AI-Driven Data Monetization: The Other Face of Data in IoT-Based Smart and Connected Health

Farshad Firouzi, *Member, IEEE*, Bahar Farahani, *Member, IEEE*, Mojtaba Barzegari, and Mahmoud Daneshmand

***Abstract*—As the trajectory of the Internet of Things (IoT) moving at a rapid pace and with the rapid worldwide development and public embracement of wearable sensors, these days, most companies and organizations are awash in massive amounts of data. Determining how to profit from data deluge can give companies an edge in the market because data have the potential to add tremendous value to many aspects of a business. The market has already seen a level of monetization across vertical domains in the form of layering connected devices with a variety of Software-as-a-Service (SaaS) choices, such as subscription plans or smart device insights. Out of this arena is evolving a "machine economy" in which the ability to correctly monetize data rather than simply hoard it, will provide a significant advantage in a competitive digital environment. The recent advent of the technological advances in the fields of big data, analytics, and artificial intelligence (AI) has opened new avenues of competition, where data are utilized strategically and treated as a continuously changing asset able to unleash new revenue opportunities for monetization. Such growth has made room for an onslaught of new tools, architectures, business models, platforms, and marketplaces that enable organizations to successfully monetize data. In fact, emerging business models are striving to alter the power balance between users and companies that harvest information. Start-ups and organizations are offering to sell user data to data analytics companies and other businesses. Monetizing data goes beyond just selling data. It is also possible to include steps that add value to data. Generally, organizations can monetize data by: 1) utilizing it to make better business decisions or improve processes; 2) surrounding flagship services or products with data; or 3) selling information to current or new markets. This article will address all important aspects of IoT data monetization with more focus on the healthcare industry and discuss the corresponding challenges, such as data management, scalability, regulations, interoperability, security, and privacy. In addition, it presents a holistic reference architecture for the healthcare data economy with an in-depth case study on the detection and prediction of cardiac anomalies using multiparty computation (MPC) and privacy-preserving machine learning (PPML) techniques.**

***Index Terms*—Artificial intelligence (AI), big data, blockchain, eHealth, healthcare, Internet of Things (IoT).**

Manuscript received July 13, 2020; revised September 16, 2020; accepted September 18, 2020. Date of publication September 30, 2020; date of current version April 7, 2022. *(Corresponding authors: Farshad Firouzi; Bahar Farahani.)*

Farshad Firouzi is with the Electrical and Computer Engineering Department, Duke University, Durham, NC 27708 USA (e-mail: farshad.firouzi@duke.edu).

Bahar Farahani is with the Cyberspace Research Institute, Shahid Beheshti University, Tehran 1983969411, Iran (e-mail: b_farahani@sbu.ac.ir).

Mojtaba Barzegari is with the Biomechanics Section, Department of Mechanical Engineering, KU Leuven, 3001 Leuven, Belgium.

Mahmoud Daneshmand is with the School of Business, Stevens Institute of Technology, Hoboken, NJ 07030 USA.

Digital Object Identifier 10.1109/JIOT.2020.3027971

I. Introduction

THE INTERACTION between technology and healthcare has a long history. However, recent years have witnessed the rapid growth and adoption of the Internet of Things (IoT) and the advent of miniature wearable biosensors generating innovative opportunities in the area of customized, smart, and connected healthcare [1]. The integration of IoT indicates a radical paradigm shift that will bring greater availability, accessibility, personalization, precision, and lower-cost delivery of healthcare services. Previously, patients were found only in clinical environments, but today's technological progress ranging from IoT to blockchain and artificial intelligence (AI)—including machine learning (ML), deep learning (DL), and big data analytics—continues to remove the barriers found in conventional models of care. The emergence of data-driven products will allow patients to support their own health and wellness using the value created by their own data [e.g., creating a clearer picture of overall health than that provided via general electronic health records (EHRs)] [2]–[5]. However, for many, technology in the healthcare field still falls far behind those used in other areas of life because the information remains siloed, blocking patient data analysis and wringing insights from data difficult. In particular, utilizing and monetizing health data—from data sharing to data analysis—raises questions around data ownership, access, personal privacy, national security, and distribution of monetization benefits. Such barriers hinder the efficient exchange of information and decision making regarding patient care. Therefore, while health data are constantly being gathered, data monetization and its supporting strategy is still limited [1].

The world of healthcare is currently changing paradigms. Digital innovations are providing large amounts of data within healthcare systems. The explosive growth of available healthcare data comes from Wearable IoT (WIoT) devices, EHR, medical imaging, lab results, social media, as well as external patient data. Understanding that data are truly a business asset that creates the opportunity for today's digitally connected patients and healthcare systems to benefit from monetizing data. An IDC and Seagate study revealed that health data's expected compound annual growth rate (CAGR) is 136% through 2025, which is greater than that of media, manufacturing, or financial services [6]. As advanced technology provides more health data, monetizing that data is a benefit that healthcare organizations cannot ignore in our growing digital environment.

Generally, there are three main approaches to monetizing data.

1) *Sharing Data Assets:* Data assets can be shared (e.g., using blockchain-based marketplaces) as raw data. For example, health insurance companies can offer a predictive analytical model to analyze for chronic conditions as a monetized service. The model could also be offered as an application programming interface (API) available for purchase by other insurance companies, enabling smaller organizations to create their own analytics.
2) *Personalization:* Data-driven products can be provided as nonessential services. For example, insurance companies can charge a fee to provide wellness recommendations based on the data gathered from an individual's health devices. In another example, a U.S. startup company has created a blockchain marketplace to allows users to develop a clearer healthcare picture than that presented via EHRs. The blockchain marketplace enables patients to monetize health data and provides customized health recommendations to improve wellness.
3) *Leveraging Data Analytics:* Perhaps one of the most widespread approaches involves using data analytics to discover data insights that enable organizations to clarify customer preferences, identify common risks, or predict adverse events. Using advanced analytics on aggregated healthcare data can provide a comprehensive view of patient groups, and analytical results can lessen the use of inessential medical procedures while supporting better coordination of care [1], [7]–[11].

There are several barriers in the area of healthcare that block data monetization (e.g., blocking date sharing among organizations that restrict health providers in their ability to collaborate when making healthcare decisions). For instance, understandably, hospitals are hesitant when it comes to sharing patient data and regulations, such as the health insurance portability and accountability act (HIPAA), general data protection regulation (GDPR), and California Consumer Privacy Act (CCPA) seriously impede the ability for organizations to combine private data [12]–[15]. HIPAA was created as the first nationwide U.S. standard to guard personal health information. In 2018, the European Union's (EU) GDPR also put clear guidelines around data security and privacy into place, stressing that the gathering of a user's data must be clearly communicated. After the GDPR was enacted, the CCPA and the China Internet Security Law, along with others, brought further scrutiny to data security. Data from diverse sources require that analysts adhere to the requirements of several privacy laws, which further complicates medical research. The problem of analyzing medical data while protecting patient privacy is a complex and pressing issue [16], [17]. In addition, the data are constantly changing; therefore, combining data must be repeated often in order to be successful, which is substantially more difficult than pooling data only once. The heterogeneous nature and dimensionality of the data in the smart and connected health sector due to complicated data types, such as medical images or clinical notes, fractured data sources, and privacy concerns are other barriers to multiorganizational research.

Although the concept of data monetization—from collapsing data silos, combining, and sharing healthcare data, to collaborative data analytics—is still new, a variety of models and architectures are becoming available. This includes data marketplaces, data sharing solutions, as well as multiparty computation (MPC) and privacy-preserving ML (PPML) which enables us, e.g., to evaluate data from several hospitals and patients of the same kind creating new services or improve a model's function while safeguarding data privacy. This article aims at advancing the understanding of data monetization models, techniques, and opportunities, and discussing how technology helps up tackle the current barriers and challenges. This article also discusses the state-of-the-art architectures and then presents a holistic reference architecture for data monetization covering both data sharing and PPML/MPC using a novel case study on detection and prediction of cardiac anomalies.

The remainder of this article is organized as follows. In Section II, we discuss how emerging smart and connected health technologies from IoT to AI and big data analytics could benefit the healthcare systems across the globe and what challenges and barriers have to be tackled to grow further. Section III presents data monetization strategies and techniques. Section IV presents a general reference architecture for data monetization. Section VI with the help of a case study demonstrates the capabilities of data sharing as well as MPC/PPML techniques for data monetization. Finally, Section VII concludes this article.

II. Interaction Between Healthcare and Technology

A. Shift From Hospital-Centric to Patient-Centric

Healthcare systems are generally grouped into three categories: 1) large healthcare organizations (i.e., hospitals); 2) pharmacies and smaller clinics; and 3) nonclinical domains (i.e., rural areas lacking healthcare, patient homes, and communities). Areas all over the world are facing large-scale healthcare challenges as they struggle to handle a rapidly aging population, individuals living with chronic illness, increased child mortality, unsanitary living conditions, increasing pollution, lack of access to clean water, and epidemics of disease. Although the need for medical care has risen greatly over the last several years, the hospital-centric model of visiting a physician when ill remains the normative practice. Managing chronic illnesses means patients must physically visit a clinic or hospital so that a doctor can assess disease progression and make changes to the treatment plan based on the in-clinic examination. In general, hospitals function within a physician and disease-centered, reactive paradigm that does not include patients as active participants in their own medical care process. The central challenges facing hospital-centered healthcare include [2]–[4] the following.

1) *Adherence Monitoring:* Physicians are often not equipped to monitor compliance with treatments, such as medication, rehabilitative exercise, or dietary guidelines. Treatment noncompliance increases the chance a

patient will require hospitalization, driving up healthcare costs, and increasing the economic burden to the patient, family members, and society.

2) *Increasing Geriatric Population:* By 2050, the number of adults aged 60 years or older is expected to increase to more than 200% from 841 million (2013) to more than 2 billion worldwide. This rapid rise will require additional healthcare resources and facilities to manage the healthcare needs of a significantly expanded geriatric population.
3) *Urbanization:* By 2015, the World Health Organization (WHO) predicted that 70% of the world's population would live in urban landscapes, suggesting that big cities would need larger healthcare infrastructures to handle the expanding populations. In addition, urban areas are more prone to be the center point of epidemics of disease because contagious diseases spread quickly in more densely occupied areas.
4) *Healthcare Workforce Shortages:* The need for surgical staff, physicians, caregivers, medical laboratory staff, and nurses to fill healthcare system roles in rural and urban areas increases as the need for healthcare rises. One possible means of addressing this challenge would be to increase the utilization of telemedicine.
5) *Rising Medical Costs:* One of the greatest medical industry challenges currently faced is the ever-rising cost of healthcare. Diabetes care costs in the U.S. expanded by 21% from 2007 to reach around $245 billion in 2016.

Patient-centered care (PCC) based on smart and connected health technologies is a newer model that centers around the healthcare needs of patients. The term PCC originated with the Picker/Commonwealth Program created in 1988 by the Picker Institute. In a 2001 watershed report, the Institute of Medicine defined PCC as "Healthcare that establishes a partnership among practitioners, patients, and their families (when appropriate) to ensure that decisions respect patient's wants, needs, and preferences and that patients have the education and support they need to make decisions and participate in their own care." Within the PCC model, patients are required to actively participate in their own healthcare as they are empowered with greater knowledge and pertinent insights that equip them to better manage medical conditions. Patients are treated as part of the care team and are able to take healthcare outside the physician's office or hospital. However, it is important to note that PCC is not intended to eliminate clinics or hospitals, but to better leverage those organizations in a shared model of care that uses IoT [4].

