This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

Secure and Robust Machine Learning for Healthcare: A Survey

Adnan Qayyum[1], Junaid Qadir[1], Muhammad Bilal[2], and Ala Al-Fuqaha[3*]

[1] Information Technology University (ITU), Punjab, Lahore, Pakistan
[2] University of the West England (UWE), Bristol, United Kingdom
[3] Hamad Bin Khalifa University (HBKU), Doha, Qatar

***Abstract*— Recent years have witnessed widespread adoption of machine learning (ML)/deep learning (DL) techniques due to their superior performance for a variety of healthcare applications ranging from the prediction of cardiac arrest from one-dimensional heart signals to computer-aided diagnosis (CADx) using multi-dimensional medical images. Notwithstanding the impressive performance of ML/DL, there are still lingering doubts regarding the robustness of ML/DL in healthcare settings (which is traditionally considered quite challenging due to the myriad security and privacy issues involved), especially in light of recent results that have shown that ML/DL are vulnerable to adversarial attacks. In this paper, we present an overview of various application areas in healthcare that leverage such techniques from security and privacy point of view and present associated challenges. In addition, we present potential methods to ensure secure and privacy-preserving ML for healthcare applications. Finally, we provide insight into the current research challenges and promising directions for future research.**

I. Introduction

We are living in the age of algorithms, in which machine learning (ML)/deep learning (DL) systems have transformed multiple industries such as manufacturing, transportation, and governance. Over the past few years, DL has provided state of the art performance in different domains—e.g., computer vision, text analytics, and speech processing, etc. Due to the extensive deployment of ML/DL algorithms in various domains (e.g., social media), such technology has become inseparable from our routine life. ML/DL algorithms are now beginning to influence healthcare as well—a field that has traditionally been impervious to large-scale technological disruptions [1]. ML/DL techniques have shown outstanding results recently in versatile tasks such as recognition of body organs from medical images [2], classification of interstitial lung diseases [3], detection of lungs nodules [4], medical image reconstruction [5], [6], and brain tumor segmentation [7], to name a few.

It is highly expected that intelligent software will assist radiologists and physicians in examining patients in the near future [8] and ML will revolutionize the medical research and practice [9]. Clinical medicine has emerged as a exciting application area for ML/DL models, and these models have already achieved human-level performance in clinical pathology [10], radiology [11], ophthalmology [12], and dermatology [13]. Some of these studies have even reported that DL models outperform human physicians on average. The aspect of better performance of DL models in comparison with humans has led to the development of computer-aided diagnosis systems—for instance, the U.S. Food and Drug Administration (FDA) in 2018 has announced the approval of an intelligent diagnosis system to detect certain diabetes-related eye problems from medical images that will not require any human intervention.[1]

The potential of ML models for healthcare applications is also benefitting from the progress in concomitantly-advancing technologies like cloud/edge computing, mobile communication, and big data technology [14]. Together with these technologies, ML/DL is capable of producing highly accurate predictive outcomes and can facilitate the human-centered intelligent solutions [15]. Along with other benefits like enabling remote healthcare services for rural and low-income zones, these technologies can play a vital role in revitalizing the healthcare industry.

Notwithstanding the impressive performance of DL algorithms, many recent studies have raised concerns about the security and robustness of ML models—for instance, Szegedy et al. demonstrated for the first time that DL models are strictly vulnerable to carefully crafted adversarial examples [20]. Similarly, various types of data and model poisoning attacks have been proposed against DL systems [21] and different defenses against such strategies have been proposed in the literature [19]. However, the robustness of defense methods is also questionable and different studies have shown that most of the defense techniques fail against a particular attack. The discovery of the fact that DL models are neither secure nor robust hinders significantly their practical deployment in security-critical applications like predictive healthcare which is essentially life-critical. For instance, researchers have already demonstrated the threat of adversarial attacks on ML-based medical systems [22], [17]. Therefore, ensuring the integrity and security of DL models and health data are paramount to the widespread adoption of ML/DL in the industry.

Before moving further, we will elaborate upon the two key terms on which this survey is focused—namely, *security* and *robustness*—particularly in the context of ML/DL models. Security is concerned with the possible threats/attacks that can be realized on an ML/DL system influencing it to get

Email:aalfuqaha@hbku.edu.qa

[1]https://tinyurl.com/FDA-AI-diabetic-eye

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

TABLE I: Comparison of this paper with existing review and survey papers on secure, private, robust ML/DL for healthcare applications. (Covered: $\surd$; Not covered: $\times$; Partially covered: $\approx$)

Year	Authors	Highlights	Type	Applications of ML in Healthcare	Conventional Challenges	Privacy Challenges	Adversarial ML	Secure & Private ML Methods	Solutions for Adversarial ML Attacks	Open Research Issues
2017	Miotto et al. [16]	Presented a review of DL applications in healthcare and the challenges and imitations in terms of ease-of-understanding of outcomes to human experts.	Review	$\approx$	$\surd$	$\approx$	$\times$	$\times$	$\times$	$\times$
2018	Papangelou et al. [17]	Provided an understanding of adversarial examples in clinical applications and introduced the concept of adversarial patients in the context of counterfactual models in clinical trials.	Position Paper	$\times$	$\times$	$\times$	$\surd$	$\times$	$\times$	$\surd$
2019	Kim et al. [18]	Provided a review of different adversarial attacks and defenses with their applications in ML based medical image analysis.	Review	$\approx$	$\times$	$\times$	$\surd$	$\times$	$\surd$	$\times$
2019	Yuan et al. [19]	Provides an overview of literature on adversarial attacks and defenses in general.	Survey	$\times$	$\times$	$\times$	$\surd$	$\times$	$\surd$	$\times$
2020	Our Paper	Presents a comprehensive survey of various security challenges associated with the application of ML/DL in healthcare systems and outlines robust solutions.	Survey	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$

intended behavior or outcome, whereas robustness defines the capability of the ML/DL system to survive under such attacks. Security is analyzed along two dimensions: (a) the attacks on ML/DL systems attempting to get the control of the system or to get the intended behavior/outcome; (b) the attacks trying to learn about the training data, i.e., privacy attacks. On the other hand, robustness is also analyzed along two axes: (a) the survivability of ML/DL systems under attacks attempting to influence them (i.e., robustness to attacks like adversarial ML attacks); (b) the resistance to privacy attacks. Note that the robustness is a relative term and the effectiveness of the system varies according to the nature of the attack, i.e., an ML/DL system might be robust under a particular attack but vulnerable to a different attack.

In this paper, we present a comprehensive survey of existing literature on the security and robustness of ML/DL models when used for building healthcare systems with a specific focus on the above-mentioned dimensions. We note here that the aim of this paper is to provide an in-depth survey of various security challenges associated with the application of ML/DL in healthcare systems and to provide a taxonomy of potential solutions to overcome these issues. Along with discussing security and robustness challenges of using ML/DL models, we also briefly elaborate on various general challenges and sources of vulnerabilities that hinder the safe and robust application of ML/DL in healthcare applications. In addition, potential solutions to address security, privacy, and robustness challenges are presented in this paper. In summary, the following are the specific contributions of this paper.

1) We present an overview of diverse literature on applications of ML/DL techniques by categorizing it to four major tasks in healthcare, i.e., prognosis, diagnosis, treatment, and clinical workflow.
2) We formulate the ML pipeline for data-driven healthcare applications and describe different sources of vulnerabilities at each stage that raises security and robustness challenges.
3) We present an overview of various security and robustness challenges associated with the adoption of ML/DL models for healthcare applications.
4) We present a taxonomy of different solutions that can be used for ensuring secure and robust application of ML/DL techniques for healthcare applications.
5) Finally, we highlight various open research issues that require further investigation.

A comparison of this paper with existing surveys and review papers on the security of ML/DL models in healthcare systems is also presented in Table I.

Organization of the Paper: The rest of the paper is organized as follows. In Section II, various applications of ML and DL techniques in healthcare are discussed. Section III presents the ML pipeline in data-driven healthcare and various sources of vulnerabilities along with different challenges associated with the use of ML. Different potential solutions to ensure secure and privacy-preserving ML are discussed in Section IV and various open research issues are outlined in Section V. Finally, we conclude the paper in Section VI.

II. ML for Healthcare: Applications

In this section, various prominent applications of ML in healthcare are discussed.

A. ML in Healthcare: The Big Picture

The major phases for developing a ML-based healthcare system are illustrated in Figure 1 and major types of ML/DL that can be used in healthcare applications are briefly described next.

1) Unsupervised Learning: The ML techniques utilizing unlabelled data are known as unsupervised learning methods. Widely used examples of unsupervised learning methods are a clustering of data points using a similarity metric and dimensionality reduction to project high dimensional data to lower-dimensional subspaces (sometimes also referred to as feature selection). In addition, unsupervised learning can be

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

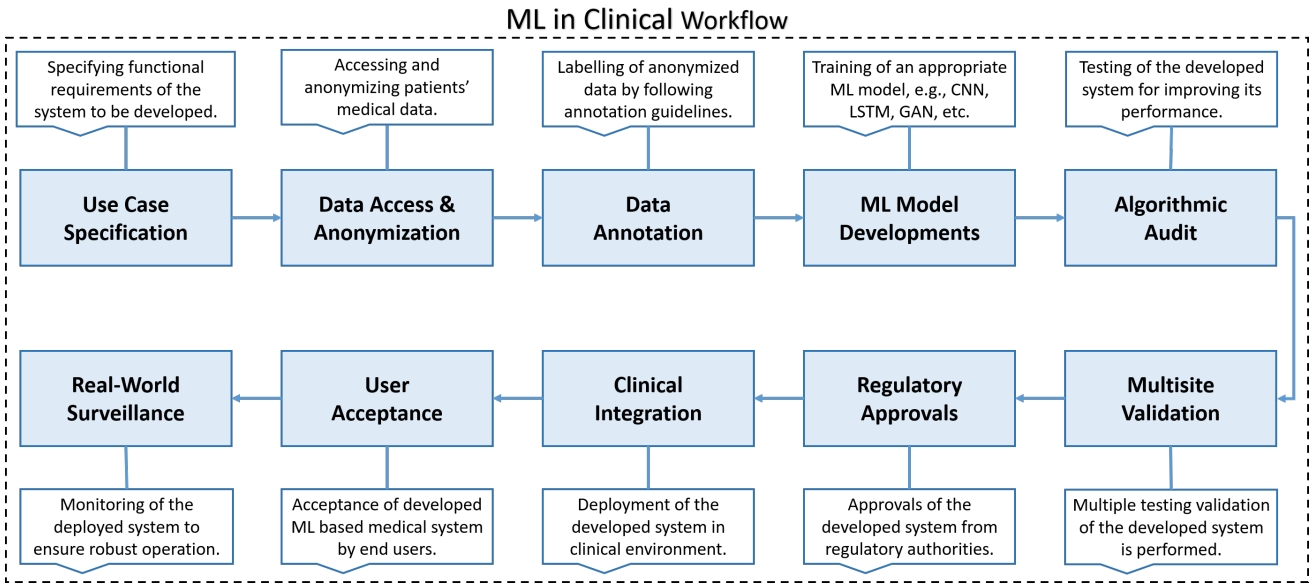

Fig. 1: The illustration of major phases for development of machine learning (ML) based healthcare systems.

used for anomaly detection, e.g., clustering [23]. Classical examples of unsupervised learning methods in healthcare include the prediction of heart diseases using clustering [24] and prediction of hepatitis disease using principal component analysis (PCA) which is a dimensionality reduction technique [25].

