a solution, robust authentication and secure aggregation techniques can ensure that only trusted clients participate. Techniques such as Secure Multiparty Computation (SMC) and blockchain can improve trust and integrity in the FL process. Robust authentication mechanisms and secure aggregation techniques in the data management and orchestration modules discussed in Section IV can ensure that only trusted entities participate in the FL process and that data integrity is maintained.

4) STATISTICAL FAIRNESS

InFL, faced particularly in the health domain, the statistical fairness for the different user data distributions is important. Simply taking the average as the combining mechanism of central server would be unfair if the data distributions are skewed. One solution is to model the target distribution based on a mixture of client distributions obtained by having each client node uniformly sample their data and share it across the network. The data sharing process can be between trusted parties or only with representative data containing mostly non-sensitive content [126]. We note that Eqn. (1) assumes homogeneity among D_k's. In other words, the optimal model parameter $\boldsymbol{w}$ would not be the best fit if D_k's have dramatic differences. This observation and heterogeneity in the real world have led researchers to explore multi-task learning (MTL) strategies in which non-identically distributed imbalanced data can be modelled based on the relationship between them. Note that non-IID (non-Independent and Identically Distributed) data can lead to biased models if not handled properly, as standard aggregation techniques may not adequately represent the different data distributions. Thus, instead of training a single global model, multiple models are trained in their graph representation based on sparsity, low-rank structure, etc.. To account for data heterogeneity, studies such as [127] have explored variations in gradient descent and shown that training multiple models can outperform the i.i.d. assumption commonly adopted for health data. Of course, the success of MTL here depends heavily on the assumptions about the relationship between different types of health data from different institutions. Recent studies such as [29] proposed to minimize the runtime gap between clients and maximise convergence gain by optimizing the number of local iterations based on the workload assignments. Techniques such as personalized federated learning, where models are tailored to individual customers or groups of clients with similar data distributions, can also mitigate the problem and enable the creation of models that are better tailored to the specific data distributions of individual clients or groups. Weighted aggregation techniques in the data analysis and processing modules, discussed in Section IV, can provide a fair distribution of data and ensure that the models accurately represent different patient populations. In addition, feature engineering and model training components can utilize federated multi-task learning to account for the distribution of non-IID data and ensure fair and accurate model training.

5) INDIRECT LEAKAGE

Although the natural setting of FL prevents direct leakage of patient data in the training phase, it does not preclude clients from learning about another client's private dataset by observing $f_i(\boldsymbol{w}, \boldsymbol{x}_i)$ or the shared predictive model parameter $\boldsymbol{w}$. There are methods such as homomorphic encryption, where each client encrypts its data with a public key, and any client with the same public key can process the private data. However, a homomorphic approach does not completely solve the problem of individual privacy. It is also very computationally intensive. An alternative theoretical method is known as DP, where the goal is to keep the statistical nature of the dataset the same (e.g., for FL) while keeping the identities of individual patients secret. Maintaining the privacy of individuals has a limited impact on the descriptive nature of the data (pattern). Because of its lossy nature, using DP may reduce predictive accuracy [128]. Therefore, combining homomorphic encryption with DP might be a good idea to get the best of both worlds [129]. The combination of homomorphic encryption and differential privacy in the data storage and management modules discussed in Section IV can protect patient data while preserving the performance of the model.

6) CROSS-VALIDATION, BIAS, AND CALIBRATION

Healthcare data can be biased and may not represent the physical world accurately. While the validity of data-driven research largely depends on the accuracy and representativeness of the used data, healthcare data collected with common mechanisms can be biased compared with the distribution of related features in the physical world. Many healthcare observations come from small datasets, and using these to make predictions for larger datasets could introduce bias due to size dependency. The choice of cross-validation method is crucial in obtaining accurate estimates, as different variants have varying levels of bias and variance depending on the problem at hand [130]. The importance of calibration in healthcare settings is highlighted in [131], which proposes a new approach that outperforms classical and modern survival analysis baselines regarding discriminative performance and calibration, particularly for minority demographics. Continuous validation and updating are recommended in [132] to account for variations in calibration and reliability of predictions when pursuing clinical implementation. While current models for hypertension prediction demonstrate acceptable to good discrimination and calibration ability, more validation and adjustment are needed before they can be applied in clinical practice, as stated in [133]. In [134], a model is developed to predict infants at risk of severe adverse neonatal outcomes, which achieves high calibration with moderate accuracy using a combination of maternal, intrapartum, and ultrasound variables. To address calibration issues, techniques such as data resampling [135] and improved calibration approaches [136] have been proposed.

FL has shown promise in developing robust predictive models for COVID-19 EHR data while maintaining patient privacy, as discussed in [137]. FL can also allow models to be trained on datasets of unprecedented size, which can significantly impact precision/personalized medicine, as noted in [138]. Reference [139] proposes optimizations to FL methods to handle heterogeneity across institutions, providing valuable guidance for real clinical applications. FL can also be used for analyzing medical images while protecting patient information, as shown in [140]. A new method called Federated-Autonomous Deep Learning (FADL) is proposed in [141], which trains part of the model using all data sources in a distributed manner and other parts using data from specific sources, outperforming traditional FL strategies. Another adaptive boosting method, LoAdaBoost FL, which increases the efficiency of federated machine learning, is proposed in [142] using intensive care unit data. Decentralized FL is shown to be superior to classic methods in [143], based on multiple numerical simulations using large real-world electronic health record databases. Finally, in [144], the authors propose a feature fusion method to address communication costs and performance drops in federated averaging, the leading optimization algorithm in FL, especially when the local data is distributed non-IID. Other solutions, such as the XGBoost model, are also discussed in [145] for obtaining the best calibration compared with other machine learning algorithms. Robust cross-validation and calibration techniques in the model evaluation module discussed in Section IV can ensure the reliability and accuracy of predictive models in clinical settings.

7) REAL-TIME DATA PROCESSING

Finally, in the context of Healthcare IoT, the size of typically generated patient data is voluminous and difficult to process. The data load may result in congestion in the network and increased latencies due to the imperfection of the network and the required computation. The increased round-trip and hop times between IoT devices and cloud servers may render healthcare data stale, irrelevant or inadequate for some users. In the case of time-sensitive healthcare sectors, the real-time data streams must be processed by ML on time to make FL useful. Particularly, if some of the client processing is delayed, the overall latency and accuracy of the trained model can be seriously affected.

There are multiple solutions proposed in the literature. An interesting load of work has been conducted in the context of *coded computation* [146] where matrix multiplications are distributed for reliability in the event of laggy, struggling and unreliable commodity hardware [147]. In addition, the same line of work has been extended to gradient computations which are encoded to avoid stragglers [148]. In such a setting, the computation (gradients in an ML context) gets encoded in a redundant fashion and distributed across multiple client nodes. The redundancy introduced in the computation enables reconstruction if a subset of these clients completes their assigned work and can communicate it with the server. That way, a straggler would not be a bottleneck to the overall iterative global model training process. Edge computing and coded computation techniques in the data connection and ingestion modules discussed in Section IV can process data in real-time, reducing latency and ensuring timely patient care.