B. Role of IoT in Healthcare

IoT provides a seamless platform to connect people, things (objects), data, and processes to one another for enriching and making our lives easier. This vision carries us from compute-based centralized schemes to a more distributed/decentralized environment offering a numerous amount of applications, such as smart wearables, smart home, and smart mobility, as well as smart and connected health. The growth of IoT and the increasing popularity of wearable biosensors has created new opportunities in the area of customized eHealth services and

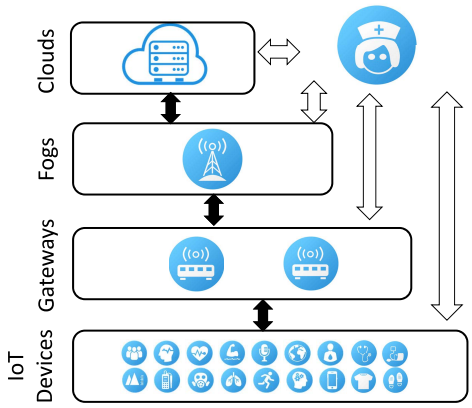


Fig. 1. Hierarchical architecture of IoT-based eHealth.

PCC. IoT serves as an umbrella term for any of the networks of actuators, sensors, computers, and processors that connect to the Internet. The evolution of IoT has created multiple benefits and new possibilities as indicated below [2]–[4].

1) *Holistic:* IoT and eHealth offers a comprehensive solution for diverse needs and can be utilized in a variety of areas, such as exercise, patient safety, beauty, and health monitoring.
2) *Customized Content or Services:* IoT in combination with big data analytics can expand the possibilities for meeting unique treatment needs, filling a vital role in each individual's health and wellbeing. For instance, ML and big data can be used to forecast health conditions, such as cancer growth, infections, or heart attack prior to the occurrence, allowing physicians and patients to act more quickly.
3) *Lifetime Monitoring:* Physicians and patients benefit from the collection of and access to all-encompassing (past, present, and future) health data.
4) *Lower Costs:* IoT eHealth integrates a variety of technologies, eliminating the need to pay for multiple technologies and better equipping patients to monitor their health. This means that patients only pay to consult a physician if and when there is a significant change in their health status.
5) *Increased Physician Involvement:* Physicians have access to the health data of patients in real time; therefore, fewer exams are required. This also enables physicians to monitor a larger number of patients if the healthcare organizations structure IT systems evolve to make use of real-time data via telemedicine.
6) *Accessibility and Availability:* Geographical barriers are eliminated as healthcare professionals, patients, and caregivers gain access to eHealth services or data at any given time.
7) *Online Assistance:* IoT eHealth enables us to have instant 24/7 online access to a wide variety of specialists—doctors, therapists, consultants, coaches, and many other experts—to address health issues as they pop up.
8) *Remote Healthcare:* The advent of telemedicine allows treating patients and connecting with professionals at

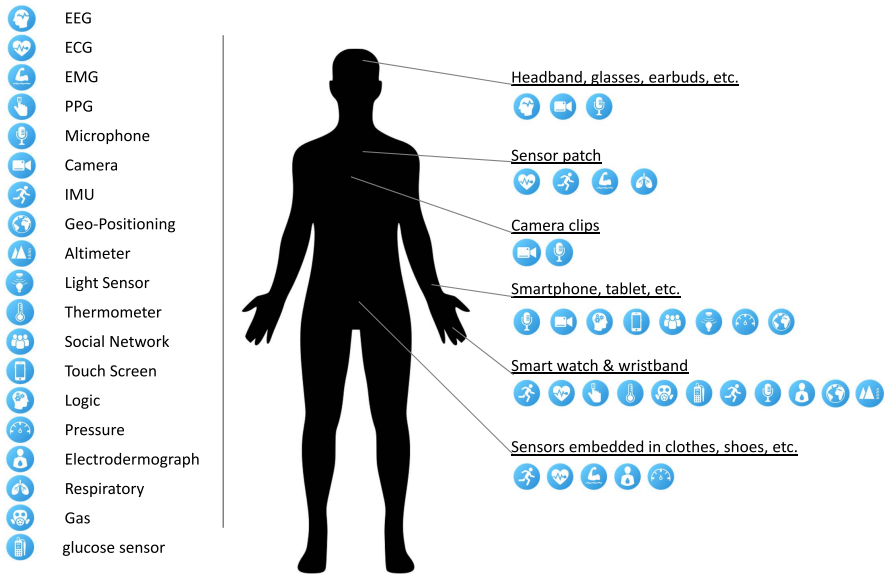


Fig. 2. Most common (wearable) IoT health devices.

a click of a button immediately no matter your location, ensuring that you stay healthy wherever you are without any geographical barriers. eHealth technologies and applications grant patients increased access to information, enabling them to better monitor their personal health. The utilization of self-aware tools and engagement strategies allows the sending and receipt of data by healthcare providers.

9) *Preventive Healthcare:* When continuous monitoring is possible, real-time data analysis can be done by care teams as needed, allowing for more immediate intervention should an adverse event arise.
10) *Improved Treatment Plans:* IoT enables the recognition of in-context behavioral or physiological patterns over long periods of time. This allows physicians and other healthcare professionals to more thoroughly understand the effects of treatment for individual patients. The creation of more checkpoints between a doctor or hospital visits supports a more appropriate review of patient response to engagement approaches, medication, or treatment plans.
11) *Life Coaching:* eHealth can provide feedback and suggest behavioral changes founded on personal data as well as population-wide data.
12) *Optimization:* Users are better able to comprehend how adjustable variables influence health outcomes.

As shown in Fig. 1, the state-of-the-art IoT-based healthcare solutions rely on a hierarchical architecture as it brings tremendous benefits and makes it possible to distribute the intelligence and computation to achieve an optimal solution while satisfying the given constraints (e.g., optimization for energy versus optimization for latency). A typical hierarchical IoT healthcare architecture consists of several processing layers, including: 1) smart and connected devices; 2) gateway and fog; and 3) cloud [2]–[4].

1) *Smart and Connected Devices:* Smart and connected health devices are becoming increasingly ubiquitous enabling individuals to monitor their health status in real time and sync the readouts with the cloud through a gateway (see Fig. 2).
2) *Cloud:* The integration of IoT and cloud technologies created another paradigm, cloud IoT, offering the following key benefits.
   a) *Management:* With the help of the cloud, IoT devices can be monitored and managed remotely from any location at any time.
   b) *Unlimited Computing and Storage:* Cloud provides unlimited computing and storage resources enabling us to address the massive amount of data coming from devices, i.e., big data challenges that are characterized by four Vs (high volume, high variety, high veracity, and high velocity).
   c) *Things as a Service (TaaS):* The integration of IoT and cloud creates a TaaS model in healthcare.
3) *Fog:* In the traditional cloud computing paradigm, the cloud used to perform almost whole computing. As the number of IoT devices is significantly increasing, the fog computing layer has been inserted between device and cloud layers to enable: a) real-time analytics and decision making to address latency-critical applications; b) traffic reduction on an overburdened network; c) transient data storage; d) local device management; and e) protocol translation.

C. Role of AI and Big Data in Healthcare

Health information technology (HIT), mobile health (mHealth), telemedicine, WIoT devices, and personalized medicine continue to provide sources of data for research and care. Using huge amounts of data extracted from patient exams, research, and clinical trials provides an unprecedented

database for data analytics and data-driven products/services. The incorporation of AI for effective manipulation of large, multiscale, multimodal, distributed, and heterogeneous data sets are just beginning to be adopted as they enable improved treatment quality. The main benefits of AI and Big Data analytics in healthcare can be summarized as follows [2]–[4].

1) *Early Detection:* When health problems occur, AI is able to detect them more quickly. For example, Microsoft is creating computers that are able to function at the molecular level to destroy cancer cells immediately upon detection. AI can also be utilized to process online search histories to recognize mental health problems.
2) *Diagnosis:* Physicians can be assisted by AI in the proper diagnosis of patients, enabling them to arrive at the correct diagnosis accurately based on 80% of health data currently not visible in healthcare systems because it is unstructured data.
3) *Decision Making:* AI tools and systems designed to support clinical decisions are able to assist physicians and patients with appropriately prioritizing needs and tasks.
4) *Treatment:* AI tools are already being used in all areas of healthcare. For example, Google DeepMind is lessening the amount of time needed to create radiotherapy treatment plans and IBM's Watson is able to provide treatment recommendations centered on global medical records.
5) *Palliative Care:* People are living longer lives than ever before and our aging societies require increased end-of-life care. Virtual assistants utilizing the AI technology and robots are predicted as possible future developments, and in fact, robots are being utilized in some instances to care for elderly individuals, e.g., in Japan.
6) *Research:* AI is able to assist in the discovery of new treatment protocols or drugs and is also useful in the exploration of actual diseases. In the future, patients may be able to vaccinate against or eradicate them. In fact, a Canadian startup called Meta is utilizing AI to rapidly analyze scientific articles to generate insights.
7) *Training:* Simulations generated by AI may assist surgeons or other healthcare professionals to sharpen their skills without additional risk to real-life patients. AI models are usually more practical and dependable. In addition, utilizing AI for training enables instructors to customize training toward the needs of diverse individuals.
8) *High-Risk Patient Care:* When patients utilize emergency care, increased costs and complications often arise. While costs may go up, patients are not necessarily benefiting from improved outcomes; therefore, putting effective changes in place in the emergency department would transform the functionality of hospitals. If all patient records are consolidated in a digital format, patient patterns can be recognized more readily. Regular monitoring of patients with higher risk issues and utilizing an effective, tailored treatment model is then possible. Lack of accurate or complete data makes the development of PCC programs more troublesome, clarifying the need for taking advantage of big data initiatives within the industry.
9) *Preventative Care:* Big data and AI enable large healthcare systems such as hospitals to move away from reactive care to focus on preventative care.
10) *Cost Reduction:* Big data analytical tools and AI can be used to gain insights, lower costs, and provide improved patient experiences.

Note that regardless of whether these tools are being used for training purposes or for improving actual patient outcomes, the data utilized must be complete and comprehensive. Health organizations and scientists must obtain complete ownership of the data processed because using limited or flawed data with ML or AI will generate inaccurate results. Unfortunately, the health data maintained by most current healthcare systems are fragmentary and incomplete. Many sources of health data maintain the information, such as financial, operational, or clinical data separated, creating a problem compounded by the special formats, key identifiers, and validation requirements of each data system. When dissimilar software systems and databases contain diverse data subsets and ownership, it becomes difficult to create a full picture of the patient as the correct analysis of disjointed information is difficult or even impossible.

III. Data Monetization Strategies

Finding ways to utilize data is a growing trend that companies are turning to as an ancillary revenue source, but it is not a completely new idea. Gartner has recently named the creation and use of data as "infonomics." In the era of IoT, as smart devices connect with the IoT and gather data, new markets comprised of data creators and data consumers/buyers are created. Driving revenue via IoT data is generally possible through either *direct data monetization* or *indirect data monetization* (see Fig. 3) [1], [19].

A. Direct Data Monetization

There are likely customers willing to pay for raw data. Although many means of selling data exist, data are primarily sold via data marketplaces. Direct monetization of data is usually broken into two categories as follows.