2) Supervised Learning: Such methods that build or map the association between the inputs and outputs using labeled training data are characterized as supervised learning methods. If the output is discrete then the task is called *classification* and for a continuous value output, the task is called *regression*. Classical examples of supervised learning methods in healthcare include the classification of different types of lung diseases (nodules) [4] and recognition of different body organs from medical images [2]. Sometimes, ML methods can be neither supervised nor unsupervised, i.e., where the training data contains both labeled and unlabelled samples. Methods utilizing such data are known as semi-supervised learning methods. A systematic review of supervised and unsupervised learning techniques can be found in [26].

3) Semi-supervised Learning: Semi-supervised learning methods are useful when both labelled and unlabelled samples are available for training, typically, a small amount of labelled data and a large amount of unlabelled data. Semi-supervised learning techniques can be particularly useful for a variety of healthcare applications as acquiring a sufficient amount of labelled data for model training is difficult in healthcare. Different facets of semi-supervised learning using different learning techniques have been proposed in the literature. For instance, a semi-supervised clustering method for healthcare data is presented in [27] and a semi-supervised ML approach for activity recognition using sensors data is presented in [28]. In [29], [30], authors applied a semi-supervised learning method to medical image segmentation.

4) Reinforcement Learning: Methods that learn a policy function given a set of observations, actions, and rewards in response to actions performed over time fall in the class of reinforcement learning (RL) [31]. RL has a great potential to transform many healthcare applications and recently, it has been used for context-aware symptoms checking for disease diagnosis [32]. Furthermore, the potential of using RL for healthcare applications can be seen through the recent example of the Go game, where a computer using RL with the integration of supervised and unsupervised learning methods defeated a human champion player [33].

B. Applications of ML in Healthcare

Healthcare service providers generate a large amount of heterogeneous data and information daily, making it difficult for the "traditional methods" to analyze and process it. ML/DL methods help to effectively analyze this data for actionable insights. In addition, there are heterogeneous sources of data that can augment healthcare data such as genomics, medical data, data from social media, and environmental data, etc. A depiction of these sources of data is shown in Figure 2. The four major applications of healthcare that can benefit from ML/DL techniques are prognosis, diagnosis, treatment, and clinical workflow, which are described next.

1) Applications of ML in Prognosis: Prognosis is the process of predicting the expected development of a disease in clinical practice. It also includes identification of symptoms and signs related to a specific disease and whether they will become worse, improve, or remain stable over time and identification of potential associated health problems, complications, ability to perform routine activities, and the likelihood of survival. As in clinical setting, multi-modal patients' data is collected, e.g., phenotypic, genomic, proteomic, pathology tests results, and medical images, etc., which can empower the ML models to facilitate disease prognosis, diagnosis and treatment [34]. For instance, ML models have been largely developed for the identification and classification of different

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

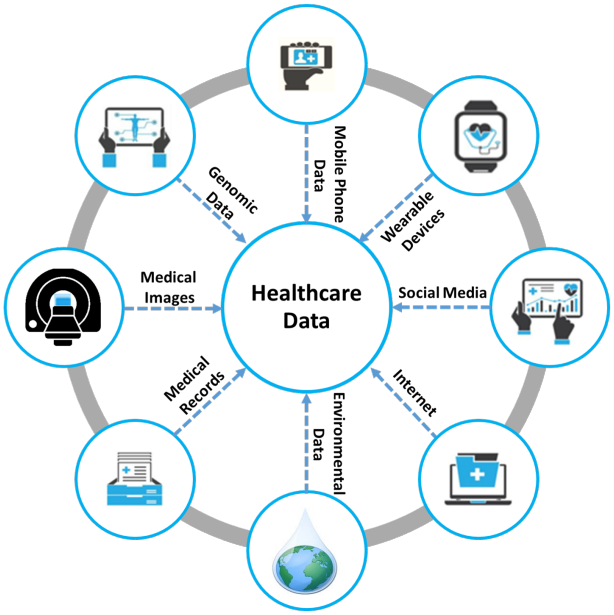

Fig. 2: Illustration of heterogeneous sources contributing to healthcare data.

types of cancers, e.g., brain tumor [35] and lung nodules [36]. However, the potential applications ML for disease prognosis, i.e., prediction of disease symptoms, risks, survivability, and recurrence have been exploited under recent translational research efforts that aim to enable personalized medicine. However, the field of personalized medicine is nascent that requires extensive development of adjacent fields like bioinformatics, strong validation strategies, and demonstrably robust applications of ML thus to achieve the huge and translational impact.

2) Applications of ML in Diagnosis:

a) Electronic Health Records (EHRs): Hospitals and other healthcare service providers are producing a large collection of electronic health records (EHRs) on a daily basis and comprise of structured and unstructured data that contains a complete medication history of patients. ML-based methods have been utilized for the extraction of clinical features for facilitating the diagnosis process [37]. For example, a semi-supervised approach for the extraction of diagnosis information from unstructured EHRs is presented in [38]. The use of ML for the diagnosis of diabetes from EHRs is presented in [39]. In [40], features robustness using EHRs data for the year of care for each record is examined for two tasks, i.e., mortality prediction and length-of-stay and authors showed that prediction performance gets degraded when ML models are trained on historical data and tested on unseen (future) data.

b) ML in Medical Image Analysis: In medical image analysis, ML techniques are used for efficient and effective extraction of information from medical images that are acquired using different imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, and positron emission tomography (PET), etc. These modalities provide important functional and anatomical information about different body organs and play a crucial role in the detection/localization and diagnosis of abnormalities. A taxonomy of key medical imaging modalities is presented in Figure 3. The key purpose of medical image analysis is to assist clinicians and radiologists for efficient diagnosis and prognosis of the diseases. The prominent tasks in medical image analysis include detection, classification, segmentation, retrieval, reconstruction, and image registration which are discussed next. Moreover, fully automated intelligent medical image diagnosis systems are expected to be part of next-generation healthcare systems.

- *Enhancement:* Enhancement of degraded medical images is an important pre-processing step that directly effects the diagnosis process. There are many sources of noise and disturbances encountered in the medical image acquisition process which degrade the quality and significance of the resultant images. For instance, generating MRI images is a quite lengthy process that typically requires several minutes to produce a good quality image and to acquire detailed soft-tissue contrast, patients have to remain still and straight as much as possible. Because movements can cause false artifacts in image acquisition, the complete process has to be repeated usually multiple times to produce significantly useful images. Also, depending on the body area being scanned and the number of images to be taken, patients might be asked to hold their breath during short scans [42]. Therefore, any movement of the subject can introduce artifacts in the acquired image. Moreover, some sort of mechanical noise is also sometimes introduced in the output image. In the literature, different DL models are used for denoising medical images such as convolutional denoising autoencoders [43] and GANs. In addition, GANs have been successfully used for cleaning motion artifacts introduced in multi-shot MRI images [14]. Super-resolution is yet another powerful and impactful enhancement technique for medical images, e.g., MRI denoising [44].
- *Detection:* The process of identifying specific disease patterns or abnormalities (e.g., tumor, cancer) in medical images is known as detection. In traditional clinical practice, such abnormalities are identified by expert radiologists or physicians that often require a lot of time and effort. Whereas, DL based methods have shown their potential for this task and various studies have been presented in the literature for the detection of diseases. For instance, a locality-sensitive approach utilizing CNN for the detection and classification of nuclei colon cancer in histopathological images is presented in [45]. A hybrid method utilizing handcrafted features and a CNN model for the detection of mitosis in breast cancer images is presented in [46].
- *Classification* DL models in particular, convolutional neural networks (CNNs) have proven to give high performance in medical image classification tasks when compared with other state-of-the-art non-learning based techniques. Modality classification, recognizing different body organs, and abnormalities from medical images

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

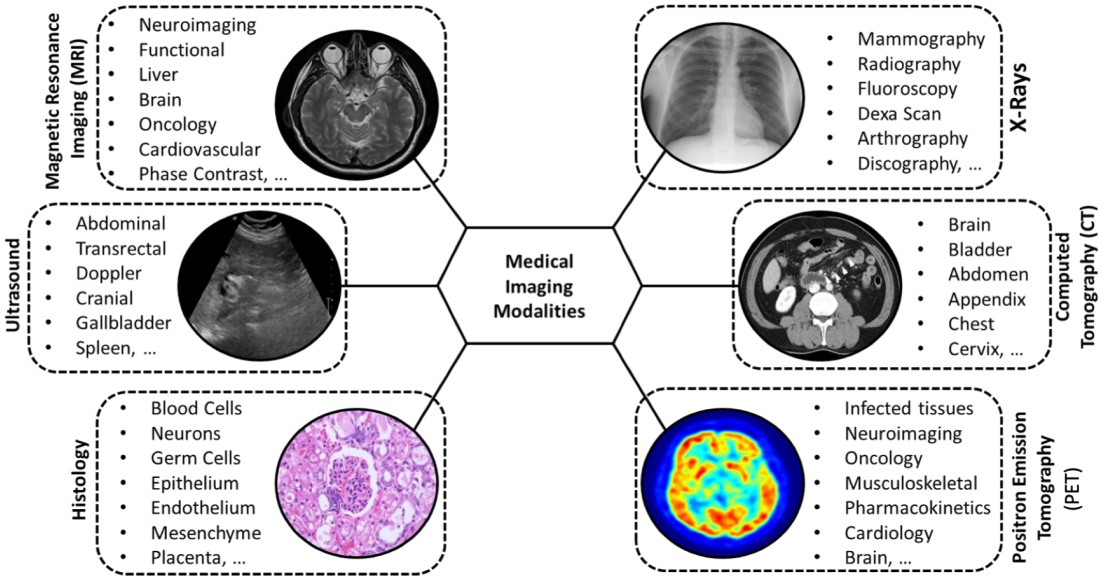

Fig. 3: A typology of commonly used medical imaging modalities (adapted from [41]).

using CNNs have been extensively studied in the literature. In [2], an approach using CNN for multi-instance recognition of different body organs is presented and a CNN based method for classification of interstitial lung diseases (ILDs) is presented in [3]. In another study, a CNN model is trained for the classification of lung nodules [4].