B. CHALLENGES OF ML LIFECYCLE MANAGEMENT AND POTENTIAL SOLUTIONS

In this section, we discuss potential challenges that may arise when applying the proposed end-to-end ML lifecycle management and data engineering pipelines in real-world scenarios. There are several challenges to overcome when applying the techniques of AI/ML to large, complex, and imperfect health data. Therefore, both data management and algorithm transparency should be systematically addressed. One of the biggest challenges in big data in healthcare is the sheer volume of data that needs to be managed, as well as the complexity of the data. This includes data from electronic health records, medical imaging, genomic sequencing, and wearable devices. Recent experiments have focused on using machine learning algorithms and AI to manage and analyze large and complex healthcare datasets.

1) DATA VOLUME AND COMPLEXITY

A fundamental problem is that all deep learning approaches require large datasets to train models (e.g., image classifiers in supervised learning models). Some complex models such as deep learning-based models, require large datasets for acceptable generalization. This is particularly difficult in healthcare, where collecting large datasets (e.g., in medical imaging, medical recommendations) is difficult, especially when patient privacy is involved. Some of the relevant solutions to this problem are listed below:

(i) *Labelling:* To quickly label the new data or relabel the existing data for a new model. Solutions such as Snorkel can be used for this purpose. Snorkel is based on labelling functions (LFs) where, for example, in healthcare, the patient's health history/report is encoded as functions, and these functions are used to label training data programmatically. If a health history/report says "malignant", the data can be labelled as "emergent" and so on.

(ii) *Transfer Learning:* Since not many samples/images of patients can be collected and separate models are trained in each hospital, a suitable approach is to use transfer learning to train deep neural networks (e.g., Convolutional Neural Networks (CNN))). Together with transfer learning, better accuracy of deep learning models can be achieved with small datasets. This is achieved by leveraging existing models trained with large datasets. For example, using the knowledge of how bacterial pneumonia is detected with X-rays taken from the patient using very large datasets of images (e.g., RestNet-50 [149]), learning to detect different types of viral pneumonia (e.g., COVID-19) is called transfer learning. This is based on using similar low-level image features. For example, the results in [150] show that the transfer learning approach (pre-trained on ImageNet [151]) can perform better in detecting COVID-19 viral pneumonia than without transfer learning (i.e., training from scratch).

(iii) *Weak Supervision:* This is based on leveraging noisy and imprecise sources to create labels.

(iv) *Semi-supervised Learning:* It is a special version of weak supervision. To use unlabeled data automatically, this learning technique aims to use a small amount of labeled data to label large amounts of unlabeled data to train models.

(v) *In active learning*, the points that are most valuable to solicit labels are estimated. For example, in computerized tomography (CT) scans, labeling can be solicited only for those close to the decision boundary.

Using techniques such as transfer learning and weak supervision, the data source layer in Section IV can effectively process diverse and large data sets and improve the overall quality of data collected from different sources. Using active learning tools and data labeling tools can ensure that data ingestion processes are efficient and capable of managing large volumes of complex data. Finally, the above solutions can enable more accurate and comprehensive data analysis in the data analysis and processing module in Section IV, leading to better insights and decisions.

2) DATA QUALITY AND BIAS

The second fundamental challenge is that medical data, like other real-world data, can be incomplete, missing, incorrect, or biased (depending on gender, weight, race, sexuality, and many other factors). This can be due to various reasons, such as incorrect calibration of the instruments used to collect patient data (e.g., heart rate monitors, pulse oximeters, etc.). Therefore, biases, delays, and errors in data and observations can lead to inaccurate diagnoses and decisions when using AI/ML algorithms that rely on these data. In addition, incorrectly assuming that AI/ML algorithms are implemented without error can negatively affect patients because of errors in the software implementation of the algorithm. The collected data must also be carefully preprocessed (data cleaning, data joining, etc.). Data augmentation (labeling, expert access) and data analysis (profiling) are other considerations during the data management process. From ML pipeline building perspective of Section IV, ensuring high-quality data ingestion processes can improve the reliability of downstream data analysis and processing. In addition, addressing data quality and bias in the data analysis and processing module can ensure more accurate and fairer model training and evaluation. Finally, continuous monitoring of the model monitoring module helps maintain model performance and ensures reliable and unbiased predictions over time.

3) CONCEPT DRIFT

The third challenge is that the model must be updated after deployment due to concept drift constraints. The model must be retrained if the data distribution has shifted. For this reason, additional AI/ML infrastructure should be put in place as to monitor and perform the continuous delivery of the model, which is normally part of the data operations. From ML pipeline building perspective of Section IV, the continuous delivery mechanisms of the model training module ensure that the models are regularly updated to reflect changes in the data distribution. Monitoring model performance and triggering retraining processes in the model monitoring module help maintain the models' relevance and accuracy.

4) DATA PRIVACY AND SECURITY

The fourth challenge is related to data privacy and secure communication. Patient data should be used to train models while maintaining patient privacy. The whole process should also be GDPR-compliant. As a solution, libraries such as PySyft* can be used for secure and private deep learning using techniques such as FL, DP, and Encrypted Computation (such as MPC and HE) within popular deep learning frameworks such as TensorFlow and PyTorch. Moreover, secure communication channels should be established via encryption algorithms (like AES, RSA) and other security measures (like HTTPS and SSH) to protect data as it is transmitted over the internet or other network [152]. From ML pipeline building perspective of Section IV, Data source layer can ensure secure data capture and transmission from various sources while maintaining patient privacy and data integrity. The Data Storage Module can provide secure storage solutions to protect patient data while complying with privacy regulations. Finally, the model serving module can provide models securely, ensuring patient data used in predictions is handled responsibly.

5) INTEROPERABILITY AND INTEGRATION

The fifth challenge is aspects of interoperability, integration and data heterogeneity. There are several data models and standards (e.g., HL7, OpenEHR, CEN/ISO, CPT4) for unstructured data, such as EHRs [153]. Since health data are large datasets with heterogeneous resources (e.g., XML, CSV, SQL etc.), different schemas, vocabularies, structures and standards, interoperability is a critical problem to solve. Processes such as missing data, same values entered in different data forms, scaling, and maintaining semantic interoperability must be performed during the data cleansing and preparation phase, which can be done, for example, in the data ingestion module of the data engineering pipeline. ETL (extract, transform, load) tools like Apache Nifi, data integration platforms like Apache Camel and programming tools Python and Java can help to manage this challenge. From ML pipeline building perspective of Section IV, effective ETL processes within the data ingestion module can ensure the seamless integration of different data sources and thus enable comprehensive data analysis. Standardized data formats in the data analysis and processing module can improve data processing and model training accuracy and efficiency.

6) NON-IID DATA

The sixth challenge is concerning Non-independent and non-identically distributed (non-IID) data [154] in healthcare. This can happen due to data collection from different hospitals with different distributions of patient populations or treatment protocols. This makes generalising models to new populations difficult since the standard statistical assumptions may no longer hold. Advanced techniques such as hierarchical models [155] or multi-level models [156] can address these challenges. From ML pipeline building perspective of Section IV, in the data source layer, Collecting and integrating different data sets using advanced modeling techniques improves the robustness of the data source layer. Personalized federated learning can ensure that models trained on non-IID data are accurate and generalizable in the data analysis and processing module. Finally, hierarchical and multi-level models improve the ability to process diverse and complex data sets in the model training module.