1) *Raw Data Sharing:* Direct data access and data sharing is allowed based on the payment of money or cryptocurrency. Raw data are usually sold in two types of marketplaces. Which one you choose depends on the company's strategy.
    a) *Centralized Marketplaces:* The platform is owned by a single party that works as a central location where diverse types of data are exchanged. Both data and metadata can be stored here.
    b) *Decentralized Marketplace:* The platform enables participants to trade data via peer-to-peer transactions. This marketplace stores only metadata that allow consumers to locate data owners.
2) *Sale of Data Analysis or Insights:* Analyzing raw data improves the information's quality. Some companies are unable to analyze data, which makes space to monetize through data quality improvement, benefiting those on both ends of a transaction. Marketplaces may offer

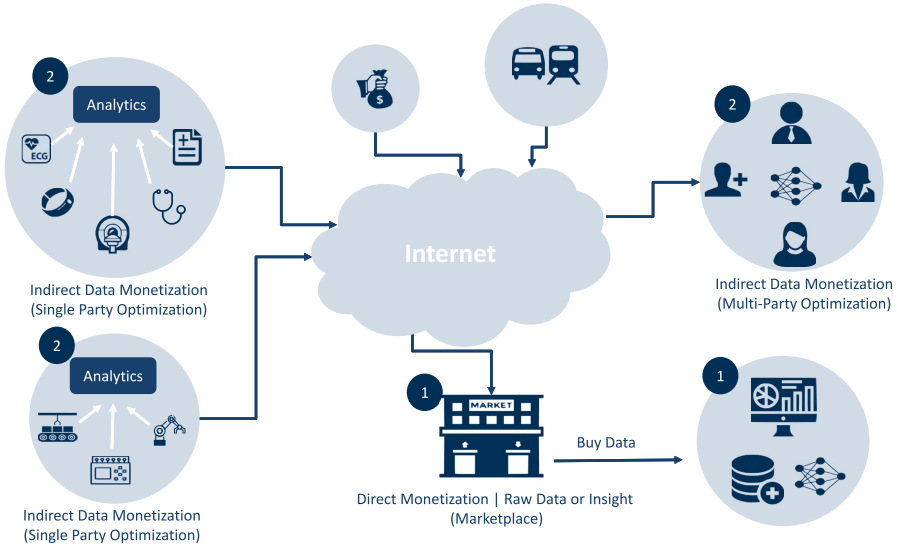


Fig. 3. Data monetization techniques.

data analysis, but it is also available through additional channels.

B. Indirect Data Monetization

Data are useful for improving business function, creating new business models or products/services. Making the most of data usually takes one of two paths.

1) *Data-Driven Optimization:* Data are used to reduce costs and increase business process efficiency by applying AI/ML and big data analytic. Note that AI/ML and analytic techniques can be conducted by only one party or it can be performed by several participants in a collaborative manner using privacy-preserving and MPC techniques. This kind of optimization applies to many industries. For example, manufacturing test periods can be shortened or data from consumers can be used to redesign a product.
2) *Data-Driven Business Models:* Product or process data are utilized to support business development or attract new customers through the creation of innovative products and services. This kind of business model allows data owners to discover completely new data-driven businesses rather than business-adjacent options. Such models are also vital to developing diverse revenue streams. For example, Bosch uses manufacturing information to develop customized subscription services that monitor hydraulic system conditions for customers.

C. Challenges Around Monetizing Data

The market for IoT data will continue to expand as organizations discover the power and revenue generated by monetizing data. The main challenges around monetizing data include the following.

1) *Guaranteeing Data Quality:* Enabling customers to trust data means making sure it is intact and of high quality. It is also important that data be accurate, recently obtained, and gathered using ethical means.
2) *Confirming Information Type:* Data providers must adapt to the needs of customers. Companies may need data in a variety of forms, including data that may not have originally seemed to fit their needs. Data customers might request extra information for data points that were not initially listed.
3) *Nontraditional Marketing and Product Management:* IoT data cannot be treated as a traditional product. Companies may need to reconsider the kinds of activities used to sell physical products, including product design, research, promotions, or marketing.
4) *Unlicensed Use Protections:* It is vital that data integrity be protected. It is not difficult to copy data, making it challenging to ensure that customers are using data as intended. Contracts should ensure that licensed users comprehend the ethical and correct handling of data, including methods of auditing data use.

D. Data Silos in Digital Health and the Challenges They Create for Monetization

While medical providers were traditionally urged to silo information, vital information went uncaptured, without analysis, and isolated. However, the incorporation of digital technology into healthcare is becoming standardized and adoption of new methods and technologies is on the rise. Rock Health, a San Francisco-based organization, recently completed a national consumer survey concerned with the adoption of digital health. Currently, 87% of those surveyed indicated that they utilize at least one digital health device, an increase of 7% since 2015. Hospitals and large healthcare organizations are continuing to strive to monetize data by incorporating previously siloed health data into valuable tools aimed at reducing healthcare costs, improving the patient experience, and supporting sound business decisions. This proves challenging because the data maintained by healthcare entities are often siloed according to purpose, including:

1) strategic planning data;

2) market development;
3) service management;
4) claims;
5) diverse clinical data including electronic medical records;
6) medical imaging:
    a) *DICOM Standard:* e.g., MRIs, X-rays, and CT scans;
    b) *Non-DICOM Standard:* Dermatology photos, surgical video, endoscopy, pathological studies, wound images (JPEGs, TIFFs, and MPEG-4s), etc.

E. Role of Data Sharing Across Data Silos

It can be very challenging for researchers to gain access to healthcare data because data are siloed in both public and private silos. In order to improve the outcomes of patient care, healthcare organizations must overcome the barrier of siloed information by developing integrated systems that collect and analyze vital real-time patient information. Achieving this goal will require enlarging the scope of care and providing clinicians with effective tools for managing care within update models.

According to Metcalfe's law, an essential principle for understanding network power, a network's value is proportionate to the squaring of the number of contributors. Therefore, adding more participants will increase a network's power exponentially [20]. A variant to Metcalfe's law that suggests the value of data siloes is very low, but when integrated data siloes provide accelerated returns rather than diminishing returns suggested by Metcalfe's network effect. There are many benefits to greater data sharing within healthcare systems including (but not limited to) [21], [22] the following.

1) *Patient Care:* Providing medical professionals with a holistic picture of the patient is highly valuable. In the bigger picture, access to the data of competing healthcare organizations can enable other organizations to pursue improvements that result in lower costs and improved patient care.
2) *Business Decision Making:* Sharing data can also support strategic business planning, such as the ability to recognize readmissions within data and develop strategies to reduce these rates and their accompanying Centers for Medicare and Medicaid Services (CMSs) penalties.
3) *Improved Analytics:* AI and ML are founded on data. Assuming access to accurate and complete data, their algorithms are able to provide greater insight into individual and population health by processing massive amounts of data by recognizing tumors, disease warning signs, or other medical problems at very early stages. If the data processed are limited to a single department, the effectiveness will be diminished.

New digital services or products are often born within business ecosystems in which organizations collaborate to better meet customer needs. These ecosystems are defined by the reality that no one organization is able to innovate entirely independently; therefore, each ecosystem member contributes to the benefit of all members. In an ideal system, the members function in a state of equilibrium that mutually benefits everyone. Data-driven ecosystems are those in which data are strategically utilized to develop valuable, innovative contributions. Sharing and mutually maintaining data are vital to the success of these ecosystems because the complete customer process can only be supported if all members work together and share data. However, conflicting goals can be the main pain point for data monetization as organizations desire to exchange data while also protecting data as its importance and value grow. This is where it is important to delineate between data sharing and data exchange.

1) *Data Sharing:* Occurs in horizontal as well as vertical collaborations among organizations in pursuit of a common goal, such as predictive maintenance in the manufacturing sector. Data sharing is also vital in creating updated business models by extracting more value from data through data marketplaces. In addition, data sharing suggests collaborating toward competition.
2) *Data Exchange:* Vertical coordination among companies in order to optimize supply or value chains (e.g., HL7 for medical settings).

F. Challenges in Sharing Healthcare Data

Breaking down healthcare data silos requires legislation that supports data sharing among privately owned and federal healthcare organizations. Regulations must also include security guidelines for the protection of personally identifiable information (PPI) as well as protected health information (PHI). Even with updated legislation, challenges remain in the areas of data quality, exchange of information, data pedigree, and data mapping. Some of the most common barriers include [23]–[29] the following.

1) *Privacy:* While data silos can be connected in order to support information sharing and analysis, the protection of patient information is vital. Sharing siloed data creates challenges around ensuring that organizations adhere to privacy regulations.
2) *Security:* Safeguarding patient confidentiality and identity is also important when sharing data. Increased utilization of virtual healthcare interactions opens the door to a greater risk of data breaches because data are transmitted electronically. If infrastructures are not secure, significant financial or legal consequences may arise around issues such as medical identity theft as virtual or telemedicine visits replace in-person clinical interactions.
3) *Competition:* Sharing data across healthcare organizations is complex because data can provide a competitive edge. Exchanging information among private and public healthcare organizations is slowly narrowing the competitive gap. Therefore, healthcare systems must find ways to improve the patient experience by sharing data while also seeking consumer strategies for keeping a competitive edge.
4) *Workflow:* It has often been indicated that technology is an obstacle to combining data silos. However,

Soon-Shiong, a South African surgeon, researcher, and UCLA professor says "The barriers [to improving healthcare with data] technologically do not exist any longer healthcare is falling behind other industries like banking and entertainment because it is not using the technology properly—and that is a workflow management problem."

5) *Culture:* Perhaps the greatest obstacle to integrating siloed data is the healthcare industry culture which assumes that holding on to data creates increased value. However, siloing healthcare data actually reduces the bottom line for the organization, provider, and patient.
6) *Low-Trust Relationships Between Organizations:* Trustworthy relationships among healthcare organizations or institutions are vital to supporting digital communication and data sharing when the data are accessible to multiple entities. While large-scale healthcare organizations (i.e., enterprise hospital systems) are often networked, smaller or private organizations may not have that advantage.
7) *Scalability:* Large amounts of data can be difficult to send electronically because of firewall settings or bandwidth limits (more common in rural areas). Scalability issues can affect the speed of data transmissions as well as total system response time.
8) *Enforcement of Interoperable Data Standards:* If interoperable data standards are not enforced (i.e., HL7's Fast Healthcare Interoperability Resources) the format and structure of health can vary so that it is difficult to interpret data or integrate the data into a variety of systems.

G. Data Sharing Architectures

The availability of accurate and high-quality data is the essential centerpiece of AI-based system development. The central idea is the creation of a complete, decentralized, data-driven ecosystem founded on population health data that is optimized for use in AI business models. To this end, current top-tier health data silos ensure quality and new interfaces are developed to support the creation of semiprofessional user databases. While adding value, small and medium-sized enterprises (SMEs) and citizen science research projects can utilize the services and interfaces provided by simplified AI analytical tools in order to efficiently create functional models for health data businesses. Additionally, a variety of stakeholder network activities support standardization and social acceptance in the areas of health and personal fitness data. Therefore, scalable and secure data sharing is vital in the provision of effective care for patients. It is likely that patients will visit multiple healthcare providers across their lifetime, and these clinicians must be empowered to exchange health data in a timely and secure manner, ensuring that each has the most current knowledge of the patient's health status. Generally, there are two methods available for the exchange of data.

H. Centralized Solutions

Centralized solutions (e.g., data marketplaces, online stores, shared repositories, or collaborative environments) are typically realized based on a cloud platform that makes it possible to store, sell, and share data via an access management system. The primary issue when it comes to collecting mass data from different data producers/providers by one organization (e.g., the marketplace owner) is that the data are kept on centralized servers that are very vulnerable to leaks and hacking, putting information at greater risk of security breaches. The risks around storing data on centralized servers were brought to the forefront in the 2017 Experian data leak that made the data of approximately 123 million American's. In addition, as long as centralization continues, there will be always a level of uncertainty with regard to data privacy because the owner of the platform has control over all collected data. When we place critical information in the hands of whoever owns the server, and when all data are going through that one point, data producer/providers have little control over how their data are used once those data are handed over to the platform owner. Unfortunately, these issues are an integral part and in the nature of the centralized solutions which makes it less attractive for main stakeholders. Vulnerability to exposure and lack of trust or competitiveness usually prevents stakeholders to store any privacy-sensitive data in centralized marketplaces as it reveals their strategic data or business secrets.