Transfer learning approaches have also been used for medical image classification [47]. In transfer learning, a pre-trained DL model (typically trained on natural images) is fine-tuned on a comparatively small dataset of medical images. The results obtained by this approach, as reported in the literature, are promising; however, a few studies have reported contradictory results. For instance, results obtained by transfer learning in [48] and [49] are contradictory.

- *Segmentation:* The segmentation of tissues and organs in medical images enables quantitative analysis of abnormalities in terms of clinical parameters, e.g., automatically measuring the volume and shape of cancer in brain images. In addition, the extraction of such clinically significant features is an important and foremost step in computer-aided detection and diagnosis systems that we discuss later in this section. The process of segmentation deals with the partitioning of an image into multiple non-overlapping parts using a pre-defined criterion such as intrinsic color, texture, and contrast, etc. Addressing the problem of segmentation utilizing various DL models (e.g, CNN and recurrent neural network (RNN) [50]) is widely studied in the literature and the common architecture used for segmentation of medical images is U-net [51]. Various DL architectures are being proposed for the segmentation of multi-modal images such as the brain, skin cancer, CT images, etc. as well as segmentation of volumetric images [52]. An overview of various DL models for segmentation of medical images is presented in [53].
- *Reconstruction:* The process of generating interpretable images from raw data acquired from the imaging sensor is known as medical image reconstruction. The fundamental problem in medical image reconstruction is to accelerate the inherently slow data acquisition process, which is an interesting ill-posed inverse problem in which we want to determine the system's input given its output. Many important medical imaging modalities require a lot of time for reconstructing an image from the raw data samples, e.g., MRI and CT. Thus in medical image reconstruction, we aim to reduce image acquisition time and storage space.

  Research on medical image reconstruction using deep models is drastically increasing and various DL models such as CNNs [54] and autoencoders [6] have been extensively used for the reconstruction of MRI and CT images. Recently, generative adversarial networks (GANs) have been widely used for the reconstruction of medical images and have produced outstanding results. For instance, a GAN based MRI reconstruction method is presented in [55] that also cleans the motion artifacts.
- *Image Registration:* Image registration is the process of mapping input images with respect to a reference image and it is the first step in image fusion. Image registration has many potential applications in medical image analysis as described in detail by El-Gamal et al. [56], however, their use in actual clinical applications is very limited [57]. To facilitate the surgical spinal screw implant or tumor removal, image registration is usually applied in spinal surgery or neurosurgery for the localization of spinal bony landmark or a tumor, respectively. Various similarity metrics and reference points are calculated to align the sensed image with the reference image. In [58], a framework for deformable image registration named as Quicksilver is proposed that uses the large deforma-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

tion diffeomorphic metric mapping (LDDMM) model for patch-wise prediction strategy. Similarly, an unsupervised learning based methods for deformable image registration is presented in . In [59], a CNN based regression approach for 2D/3D image registration is presented that addresses two fundamental limitations of existing intensity-based image registration methods, i.e., small capture range and slow computation.

- *Retrieval:* The recent era has witnessed the revolution of digital interventions from the large-scale image and video collections to big data. This trend is true for medical imaging as well, every hospital and clinic having radiology services are producing thousands of medical images daily in diverse modalities, resulting in the growth of large-scale multi-modal medical image repositories. Thus making it difficult to manage and query such huge databases. In particular, it is more challenging for multi-modal medical data. To facilitate the production and management of multi-modal medical data, traditional methods are not sufficient and various ML/DL techniques are proposed in the literature [60], [61].

  In routine practice, clinicians usually compare the current cases with the previous ones, mainly to effectively plan the diagnosis and treatment of the patient being examined. In this regard, identifying modality (i.e., modality classification discussed above) is of great significance as it serves as an initial tool to facilitate the process of comparison and an efficient modality classification system will reduce the search space by only looking for relevant images in the collections of the desired modality.

3) Applications of ML in Treatment:

a) Image Interpretation: As discussed above, medical images are widely used in the routine clinical practice and the analysis and interpretation of these images are performed by expert physicians and radiologists. To narrate the findings regarding images being studied, they write textual radiology reports about each body organ that was examined in the conducted study. However, writing such reports is very challenging in some scenarios, e.g., less experienced radiologists and healthcare service providers in rural areas where the quality of healthcare services is not up to the mark. On the other side, for experienced radiologists and pathologists, the process of preparing high-quality reports can be tedious and time-consuming which can be exacerbated by a large number of patients visiting daily. Therefore, various researchers have attempted to address this problem using natural language processing (NLP) and ML techniques. In [62], a natural language processing based method is proposed for annotating clinical radiology reports. A multi-task ML based framework is proposed for automatic tagging and description of medical images [63]. In a similar study [64], an end-to-end architecture developed with the integration of CNN and RNN is presented for thorax disease classification and reporting in chest X-rays. In [65], a novel multi-modal model utilizing CNN and long short term memory (LSTM) network is developed for automatic report generation.

b) ML in Real-time Health Monitoring: Real-time monitoring of critical patients is crucial and is a key component of the treatment process. Continuous health monitoring using wearable devices, IoT sensors, and smartphones is gaining interest among people. In a typical setting of continuous health monitoring, health data is collected using a wearable device and smartphone and then transmitted to the cloud for analysis using an ML/DL technique. Then the outcomes are transmitted back to the device for appropriate action(s). For instance, a framework having a similar system architecture is presented in [66]. The system is developed by integrating mobile and cloud for monitoring of heart rate using PPG signals. Similarly, a review of different ML techniques for human activity recognition with application to remote monitoring of patients using wearable devices is presented in [67]. The sharing of health data with clouds for further analysis raises many privacy and security challenges that we discuss in the next section.

4) Applications of ML in Clinical Workflows:

a) Disease Prediction and Diagnosis: The early prediction and diagnosis of diseases from medical data are one of the exciting applications of ML. Various studies have highlighted the potential of using predictive healthcare for the timely treatment of diseases. For instance, the case of cardiovascular risk prediction using different ML algorithms with clinical data is studied in [68] and the study concluded that ML techniques improved the prediction efficacy. A survey of various ML techniques for the detection and diagnosis of different diseases (such as diabetes, dengue, hepatitis, heart, and liver) is presented in [69]. The potential of using ML-based methods for prediction and prognosis of cancer is highlighted in [70].

b) ML in Computer-Aided Detection or Diagnosis: The computer-aided detection (CADe) or computer-aided diagnosis (CADx) systems are being developed mainly for the automatic interpretation of medical images that would assist the radiologist in their clinical practice. The system works by utilizing different functionalities including ML/DL, traditional computer vision and image processing techniques and relies heavily on the performance of these techniques. IBM's Watson is a classical example of CADx system developed by integrating various techniques including ML. However, any task in medical image and signal analysis automated by the application of ML/DL models can be deemed as a CADe or CADx systems, e.g., automation detection of fatty liver in ultrasound kurtosis imaging [71].

c) Clinical Reinforcement Learning: In reinforcement learning, the key objective is to learn a policy function for making precise decisions in an uncertain environment to maximise accumulated reward. In clinical medicine, RL can be used for providing optimal diagnosis and treatment for patients with distinct characteristics [72]. The performance evaluation of different RL techniques (i.e., Q-value iteration, tabular Q-learning, fitted Q-iteration (FQI), and deep Q-learning) for the treatment of sepsis in ICU using real-world medical dataset is presented in [73]. Sepsis is a severe infection involving organ dysfunction and is a leading cause of mortality due to expensive and suboptimal treatment. The dataset contains trajectories of a patient's physiological state and the provided treatments by clinicians at each time, along with the outcome (i.e., survival or mortality). The study concluded that simple

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

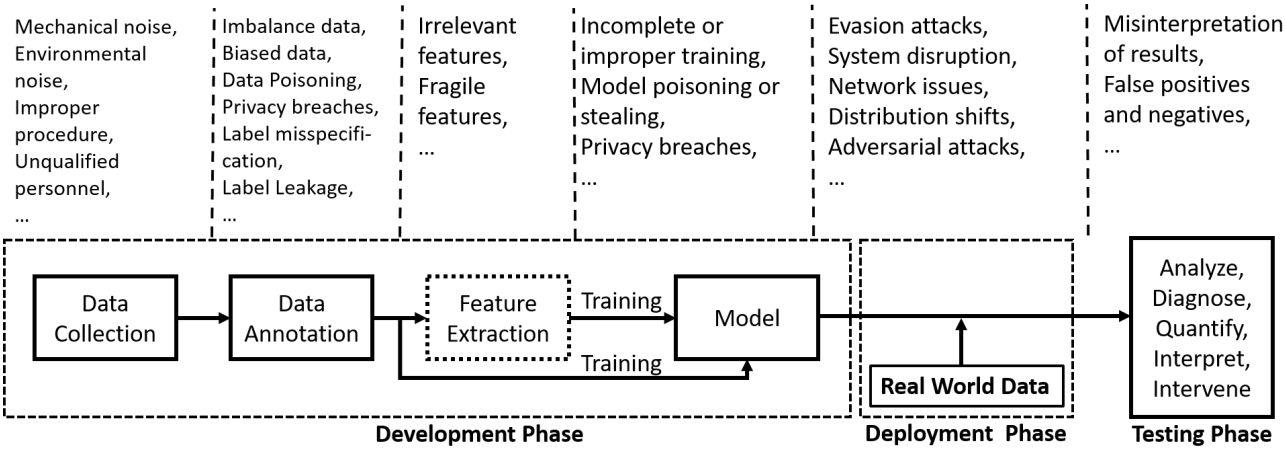

Fig. 4: The pipeline for data-driven predictive clinical care and various sources of vulnerabilities at each stage.

and tabular Q-learning can learn effective policies for sepsis treatment and their performance is comparable with a complex continuous state-space method, i.e., deep Q-learning.