VII. LESSONS LEARNED AND FUTURE WORK

In this section, we reflect on the key lessons learned from our study on FL in healthcare and highlight the main challenges and potential solutions for effective FL implementation. We also outline future research directions that can further improve the applicability and robustness of FL in real-world healthcare scenarios. The convergence of data engineering and advanced machine learning approaches such as FL and DL offers numerous opportunities, but also brings significant challenges that need to be overcome to fully realize the potential of these technologies in healthcare.

A. LESSONS LEARNED

While FL holds immense potential to revolutionize healthcare by enabling collaborative model training while maintaining data privacy, there are also some challenges that need to be overcome.

- **Data Privacy Concerns:** Healthcare data is sensitive and subject to strict privacy regulations such as HIPAA. FL solves this problem by localizing the data and enabling model training without sharing the raw data. However, ensuring robust privacy mechanisms remains

[*] Online: https://github.com/OpenMined/PySyft, [accessed June-2022]

paramount. The use of secure learning frameworks and secure communication protocols within the ML pipeline ensures compliance with privacy regulations and, thus, patient trust and data integrity.

- **Data Heterogeneity:** Healthcare data differs from institution to institution in terms of format, quality and scope. FL must accommodate this diversity while ensuring the performance of the model across all participating facilities. Techniques such as transfer learning and model aggregation strategies can mitigate these challenges. Advanced modeling techniques and personalized federated learning in the data analysis and processing module in Section IV can help to consider different data sources and improve model accuracy and generalizability.
- **Communication Overhead:** FL requires frequent communication between the central server and local devices, which can be resource-intensive and prone to network latency. Optimizing communication protocols and model compression techniques can help mitigate this overhead. Efficient communication protocols and asynchronous FL within the data connection and ingestion modules in Section IV can ensure seamless and timely data processing.
- **Model Aggregation Challenges:** Aggregating local model updates while preserving data privacy and model performance is non-trivial. Federated averaging and secure aggregation methods are commonly used but may need further optimisation for healthcare applications. Secure aggregation techniques and federated learning in the data analysis and processing module in Section IV can ensure robust and accurate model updates, improving overall system reliability.
- **Bias and Generalization Issues:** FL models can suffer from bias due to non-representative local data sets, resulting in poor generalization performance. These issues can be addressed by incorporating bias detection and mitigation techniques, such as federated meta-learning or personalized FL. Bias detection and fairness-aware learning methods within the model evaluation module can ensure that models are equitable and generalizable across diverse patient populations.
- **Regulatory Compliance:** Healthcare regulations pose significant challenges for using FL. Ensuring compliance with privacy laws and regulatory standards requires close collaboration between stakeholders, including clinicians, data scientists and policymakers. The data governance and compliance mechanisms of ML pipeline can ensure regulation adherence, fostering stakeholder trust.

B. FUTURE WORK

Future research should focus on various techniques, such as improving privacy mechanisms, enhancing the robustness and fairness of the model, optimizing scalability and efficiency, validating clinical effectiveness, promoting interoperability, and ensuring long-term sustainability and governance, as listed below. Collaboration between researchers, clinicians, policymakers and industry stakeholders is essential to address these challenges and realize the full benefits of FL in healthcare.

- **Enhanced Privacy Mechanisms:** Future research should focus on developing advanced privacy-preserving techniques, such as differential privacy and homomorphic encryption, to strengthen the privacy guarantees of FL in healthcare. Incorporating these techniques into the data storage and model serving modules will enhance patient data protection.
- **Robustness and Fairness:** Dealing with bias and fairness in FL models requires ongoing research. Bias detection techniques, fairness-aware learning and interpretability should be further developed to ensure equitable healthcare outcomes. Implementing these methods in the model training and evaluation modules will ensure that the models are fair and interpretable.
- **Scalability and Efficiency:** Improving the scalability and efficiency of FL frameworks is essential for widespread use in healthcare. This includes the optimization of communication protocols, the development of lightweight model architectures and the use of edge computing resources. Enhancing the data connection and ingestion modules to handle large-scale data efficiently and leverage edge computing for real-time processing.
- **Clinical Validation and Real-World Deployment:** Conducting large-scale clinical trials and validation studies is crucial for evaluating the real-world effectiveness of FL in healthcare. Collaboration between researchers, clinicians and healthcare institutions is essential for successful implementation and integration into clinical workflows. Integrating clinical validation processes within the data analysis and processing modules can ensure the practical applicability of developed models.
- **Interoperability and Standardization:** To ensure seamless data exchange and model compatibility, interoperability standards and best practices for FL implementation across healthcare systems need to be established. This requires collaboration between industry stakeholders, standardization bodies and regulators. Standardizing data formats and integration practices within the data ingestion and management modules can facilitate interoperability.
- **Long-Term Sustainability:** Addressing long-term sustainability and governance challenges of FL in healthcare is critical. Developing sustainable business models, ensuring data governance and fostering trust between stakeholders are critical to the continued success of FL initiatives. Implementing robust data management and orchestration frameworks can ensure long-term operational efficiency and stakeholder trust.

Data engineering and recent approaches such as FL and DL are nascent fields and are likely to shape future developments in healthcare. However, there are also major challenges to be overcome in the convergence of healthcare and data engineering. High velocity, huge datasets with high dimensions or diversity of healthcare data (e.g., radiology, genomics, etc) present additional challenges for healthcare analytics platforms. In addition, data can be of poor quality and trustworthiness, imbalanced and subject to noise. Additional stringent clinical guidelines, data ownership, administrative compliance rules, and regulatory submission standards further complicate the work of healthcare platform teams. Processing segmented or siloed data across multiple hospitals is also no easy task. For these reasons, healthcare data engineering teams face unparalleled challenges compared to other industries. On the other hand, advances in next-generation high-performance computing platforms, paradigms and technologies based on concepts such as AI/ML/DL, cloud/fog/edge, robotics, blockchain, serverless and quantum computing can help mitigate these challenges and improve the connection between the data science/engineering and healthcare fields [17].

VIII. CONCLUSION

The ecosystem of data engineering and data science will inevitably play an important role in the healthcare sector. In this paper, we have explored recent advances in data engineering-based approaches tailored to healthcare needs. In particular, we have analyzed and proposed a method to build an end-to-end data engineering pipeline and focused on efficient end-to-end ML lifecycle management of healthcare data. Our key findings highlight the significant potential of integrating AI/ML frameworks and cloud infrastructures in developing scalable, secure and efficient healthcare solutions. By leveraging FL, we can ensure data privacy and compliance with stringent healthcare regulations while enabling a collaborative model for training across multiple institutions. We have outlined the architecture of an end-to-end data engineering pipeline that includes data source integration, ingestion, analysis, storage and visualization. Each module is designed to address specific healthcare data challenges, such as heterogeneity and volume of data. We have addressed critical challenges in ML lifecycle management, including data privacy, data quality, communication overhead and interoperability. Our proposed solutions, such as secure aggregation techniques and advanced modeling methods, provide practical approaches to these challenges. In summary, integrating advanced data engineering and AI/ML technologies holds immense potential to revolutionize healthcare. We can develop robust, scalable and patient-centric healthcare solutions by addressing the challenges outlined and focusing on future research directions. The findings and proposals in this paper pave the way for significant advances in the effective use of healthcare data to improve patient and healthcare outcomes ultimately.