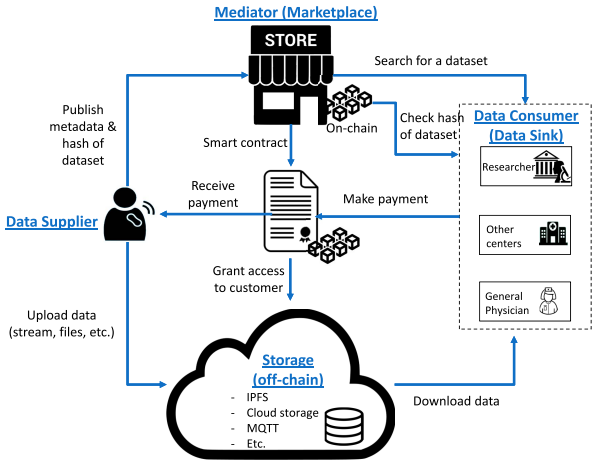


Fig. 4. Blockchain-based data monetization.

I. Decentralized Solutions

Decentralized solutions have been proposed to put the power back into the hands of users and data producers/providers by allowing them to have complete control over the information they generate and share in a trusted environment. In decentralized solutions, the data producer and data consumers can communicate, interact, share, and exchange their data in a peer-to-peer manner without requiring any central administration. This approach is also fault tolerant, meaning that there is no single point of failure in the system. Generally, decentralized marketplaces (e.g., Ocean Protocol marketplace, Enigma Data marketplace, IOTA marketplace, and Streamer marketplace) are implemented based on secure smart contract and distributed ledger technologies (DLTs) and blockchains that enable data producers to transact with consumers directly while maintaining full control over their data [30]–[32]. Fig. 4

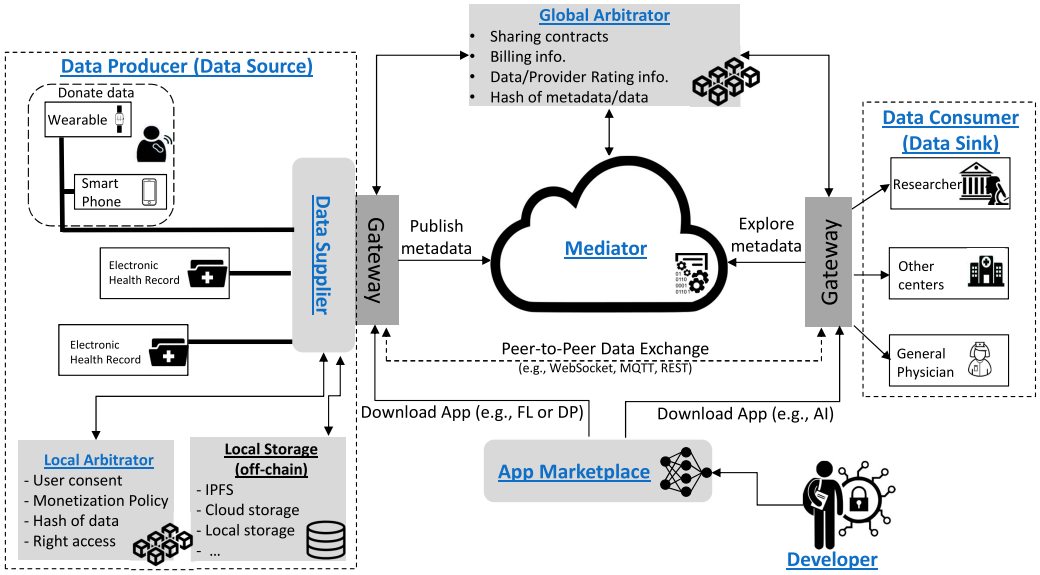


Fig. 5. Reference architecture for data monetization for smart and connected health.

illustrates the functional overview of a typical decentralized solution [33]–[39]. It should be noted that there are two general techniques for storing data in blockchain-based solutions.

1) *On-Chain Solutions:* In this approach, producers store the raw (encrypted) data directly on-chain in the blockchain. This approach is not usually scalable due to the transaction process capability and throughput of blockchain networks.
2) *Off-Chain Solutions:* In this approach, raw data are stored off-chain, while reference pointers to the raw data are kept on-chain. Note that raw data are not stored on-chain because they are more secure and scalable to maintain and exchange metadata that reference raw data (i.e., reference pointer to the data set). Providers can keep the data ownership and choose whether or not to share data at will by exchanging reference pointers. This method also keeps anyone who intercepts the pointers from achieving unauthorized access to the data.

IV. Reference Architecture

Fig. 5 demonstrates a general decentralized architecture that supports the secure sharing of sensitive data among core participants through a peer-to-peer network of data producers and consumers [10], [33]–[39]. As shown in this figure, they are as follows.

1) *Data Producer (Data Source):* Users (e.g., patients) that serve as the source of data, able to generate, collect, monitor, and control the flow of data while receiving incentives for producing data.
2) *Data Supplier:* While the data producer is the person or legal entity that is generating data or controlling it through usage policies, the data supplier (e.g., hospitals) engages in the following tasks: a) generates data source descriptions registered with the mediator (broker or dealer) used by other participants (data consumer) in the ecosystem to obtain data and b) chooses data available from internal systems and different data producers for availability in the ecosystem, processes and integrates data, transforms it into a target data model, and includes data usage terms and conditions on behalf of data producers.
3) *Data Consumer (Data Sink):* Legal organization with the right to use data owner's data in accordance with the usage guidelines while also: a) obtains data from contractors and b) chooses data from a variety of sources (i.e., data providers), processes data, and integrates information, changing it into a target data model.
4) *Mediator:* The mediator functions as a broker able to manage and store information about available data sources within the ecosystem. While this role is vital but not exclusive, many mediators can exist simultaneously. It mainly serves to provide a publication and discovery registry of data sources as well as all usage terms and legal information. Therefore, the mediator manages the data sources' metadata and contracts, which are generally considered public data that are available for discovery by any interested party.
5) *Identity Provider:* An entity that creates, maintains, and manages identity information for participants in the ecosystem.
6) *Arbitrator:* The arbitrator serves as a mediator responsible for settling and clearing financial transactions and the exchange of data. The arbitrator records all activities involved in the exchange of data. Once completed, the data supplier and data consumer are able to validate the transfer of data by recording transaction details with the arbitrator. That information is then used to bill the transaction. Logging information is also useful for resolving conflicts and disputes such as confirming whether or not a package of data was received by a data

consumer. Both data transmission and receipt must be confirmed by logging the information in the arbitrator transaction record. Other major tasks of the arbitrator include: a) *accounting:* billing details, invoices, and transactions can be stored in arbitrator's record; b) *dispute resolution:* information stored in the arbitrator is utilized to verify data transactions, quality of data, and the receipt of data by the data consumer; and c) *rating management:* the arbitrator can also be utilized to store and publish crowdsourced reviews about the quality of data sets as well as participants (data suppliers and data consumers). Note that data quality is a complex and multidimensional concept comprising accessibility (i.e., the quality of being available when needed), the quantity of data (i.e., amount of data), completeness (i.e., how comprehensive the data are), representation (data format), ease of manipulation, error free (i.e., the number of error or missing values), interoperability, current (up-to-date), value added (i.e., how data adds value to the business), and reputation [40].

7) *App Marketplace and App Developers:* App marketplace enables application developers to create data and AI services and then make them available to participants.
8) *Gateway:* Gateways are an integral component for participation in the ecosystem facilitating peer-to-peer communication for participants by allowing data suppliers and data producers to exchange or share data with data consumers. Gateways can be also connected to app marketplaces in order to purchase and download apps and source codes to extend their functionality. For instance, gateways can be utilized for data cleaning or running AI algorithms, such as MPC and PPML models.

A. Identity Provider

Generally, identity is based on the ten principles outlined as follows.

1) *Existence:* Users are never completely digital.
2) *Control:* Users can manage privacy and control their identities according to their own preferences.
3) *Access:* Individuals can access their own data without gatekeepers.
4) *Transparency:* All algorithms and systems require transparency and openness.
5) *Persistence:* Identities exist for as long as the user wants them to.
6) *Portability:* Identity services and information can be transported by the user.
7) *Interoperability:* User identities are usable across multiple platforms and borders.
8) *Consent:* Users are able to agree with the manner in which identity information is used.
9) *Minimization:* In interactions, the information provided about the identity owner must be minimal.
10) *Protection:* The rights of individuals are protected from abuse by those with more power.

Identity management is difficult within the current Internet environment because personal information is required in order to interact with online services. Our personal information is then stored in data silos that are often the target of hacking. In addition, online services often gather more information that is actually necessary for interaction. Additional problems include the following.

1) *Unwanted Correlation:* Information about identity is associated across many systems without consent. The growth of Internet correlation is mainly driven by marketing and has resulted in a loss of privacy by the majority of Internet users. Correlation occurs because of the main identifiers, such as email addresses, which we use each day. While email is the biggest factor, using the same account name on multiple sites also increases correlation. When additional identity information, such as an address, phone number, identification cards, etc., are utilized online, that data are associated across sites as well. Cookies utilized by websites empower the association of identities across multiple platforms. While, for example, the GDPR regulations are meant to stop this from occurring in Europe, those protections do not exist everywhere. Because it is a legal solution rather than a technical one, it is open to those with nefarious intentions.
2) *Centralized Identifiers:* Most identifiers utilized today are based on a centralized entity, such as the government (i.e., driver's license and tax ID number) or a company (i.e., website login ID). The main issue with this method is that the centralized entity can take away the identifier at will. This is an issue of particular concern for a critic of the entity regardless of whether the entity is a public or private institution. Another critical concern is the misuse of identity information in the event that the controlling entity is compromised.
3) *Hierarchical Architecture:* Currently, managing certificates requires a hierarchical architecture where a ladder of certificate authorities must independently verify the lower certificates in the hierarchy until the root certificates embedded within devices is reached.
4) *Data Breaches:* Identity information is highly valuable, so data silos are clear targets for hackers. The coveted information includes user IDs, passwords, names, email addresses, etc., because this information enables hackers to gain unauthorized access to accounts. The high availability of identity data makes it impossible to use it for high-value interactions because of the risk that someone other than the owner is utilizing the data.

Current centralized identity management solutions generally require public-key cryptography [i.e., public-key infrastructure (PKI)], but PKI is centralized and expensive. Service will fail if a certificate authority (CA) makes an error on a digital certificate. To tackle this issue, recently, distributed identity management (DIM) solutions, such as Hyperledger Indy have proposed as a method for utilizing cryptography to build trust without requiring a PKI, resulting in a more available, decentralized system with fewer control points with the ability to become points of failure. In short, DIM solutions are built on a user-centric philosophy when it comes to identity (see Fig. 6). Credentials are, generally, items, such as passports,

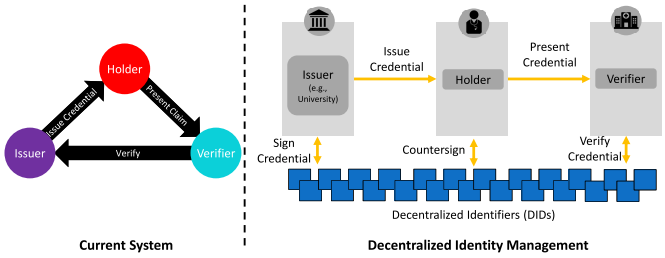


Fig. 6. Identity provider.

driver's licenses, or educational degrees conferred by an issuing entity. Virtual credentials (VCs) are the digital counterparts of real-life credentials that are cryptographically processed so that when a person proves the claims (data on the credentials), the receiver can be sure:

1) who is responsible for issuing a claim;
2) claims were issued to the presenting identity;
3) claims have not been forged or tampered with.

The basic requirements of DIM include the following.