d) ML for Clinical Time-Series Data: One of the tasks in clinical workflows is the modeling of clinical time-series data. Applications of clinical time-series modeling include prediction of clinical interventions in intensive care units (ICUs) using CNN and LSTM [74], mortality prediction in patients with traumatic brain injury (TBI) [75], and estimation of mean arterial blood pressure (ABP) and intracranial pressure (ICP) which are important indicators cerebrovascular autoregulation (CA) in TBI patients. In a recent study, attention models are used for the management of ICUs forecasting tasks (such as diagnosis, estimation, and prediction, etc.) by integrating clinical notes with multivariate and time-series measurements data [76]. In a similar study, the problem of unexpected respiratory decompensation using ML techniques is investigated in [77].

e) Clinical Natural Language Processing: Clinical notes are a widely used tool by the clinicians to communicate patient state. The use of clinical text is crucial as it often contains the most important information. The progress in clinical NLP techniques is envisioned to be incorporated in future clinical software for extracting relevant information from unstructured clinical notes for improving clinical practice and research [78]. Clinical NLP offers unique challenges such as the use of acronyms, language disparity, partial structure, and quality variance, etc. The challenges and opportunities of clinical NLP for languages other than English along with a review of clinical NLP techniques is presented in [79]. In [80], authors presented a toolkit named CLAMP that provides different state of the art NLP techniques for clinical text analysis.

f) Clinical Speech and Audio Processing: In the clinical environment, clinicians have to do a lot of documentation, i.e., preparing clinical notes, discharge summaries, and radiology reports, etc. According to Dr. Simon Wallace, clinicians spend 50% of their time on clinical documentation and are highly demotivated due to clinical workload, administrative tasks, and lack of leisure time [81]. Typically, they spend more time in preparing clinical documentation as compared to interacting directly with patients. To overcome such challenges, clinical speech and audio processing offer new opportunities such as speech interfaces for interaction less services, automatic transcription of patient conversations, and synthesis of clinical notes, etc. There are many benefits for using speech and audio processing tools in the clinical environment for each stakeholder, i.e., patients (speech is a new modality for determining patient state), clinicians (efficiency and time-saving), and healthcare industry (enhance productivity and cost reduction). In the literature, speech processing has been used for the identification of disorders related to speech, e.g., vocal hyperfunction [82] and as well as disorders that manifest through speech, e.g., dementia [83]. Alzheimer's disease identification using linguistic features is presented in [84]. In clinical speech processing, disfluency and utterance segmentation are two well-known challenges of clinical speech processing.

III. Secure, Private, and Robust ML for Healthcare: Challenges

In this section, we analyze the security and robustness of ML/DL models in healthcare settings and present various associated challenges.

A. Sources of Vulnerabilities in ML Pipeline

ML application in healthcare settings suffers from various privacy and security challenges that we will thoroughly discuss in this section. In addition, the three major phases of ML model development along with different potential sources of vulnerabilities causing such challenges in each step of the ML pipeline are depicted in Figure 4.

1) Vulnerabilities in Data Collection: Training of ML/DL models for clinical decision support requires the collection of a large amount of data (in formats such as EHRs, medical images, radiology reports, etc.), which is in general often time-consuming and requires significant human efforts. Although in practice, medical data is carefully collected to ensure the effectiveness of the diagnosis, however, there can be many sources of vulnerabilities that can affect the proper (expected)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

functionality of the underlying ML/DL systems, a few of them are described next.

Instrumental and Environmental Noise: The collected data often contains many artifacts that arise due to instrumental and environmental disturbances. Let's consider the example of one of the widely used imagining modalities used to acquire high-resolution medical images, i.e., multishot MRI. This modality is highly sensitive to motion, and even slight movement of the subject's head or respiration can cause undesirable artifacts in the resultant image [14], thereby increasing the risk of misdiagnosis [85].

Unqualified Personnel: Healthcare ecosystems are extremely interdisciplinary and comprise of technical and non-technical personnel and often lack qualified workers that can develop and maintain ML/DL systems. As for the efficient application of data-driven healthcare, workers with strong statistical and computational backgrounds are required, e.g., engineers and data scientists. On the contrary, the clinical usability of ML/DL based systems is extremely important. Considering this aspect, hospitals tend to rely solely on physician-researchers who lack computational expertise to develop such systems [86].

2) Vulnerabilities Due to Data Annotation: Most applications of ML/DL in healthcare systems are supervised ML tasks which require an abundance of labelled training data. The process of assigning labels to each data sample (e.g., medical image) is known as data annotation. Ideally, this task shall mostly be performed by experienced clinicians (physicians or radiologists) to prepare domain-enriched datasets which are crucial to the development of useful ML/DL models in healthcare systems. The literature has revealed that training ML/DL models without a sound grip of the domain could be disastrous [87]. However, clinicians like expert radiologists are rare professionals and hard to engage in secondary tasks like data annotation. As a result, trainee staff (with little domain expertise) or ML/DL automated algorithms are usually employed during data labelling, which often leads to many problems such as coarse-grained labels, class imbalance, label leakage, and misspecification. Some specific data annotation-based vulnerabilities are discussed as below:

Ambiguous Ground Truth: In medical datasets, the ground truth is often ambiguous, e.g., medical image classification task [22] and even expert clinicians disagree on well-defined diagnostic tasks [88]. This problem becomes more adverse with the presence of malicious users who want to perturb data, making the diagnosis difficult and causing difficulties in detecting its influence even with a human expert review.

Improper Annotation: The annotation of data samples process for life-critical healthcare applications should be informed by proper guidelines and various privacy and legal considerations [89]. Most widely used healthcare datasets are annotated for coarse-grained labels whereas real-life utility of ML/DL is to highlight rare, fine-grained and hidden strata within the clinical environment. This inability to perform labelling appropriately can lead to various efficiency challenges that are discussed next.

Efficiency Challenges: The collections of healthcare data on which ML/DL models are built suffer from various issues that arise several efficiency challenges. A few major problems impacting the quality of data are described next.

(a) *Limited and Imbalanced Datasets:* The size of datasets used for training ML/DL models is not up to the required scale. In particular, one major limitation of the efficient application of DL approaches in healthcare is the unavailability of large-scale datasets, as health data is often small in size. Notably, most life-threating health conditions are naturally rare and diagnosed once in many (thousands to millions) patients. Therefore, most ML/DL algorithms can not be efficiently trained and optimized for such life-threatening healthcare task.

(b) *Data Augmentation:* To circumvent the problem of availability of large scale medical datasets, one commonly followed method is data augmentation in which various techniques (such as cropping, filliping, rotation, and translation, etc.) are used for diversifying the training data and increasing its size. In addition, different transformation techniques are used for augmenting training datasets, e.g., use of Gaussian for data augmentation [90], [91]. However, the use of data augmentation might reduce the robustness of the developed ML/DL based system, for example, it is highly likely that the distribution of transformed data diverges from the underlying actual distribution of the training data which is unknown generally and there are no statistical and probabilistical guarantees for having same distribution of the training data. The literature suggest that Guassain data augmentation does not improves the adversarial robustness of the models against iterative attacks [92].

(c) *Class Imbalance and Bias:* Class imbalance is yet another problem that arises in the supervised ML/DL which refers to the fact that the distribution of samples among classes is not uniform. If a class imbalanced dataset is used for training of the model then it will be reflected in the model's outcomes in terms of bias to certain categories. Biases in models' predictions in healthcare settings will have profound consequences and should, therefore, be mitigated. Various approaches have been proposed in the literature to address class imbalance problems. These approaches are discussed in the next section.

(d) *Sparsity:* Data sparsity, i.e., missing values are common in real-world data that arise due to various reasons (e.g., unmeasured and unreported samples, etc.). Missing values and observations significantly affect the performance of ML/DL techniques.

3) Vulnerabilities in Model Training: The vulnerabilities regarding model training include improper or incomplete training, privacy breaches, model poisoning and stealing. Improper or incomplete training refers to the situations when the ML/DL model is trained with improper parameters, e.g., learning rate, epochs, batch size. Moreover, ML/DL models have been found strictly vulnerable to various security and privacy threats such as adversarial attacks [20], model [93] and data poisoning attacks [94], etc. The vulnerabilities of ML/DL systems hinder their efficient deployment for security-critical applications (such as digital forensic, bio-metrics, etc.) and as well as life-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

critical applications (such as self-driving cars and healthcare, etc.). Therefore, ensuring the security and integrity of the ML/DL systems is of paramount importance for such critical applications. Various security threats associated with ML/DL systems are thoroughly described in the next section.

4) Vulnerabilities in Deployment Phase: The deployment of ML/DL techniques in a clinical environment essentially involves human-centric decisions. Therefore, ensuring the robustness of the system while considering fairness and accountability is necessary for the deployment phase. The following are the major vulnerabilities that can be encountered in the deployment phase of ML/DL systems. Whereas, security issues (e.g., adversarial attacks) are discussed in the next section.

Distribution Shifts: Distributions shifts are very much expected in realistic healthcare settings, for example, let's consider different imaging centers and DL models trained on images of one domain (imaging center) are subsequently deployed on different domain images. In such settings, the performance of the underlying DL model degrades significantly. Moreover, in predictive healthcare, ML models are developed using historical patient data and are usually tested on the new patients which raise questions about the efficacy of the ML predictions. Moreover, such differences can be exploited for generating adversarial examples [95].

Incomplete Data: In realistic settings, data collected for providing patient care may contain missing observations or variables, e.g., EHRs. The simplest way to avoid missing values is to ignore them completely while doing analysis but it cannot be done without knowing their relationships with already observed or unobserved data. Using the missing observations for training ML/DL models, on the other hand, leads to two well-known problems, i.e., false positives (a healthy person is diagnosed with a disease) and false negatives (a patient is identified as healthy). Both problems can have severe outcomes in actual healthcare settings, therefore, the healthcare data should be complete and compact in all aspects to ensure accurate predictions of outcomes.

5) Vulnerabilities in Testing Phase: Vulnerabilities in the testing phase are concerned with the interpretation of the results from the underlying ML/DL systems that include misinterpretation, false positive, and false-negative outcomes. False-positive and false-negative outcomes are due to incomplete/inefficient training of the model or due to incomplete data fed for the inference that we have discussed in the earlier section. Finally, the true essence of ML empowered healthcare is not just about turning a crank but it demands the cautious application of analytical methods [96].

B. The Security of ML: An Overview

In this section, we provide an overview of ML security particularly from the perspective of healthcare and highlight various associated security challenges with the use of ML.