REFERENCES

[1] R. Agius, C. Brieghel, M. A. Andersen, A. T. Pearson, B. Ledergerber, A. Cozzi-Lepri, Y. Louzoun, C. L. Andersen, J. Bergstedt, J. H. von Stemann, M. Jørgensen, M.-H.-E. Tang, M. Fontes, J. Bahlo, C. D. Herling, M. Hallek, J. Lundgren, C. R. MacPherson, J. Larsen, and C. U. Niemann, "Machine learning can identify newly diagnosed patients with CLL at high risk of infection," *Nature Commun.*, vol. 11, no. 1, pp. 1–17, Jan. 2020.

[2] T. Carlson and J. del R. Millan, "Brain-controlled wheelchairs: A robotic architecture," *IEEE Robot. Autom. Mag.*, vol. 20, no. 1, pp. 65–73, Mar. 2013.

[3] D. H. Freedman, "Hunting for new drugs with AI," *Nature*, vol. 576, no. 7787, pp. S49–S53, Dec. 2019.

[4] C. Jacobs and B. van Ginneken, "Google's lung cancer AI: A promising tool that needs further validation," *Nature Rev. Clin. Oncol.*, vol. 16, no. 9, pp. 532–533, Sep. 2019.

[5] R. Hodson, "Precision medicine," *Nature*, vol. 537, no. 7619, pp. S49–S49, 2016.

[6] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Explainable AI in healthcare," in *Proc. Int. Conf. Cyber Situational Awareness, Data Analytics Assessment (CyberSA)*, Jun. 2020, pp. 1–2.

[7] J. Passerat-Palmbach, T. Farnan, M. McCoy, J. D. Harris, S. T. Manion, H. L. Flannery, and B. Gleim, "Blockchain-orchestrated machine learning for privacy preserving federated learning in electronic health data," in *Proc. IEEE Int. Conf. Blockchain (Blockchain)*, 2020, pp. 550–555.

[8] S. S. Arslan, R. Jurdak, J. Jelitto, and B. Krishnamachari, "Advancements in distributed ledger technology for Internet of Things," *Internet Things*, vol. 9, Mar. 2020, Art. no. 100114.

[9] A. Rahman, M. Rahman, D. Kundu, M. R. Karim, S. S. Band, and M. Sookhak, "Study on IoT for SARS-CoV-2 with healthcare: Present and future perspective," *Math. Biosciences Eng.*, vol. 18, no. 6, pp. 9697–9726, 2021.

[10] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny, "A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues," *J. Biomed. Informat.*, vol. 113, Jan. 2021, Art. no. 103627.

[11] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, "Multi-label active learning-based machine learning model for heart disease prediction," *Sensors*, vol. 22, no. 3, p. 1184, Feb. 2022.

[12] G. Turk, M. Ozdemir, R. Zeydan, Y. Turk, Z. Bilgin, and E. Zeydan, "On the identification of thyroid nodules using semi-supervised deep learning," *Int. J. Numer. Methods Biomed. Eng.*, vol. 37, no. 3, p. e3433, Mar. 2021.

[13] A. Afrasiabi, J. T. Keane, J. I.-T. Heng, E. E. Palmer, N. H. Lovell, and H. Alinejad-Rokny, "Quantitative neurogenetics: Applications in understanding disease," *Biochem. Soc. Trans.*, vol. 49, no. 4, pp. 1621–1631, Aug. 2021.

[14] S. Hiriyannaiah and K. G. Srinivasa, "A comparative study and analysis of LSTM deep neural networks for heartbeats classification," *Health Technol.*, vol. 11, no. 3, pp. 663–671, May 2021.

[15] M. Türk, R. Ertaş, E. Zeydan, Y. Türk, M. Atasoy, A. Gutsche, and M. Maurer, "Identification of chronic urticaria subtypes using machine learning algorithms," *Allergy*, vol. 77, no. 1, pp. 323–326, Jan. 2022.

[16] R. Harper and J. Southern, "A Bayesian deep learning framework for end-to-end prediction of emotion from heartbeat," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 985–991, Apr. 2022.

[17] A. A. Abdullah, M. M. Hassan, and Y. T. Mustafa, "A review on Bayesian deep learning in healthcare: Applications and challenges," *IEEE Access*, vol. 10, pp. 36538–36562, 2022.

[18] G. S. Chilla, L. Y. Yeow, Q. H. Chew, K. Sim, and K. N. B. Prakash, "Machine learning classification of schizophrenia patients and healthy controls using diverse neuroanatomical markers and ensemble methods," *Sci. Rep.*, vol. 12, no. 1, pp. 1–11, Feb. 2022.

[19] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time series prediction using deep learning methods in healthcare," *ACM Trans. Manage. Inf. Syst.*, vol. 14, no. 1, pp. 1–29, Mar. 2023.

[20] H. Kamimura, H. Nonaka, M. Mori, T. Kobayashi, T. Setsu, K. Kamimura, A. Tsuchiya, and S. Terai, "Use of a deep learning approach for the sensitive prediction of hepatitis B surface antigen levels in inactive carrier patients," *J. Clin. Med.*, vol. 11, no. 2, p. 387, Jan. 2022.

[21] J. Lu, L. Wang, M. Bennamoun, I. Ward, S. An, F. Sohel, B. J. W. Chow, G. Dwivedi, and F. M. Sanfilippo, "Machine learning risk prediction model for acute coronary syndrome and death from use of non-steroidal anti-inflammatory drugs in administrative data," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, Sep. 2021.

[22] D. S. Prasad, S. R. Chanamallu, and K. S. Prasad, "Optimized deformable convolution network for detection and mitigation of ocular artifacts from EEG signal," *Multimedia Tools Appl.*, vol. 81, no. 21, pp. 30841–30879, Sep. 2022.

[23] X. Liu, S. Xiong, X. Wang, T. Liang, H. Wang, and X. Liu, "A compact multi-branch 1D convolutional neural network for EEG-based motor imagery classification," *Biomed. Signal Process. Control*, vol. 81, Mar. 2023, Art. no. 104456.

[24] X. Lin, J. Chen, W. Ma, W. Tang, and Y. Wang, "EEG emotion recognition using improved graph neural network with channel selection," *Comput. Methods Programs Biomed.*, vol. 231, Apr. 2023, Art. no. 107380.

[25] V. Mani, C. Kavitha, S. S. Band, A. Mosavi, P. Hollins, and S. Palanisamy, "A recommendation system based on AI for storing block data in the electronic health repository," *Frontiers Public Health*, vol. 9, pp. 1–12, Jan. 2022.

[26] O. Samuel, A. B. Omojo, A. M. Onuja, Y. Sunday, P. Tiwari, D. Gupta, G. Hafeez, A. S. Yahaya, O. J. Fatoba, and S. Shamshirband, "IoMT: A COVID-19 healthcare system driven by federated learning and blockchain," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 2, pp. 823–834, Feb. 2023.