1) *Governance:* All stakeholders should trust the solution.
2) *Performance (Scalability):* There are issues with current blockchain systems, such as Bitcoin due to its low throughput. Ethereum is also very slow in support of decentralized identifiers (DIDs). Hyperledger Invy is a very promising solution, but still it is evolving and cannot yet support the required performance for a worldwide system.
3) *Accessibility:* Data should not be controlled by a centralized entity, and identity information should be available to all stakeholders.
4) *Privacy:* Personal data should not be stored in a centralized database or on-chain distributed ledger.

The foundational concept behind DIM solutions is the use of a DID and a self-sovereign identity (SSI). Simply, SSI refers to the idea that individuals that a single user can regain ownership of her own data. DID requirements guarantee that both the issuer and verifiers can confirm necessary public keys via public blockchain whether they are part of an identity federation or the same organization. DID has evolved from previously disconnected "islands of identity" that required individual PKIs to a worldwide identity network based on distributed PKI (DPKI), much like evolving from a "local-area network" to the "global Internet." A DID consists of two main components: 1) a unique identifier and 2) an associated DID Document. Unique identifiers are produced by the owner, without the oversight of a centralized entity. DID are used for looking up a DID document. DID documents are typically expressed using JSON-LD and contain at least one public key created by the owner, a list of ways that DID can be used to authenticate, and a list of services that DID can be used and a list of "verifiable claims." Note that each DID service is also associated with one or more addresses (i.e., endpoints that are represented by a URL) where messages are delivered for a particular service. Examples of services include discovery services or verifiable claim repository services. DIDs can be resolved like a URL by returning data (i.e., endpoints and public keys) associated with that particular DID. A verifiable claim typically consists of three pieces: 1) subject of the claim, such as a person, organization, animals, and things; 2) claim issuer, e.g., an organization such as a university, a bank, etc.; 3) claim, i.e., a statement—such as the entity's background or quality, such as name, ID, address, university degree—that one subject makes about itself or another one. Note that the claim is tamperproof and verifiable because it should be cryptographically signed by the issuer. There are also two important roles supported by verifiable claims: 1) *claim holders:* an entity that receives and holds the claim and 2) *verifier:* an entity, e.g., a service provider (such as platforms or websites) that verifies a claim about a subject [41]–[43].

In DIM solutions, such as Hyperledger Indy, DID is recorded in a decentralized area such as a blockchain network. A person, device, or organization (i.e., identity) is able to create and publish a DID to an immutable public ledger. For the sake of privacy, each entity may have many DIDs, useful in a variety of contexts. This lessens the likelihood of correlation because each party only holds partial information about the user. DID can be accessed via the public ledger and the information attached to the DID (i.e., DID document) such as endpoints, public keys, as well as associated services and claims, to interact with the entity are returned. DIDs are utilized to create connections between two identities, such as a platform and a user, to facilitate secure communication. It is expected that an entity will have multiple DIDs (one for every relationship with any other entity). You can think of a DID as a user ID and password pair supported by cryptography made up of public/private key pairs. In addition, both ends of a relationship create a DID for the other in order to support secure communication. Because users create their own DIDs and provide a different DID to each service, the identity cannot be correlated across multiple services. Note that PPI is not saved in the ledger; therefore, it is not easily accessed by others because PPIs are sent using peer connections. These connections are secure due to authenticated encryption. Any information shared is known only to two entities so no PPI is exposed to the public. Information located in the public ledger provides the trust of the proof. In addition, neither credentials nor private data are not located in the ledger. The only data located in the ledger are that which you explicitly want to provide as all private data are traded directly between entities and held in secure wallets. Distributed solutions (e.g., Indy) also include selective disclosure and zero-knowledge proofs (ZKPs). This means that claims can be selectively disclosed with only limited data elements provided as proof. Also, ZKPs allow one to prove information without providing supportive data. For example, proof that someone is older than an indicated age based on the date of birth on a VC driver's license without providing additional information like a name.

B. Arbitrator

As mentioned earlier, one of the main important tasks of the Arbitrator is to store and manage user consents. Note that according to recent regulations, such as GDPR, monetizing any personal data without user consent is illegal. In general, developing a modern and adaptable consent management system

requires the consideration of the six different areas as outlined as follows.

1) *Disclosure:* Individuals (e.g., patients) must be informed regarding the potential benefits and/or harm that could arise from data monetization using language that is simple, clear, and not overly technical. It is also vital that the main values, interests, and needs of the user are considered and that any incorrect beliefs be addressed. In particular, the following points must be explicating stated:
   a) What specific information will be gathered?
   b) Who will be able to access the collected data?
   c) How long will the collected data be stored?
   d) For what purpose will the data be used?
   e) How will the user's identity be protected?
2) *Comprehension:* The user (e.g., patient) fully understands what data are being made known. For example, an eHealth company could utilize an AI-powered system that makes recommendations to alleviate mental stress and anxiety based on the behavior and historical data of other human subjects with similar health profiles. After reviewing the disclosure, the user should be able to answer the questions below regarding the collection and use of data.
   a) Will information regarding the user's previous five stressful situations and the corresponding health data and vital signs (e.g., heart rate) be incorporated into the recommendation system?
   b) Will the other user of the system be able to figure out what another customer has previously experienced?
   c) Will the user's past stress and health information be included in the recommendation system two years from now?
3) *Voluntariness:* Each individual can choose whether or not to take considered action. The user must not be unduly influenced, manipulated, or coerced by altering the user's perception of the actual choices available. An example of manipulation would be leading a user to believe that a particular choice must be made (i.e., giving a company data) in order to complete an action, even though that choice is not actually necessary.
4) *Competence:* Providing valid consent requires that the user be physically, mentally, and emotionally capable to do so. A user's competency must be verified by the party seeking consent.
5) *Agreement:* The user must make a clear decision to decline or accept taking a certain action. The manner of acceptance or decline must be clearly visible and easily accessed. If the agreement is ongoing, the user must be able to remove consent at any point without providing a reason for withdrawal.
6) *Minimal Distraction:* Users should not be distracted by unnecessary information while giving consent. This may seem counterintuitive because the concept of the disclosure involves providing all information to the user. However, it is important to only provide the information needed for comprehension, voluntariness, competence, and agreement without inundating the user with unnecessary information.

In today's digital world, people are becoming increasingly more aware of the threats posed by data breaches as well as the misuse of personal information by commercial entities. For example, GDPR is working to coordinate data protection regulations all across the EU and hopes to empower users to retain control of their personal data. It also seeks to address the movement of data outside the EU or European economic area (EEA). The GDPR is also working to create regulations that ensure less complicated access to a person's own data, rights to consent, erasure, and rectification, as well as the right to data portability and to be made aware of any data breach with the potential to impact a user's rights. The GDPR identifies three roles as described as follows [13].

1) *Data Subjects:* Any individual whose personal data are being gathered, stored, or processed.
2) *Data Controller:* The entity, authority, agency, or body that is either independently or in collaboration with others, deciding how data will be processed and for what purposes. Because only data controllers may obtain personal data from subjects, they are also responsible for determining the legality of collecting the data. Data controllers should establish the legal precedent for gathering data utilizing at least one of the six bases for collecting data as indicated in the GDPR. The data collector must also provide a transparent privacy policy that clearly indicates the following.
   a) What data will be gathered.
   b) How the data will be stored.
   c) How the information will be used.
   d) Who the data will be shared with.
   e) If the data will be shared with any third parties.
   f) When and how the data will be deleted.
3) *Processor:* A legal or natural person, agency, public authority, or other body that processes personal data for the data controller. If a data processor is involved in any part of data collection, the processor becomes subject to all of the data controller responsibilities listed above. When a data processor and data controller work together, a contract must be utilized to clarify roles. Such a contract must include the following information in accordance with GDPR regulations.
   a) A complete processing plan that outlines the subject, nature, purpose, and total processing timeline.
   b) Data controller obligations and rights.
   c) Kinds of data to be collected.
   d) Grouping of data subjects.
   e) Agreement to follow instructions.
   f) Issues of confidentiality.
   g) Subprocessor terms of hiring.
   h) Deleting or return of data.

The above-mentioned roles can be best explained using an example. Assume that a healthcare application (e.g., stress management) gathers personal information of customers (e.g., the pattern of walking) and then transfers the collected data to an AI consulting company to train the corresponding AI models. In this scenario, if the healthcare company provides both

the collected data and the instructions for processing data to the AI consulting company, the data controller is the healthcare company, whereas the AI consulting company is only the data processor. On the other hand, if the healthcare company is responsible for providing only data and the AI consulting company must determine the processing means, then both parties are data controllers and the AI consulting company also functions as the data processor.

One promising means of assuaging the growing distrust of data collection is utilizing decentralized, immutable ledger such as blockchain for consent management. When the consent information is stored in a blockchain ledger, it minimizes the risk of fraud or the misuse of information. The central idea is to establish a legal contract using a template that includes terms and conditions, natural language, and markup tags. When the template has been created, a contract draft can be electronically signed by all contracted parties (e.g., data producer, data supplier, and data consumer). The general flow for a typical blockchain-based consent management system can be broken down into the following steps.

1) When a data subject (e.g., data producer) consents to the processing of PPI and the data are collected and stored by the controller, a digital copy of the consent is created and registered on the blockchain.
2) The PII is then stored off-chain in the controller's database. PII is not generally stored on-chain because the data cannot be deleted or modified later, which is not in alignment with the data subject's right to modification or the right to erasure the collected data.
3) When the data controller (e.g., data supplier) sends the PII to an external processor (e.g., data consumer), a new transaction occurs on the blockchain (e.g., in the arbitrator). The transaction contains data and processing details such as data categories, who and for what reason the data are being transferred. The transaction also indicates the conditions and timeframe of processing. In addition, the processor registers processing steps in the blockchain.
4) The data subject is given blockchain access so that the subject can view consent transaction history as well as the processor and controller's activities as registered in the network. For example, a data subject could view the list of processors (including contact data) handling the subject's data in accordance with specific viewable conditions. If the subject believes that the processor is undertaking activities that are outside the boundaries of consent or wishes to withdraw consent, the subject can request restricted processing or remove consent. Requesting the deletion of data, access to data, or correction of data should also be guaranteed in the ecosystem.

C. Gateway

Gateway is a logical building block that facilitates the peer-to-peer connection among participants. A gateway can be a server, a virtual machine, or a service that provides a common interface and protocol (e.g., REST, MQTT, and WebSocket) enabling a participant to communicate and exchange data on demand with another one in the ecosystem based on a predetermined data model and semantic. Aggregating diverse data from several data suppliers can reveal unforeseen insights, improve ML models, and drive improved decision making while also generating new or improved services and products. The functionality of gateways can be extended and enriched by deploying and installing applications, which are offered by app developers or data scientists in the marketplace. Apps can play a valuable role in assisting in the processing of data, such as data anonymization, data cleansing, data transformation, and ML applications. Gateways can be also significantly enhanced by deploying MPC and PPML applications.

Training AI algorithms, including those utilized by DL, data mining, data processing, ML, business intelligence, and big data analytics, requires that models receive a massive amount of data in order to generate valuable insights. In addition, ML generally requires the use of raw data, which means addressing privacy concerns. Applying conventional ML techniques is possible, but data privacy is lost. In order to address this issue, the idea of PPML has evolved. PPML is a subcategory of ML that strives to train models while protecting data privacy. Preserving privacy is possible by permitting several input parties to work together to train an ML model without granting other parties access to private data. In addition, MPC enables organizations to analyze private data owned by other institutions without allowing access to inputs. MPC does not allow groups to obtain information about another group's inputs, except what is obtainable from the publicly available output. In other words, MPC and PPML allow multiple parties to work together to analyze data as if they shared a database without obtaining any other parties' sensitive data.