1) Security Threats: The security threats on ML systems can be broadly categorized into three dimensions, i.e., influence attacks, security violations, and attack specificity [97]. A taxonomy of these security threats on ML systems is depicted in Figure 5.

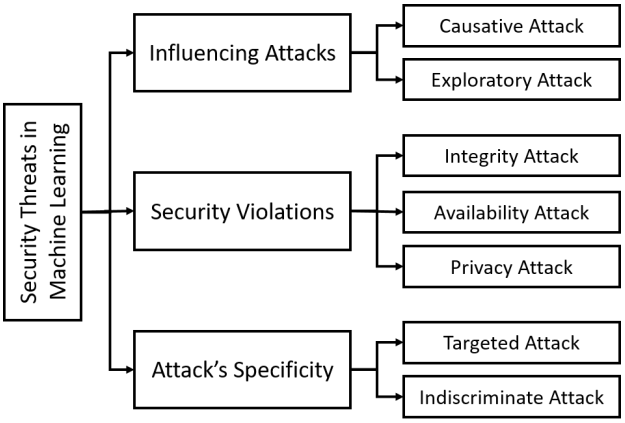

Fig. 5: A taxonomy of different security threats on ML/DL models.

(a) *Influence*: Influence attacks can be of two types: (1) causative: the one that attempts to get control over training data; (2) exploratory: the one that exploits the miss-classification of the ML model without intervening the model training.
(b) *Security Violation*: It is concerned with the availability and integrity of the services and can be categorized into three types: (1) integrity attack: It attempts to increase the false-negative rate of the deployed model (classifier) when the model is given harmful inputs; (2) availability attack: Unlike integrity attack, it tries to achieve an increase in the false-positive rate of the classifier in response to benign inputs; (3) privacy violation attack: It is concerned with the unveiling of sensitive and confidential information of the training data, trained model or both.
(c) *Attack Specificity*: The specificity of an attack can be defined in two ways: (1) targeted attack: whether the attack is intended for a specific input sample or a group of samples; (2) indiscriminate attack: it causes the ML model to fail indiscriminately.

The first axis in the taxonomy of the attacks on ML/DL systems (as shown in Figure 5) defines the capabilities of the adversaries, e.g., whether they are able to modify training process by injecting poisoned data or not (i.e., attempting access to training data). If the attacker does not have access to the training data, the attacker can realize an exploratory attack, e.g., consider a disease classification problem, the adversary can exploit query-response pairs to get intended behavior (i.e., misclassification in this case). The second dimension of attacks is concerned with the type of security violations that an adversary can perform, e.g., trying to learn about the privacy of users in training data or attempting to increase the false-negative or false-positive rate of the classifier. Each type of security violation is severely problematic for healthcare applications, i.e., preserving the privacy of users is a matter of high concern, and models with minimum uncertainty are highly desirable. The third dimension describes the specific objectives of the adversary. The attacker might be interested in attempting a targeted attack, e.g., forcing the classifier to classify a given input sample to a target class (e.g., bypassing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

disease detection system by influencing the detector to identify the input as benign), or, he might intend to break down the classifier in an indiscriminate manner.

2) Adversarial Machine Learning (ML): Adversarial attacks are the result of recent efforts for identifying vulnerabilities in ML/DL models training and inference. Adversarial attacks have appeared as one of the biggest security threats to ML/DL systems [20], [98], [99], [100], [101]. In adversarial attacks, the key goal of an adversary is to generate adversarial examples by adding small carefully crafted (unnoticeable) perturbation into the actual (non-modified) input samples to evade the integrity of the ML/DL system. In general, there are two types of adversarial attacks that are described next.

(a) *Poisoning Attacks:* Adversarial attacks affecting the model training, i.e., manipulating the training data to mislead the learning of ML/DL model are known as poisoning attacks [93].

(b) *Evasion Attacks:* Adversarial attacks on the inference phase of the training process are known as evasion attacks [102]. In such attacks, an attacker manipulates the test data to compromise the integrity of the ML/DL model to harmful inputs.

In healthcare applications, poisoning attacks are highly relevant because direct manipulation of the training data may be difficult or even impossible in some cases. Alternatively, the addition of new samples might be relatively easy, however, any such consequences hinder the applicability of the ML/DL systems. Therefore, the detection of poisoning attacks is critical for the robust application of ML/DL in healthcare applications. For instance, systematic poisoning attacks against six conventional ML models that were developed for hypothyroid diagnosis are presented in [103], where the objective of the attacker was to prevent hypothyroid diagnosis.

Similarly, a few researchers have highlighted the threat of these attacks to ML/DL models in healthcare settings and we provide insights from such articles in this section. Unlike adversarial examples created for evading ML/DL models in other settings, the concept of *adversarial patients* for healthcare applications is introduced in [17]. The authors argue that rather than intentional adversarial examples, the caution should be for unintentional adversarial patients that can lead to severe ethical issues. They identified a subgroup of adversarial patients and empirically validated that patients with identical predictive features can have significantly different individual treatment effects. In recent studies, white box and black box adversarial attacks have been demonstrated against three clinical applications; namely, fundoscopy, dermoscopy, and chest X-ray analysis [22], [104]. Furthermore, in [104], authors highlighted various potential incentives for adversaries via adversarial attacks in clinical trials that will rise with the increasing use of ML in the future, particularly, with the emergence of computer-aided diagnosis and decision support systems.

Adversarial ML is a major dilemma for the security and privacy of ML/DL models deployed in healthcare biometrics applications and can lead to sever unintended circumstances. Biometrics can provide many advantages, e.g., fraud detection, protection of confidential medical records, and securing medical facilities and equipment, etc. In this regard, different biometrics technologies such as palm vein readers, fingerprint, ECG, and iris scanners [105], and face recognition have great potential to be deployed in healthcare systems. It is very common to use ML/DL techniques for building healthcare biometric systems, which are themselves vulnerable to security and privacy attacks [106], [107], [108]. For example, an adversary can easily evade a face recognition system that is deployed in a restricted area to restrain unintended access for security purposes.

C. ML for Healthcare: Challenges

In this section, we discuss various challenges which hinders the applicability of ML/DL systems in practical healthcare applications.

1) Safety Challenges: Excellent performance in a controlled lab environment (which is a common ML community practice) is not evidence of safety. Safety of ML/DL is the determination of how safe the ML/DL system is for patients. There should be a constant thought of safety throughout the ML/DL lifecycle. Majority of routine clinicians tasks are mundane, and patients they encounter have common health conditions. It is their role of diagnosing rare, subtle, and hidden health conditions which occur once in millions. Enabling ML/DL to performing well on hidden strata, outliers, edge, and subtle cases is key to ensure the safety of current AI systems.

2) Privacy Challenges: Privacy is one of the major challenges in data-driven healthcare which is concerned with the use of users' data by the ML/DL systems for making predictions. The users (i.e., patients) expect that their healthcare service providers are following necessary safety measures to safeguard their inherent right to the privacy of their confidential information, e.g., age, sex, date of birth, and health data. Potential privacy threats can be of two types, i.e., unveiling confidential information and malicious use of data (potentially by unauthorized agents).

Privacy depends upon the characteristics and nature of the data being collected, the environment it has been created in, and patients' demographics. Therefore, mitigation of privacy breaches using the appropriate technique(s) is critical as such breaches can directly harm the patients. The confidential data should be anonymized to prevent privacy breaches such as (re-)identification of the individuals [109]. Moreover, necessary attention should be paid to understand privacy concerns at each stage of data processing and the transfer of data among different departments within a hospital should be communicated in a secure environment.

Privacy challenges also arise with adoption of ML/DL techniques for building biometric healthcare systems either offline (e.g., face or fingerprint recognition based system to protect medical facilities and equipment [110]) or online systems, e.g., real-time medical systems [111] and use of biometrics for authentication of medical IoT devices [112], etc. The security and privacy of such systems are of utmost importance; therefore, worst-case robustness test should be performed for biometrically secure healthcare systems. Worst-case testing is a powerful tool that can provide enough evidence about

systems robustness and can distinguish from a system that never fails and a system that fails once in billion trails.

3) Ethical Challenges: In user-centric applications of ML such as healthcare, it is important to ensure the ethical use of data. Explicit measures should be taken to understand the targeted user population and their sociological aspects before collecting data for building ML models. Moreover, understanding how data collection can harm a patient's well-being and dignity is an important consideration in this regard. If ethical concerns are not taken into account then the application of ML in realistic settings will have adverse results. Furthermore, to ensure fair and ethical operation of automated systems, it is imperative to have a clear understanding of the AI system in uncertain and complex scenarios [113].

4) Causality is Challenging: Understanding causality is important in healthcare because most of the crucial healthcare problems require causal reasoning, i.e., *"what if?"* [114]. For example, asking a question about what will happen if a doctor prescribed treatment A instead of treatment B. Such questions cannot be exploited through classical learning algorithms and to answer them we need to analyze the data from the lens of causality [115]. In healthcare, learning is often solely based on observational data and asking causal questions by learning from observational data is quite challenging which requires building causal models.

DL models are black-box which lacks fundamental underlying theory and these models essentially work by exploiting patterns and correlations without considering any causal link [116]. In general, this cannot be deemed as a limitation since prediction does not require any causal relation. In predictive healthcare, the absence of causal relation can raise questions about the conclusions that can be drawn from outcomes of DL models. Furthermore, fairness in decision making can better be enforced through the lens of causal reasoning [117], [118]. The estimation of the causal effect of some variable(s) on a target output (e.g., target class in multi-class classification problem) is important to ensure fair predictions.

5) Regulatory and Policy Challenges: The full potential of ML/DL systems (which essentially constitutes software as a medical device) in actual healthcare settings can only be realized by addressing regulatory and policy challenges. The literature suggests that the regulatory guidelines are needed for both medical ML/DL systems and their integration in actual clinical settings [131]. Therefore, the integration of AI-empowered ML/DL systems in the actual clinical environment should be in compliance with the policies and regulations defined by the government and regulatory agencies. However, existing regulations are not suitable for certifying systems which are ever-evolving such as ML/DL empowered systems because yet another key challenge with the use of ML/DL algorithms in clinical practice is to determine how these models should be implemented and regulated since these models will incorporate learning from the new patient data [132]. In addition, the objective clinical evaluation of ML/DL systems for particular clinical settings is crucial to ensure safe, effective, and robust operation that does not harm the patients in either way. Data scientist and AI engineers should be employed in hospitals for assessing AI systems regularly to ensure it is still safe, relevant, and working fine.