[27] S. Lau, L. Flemming, and J. Haueisen, "Magnetoencephalography signals are influenced by skull defects," *Clin. Neurophysiol.*, vol. 125, no. 8, pp. 1653–1662, Aug. 2014.

[28] J. Á. Morell and E. Alba, "Dynamic and adaptive fault-tolerant asynchronous federated learning using volunteer edge devices," *Future Gener. Comput. Syst.*, vol. 133, pp. 53–67, Aug. 2022.

[29] J. Zhang, X. Cheng, C. Wang, Y. Wang, Z. Shi, J. Jin, A. Song, W. Zhao, L. Wen, and T. Zhang, "FedAda: Fast-convergent adaptive federated learning in heterogeneous mobile edge computing environment," *World Wide Web*, vol. 25, no. 5, pp. 1971–1998, Sep. 2022.

[30] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, and W.-J. Hwang, "Federated learning for smart healthcare: A survey," *ACM Comput. Surveys*, vol. 55, no. 3, pp. 1–37, Mar. 2023.

[31] E. Zeydan and J. Mangues-Bafalluy, "Recent advances in data engineering for networking," *IEEE Access*, vol. 10, pp. 34449–34496, 2022.

[32] L. Wang and C. A. Alexander, "Big data analytics in medical engineering and healthcare: Methods, advances and challenges," *J. Med. Eng. Technol.*, vol. 44, no. 6, pp. 267–283, Aug. 2020.

[33] C. J. Cremin, S. Dash, and X. Huang, "Big data: Historic advances and emerging trends in biomedical research," *Current Res. Biotechnol.*, vol. 4, no. 4, pp. 138–151, Mar. 2022.

[34] S. Shafqat, S. Kishwer, R. U. Rasool, J. Qadir, T. Amjad, and H. F. Ahmad, "Big data analytics enhanced healthcare systems: A review," *J. Supercomput.*, vol. 76, pp. 1754–1799, Feb. 2020.

[35] Z. F. Khan and S. R. Alotaibi, "Applications of artificial intelligence and big data analytics in m-health: A healthcare system perspective," *J. Healthcare Eng.*, vol. 2020, pp. 1–15, Sep. 2020.

[36] L. M. Dang, M. J. Piran, D. Han, K. Min, and H. Moon, "A survey on Internet of Things and cloud computing for healthcare," *Electronics*, vol. 8, no. 7, p. 768, Jul. 2019.

[37] V. S. Naresh and M. Thamarai, "Privacy-preserving data mining and machine learning in healthcare: Applications, challenges, and solutions," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 13, no. 2, pp. 1–42, 2023, Art. no. e1490.

[38] Q. P. He and J. Wang, "Application of systems engineering principles and techniques in biological big data analytics: A review," *Processes*, vol. 8, no. 8, p. 951, Aug. 2020.

[39] A.-T. Shumba, T. Montanaro, I. Sergi, L. Fachechi, M. De Vittorio, and L. Patrono, "Leveraging IoT-aware technologies and AI techniques for real-time critical healthcare applications," *Sensors*, vol. 22, no. 19, p. 7675, Oct. 2022.

[40] M. Yeo, H. K. Kok, N. Kutaiba, J. Maingard, V. Thijs, B. Tahayori, J. Russell, A. Jhamb, R. V. Chandra, and M. Brooks, "Artificial intelligence in clinical decision support and outcome prediction–applications in stroke," *J. Med. Imag. Radiat. Oncol.*, vol. 65, no. 5, pp. 518–528, 2021.

[41] G. Currie, K. E. Hawk, E. Rohren, A. Vial, and R. Klein, "Machine learning and deep learning in medical imaging: Intelligent imaging," *J. Med. Imag. Radiat. Sci.*, vol. 50, no. 4, pp. 477–487, Dec. 2019.

[42] J. Scherer, M. Nolden, J. Kleesiek, J. Metzger, K. Kades, V. Schneider, M. Bach, O. Sedlaczek, A. M. Bucher, and T. J. Vogl, "Joint imaging platform for federated clinical data analytics," *JCO Clin. Cancer Informat.*, vol. 4, pp. 1027–1038, Nov. 2020.

[43] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–9, Jan. 2021.

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[45] I. Sechopoulos, J. Teuwen, and R. Mann, "Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art," *Seminars Cancer Biol.*, vol. 72, pp. 214–225, Jul. 2021.

[46] A. K. Saxena, R. R. Dixit, and A. Aman-Ullah, "An LSTM neural network approach to resource allocation in hospital management systems," *Int. J. Appl. Health Care Analytics*, vol. 7, no. 2, pp. 1–12, 2022.

[47] A. K. Teng and A. B. Wilcox, "A review of predictive analytics solutions for sepsis patients," *Appl. Clin. Informat.*, vol. 11, no. 3, pp. 387–398, May 2020.

[48] P. J. Kennel, H. Rosenblum, K. M. Axsom, S. Alishetti, M. Brener, E. Horn, A. J. Kirtane, E. Lin, J. M. Griffin, and M. S. Maurer, "Remote cardiac monitoring in patients with heart failure: A review," *JAMA Cardiology*, vol. 7, no. 5, pp. 556–564, 2022.

[49] D. Borthakur, "HDFS architecture guide," *Hadoop Apache Project*, vol. 53, nos. 1–13, p. 2, 2008.

[50] A. Darwish, G. I. Sayed, and A. E. Hassanien, "The impact of implantable sensors in biomedical technology on the future of healthcare systems," in *Intelligent Pervasive Computing Systems for Smarter Healthcare*. USA: Wiley, 2019, pp. 67–89.

[51] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, and K. Shpanskaya, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 590–597.

[52] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3462–3471.

[53] C. K. Reddy and C. C. Aggarwal, *Healthcare Data Analytics*, vol. 36. Boca Raton, FL, USA: CRC Press, 2015.

[54] N. Garg, *Apache Kafka*. Birmingham, U.K.: Packt, 2013.

[55] S. Hoffman, *Apache Flume: Distributed Log Collection for Hadoop*. Birmingham, U.K.: Packt, 2013.

[56] R. Bhartia, "Amazon kinesis and apache storm," Amazon Web Services, Inc., USA, Tech. Rep. 1, 2014.

[57] X. Zhang, L. Yao, S. Zhang, S. Kanhere, M. Sheng, and Y. Liu, "Internet of Things meets brain–computer interface: A unified deep learning framework for enabling human-thing cognitive interactivity," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2084–2092, Apr. 2019.

[58] Y. Wang, L. Kung, and T. A. Byrd, "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations," *Technological Forecasting Social Change*, vol. 126, pp. 3–13, Jan. 2018.

[59] D. A. Davis, N. V. Chawla, N. Blumm, N. Christakis, and A.-L. Barabási, "Predicting individual disease risk based on medical history," in *Proc. 17th ACM Conf. Inf. Knowl. Manage.*, 2008, pp. 769–778.

[60] S. He, L. G. Leanse, and Y. Feng, "Artificial intelligence and machine learning assisted drug delivery for effective treatment of infectious diseases," *Adv. Drug Del. Rev.*, vol. 178, Nov. 2021, Art. no. 113922.