Note that the presented decentralized peer-to-peer data sharing model can potentially balance monetization options with compliance to data security and privacy regulations to some extent. However, still, there are several use cases in the healthcare domain in which utilizing traditional ML methods (e.g., combining data from multiple hospitals to create a single model) can be very challenging. Data owners (e.g., hospitals) are often aware that analyzing data in collaboration with data from others offers great monetary and technical potential. Unfortunately, for some specific use cases, it is almost impossible to allow access to patient data (even in a peer-to-peer data sharing system) without undertaking a long process to de-anonymize information. Even after completing the process, data security/privacy can be breached by advanced inference attacks. Therefore, hospitals may rely mainly on their own data to train models. The focus of MPC/PPML and their integration with the gateway is the creation of a framework that enables research organizations or hospitals to analyze data without sacrificing privacy. Consider the following example. Three hospitals spread across three different geographic areas collect data regarding readmission within 30 days of discharge. While each hospital is gathering similar data, they are serving different populations. Within the current framework, each of the three hospitals is only able to train a readmission model with their patient population's data. Utilization of MPC/PPML framework inside the gateways (see Fig. 7) enables each of

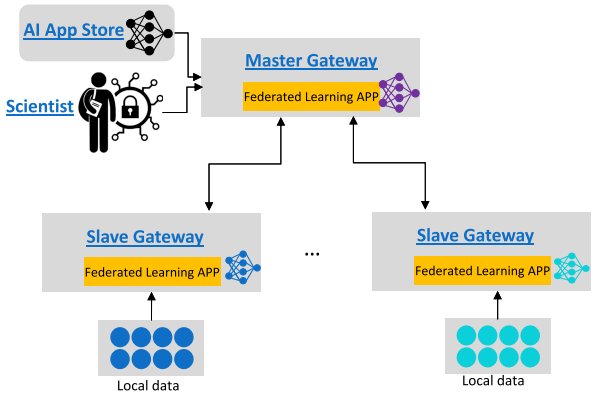


Fig. 7. Deployment of one PPML app (FL app) inside the gateways.

the three hospitals to work together to train a common model, potentially solving a problem worth $41 billion each year.

V. Privacy-Preserving Machine Learning and Multiparty Computation Techniques

Over the past few years, some privacy-enhancing techniques have been proposed to bridge the gap between ML and privacy (see Fig. 8).

A. Anonymization

The simplest and common approach to address the privacy issue is anonymization. Anonymization is defined as the process of removing PPI (e.g., name) before allowing public access to data. There are some organizations that strive to protect individual privacy by allowing access only to anonymized data sets. For instance, a few years ago, Netflix released a superior movie recommendation data set with anonymous film ratings to help contestants compete for a million-dollar prize. Regardless of anonymization, contestants used the data set in conjunction with IMDB information to pinpoint Netflix records for known users and extrapolating the political preferences of users, suggesting that anonymization is not a reliable means of protecting privacy when faced with substantial opponents.

B. Secret Sharing

In 1979, Shamir introduced the concept of secret sharing. In this technique, one party shares pieces of a secret (privacy-sensitive data) with others in such a way that no single party is able to obtain the secret individually. The information from at least "t" (a preset level) of "n" parties must be combined to obtain the secret. If fewer than t parties try to obtain the secret, they fail. When utilizing secret sharing, inputs are broken into pieces and issued to multiple parties. Computations are completed at the local level, and no party has access to complete inputs. When the results are combined, a complete and correct output is disclosed to all parties [44]. Consider the following example. A secret ($x = 32$) is shared between two different servers. If random numbers are chosen ($x1$ and $x2$ so that $x = x1 + x2$ or $x1 = 40$ and $x2 = -8$) and $x1$ is provided to one server while $x2$ is provided to the second server, neither server knows what the secret "x" is. One party knows the number "40" but the other number could be anything. In a similar fashion, if another number ($y = 43$) is shared with two servers (i.e., $y1 = 40$ and $y2 = 3$), the servers are able to determine $x+y$ and generate a public result without the servers learning x or y because the first server calculates $x1 + y1$ and then transmits the result to the second server, which calculates $(x1 + y1) + (x2 + y2) = x + y$. This is accomplished without either server learning about x or y, except for the end sum. The secret sharing approach suffers from several shortcomings and, as a result, its application in different industries is limited. Importantly, the computation of secret sharing takes significantly more time than if it would be performed on a plaintext data set because secret sharing requires each party to go through several steps and multiple back-and-forth between participants. Moreover, secret sharing usually cannot support categorical data. Finally, although addition and multiplication are usually supported out of the box, the implementation of complex operations (e.g., nonlinear functions) is not straightforward. Usually, those functions should be approximated, which reduces the accuracy of ML.

C. Homomorphic Encryption

Homomorphic encryption (HE) enables the computation of ciphertexts and creates an encrypted result. When the result is decrypted, it is identical to the computation result, just as if the computation had been completed using plaintext. The term homomorphic references Algebra's homomorphism. Encrypting and decrypting of algebraic functions are similar to the homeomorphisms of ciphertext and plaintext. These computations are made up of the Boolean or arithmetic circuits and include several types of encryption able to complete different computations using encrypted data. The most common HE techniques include "partially homomorphic," "somewhat homomorphic," "leveled fully homomorphic," or "fully homomorphic" [45]–[48]. The main issue of HE is the lack of scalability. Indeed, computations performed on the encrypted data are significantly slower than performing the same operations on the plaintext data. In addition, many HE techniques only allow for one type of operation (e.g., addition or multiplication, but not both). This technique is also incompatible with complex operations (e.g., nonlinear functions). Generally, those functions should be approximated by addition or multiplication. Therefore, with the current technology, training holistic ML models using HE techniques is almost infeasible.

D. Differential Privacy

The main idea behind differential privacy (DP) is adding random noise to raw, plaintext data. For example, adding several beats to the number of times your heart nominally beats per minute or adding five years to your age while answering a survey. In MPC, the noise is averaged out, and thus the result of analytics can be close to the original one. In general, applying more noise results in better privacy, but potentially it can lead to less accurate results. In other words, in DP, there is a tradeoff between precision and privacy, which might make

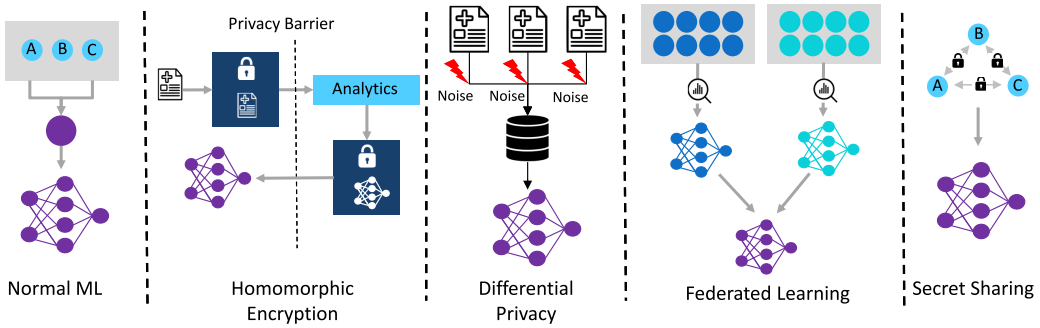


Fig. 8. PPML and MPC techniques.

it inappropriate for some specific use cases, such as pattern matching (e.g., optical character recognition). DP can be classified into two categories, namely, global privacy and local privacy. In the first approach, one trusted entity, known as a curator (e.g., hospital), collects the raw, privacy-sensitive data from lots of different parties or individuals (e.g., patients). The curator analyzes and adds noise to the collected data before sharing them with internal/external data scientists, researchers, and interested parties to simultaneously address the utilization of AI-driven IoT and the demands for data protection. In the local privacy method, there is no curator. Thereby, each and every party is responsible for adding an appropriate amount of noise to its own sensitive data before sharing it with other parties. Usually, the noise injection logic in both global and local techniques is implemented based on a probability distribution, such as the Laplacian distribution [49]–[51].

E. Federated Learning

Also known as collaborative learning, federated learning (FL) is a generalized framework for "bringing code to the data, instead of data to the code." This technique works based on a central server and multiple client servers. In this technique, private data are not transmitted from its origin. Instead, the central server trains a model using proxy data made available in advance. The model is then provided to each client for use in local model training. Each separate client uses its local private data to train the initial model. Each client receives the initial model's parameters and once the model has been trained with local data, the client returns the model's calculated coefficients or gradients to the central server. Then, the updated coefficients/gradients are averaged in the server to create a global model and then the server returns the refined global model to each client contributing to the model training. These steps are then repeated until the appropriate accuracy level is obtained (see Fig. 7). Note that FL might suffer from network overhead. For instance, if the size of ML models is huge, there are lots of coefficients/gradients that should be exchanged between the central server and the client's servers. This implies that the communication costs could be increased [52]–[55]. In general, there are two types of FL as follows.

1) *Horizontal FL:* Also known as "sample-based" FL, is utilized in situations where data sets share feature space but have different sample space. For example, if two regional hospitals serve very different users in their areas, the convergence of each hospital's users is minimal. However, because their business is alike, the feature spaces do not differ. In other words, in horizontal FL, all participants have access to the whole feature set and labels needed to train a local model [56].
2) *Vertical FL:* Also known as "feature-based" FL, is useful in cases where two data sets have the same sample space but different features spaces. Consider the following example. Two companies exist in the same city. One company is a healthcare provider (e.g., it offers mental stress management) and the other company is focused on social networking. Users of both are likely to be located in the same geographical area, therefore, the convergence of user space is big. However, because the healthcare company captures the user behavior and vital signs and the other company captures browsing history, the feature spaces are not similar. Vertical FL allows parties to collect data from different feature spaces while a single party can view the label. This means that neither party can train a model independently and creates another layer of complexity for aggregation. Sample spaces have to be aligned so that their training process can exchange only partial model updates. For example, the healthcare company and the social networking company may both desire to create an ML model to predict the social interaction behavior of users under stress. Vertical FL enables both companies to establish a feature vector utilized to train a model that reduces the stress level and improves the user experience [56].

VI. Case Study: Collaborative Cardiac Anomaly Detection

A. Problem Statement and Overall Flow

Electrocardiography (ECG) signals are widely used by healthcare professionals as a reliable way to monitor and evaluate the health of the cardiovascular system. ML techniques play an important role in automating this process to recognize any arrhythmia by accurate classification of ECG signals based on the AAMI EC57 standard [57], [58]. This case study aims at demonstrating how healthcare providers can collapse data (in this study, ECG signals) silos and monetize their data by training a cutting-edge convolutional neural network (CNN)

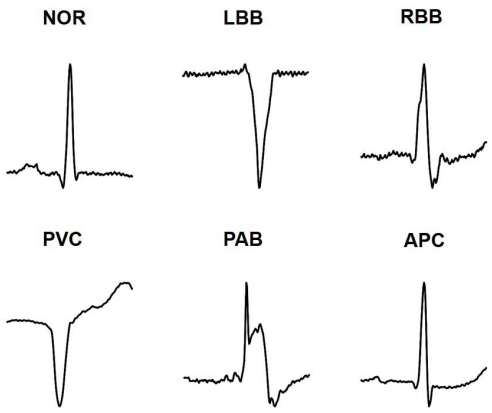


Fig. 9. Examples of different ECG waveforms.

collaboratively to detect anomalies in ECG. In particular, we demonstrate how different data owners can create a consortium to train a CNN based on DP and FL for the classification of ECG signals. To evaluate the performance of these techniques, we compare the results (e.g., the accuracy of the trained ML models) of DP/FL against the traditional method in which we assume that all ECG signals from different owners (e.g., patients or hospitals) can be combined together in a single database without any privacy issue.