6) Availability of Good Quality Data: The availability of representative, diverse and high-quality data is one of the major challenges in healthcare. For instance, the amount of data available to the research community is very small in size and limited in scope as compared to the heterogeneous collections of large-scale multi-modal patient data being generated on daily basis by different small and large size healthcare institutions. However, the development of good quality data that resembles real clinical settings is on the other very challenging and requires resources for management and maintenance. The availability of high-quality data can effectively serve the intended purpose of disease prediction and decision making for planning treatment.

The data collected in practice suffer from different issues such as subjectivity, redundancy, and bias. As the ML/DL models perform inferences by solely learning the latent factors of the data on which they are trained, therefore, the effect of data generated by the undesirable past practices of hospitals will be reflected in the outcomes of the algorithm. For example, most people with no health insurance are denied healthcare services and if AI learns from that data, it will do the same. It has been shown that a model could depict racial bias by producing varying outcomes for different subpopulations [133] and the training data can also introduce its own modeling challenges [134], [135].

7) Lack of Data Standardization and Exchange: Medical ML/DL system shall facilitate a deep understanding of the underlying healthcare task, which (in most cases) can only be achieved by utilising other forms of patients data. For example, radiology is not all about clinical imaging. Other patient EMR data is crucial for radiologists to derive the precise conclusion for an imaging study. This calls for the integration and data exchange between all healthcare systems. Despite extensive research on data exchange standards for healthcare, there is a huge ignorance in following those standards in healthcare IT systems which broadly affects the quality and efficacy of healthcare data, accumulated through these systems. There are numerous guidelines to perform specific medical interventions like imaging studies (i.e., with define exposure and positioning) to ensure the significance of the data clinically. However, current healthcare IT systems largely ignore standards and clinicians barely follow well-established guidelines. As a result, data integration and exchange efforts across different specialities and organisations fail. Data integration to match diverse patients' medical records is crucial to deliver high-value patient care. The lack of appetite to implement data exchange standards in wider healthcare industry hinders the efficacy of ML/DL systems as multi-modal data is vital to ensure the deep understanding of algorithms, and will undoubtedly enhance the performance of physicians towards clinical decisions using data driven insights.

8) Distribution Shifts: The problem of data distribution shifts is yet another major challenge and perhaps one of the most challenging problems to solve [136]. In clinical practice, training and testing data distributions can diverge due to many reasons, e.g., medical data is generated by different institutions using different devices for patients having complicated cases.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

TABLE II: Summary of the state of the art data secure and privacy preserving methods in healthcare settings.

Authors	Goal	Method	ML Model(s)	Medical Dataset(s)
David et al. [119]	Security & Privacy	Commodity based cryptography.	Hyperplane decision and Naive Bayes classifiers.	N/A
Zhu et al. [62]		Polynomial aggregation and multi-party random masking.	SVM with nonlinear kernel.	N/A
Jagielski et al. [120]		Proposed an algorithm names as TRIM to defend poisoning attacks.	Linear Regression	Anticoagulant drug Warfarin
Liu et al. [121]		XMPP server and several mobile devices.	Proposed a DL framework,	Human Activity Recognition
Malathi et al. [122]		Paillier homomorphic encryption.	Naïve Bayes, SVM, Neural Network, and FKNN–CBR	Indian Liver Patient
Takabi et al. [123]		Homomorphic encryption.	DNN	15 datasets from UCI repository.
Kim et al. [124]		Homomorphic encryption based secure logistic regression.	Logistic Regression	Five medical datasets having binary classes.
Bogdanov et al. [125]		Multi-party computation	Principal component analysis (PCA)	Genomics data
Min et al. [126]		Reinforcement learning (RL) based privacy aware offloading method.	Reinforcement learning (RL)	Data from medical IoT sensors.
Beaulieu et al. [127]		Distributed ML using differential privacy.	N/A	The eICU and The Cancer Genome Atlas databases.
Zhu et al. [128]		Encryption using random masking technique.	Non-linear support vector machine (SVM)	Pima Indians diabetes database.
Choudhury et al. [129]		Differential privacy and federated learning.	SVM, Perceptron, and logistic regression.	MIMIC III database
Liu et al. [130]		Federated learning.	Three layer neural network.	The eICU database.

Due to this issue, ML/DL models developed using available public databases (by the scientific community and academicians) do not give expected performance when deployed in an actual clinical environment. Distribution shifts are frequent in the medical domain, in particular, medical imagining where different protocols and parameter choices can result in images of significantly different distributions. ML models are typically trained under the principle of empirical risk minimization (ERM) which provides good learning bounds and guarantees if its assumptions are satisfied. For instance, one of the foremost and strong assumptions is that both the training and test datasets are derived from a similar domain (i.e., data distributions). However, this assumption is not valid in practice, and models trained under such an assumption fail to generalize to other domains In contrast, the life-critical nature of clinical applications demands a smooth and safe operation of ML/DL techniques.

9) Updating Hospital Infrastructure is Hard: Healthcare IT systems are mostly proprietary and operate in silos, which results in the revision, fixing, and update of software being costly and time-consuming. It has been reported in the literature that in 2013, the majority of hospitals were using the ninth version of the international classification of disease (ICD) system—even though a revised version (i.e., ICD-10) was released as early as 1990 [22]. The difficulties in updating hospital software infrastructure can raise many vulnerabilities with the use of modern tools like ML/DL systems.

IV. Secure, Private, and Robust ML for Healthcare: Solutions

In this section, we present an overview of various proposed methods to ensure secure, private, and robust ML for healthcare applications. A summary of articles focused on the topic of "secure and privacy-preserving ML for healthcare" is presented in Table II and various approaches for secure, private, and robust ML are described next. In addition, a taxonomy of commonly used approaches for secure, private, and robust ML is presented in Figure 6 and described individually next.

A. Privacy-Preserving ML

Preserving the privacy of the user in healthcare is paramount, as it is a user-centric application and involves the collection of personal data and any breach of privacy can lead to unavoidable consequences. Preserving privacy means that ML model training and inference should not reveal any additional information about the subjects from whom data was collected. In general, ML/DL requires training data stored on a central repository (e.g., cloud) that may include the users' private data which raises various threats and to address such concerns data anonymization techniques are used. However, it has been reported in the literature that meaningful information can be inferred about individuals' private data even when the data is anonymized [137].

Various efforts in the literature have addressed the privacy issues with the use of ML. Three different protocols for the two-server model are presented in [138], where the private data is distributed among two non-colluding servers by the data owners and then those servers train the ML models on the joint data by following secure two-party computation (2PC). Furthermore, different techniques have been proposed to perform secure arithmetic operations in the secure multi-party computational environment and alternatives to nonlinear activation functions used in ML models such as softmax and sigmoid are also proposed. Similarly, various techniques for privacy-preserving ML such as cryptographic and differential privacy approaches are discussed in [109]. Here we briefly discuss the widely used methods for preserving privacy.

1) Cryptographic Approaches: Cryptographic approaches are used in the scenarios where the ML model requires encrypted data (for training and testing purposes) from multiple

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

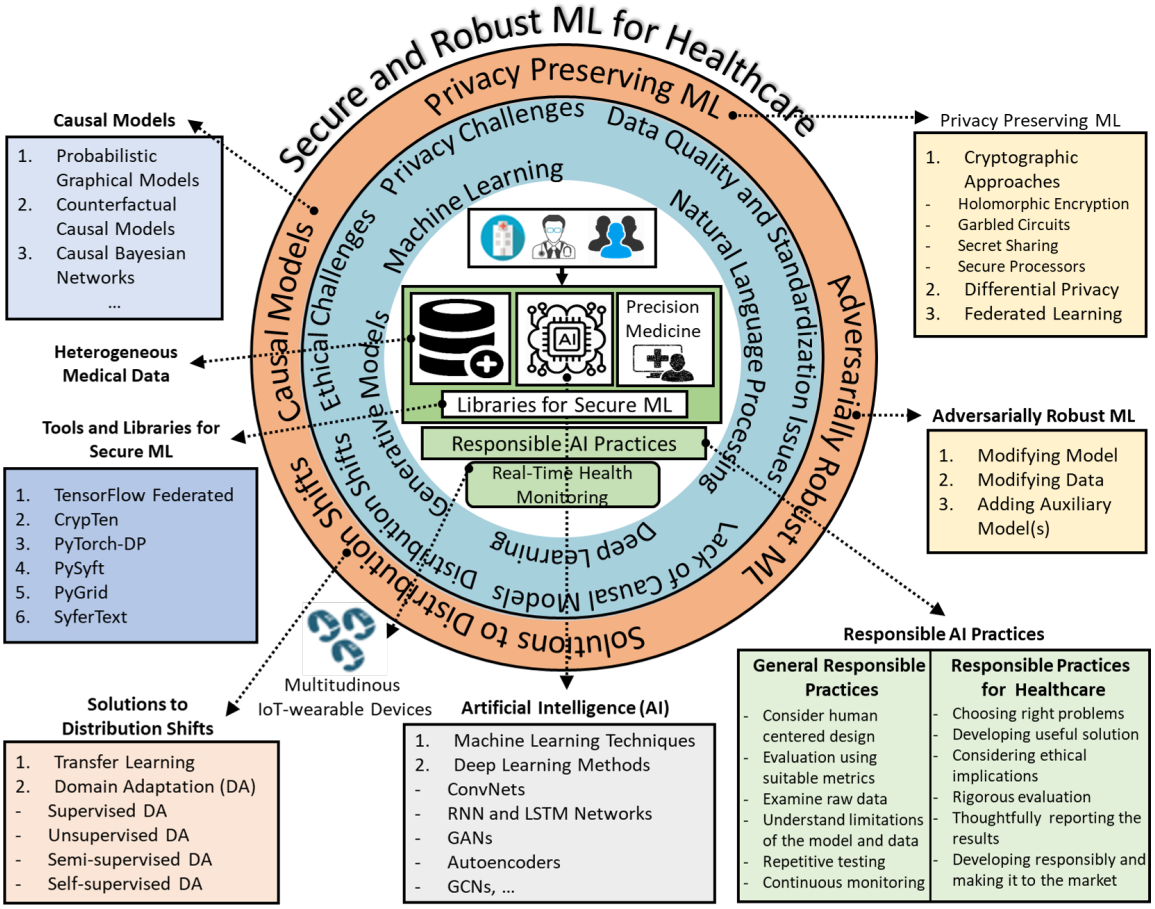

Fig. 6: A taxonomy of commonly used approaches for secure, private, and robust ML.

parties. The widely used methods include homomorphic encryption, secret sharing, garbled circuits, and secure processors which are briefly described next.