[61] A. Burkov, *Machine Learning Engineering*. Montreal, QC, Canada: True Positive Incorporated, 2020.

[62] A. Kishor, C. Chakraborty, and W. Jeberson, "A novel fog computing approach for minimization of latency in healthcare using machine learning," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 1, no. 1, pp. 7–17, 2020.

[63] G. Fu, Y. Zhang, and G. Yu, "A fair comparison of message queuing systems," *IEEE Access*, vol. 9, pp. 421–432, 2021.

[64] A. Ashfaq, A. Sant'Anna, M. Lingman, and S. Nowaczyk, "Readmission prediction using deep learning on electronic health records," *J. Biomed. Informat.*, vol. 97, Sep. 2019, Art. no. 103256.

[65] F. S. Ahmad, L. Ali, Raza-Ul-Mustafa, H. A. Khattak, T. Hameed, I. Wajahat, S. Kadry, and S. A. C. Bukhari, "A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs)," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 3, pp. 3283–3293, Mar. 2021.

[66] L. Huang, A. L. Shea, H. Qian, A. Masurkar, H. Deng, and D. Liu, "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records," *J. Biomed. Informat.*, vol. 99, Nov. 2019, Art. no. 103291.

[67] V. Huddar, B. K. Desiraju, V. Rajan, S. Bhattacharya, S. Roy, and C. K. Reddy, "Predicting complications in critical care using heterogeneous clinical data," *IEEE Access*, vol. 4, pp. 7988–8001, 2016.

[68] A. Fabregat, M. Magret, J. A. Ferré, A. Vernet, N. Guasch, A. Rodríguez, J. Gómez, and M. Bodí, "A machine learning decision-making tool for extubation in intensive care unit patients," *Comput. Methods Programs Biomed.*, vol. 200, Mar. 2021, Art. no. 105869.

[69] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton, D. J. Wales, and R. Das, "Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach," *JMIR Med. Informat.*, vol. 4, no. 3, p. e28, Sep. 2016.

[70] S. M. Lauritsen, M. E. Kalør, E. L. Kongsgaard, K. M. Lauritsen, M. J. Jørgensen, J. Lange, and B. Thiesson, "Early detection of sepsis utilizing deep learning on electronic health record event sequences," *Artif. Intell. Med.*, vol. 104, Apr. 2020, Art. no. 101820.

[71] N. S. Artzi, S. Shilo, E. Hadar, H. Rossman, S. Barbash-Hazan, A. Ben-Haroush, R. D. Balicer, B. Feldman, A. Wiznitzer, and E. Segal, "Prediction of gestational diabetes based on nationwide electronic health records," *Nature Med.*, vol. 26, no. 1, pp. 71–76, Jan. 2020.

[72] S. Qiu, Z. Wang, H. Zhao, L. Liu, J. Li, Y. Jiang, and G. Fortino, "Body sensor network-based robust gait analysis: Toward clinical and at home use," *IEEE Sensors J.*, vol. 19, no. 19, pp. 8393–8401, Oct. 2019.

[73] X. Qian, H. Chen, Y. Cai, K.-C. Chu, W. Xu, and M.-C. Huang, "Transfer learning model knowledge across multi-sensors locations over body sensor network," *IEEE Sensors J.*, vol. 22, no. 11, pp. 10663–10670, Jun. 2022.

[74] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX symposium on operating systems design and implementation (OSDI)*, 2016, pp. 265–283.

[75] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 11–12.

[76] F. Chollet, *Deep Learning With Python*, vol. 361. New York, NY, USA: Manning, 2018.

[77] S. Naidu and J. Tigani, *Google BigQuery Analytics*. Hoboken, NJ, USA: Wiley, 2014.

[78] F. Firouzi, B. Farahani, M. Barzegari, and M. Daneshmand, "AI-driven data monetization: The other face of data in IoT-based smart and connected health," *IEEE Internet Things J.*, vol. 9, no. 8, pp. 5581–5599, Apr. 2022.

[79] A. Pulsar. (Mar. 2020). *Distributed Pub-Sub Messaging System*. [Online]. Available: https://pulsar.apache.org/

[80] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, and M. J. Franklin, "Apache spark: A unified engine for big data processing," *Commun. ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[81] J. Turnbull, *Monitoring With Prometheus*. New York, NY, USA: Turnbull Press, 2018.

[82] J. Ladley, *Data Governance: How To Design, Deploy, and Sustain an Effective Data Governance Program*. New York, NY, USA: Academic, 2019.

[83] F. K. Dankar and K. El Emam, "Practicing differential privacy in health care: A review," *Trans. Data Privacy*, vol. 6, no. 1, pp. 35–67, 2013.

[84] J. Xu, K. Xue, S. Li, H. Tian, J. Hong, P. Hong, and N. Yu, "Healthchain: A blockchain-based privacy preserving scheme for large-scale health data," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8770–8781, Oct. 2019.

[85] I. Hegedűs, G. Danner, and M. Jelasity, "Gossip learning as a decentralized alternative to federated learning," in *Proc. IFIP Int. Conf. Distrib. Appl. Interoperable Syst.* New York, NY, USA: Springer, 2019, pp. 74–90.

[86] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016, *arXiv:1610.05492*.

[87] L. O. Gostin, "National health information privacy: Regulations under the health insurance portability and accountability act," *JAMA*, vol. 285, no. 23, pp. 3015–3021, 2001.

[88] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[89] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *Proc. Int. MICCAI Brainlesion Workshop*. New York, NY, USA: Springer, 2018, pp. 92–104.

[90] Apple, "Private federated learning," in *Proc. NeurIPS Expo Talk Abstract*, 2019, p. 1. [Online]. Available: https://nips.cc/ExpoConferences/2019/schedule?talk_id=40

[91] W. de Brouwer, "The federated future is ready for shipping," doc.ai, USA, Tech. Rep. 1, 2019.

[92] FeatureCloud. (2019). *Featurecloud: Our Vision*. [Online]. Available: https://featurecloud.eu/about/

[93] S. Deorowicz and S. Grabowski, "Compression of DNA sequence reads in FASTQ format," *Bioinformatics*, vol. 27, no. 6, pp. 860–862, Mar. 2011.

[94] H. Zhu and Y. Jin, "Multi-objective evolutionary federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 4, pp. 1310–1322, Apr. 2020.

[95] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-IID data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, Aug. 2019.

[96] Y. Chen, X. Sun, and Y. Jin, "Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 4229–4238, Oct. 2020.

[97] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[98] S. Caldas, J. Konečny, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," 2018, *arXiv:1812.07210*.

[99] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, 2015, pp. 1310–1321.

[100] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.

[101] M. Kamp, L. Adilova, J. Sicking, F. Hüger, P. Schlicht, T. Wirtz, and S. Wrobel, "Efficient decentralized deep learning by dynamic model averaging," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. New York, NY, USA: Springer, 2018, pp. 393–409.

[102] T. D. Bui, C. V. Nguyen, S. Swaroop, and R. E. Turner, "Partitioned variational inference: A unified framework encompassing federated and continual learning," 2018, *arXiv:1811.11206*.

[103] Y. Zhou, G. Pu, X. Ma, X. Li, and D. Wu, "Distilled one-shot federated learning," 2020, *arXiv:2009.07999*.