ECG signals can be captured from 12 leads consisting of three bipolar limb leads (I, II, and III), three unipolar limb leads (AVR, AVL, and AVF), and six unipolar chest leads, also called precordial or V leads ($V1$, $V2$, $V3$, $V4$, $V5$, and $V6$) attached to multiple regions of the body to monitor the electrical activity of the heart. The ECG signal contains different waves (P, Q, R, S, and T), and as the bipolar limb lead II captures these waves in an acceptable resolution, it is commonly used for the data entry in ECG signal processing studies [57]. Generally, ECE signals are classified into 17 different classes, out of which six classes are very important for anomaly detection: 1) NOR (normal beat); 2) LBB (left bundle branch block beat); 3) RBB (right bundle branch block beat); 4) PVC (premature ventricular contraction beat); 5) PAB (paced beat); and 6) APC (atrial premature contraction beat). Fig. 9 demonstrates the waveform of these six classes in separate graphs.

B. CNN Architecture

In this study, we rely on CNN to detect arrhythmia. As shown in Fig. 10, our proposed CNN consists of five 2-D convolution layers, three max-pooling layers, and three fully connected layers. Note that the proposed CNN is very lightweight that enables us to deploy it either at the edge or in the cloud.

C. Data Preprocessing

1) Data Set: In this study, we use the MIT-BIH data set to train and test our ML models [59]. MIT-BIH data set contains ECG records of 48 different patients. Each record has a length of approximately 30 min, sampled at 360 Hz with 11-b resolution.

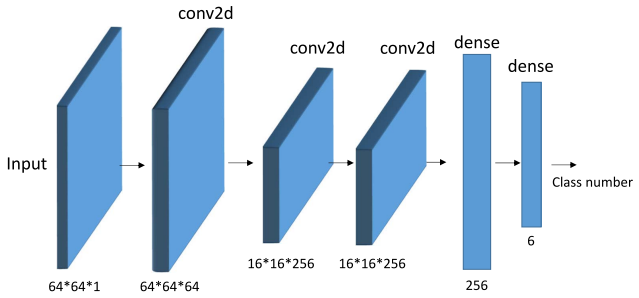


Fig. 10. Architecture of the proposed CNN model (max pooling and dropout layers are not shown).

TABLE I
ACCURACY AND TRAINING TIME OF FL, DP MODELS, AND NORMAL ML

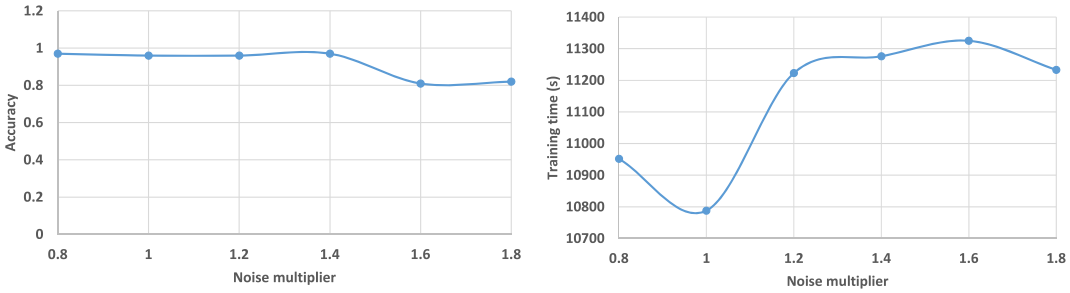
	Total elapsed time (s)	Accuracy
Normal ML	10618	0.97
DP (noise=0.8)	10952	0.97
DP (noise=1.0)	10788	0.96
DP (noise=1.6)	11325	0.81
FL	155263	0.84

2) Traditional ML and Differential Privacy: For evaluating the accuracy of the proposed model, the classifier was trained on the MIT-DIH data set [59], which contains approximately 110 000 beats of 48 patients. Each beat was properly windowed into a raw signal image with the size of (64, 64). As the database is imbalanced and includes more normal beats than the other types, the non-normal beat signal images were augmented using the cropping technique. In this technique, each image was cropped in nine different ways, producing ten images for each beat in total. As a result, for 45 studied patients, 353 690 samples were prepared as the input to the CNN model. In addition to augmentation, with a probability of 0.3, random rotation and/or flipping operations were applied to the input images in memory, and all the samples were shuffled prior to the training phase. Finally, the data were split into train, validation, and test data sets with a ratio of 0.7, 0.15, and 0.15, respectively.

3) Federated Learning: To test and evaluate the FL environment, we also used the MIT-DIH database. The database contains the data of 48 patients, but due to the computational costs of performing the FL workflow, ten patients were selected randomly to be investigated. The classification was performed on the six classes out of 17 available ones in the data set: 1) NOR (normal beat); 2) LBB (left bundle branch block beat); 3) RBB (right bundle branch block beat); 4) PVC (premature ventricular contraction beat); 5) PAB (paced beat); and 6) APC (atrial premature contraction beat). Raw signal images with the size of (64, 64) were generated from the windowed beats and appropriate augmentation techniques (cropping nine different regions of the images) were performed to reduce the lack of balance in the data set. As a result, 83 101 samples were prepared and split into separate train and test data sets for each patient with a ratio of 0.7 and 0.3, respectively. Additionally, a random rotation and flipping were applied to the images to reduce the overfitting during the training process. After this, the data were repeated to the number

TABLE II
PERFORMANCE OF FL AND DP MODELS AGAINST NORMAL ML

	Normal ML			DP (noise=0.8)			FL		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Class 0	0.98	0.98	0.98	0.99	0.99	0.99	0.85	0.71	0.77
Class 1	0.98	0.99	0.99	0.97	9.99	0.98	0.22	0.4	0.29
Class 2	0.96	0.98	0.97	9.95	0.99	0.97	0	0	0
Class 3	0.97	0.95	0.96	0.98	0.95	0.96	0.77	0.29	0.42
Class 4	0.99	0.99	0.99	1	0.98	0.99	0	0	0
Class 5	0.96	0.84	0.9	0.96	0.84	0.89	0.09	0.61	0.16



(a) (b)

Fig. 11. Impact of noise on accuracy and training time in DP. a) Impact of noise on accuracy. b) Impact of noise on training time.

of epochs to match the input specifications of the Tensorflow FL framework. Data of each patient were also shuffled in this step.

D. Performance Evaluation

Table I compares the accuracy as well as the training time of the normal ML (i.e., in which we combine all patient's data in a single DB without considering the privacy issue) against DP and FL. The results show that the training time of normal ML (10 618 s) and DP (11 325 s) are almost in the same range. On the other hand, the training time of FL is significantly higher (almost 14X) than the other two techniques, mainly due to the fact that FL relies on an iterative process. In terms of accuracy, normal ML shows the best performance followed by DP and FL. Note that the performance of DP highly depends on the amount of noise, which is injected into ECG data. For example, when the amount of noise increases from 0.8 to 1.6, the accuracy drops from 0.97 to 0.81. Fig. 11 also shows the same trend and results. The results indicate that the training time of the DP technique slightly increases when more noises are injected into the data. Table II illustrates the corresponding precision, recall, and $F1$-score. As shown in the table, the performance of DP (when noise $= 0.8$) is almost comparable to normal ML, but as mentioned earlier, it decreases with more noises. In addition, the results show that DP does not perform very well for class 5. This is mainly rooted in the imbalanced data set as there is less data for class 5 in the data set. Finally, we understand that FL does not perform very well. Indeed, the precision and the recall of FL for class 2 and class 4 are 0, meaning that FL cannot classify them.

VII. CONCLUSION

Healthcare is one of the world's largest industries as well as one of the most complex ones as patients require increasingly improved care. The field continues to evolve and progress rapidly as specialists look for innovative technologies able to provide effective solutions. As such, IoT, AI, big data, and blockchain continue to leave their mark on the industry. Within the healthcare industry, massive amounts of data are generated by healthcare records, diagnostic testing, and WIoT devices. Therefore, collecting, merging, and examining data from a variety of sources is gaining importance in smart and connected health. However, utilizing data is a complex process as it is often owned by several parties that do not wish to share with third parties, inadvertently sharing proprietary business information or losing control over the usage of data. It is clear that monetizing data offers a diverse array of opportunities, but companies need to stay aware of the legal issues surrounding such opportunities, particularly, in the healthcare industry, which is strictly regulated. To address these issues, this article discussed all important aspects of health data monetization from business models to challenges and solutions. We also presented a holistic reference architecture enabling us to balance monetization options with compliance to data security and privacy regulations. Finally, we discussed the fundamentals of PPML and MLC, including FL and DP, using a novel case study (i.e., ECG-based collaborative cardiac anomaly detection) as a promising solution for data monetization in the healthcare industry.

REFERENCES

[1] F. Firouzi, K. Chakrabarty, and S. Nassif, *Intelligent Internet of Things: From Device to Fog and Cloud*. Heidelberg, Germany: Springer, 2020.
[2] B. Farahani, F. Firouzi, and K. Chakrabarty, "Healthcare IoT," in *Intelligent Internet of Things*. Heidelberg, Germany: Springer, 2020, pp. 515–545.
[3] F. Firouzi, B. Farahani, M. Ibrahim, and K. Chakrabarty, "Keynote paper: From EDA to IoT eHealth: Promises, challenges, and solutions," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 12, pp. 2965–2978, May 2018.

[4] B. Farahani, F. Firouzi, V. Chang, M. Badaroglu, N. Constant, and K. Mankodiya, "Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare," *Future Gener. Comput. Syst.*, vol. 78, pp. 659–676, Jan. 2018.

[5] F. Firouzi *et al.*, "Internet-of-Things and big data for smarter healthcare: From device to architecture, applications and analytics," *Future Gener. Comput. Syst.*, vol. 78, pp. 583–586, Jan. 2018.

[6] *Business Wire*. Accessed: May 15, 2020. [Online]. Available: https://www.physionet.org/physiobank/database/mitdb/

[7] F. Firouzi, B. Farahani, M. Weinberger, G. DePace, and F. S. Aliee, "IoT fundamentals: Definitions, architectures, challenges, and promises," in *Intelligent Internet of Things*. New York, NY, USA: Springer, 2020, pp. 3–50.

[8] *Framing a Winning Data Monetization Strategy*. Accessed: May 15, 2020. [Online]. Available: https://home.kpmg/content/dam/kpmg/pdf/2015/10/framing-a-winning-data.pdf

[9] *Data Monetization: New Value Streams You Need Right Now*. Accessed: May 15, 2020. [Online]. Available: https://www.forbes.com/sites/douglaslaney/2020/06/09/data-monetization-new-value-streams-you-need-right-now/

[10] *International Data Spaces Association*. Accessed: May 15, 2020. [Online]. Available: https://www.internationaldataspaces.org/

[11] *New Revenue Streams in Health Data Monetization*. Accessed: May 15, 2020. [Online]. Available: https://www.tcs.com/content/dam/tcs/pdf/Industries/life-sciences-and-healthcare/insights/monetizing-healthcare-data-for-revenue-generation.pdf

[12] *Health Insurance Portability and Accountability Act*. Accessed: May 15, 2020. [Online]. Available: https://www.hhs.gov/hipaa/index.html

[13] *General Data Protection Regulation*. Accessed: May 15, 2020. [Online]. Available: https://gdpr-info.eu/

[14] *California Consumer Privacy Act*. Accessed: May 15, 2020. [Online]. Available: https://oag.ca.gov/privacy/ccpa

[15] *Federated Learning for Medical AI*. Accessed: May 15, 2020. [Online]. Available: https://medium.com/deeptek/federated-learning-for-medical-ai-65d0daeafe8b

[16] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, "Big healthcare data: Preserving security and privacy," *J. Big Data*, vol. 5, no. 1, p. 1, 2018.

[17] A. D. Dwivedi, G. Srivastava, S. Dhar, and R. Singh, "A decentralized privacy-preserving healthcare blockchain for IoT," *Sensors*, vol. 19, no. 2, p. 326, 2019.