(a) *Homomorphic Encryption:* It enables computations on encrypted data with operations such as addition and multiplication which can be used as a basis for computing complex functions. Typically, the data is encrypted using ciphertext and public keys of the original data owners.

(b) *Garbled Circuits:* Garbled circuits are used in cases where two parties (let's assume Alice and Bob) want to get results computed using their private data. Alice will send the function in the form of the garbled circuit along with her input. After obtaining the garbled version of his input from Alice in oblivious fashion, Bob will use his garbled input with the garbled circuit to get the result of the required function and can share it with Alice, if required. The use of homomorphic encryption and garbled circuits to build cryptographic blocks for developing three classification techniques; namely, Naïve Bayes, decision trees, and hyperplane decision is presented in [144], where the goal is to protect ML models and new samples submitted for inference.

(c) *Secret Sharing:* The strategy of distributing secret among multiple parties while holding a "share" of the secret is known as secret sharing. The secret can only be reconstructed when all individual shares are combined; otherwise, they are unuseful. In some settings, the secret is reconstructed using t shares (where t is a threshold value) that will not require all shares to be combined. A secret sharing paradigm for computing privacy-preserving parallelized principal component analysis (PCA) is presented in [125]. In a similar study [142], a protocol is developed using the "secret sharing" strategy for aggregating model updates received from multiple input parties, the updates are used for training of the ML model. A privacy-preserving emotion recognition framework is presented in [143]. Authors used a multi-secret sharing scheme for transmitting audio-visual data collected from users using edge devices to the cloud where a CNN and sparse autoencoder were applied for feature extraction and support vector machine (SVM) was used for emotion recognition.

(d) *Secure Processors:* Secure processors were originally developed by rogue software to ensure the confidentiality and integrity of sensitive code from unauthorized access at higher privilege levels. However, these processors are being utilized in privacy-preserving computation, e.g., Intel SGXprocessor. For instance, Ohrimenko et al. developed an SGX-processor-based data oblivious system for k-mean clustering, decision trees, SVM, and matrix factorization [146]. The key idea was to enable collaboration between multiple data owners running the ML task on an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

TABLE III: The comparisons of different techniques that can be used for privacy-preserving machine learning (ML).

Technique	Methodology	Papers	Advantage (s)	Limitation (s)
Homomorphic Encryption	Computations are performed on encrypted data that is encrypted using different cryptographic approaches.	[139], [140], [141]	Can be used for outsourcing computations on private data.	Slow and computationally intensive and can only receive input from one entity.
Multi-Party Computation	Computations are performed on secret inputs from multiple parties.	[125], [142], [143]	Fast and less overhead and can receive input from multiple parties and ensures the correctness of input and privacy.	It becomes slow with a large number of participating parties.
Garbled Circuits	Garbled circuits are used in cases where two parties want to get results computed using their private data.	[144], [145], [138]	Secure computation on the private data of multiple parties.	Low latency, as it requires the computation of expensive operations.
Secure Processors	Collaboration between multiple data owners is performed through an SGX-enabled data center.	[146], [147], [148]	Adversaries can get control over data and software except the SGX-processor being used for computations.	Adversaries can get control over data and software.
Differential Privacy	Random statistical noise is added to each attribute, to protect privacy.	[127], [149], [150]	Highly practical, as no computational overhead is involved because no encryption is performed.	Addition of noise effects precision and has some limitations from security perspectives.
Federated Learning	A shared model is collaboratively trained from distributed data without sharing the data itself.	[151], [152], [130]	Less communication overhead, as local data is not required to be transmitted and enables collaborative learning.	Parameters optimization in federated learning is challenging.

SGX-enabled data center. All types of communications between the data owners and the enclave were performed by establishing independently a secure channel (i.e., an individual channel for each data owner).

2) Differential Privacy: Differential privacy refers to the mechanism of adding perturbation into the datasets to protect private data. The idea of adding adequate noise in the database for preserving privacy was first introduced by C. Dwork in 2006 [153]. Differential privacy constitutes a strong standard for guaranteeing privacy for algorithms performing analysis on aggregate databases and it is defined in terms of the application-specific concept of neighbor datasets [154]. Differential privacy is particularly useful for applications like healthcare due to its several properties such as group privacy, composability, and robustness to auxiliary information. Group privacy implies elegant degradation of privacy guarantees when datasets contain correlated samples. Whereas, composability enables modularity of the algorithmic design, i.e., when individual components are differentially private. Robustness to auxiliary information means that the privacy of the system will not be affected by the use of any side's information that is known to the adversary. To avoid privacy breaches, the researchers can also explore encrypted and noisy datasets for building ML empowered healthcare applications [155].

Various approaches for differential privacy have been proposed in the literature, e.g., private aggregation of teacher ensembles (PATE) for private ML [156], differentially private stochastic gradient descent (DP-SGD) algorithm [154], moments accountant [157], hyperparameter selection [158], Laplace [159] and exponential noise differential privacy mechanisms [160], [161]. For instance, privacy-preserving distributed DL for clinical data using differential privacy that incorporates the idea of cyclical weight transfer is presented in [127].

3) Federated Learning: The idea of federated learning (FL) has been recently proposed by Google Inc. [162]. In FL, a shared ML model is built using distributed data from multiple devices where each device trains the model using its local data and then shares the model parameters with the central model without sharing its actual data. An FL-based decentralized scheme using iterative cluster primal-dual splitting (cPDS) algorithm to predict hospitalization requiring patients using large-scale EHR of heart-related diseases is presented in [151]. In [152], simple vanilla, U-shaped, and vertically partitioned data-based configurations for split learning DL models are presented. The proposed framework is named SplitNN that does not require sharing of patients' critical data with the server. A framework of federated autonomous deep learning (FADL) using distributed EHR is presented in [130]. A comparison of different privacy preserving techniques discussed above is presented in Table III.

B. Countermeasures Against Adversarial Attacks

In the recent literature, countermeasures against adversarial attacks are categorized into three classes: (1) modifying model; (2) modifying data; and (3) adding an auxiliary model(s) [163]. A taxonomy of such methods is presented in Figure 7 and are discussed next.

1) Modifying Model: The modifying model includes methods that modify the parameters or features of the trained ML model, widely used methods include the following:

- *Defensive Distillation:* The distillation of neural networks was first introduced by Hinton et al. as a method for transferring the knowledge from a larger model to a smaller one [164]. The notion of network distillation was then adopted by Papernot et al. to defend against adversarial attacks, also known as defensive distillation [165]. The authors used the predicted labels of the first model as the labels of the input sample to the original DL model. This strategy increases the robustness of the DL model to considerably small perturbations. However, in a later study, Carlini and Wagner demonstrated that their proposed adversarial attack (named as C&W attack) evaded the defensive distillation method [166].
- *Network Verification:* The techniques verifying certain properties of DL models in response to input samples are known as network verification methods. The key goal is to restrain adversarial examples while checking whether the input satisfied or violated certain properties. In [167], such a method is proposed that uses ReLU activation and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

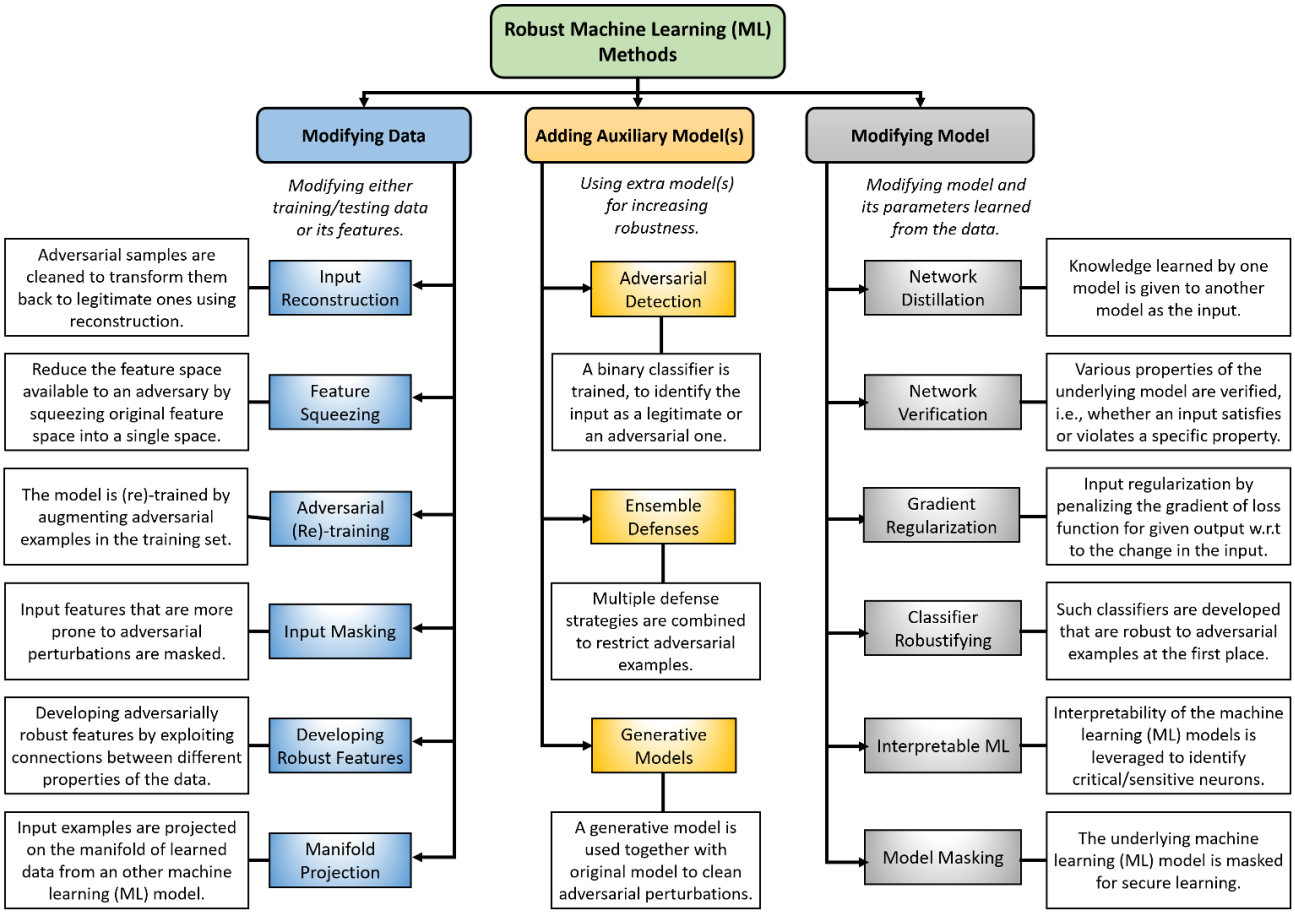

Fig. 7: Taxonomy of Adversarial Defenses (Modified from [163]). Defenses are categorized into three categories: (1) Modifying Data; (2) Modifying Model; and (3) Adding Auxiliary Model(s).

satisfiability modulo theory (SMT) to make deep models resilient against adversarial attacks.