[104] A. Guha Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "BrainTorrent: A peer-to-peer environment for decentralized federated learning," 2019, *arXiv:1905.06731*.

[105] J. Passerat-Palmbach, T. Farnan, R. Miller, M. Gross, H. Flannery, and B. Gleim, "A blockchain-orchestrated federated learning architecture for healthcare consortia," 2019, *arXiv:1910.12603*.

[106] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Blockchain and federated learning for privacy-preserved data sharing in industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 4177–4186, Jun. 2020.

[107] S. Lloyd, M. Mohseni, and P. Rebentrost, "Quantum principal component analysis," *Nature Phys.*, vol. 10, no. 9, pp. 631–633, 2014.

[108] X. Gao, Z.-Y. Zhang, and L.-M. Duan, "A quantum machine learning algorithm based on generative models," *Sci. Adv.*, vol. 4, no. 12, Dec. 2018, Art. no. eaat9004.

[109] V. Scarani, H. Bechmann-Pasquinucci, N. J. Cerf, M. Dušek, N. Lütkenhaus, and M. Peev, "The security of practical quantum key distribution," *Rev. Modern Phys.*, vol. 81, no. 3, p. 1301, 2009.

[110] S. Y.-C. Chen and S. Yoo, "Federated quantum machine learning," *Entropy*, vol. 23, no. 4, p. 460, Apr. 2021.

[111] A. Broadbent, J. Fitzsimons, and E. Kashefi, "Universal blind quantum computation," in *Proc. 50th Annu. IEEE Symp. Found. Comput. Sci.*, 2009, pp. 517–526.

[112] W. Li, S. Lu, and D.-L. Deng, "Quantum federated learning through blind quantum computing," 2021, *arXiv:2103.08403*.

[113] B. S. Glicksberg, K. W. Johnson, and J. T. Dudley, "The next generation of precision medicine: Observational studies, electronic health records, biobanks and continuous monitoring," *Human Mol. Genet.*, vol. 27, no. R1, pp. R56–R62, May 2018.

[114] *OHDSI*. Accessed: Mar. 14, 2021. [Online]. Available: http://ohdsi.org/

[115] S. S. Arslan, T. Goker, and R. Wideman, "A joint dedupe-fountain coded archival storage," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[116] G. J. Annas, "HIPAA regulations—A new era of medical-record privacy?" *New England J. Med.*, vol. 348, no. 15, pp. 1486–1490, Apr. 2003.

[117] Y. Zhu, "Smart remote personal health monitoring system: Addressing challenges of missing and conflicting data," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2022.

[118] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad, "Reviewing federated machine learning and its use in diseases prediction," *Sensors*, vol. 23, no. 4, p. 2112, Feb. 2023.

[119] Prayitno, C.-R. Shyu, K. T. Putra, H.-C. Chen, Y.-Y. Tsai, K. S. M. T. Hossain, W. Jiang, and Z.-Y. Shae, "A systematic review of federated learning in the healthcare area: From the perspective of data properties and applications," *Appl. Sci.*, vol. 11, no. 23, p. 11191, Nov. 2021.

[120] K. Farooq, H. J. Syed, S. O. Alqahtani, W. Nagmeldin, A. O. Ibrahim, and A. Gani, "Blockchain federated learning for in-home health monitoring," *Electronics*, vol. 12, no. 1, p. 136, Dec. 2022.

[121] W. Oh and G. N. Nadkarni, "Federated learning in health care using structured medical data," *Adv. Kidney Disease Health*, vol. 30, no. 1, pp. 4–16, Jan. 2023.

[122] L. Zhou, M. Wang, and N. Zhou, "Distributed federated learning-based deep learning model for privacy MRI brain tumor detection," 2024, *arXiv:2404.10026*.

[123] S. Pati, U. Baid, B. Edwards, M. Sheller, S.-H. Wang, G. A. Reina, P. Foley, A. Gruzdev, D. Karkada, and C. Davatzikos, "Federated learning enables big data for rare cancer boundary detection," *Nature Commun.*, vol. 13, no. 1, p. 7346, 2022.

[124] A. Chowdhury, H. Kassem, N. Padoy, R. Umeton, and A. Karargyris, "A review of medical federated learning: Applications in oncology and cancer research," in *Proc. Int. MICCAI Brainlesion Workshop*. New York, NY, USA: Springer, 2021, pp. 3–24.

[125] S. Chen, D. Xue, G. Chuai, Q. Yang, and Q. Liu, "FL-QSAR: A federated learning-based QSAR prototype for collaborative drug discovery," *Bioinformatics*, vol. 36, nos. 22–23, pp. 5492–5498, Apr. 2021.

[126] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," 2018, *arXiv:1806.00582*.

[127] H. Eichner, T. Koren, B. McMahan, N. Srebro, and K. Talwar, "Semi-cyclic stochastic gradient descent," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1764–1773.

[128] K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, D. Papadopoulos, and Q. Yang, "SecureBoost: A lossless federated learning framework," *IEEE Intell. Syst.*, vol. 36, no. 6, pp. 87–98, Nov. 2021.

[129] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou, "A hybrid approach to privacy-preserving federated learning," in *Proc. 12th ACM Workshop Artif. Intell. Secur.*, Nov. 2019, pp. 1–11.

[130] H. van Hasselt, "Estimating the maximum expected value: An analysis of (Nested) cross validation and the maximum sample average," 2013, *arXiv:1302.7175*.

[131] C. Nagpal, S. Yadlowsky, N. Rostamzadeh, and K. Heller, "Deep cox mixtures for survival regression," in *Proc. Mach. Learn. Healthcare Conf.*, 2021, pp. 674–708.

[132] S. A. Dijkland, K. A. Foks, S. Polinder, D. W. J. Dippel, A. I. R. Maas, H. F. Lingsma, and E. W. Steyerberg, "Prognosis in moderate and severe traumatic brain injury: A systematic review of contemporary models and validation studies," *J. Neurotrauma*, vol. 37, no. 1, pp. 1–13, Jan. 2020.

[133] D. Sun, J. Liu, L. Xiao, Y. Liu, Z. Wang, C. Li, Y. Jin, Q. Zhao, and S. Wen, "Recent development of risk-prediction models for incident hypertension: An updated systematic review," *PLoS ONE*, vol. 12, no. 10, Oct. 2017, Art. no. e0187240.

[134] C. Flatley, K. Gibbons, C. Hurst, V. Flenady, and S. Kumar, "Cross-validated prediction model for severe adverse neonatal outcomes in a term, non-anomalous, singleton cohort," *BMJ Paediatrics Open*, vol. 3, no. 1, Mar. 2019, Art. no. e000424.

[135] Z. Wang, Z. Yu, R. Fan, and B. Guo, "Correcting biases in online social media data based on target distributions in the physical world," *IEEE Access*, vol. 8, pp. 15256–15264, 2020.

[136] X. Wang, S. Wang, and W. Kindzierski, "Eliminating systematic bias from case-crossover designs," *Stat. Methods Med. Res.*, vol. 28, nos. 10–11, pp. 3100–3111, Nov. 2019.