[18] M. Mozafari, F. Firouzi, and B. Farahani, "Towards IoT-enabled multimodal mental stress monitoring," in *Proc. IEEE Int. Conf. Omni Layer Intell. Syst. (COINS)*, 2012, pp. 1–8.

[19] *A Guide to Data Monetization*. Accessed: May 15, 2020. [Online]. Available: https://blog.bosch-si.com/business-models/a-guide-to-data-monetization/

[20] J. Hendler and J. Golbeck, "Metcalfe's law, Web 2.0, and the semantic Web," *J. Web Semant.*, vol. 6, no. 1, pp. 14–20, 2008.

[21] P. Fasano, *Transforming Health Care: The Financial Impact of Technology, Electronic Tools and Data Mining*. New York, NY, USA: Wiley, 2013.

[22] L. L. Siu *et al.*, "Facilitating a culture of responsible and effective sharing of cancer genome data," *Nat. Med.*, vol. 22, no. 5, pp. 464–471, 2016.

[23] S. Bhartiya and D. Mehrotra, "Challenges and recommendations to healthcare data exchange in an interoperable environment," *Electron. J. Health Informat.*, vol. 8, no. 2, p. 16, 2014.

[24] M. A. Cyran, "Blockchain as a foundation for sharing healthcare data," *Blockchain Healthcare Today*, vol. 11, p. 1, Mar. 2018.

[25] B. Fabian, T. Ermakova, and P. Junghanns, "Collaborative and secure sharing of healthcare data in multi-clouds," *Inf. Syst.*, vol. 48, pp. 132–150, Mar. 2015.

[26] T. Schultz, "Turning healthcare challenges into big data opportunities: A use-case review across the pharmaceutical development lifecycle," *Bull. Amer. Soc. Inf. Sci. Technol.*, vol. 39, no. 5, pp. 34–40, 2013.

[27] D. Tse, C.-K. Chow, T.-P. Ly, C.-Y. Tong, and K.-W. Tam, "The challenges of big data governance in healthcare," in *Proc. 17th IEEE Int. Conf. Trust Security Privacy Comput. Commun. 12th IEEE Int. Conf. Big Data Sci. Eng. (TrustCom/BigDataSE)*, 2018, pp. 1632–1636.

[28] T. McGhin, K.-K. R. Choo, C. Z. Liu, and D. He, "Blockchain in healthcare applications: Research challenges and opportunities," *J. Netw. Comput. Appl.*, vol. 135, pp. 62–75, Jun. 2019.

[29] *Data Silos: Together, We Can Bust Them*. Accessed: May 15, 2020. [Online]. Available: https://www.initiatesolutions.com/single-post/2017/11/01/Data-Silos-Together-We-Can-Bust-Them

[30] *Ocean Protocol*. Accessed: May 15, 2020. [Online]. Available: https://oceanprotocol.com/

[31] *Enigma Data Marketplace*. Accessed: May 15, 2020. [Online]. Available: https://www.enigma.co/marketplace/

[32] *Streamr Marketplace*. Accessed: May 15, 2020. [Online]. Available: https://streamr.network/

[33] K. Fan, S. Wang, Y. Ren, H. Li, and Y. Yang, "MedBlock: Efficient and secure medical data sharing via blockchain," *J. Med. Syst.*, vol. 42, no. 8, p. 136, 2018.

[34] Q. Xia, E. B. Sifah, K. O. Asamoah, J. Gao, X. Du, and M. Guizani, "MedShare: Trust-less medical data sharing among cloud service providers via blockchain," *IEEE Access*, vol. 5, pp. 14757–14767, 2017.

[35] X. Liang, J. Zhao, S. Shetty, J. Liu, and D. Li, "Integrating blockchain for data sharing and collaboration in mobile healthcare applications," in *Proc. IEEE 28th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, 2017, pp. 1–5.

[36] Q. Xia, E. B. Sifah, A. Smahi, S. Amofa, and X. Zhang, "BBDS: Blockchain-based data sharing for electronic medical records in cloud environments," *Information*, vol. 8, no. 2, p. 44, 2017.

[37] S. Wang, Y. Zhang, and Y. Zhang, "A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems," *IEEE Access*, vol. 6, pp. 38437–38450, 2018.

[38] H. Shafagh, L. Burkhalter, A. Hithnawi, and S. Duquennoy, "Towards blockchain-based auditable storage and sharing of iot data," in *Proc. ACM Cloud Comput. Security Workshop*, 2017, pp. 45–50.

[39] T. Sultana, A. Almogren, M. Akbar, M. Zuair, I. Ullah, and N. Javaid, "Data sharing system integrating access control mechanism using blockchain-based smart contracts for iot devices," *Appl. Sci.*, vol. 10, no. 2, p. 488, 2020.

[40] L. L. Pipino, R. Y. Wang, J. D. Funk, and Y. W. Lee, *Journey to Data Quality*. Cambridge, MA, USA: MIT Press, 2016.

[41] *Hyperledger Indy*. [Online]. Available: https://www.hyperledger.org/use/hyperledger-indy

[42] A. Mühle, A. Grüner, T. Gayvoronskaya, and C. Meinel, "A survey on essential components of a self-sovereign identity," *Comput. Sci. Rev.*, vol. 30, pp. 80–86, Jul. 2018.

[43] G. Fedrecheski, J. M. Rabaey, L. C. Costa, P. C. C. Ccori, W. T. Pereira, and M. K. Zuffo, "Self-sovereign identity for IoT environments: A perspective," 2020. [Online]. Available: arXiv:2003.05106.

[44] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, 1979.

[45] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A survey on homomorphic encryption schemes: Theory and implementation," *ACM Comput. Surveys*, vol. 51, no. 4, pp. 1–35, 2018.

[46] I. Damgard, M. Geisler, and M. Kroigard, "Homomorphic encryption and secure comparison," *Int. J. Appl. Cryptography*, vol. 1, no. 1, pp. 22–31, 2008.

[47] Y. Aono *et al.*, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 5, pp. 1333–1345, Dec. 2017.

[48] P. Li *et al.*, "Multi-key privacy-preserving deep learning in cloud computing," *Future Gener. Comput. Syst.*, vol. 74, pp. 76–85, Sep. 2017.

[49] Z. Ji, Z. C. Lipton, and C. Elkan, "Differential privacy and machine learning: A survey and review," 2014. [Online]. Available: arXiv:1412.7584.

[50] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.*, 2008, pp. 1–19.

[51] M. Al-Rubaie and J. M. Chang, "Privacy-preserving machine learning: Threats and solutions," *IEEE Security Privacy*, vol. 17, no. 2, pp. 49–58, Apr. 2019.

[52] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, 2019.

[53] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," 2019. [Online]. Available: arXiv:1902.01046.

[54] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, Aug. 2020.

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

[56] A. Nagar, "Privacy-preserving blockchain based federated learning with differential data sharing," 2019. [Online]. Available: arXiv:1912.04859.

[57] B. Farahani, M. Barzegari, F. S. Aliee, and K. A. Shaik, "Towards collaborative intelligent IoT ehealth: From device to fog, and cloud," *Microprocess. Microsyst.*, vol. 72, May 2020, Art. no. 102938.

[58] B. Farahani, M. Barzegari, and F. S. Aliee, "Towards collaborative machine learning driven healthcare Internet of Things," in *Proc. Int. Conf. Omni-Layer Intell. Syst.*, 2019, pp. 134–140.
[59] *MIT-BIH*. [Online]. Available: https://www.businesswire.com/news/home/20181126005585/en/Seagate-Launches-New-Data-Readiness-Index-Revealing-Impact

Farshad Firouzi (Member, IEEE) is an Adjunct Assistant Professor with the Electrical and Computer Engineering Department, Duke University, Durham, NC, USA. He is a top-producing expert and a Technical Leader with over ten years experience offering strong performance in all aspects of AI/ML, smart data, computer architecture, VLSI, and IoT including R&D, consulting services, strategic planning, and technology solutions, across vertical industries, e.g., semiconductor, automotive, finance, manufacturing, logistics, and eHealth. He has authored over 45 conference/journal papers.

Dr. Firouzi served as a Guest/Associate Editor of several well-known journals, such as IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION, IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, *Future Generation Computer Systems* (Elsevier), and *Microprocessors and Microsystems* (Elsevier), *Journal of Network and Computer Applications* (Elsevier), and *Information Systems* (Elsevier) as well as chair of over ten international conferences/workshops on AI/IoT/eHealth, e.g., in the USA, Portugal, Greece, Czech Republic, Spain, and Germany.

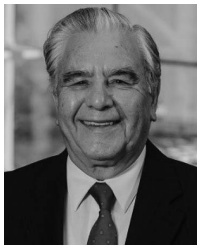
Bahar Farahani (Member, IEEE) received the Ph.D. degree in computer engineering from the University of Tehran, Tehran, Iran, and the Postdoctoral degree in computer engineering from Shahid Behesh University, Tehran.

She is an Assistant Professor with the Cyberspace Research Institute, Shahid Beheshti University. She has authored several peer-reviewed conference/journal papers as well as book chapters on IoT, Big Data, and AI.

Dr. Farahani has served as a Guest Editor of several journals, such as IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS, IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, *Future Generation Computer Systems* (Elsevier), *Microprocessors and Microsystems* (Elsevier), *Journal of Network and Computer Applications* (Elsevier), and *Information Systems* (Elsevier). Besides, she has also served in the Technical Program Committee (TPC) of many international conferences/workshops on AI/IoT/eHealth as well as the Technical Chair of the IEEE COINS conference.

Mojtaba Barzegari received the B.S. degree in materials science and engineering from the Amirkabir University of Technology, Tehran, Iran, in 2011, and the M.S. degree in biomedical engineering from the University of Tehran, Tehran, in 2014. He is currently pursuing the Ph.D. degree in computational biomedical engineering with KU Leuven, Leuven, Belgium.

He has published several peer-reviewed journal papers, conference papers, and book chapters. His research is mainly focused on developing mathematical models and high-performance numerical simulations of tissue engineering systems. His research interests include scientific computing, computational tissue engineering, machine learning, and high-performance computing.



Mahmoud Daneshmand received the B.S. and M.S. degrees in mathematics from the University of Tehran, Tehran, Iran, and the M.S. and Ph.D. degrees in statistics from the University of California at Berkeley, Berkeley, CA, USA.

He is the Co-Founder and a Professor with the Department of Business Intelligence & Analytics as well as the Data Science Ph.D. Program, and a Professor with the Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ, USA. He has more than 40 years of Industry & University experience as the Executive Director, an Assistant Chief Scientist, a Professor, a Researcher, a Distinguished Member of Technical Staff, a Technology Leader, the Founding Chair of Department, and the Dean of School with Bell Laboratories, Murray Hill, NJ, USA; AT&T Shannon Labs Research, Florham Park, NJ, USA; the University of California, at Berkeley; the University of Texas at Austin, Austin, TX, USA; the New York University, New York, NY, USA; the Sharif University of Technology, Tehran; University of Tehran; and the Stevens Institute of Technology. He is a Data Scientist, expert in big data analytics, artificial intelligence, and machine learning with extensive industry experience, including with the Bell Laboratories as well as the Info Lab of the AT&T Shannon Labs Research. He has published more than 250 journal and conference papers; authored/coauthored three books, has graduated more than 2500 Ph.D. and M.S. students.

Dr. Daneshmand holds key leadership roles in IEEE journal publications, IEEE major conferences, Industry—IEEE Partnership, and IEEE Future Direction Initiatives. He has served as the general chair, the keynote chair, the panel chair, the executive program chair, and the technical program chair of many IEEE major conferences. He has given many keynote speeches in major IEEE as well as international conferences.