[137] A. Vaid, S. K. Jaladanki, J. Xu, S. Teng, A. Kumar, S. Lee, S. Somani, I. Paranjpe, J. K. De Freitas, and T. Wanyan, "Federated learning of electronic health records improves mortality prediction in patients hospitalized with COVID-19," *MedRxiv*, vol. 1, no. 9, pp. 1–21, Aug. 2020.

[138] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen, and S. Bakas, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Jul. 2020.

[139] L. Qu, N. Balachandar, and D. L. Rubin, "An experimental study of data heterogeneity in federated learning methods for medical imaging," 2021, *arXiv:2107.08371*.

[140] H. Lee, Y. J. Chai, H. Joo, K. Lee, J. Y. Hwang, S.-M. Kim, K. Kim, I.-C. Nam, J. Y. Choi, H. W. Yu, M.-C. Lee, H. Masuoka, A. Miyauchi, K. E. Lee, S. Kim, and H.-J. Kong, "Federated learning for thyroid ultrasound image analysis to protect personal information: Validation study in a real health care environment," *JMIR Med. Informat.*, vol. 9, no. 5, May 2021, Art. no. e25869.

[141] D. Liu, T. Miller, R. Sayeed, and K. D. Mandl, "FADL: Federated-autonomous deep learning for distributed electronic health record," 2018, *arXiv:1811.11400*.

[142] L. Huang, Y. Yin, Z. Fu, S. Zhang, H. Deng, and D. Liu, "LoAdaBoost: loss-based AdaBoost federated machine learning with reduced computational complexity on IID and non-IID intensive care data," *PLoS ONE*, vol. 15, no. 4, Apr. 2020, Art. no. e0230706.

[143] S. Lu, Y. Zhang, Y. Wang, and C. Mack, "Learn electronic health records by fully decentralized federated learning," 2019, *arXiv:1912.01792*.

[144] X. Yao, T. Huang, C. Wu, R. Zhang, and L. Sun, "Towards faster and better federated learning: A feature fusion approach," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 175–179.

[145] J. Qiu, P. Li, M. Dong, X. Xin, and J. Tan, "Personalized prediction of live birth prior to the first in vitro fertilization treatment: A machine learning method," *J. Translational Med.*, vol. 17, no. 1, pp. 1–8, Dec. 2019.

[146] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1514–1529, Mar. 2018.

[147] S. S. Arslan, "Array BP-XOR codes for hierarchically distributed matrix multiplication," *IEEE Trans. Inf. Theory*, vol. 68, no. 3, pp. 2050–2066, Mar. 2022.

[148] R. Tandon, Q. Lei, A. Dimakis, and N. Karampatziakis, "Gradient coding: Avoiding stragglers in distributed learning," in *Proc. 34th Int. Conf. Mach. Learn.*, Aug. 2017, pp. 3368–3376.

[149] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[150] I. Katsamenis, E. Protopapadakis, A. Voulodimos, A. Doulamis, and N. Doulamis, "Transfer learning for COVID-19 pneumonia detection and classification in chest X-ray images," in *Proc. 24th Pan-Hellenic Conf. Informat.*, Nov. 2020, pp. 170–174.

[151] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[152] S. M. Karunarathne, N. Saxena, and M. K. Khan, "Security and privacy in IoT smart healthcare," *IEEE Internet Comput.*, vol. 25, no. 4, pp. 37–48, Jul. 2021.

[153] M. Ganzha, M. Paprzycki, W. Pawłowski, P. Szmeja, and K. Wasielewska, "Semantic interoperability in the Internet of Things: An overview from the INTER-IoT perspective," *J. Netw. Comput. Appl.*, vol. 81, pp. 111–124, Mar. 2017.

[154] L. Cao, "Beyond i.I.d.: Non-IID thinking, informatics, and learning," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 5–17, Jul. 2022.

[155] J. H. Yoo, H. M. Son, H. Jeong, E.-H. Jang, A. Y. Kim, H. Y. Yu, H. J. Jeon, and T.-M. Chung, "Personalized federated learning with clustering: Non-IID heart rate variability data application," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, 2021, pp. 1046–1051.

[156] J. Kim, G. Kim, and B. Han, "Multi-level branched regularization for federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 11058–11073.

ENGIN ZEYDAN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Department of Electrical and Electronics Engineering, Middle East Technical University, Ankara, Türkiye, in 2004 and 2006, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA, in February 2011. He was a Research and Development Engineer of Avea, a mobile operator in Türkiye, from 2011 to 2016. He was also a part-time Instructor with the Electrical and Electronics Engineering Department, Özyeğin University, from 2015 to 2018. He was a Senior Research and Development Engineer with Turk Telekom Laboratories, from 2016 to 2018. He was the Project Coordinator of the H2020 MonB5G European Project, from 2021 to 2023. He is currently with the Services and Networks Research Unit, Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), as a Senior Researcher. His research interests include telecommunications, data engineering, and network security.

SUAYB S. ARSLAN (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of California at San Diego, San Diego, CA, USA, in 2009 and 2012, respectively. He was a Research and Development Engineer with MERL, Cambridge, MA, USA, and a Senior Researcher with Quantum Corporation, Irvine, CA, USA, from 2012 to 2016. He is currently affiliated with the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Boston, MA, USA, and a Professor with the Department of Computer Engineering, Boğaziçi University, İstanbul, Türkiye. His research interests include information theory, neuroscience, digital communication, networking and storage, cloud and quantum computing, reliability/system theory, and image/video processing. He received numerous recognitions, including a Fulbright Grant and Outstanding Research Awards. He is an Associate Editor of IoT Journal (Elsevier).

MADHUSANKA LIYANAGE (Senior Member, IEEE) received the Doctor of Technology degree in communication engineering from the University of Oulu, Finland, in 2016. He is currently an Associate Professor/an Ad Astra Fellow and the Director of Graduate Research of the School of Computer Science, University College Dublin, Ireland. He is leading Network Softwarization and Security Labs (NetsLab), UCD School of Computer Science, which is a dynamic research group leading the charge in enhancing the security and privacy of next-generation mobile networks, including 5G and 6G (https://netslab.ucd.ie/). He is also an Adjunct Professor with the University of Oulu; the University of Ruhuna, Sri Lanka; and the University of Sri Jayewardenepura, Sri Lanka. His research interests include 5G/6G, blockchain, network security, artificial intelligence (AI), explainable AI, federated learning (FL), network slicing, the Internet of Things (IoT), and multi-access edge computing (MEC). He received the prestigious Marie Skıodowska-Curie Actions Individual Fellowship and the Government of Ireland Postdoctoral Fellowship, from 2018 to 2020. In 2020, he received the "2020 IEEE ComSoc Outstanding Young Researcher" award by IEEE ComSoc EMEA. In 2021, 2022, and 2023, he was ranked among the World's Top 2% Scientists (2020, 2021, and 2022) in the List prepared by Elsevier BV, Stanford University, USA. Also, he was awarded an Irish Research Council (IRC) Research Ally Prize as part of the IRC Researcher of the Year 2021 awards for his positive impact as a supervisor. In 2022, he received "the 2022 Tom Brazil Excellence in Research Award" from the SFI CONNECT Center. For more information, visit his website: www.madhusanka.com.