- *Gradient Regularization:* The idea of using input gradient regularization for defending adversarial examples was proposed by Ross et al. [168]. They trained the differentiable models by regularizing the variation in the results with respect to the change in the input due to which small adversarial perturbations were not able to affect the output of DL models. However, this method increases the complexity of the training process by a factor of two.
- *Classifier Robustifying:* In this method, classification models are developed that are robust to adversarial attacks rather than building a detection strategy for such attacks. In [169], authors exploited the uncertainty around the adversarial examples and proposed a hybrid model by utilizing Gaussian processes (GPs) with RBF kernels on top of DNNs to make them robust against adversarial attacks. In a similar study, a robust model is proposed for MNIST classification that uses analysis by synthesis through learned class-conditional data distribution.
- *Interpretable ML:* It includes those methods that aim at explaining and interpreting the outcomes of ML/DL models for robustifying them against adversarial attacks. An approach utilizing the interpretability of deep models for the detection of adversarial examples for face recognition task is presented in a recent study [170]. The key aspect of this method is that it identifies critical neurons for the individual task by initiating a bi-directional correspondence reasoning between the model's parameters and its attributes. The activation values of the identified neurons are then increased to augment the reasoning part and activation values of other neurons are decreased to mask the uninterpretable part. However, Nicholas Carlini demonstrated that the aforementioned method utilizing the interpretability of deep models is not resilient to untargeted adversarial examples generated using L_∞ norm [171].
- *Masking ML Model:* In a recent study [172], a method for secure learning is presented in which the problem of adversarial ML is formulated as learning and masking problem. The masking of the deep model was performed by introducing noise in the logit output which successfully deafened attacks with low distortions.

2) Modifying Data: It includes those methods that aim at either modifying the data or its features, commonly used methods are described next:

- *Adversarial (Re-)training:* This is a very basic method that was originally proposed by Goodfellow et al. for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

making deep models robust to adversarial examples [98]. In this method, the ML/DL models are trained (or re-trained) using an augmented training set that includes adversarial examples. Various studies have used this method for evaluating the robustness of DL classifiers using different datasets, e.g., MNIST [173] and ImageNet [167]. However, it has been reported in the literature that this method fails to defend against iterative adversarial perturbation generation methods like basic iterative method (BIM) [174].

- *Input Reconstruction:* The method of transforming adversarial examples into legitimate ones by cleaning the adversarial noise is known as input reconstruction. The transformed samples have no harmful effect on the inference of deep models. In [175], denoising autoencoder is used for the cleaning of adversarial examples.
- *Feature Squeezing:* Xu et al. [176] proposed feature squeezing as a defense method against adversarial examples by squeezing the input feature space that an adversary can exploit to construct adversarial examples. To reduce the available feature space to an adversary, authors combined heterogeneous feature vectors in the original feature space into a single space. The feature squeezing was performed at two levels: (1) smoothing the spatial domain using local and non-local operations and (2) minimizing color bit depth. Moreover, the performance evaluation of the proposed defense was performed using eleven state of the art adversarial perturbation generation methods using three benchmark datasets (i.e., CIFAR10, MNIST, and ImageNet). However, in a later study, the aforementioned defense method was found to be less effective [177].
- *Features Masking:* The method of feature masking was proposed by Gao et al. [178] that aims at masking the most sensitive features of the input that are susceptible to adversarial perturbations. The authors added a masking layer right before the classification layer (i.e., softmax) that sets the corresponding weights of the sensitive neurons to zero.
- *Developing Adversarially Robust Features:* To develop adversarially robust features, the connections between the metric of interest and natural spectral geometrical property of the dataset has been leveraged in [179]. Furthermore, the authors provided empirical evidence about the effectiveness of using a spectral approach for developing adversarially robust features.
- *Manifold Projection:* The method of projecting input samples on the manifold learned by the generative models is known as manifold projection. Song et al. [180] used generative models to clean adversarial noise (perturbations) from the adversarial images then the cleaned images are used as the input to the non-modified model. In a similar study [181], generative adversarial networks (GANs) are used for cleaning of adversarial noise.

3) Adding Auxiliary Model(s): In these methods, additional auxiliary ML/DL models are integrated to robustify the mainstream model, commonly used methods that fall into this class are described in the following paragraphs:

- *Adversarial Detection:* In this method, an additional binary classifier is trained to distinguish between the adversarial and original samples that can be regarded as the detector model [182], [183]. In [184], a simple DNN based detector model is used for the detection of adversarial examples. Similarly, an outlier class has been introduced during the training of a deep model that helps the model to detect the adversarial examples belonging to the outlier class.
- *Ensembling Defenses:* The literature suggests that adversarial examples can be constructed in multi-faceted fashion. Therefore, to develop an efficient defense method against such adversarial examples, multiple defense strategies can be integrated sequentially or in parallel [185]. The PixelDefend method is an excellent example of an ensemble defense method in which authors used an ensemble of two methods, i.e., adversarial detection and input reconstruction [180]. However, it has been shown that the ensemble of weak defenses does not necessarily increase the robustness of DL models to adversarial attacks [177].
- *Using Generative ML Models:* The idea of defending against adversarial attacks by utilizing generative models was firstly presented by Goodfellow et al. [98], however, in the same study the authors presented an alternative hypothesis of ensemble training and articulated that generative training is not sufficient. In [186], adversarial examples are cleaned using GAN that was trained on the same dataset. In a similar study [187], a framework named Defense-GAN is presented that is trained on the distribution of legitimate samples. Defense-GAN finds similar output during the testing phase without adversarial perturbations that are given as input to the original DL model. A summary of the state of the art defense methods for making DL models resilient to adversarial attacks is presented in Table IV.

C. Causal Models for Healthcare

Asking causal questions in healthcare is a very challenging yet important approach and ideally, causal inferences require experiments. But it in healthcare this not always possible, e.g., if we want to figure out what will happen if a person takes drug A instead of B, we can not experiment it directly on the patient which is unethical and can have unintended consequences. Alternatively, retrospective observational data is leveraged to train models for making counterfactual predictions of what we would have observed if we had run an experiment [189]. Causality can be deemed in two foundational ways, i.e., potential outcomes and causal graphical models that require manipulating reality. In predictive healthcare, potential outcomes can be treatment, action, and interventions. If the total number of possible treatments is T then we can have T possible outcomes and the unit of observation will be a patient who gets one of the T treatments.

In the literature, different approaches have been presented for providing causal inferences and reasoning in healthcare

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

TABLE IV: Summary of state-of-the-art defense methods for mitigating adversarial attacks.

Author	Proposed Defense Methodology	Attack Method(s)	Dataset(s)	Defense Accuracy
Gu et al. [175]	Proposed the use of denoising autoencoders (DAEs) for removing adversarial noise.	To construct adversarial examples, added additive Gaussian noise into original images.	MNIST	99.1%
Xu et al. [176]	Proposed feature space reduction which is available to an adversary.	Evaluated the proposed defense against different adversarial examples crafting methods.	MNIST, CIFAR-10, and ImageNet	MNIST (62.7%), CIFAR-10 (77.27%), and ImageNet (68.11%)
Gao et al. [178]	Proposed the masking of unnecessary neurons in the model.	Fast Gradient Search Method (FGSM)	CIFAR-10	10% increase in accuracy under adversarial attack.
Papernot et al. [165]	Proposed defense distillation for improving adversarial robustness.	Gradient based adversarial example generation method.	MNIST & CIFAR10	2.56% increase in robustness with distiallation temperature of 50.
Garg et al. [179]	Used spectral property for generating adversarially robust features.	Considered L_2 minimization based adversarial perturbations.	MNIST	N/A
Song et al. [180]	Proposed to recover adversarial examples by projecting them back to the manifold of original training data.	Evaluated five different adversarial examples generation methods.	Fashion MNIST and CIFAR-10	Achieved the increase in accuracy under adversarial attack: 21% for Fashion MNIST and 38% for CIFAR-10.
Goodfellow et al. [98]	Trained the model by adding using both original images and adversarial examples.	Fast Gradient Sign (FGSM)	MNIST	17.9% fall in model error.
Metzend et al. [184]	Trained a deep neural network (DNN) for detection of adversarial examples, i.e., binary classification into normal and adversarial examples.	FGSM, BIM, and DeepFool.	CIFAR-10	80% adversarial detectability for all attacks.
Schott et al. [188]	Variational autoencoder (VAE) for generating clean images.	Used four different adversarial example generation methods that uses $L_2(\epsilon = 1.5)$.	MNIST	80%
Ross et al. [168]	Proposed input gradient regularization for training a model that is resilient to adversarial attacks.	FGSM, TGSM, and JSMA.	Used three datasets, i.e., MNIST, SVHN, and notMNIST	MNIST (100%), SVHN ($\sim$90%), and notMNIST (100%).

using classical models. For instance, the Gaussian processes based counterfactual causal model has been presented in [189] and in a similar study, authors introduced the counterfactual Gaussian process (CGP) for predicting counterfactual future progression and argued that counterfactual model can provide reliable decision support [114]. The use of probabilistic graphical models to analyze causality in health conditions for identification sleep apnea, Alzheimers disease, and heart diseases is presented in [190]. A comprehensive review of graphical causal models can be found in this recent study [191].

D. Solutions to Address Distribution Shifts

To cater with data distribution shift problem various techniques have been proposed in the literature (e.g., transfer learning and domain adaptation), which are described next.

1) Transfer Learning: The requirement of the availability of a large-scale dataset for training DL models capable of providing high performances can be partially mitigated using transfer learning. Transfer learning is a technique in which a model trained on a larger dataset is re-trained (fine-tuned) on the application-specific dataset (relatively smaller in size to the first one). The aim is to transfer knowledge learned by the model from one domain (data distribution) to the other domain [192]. However, transfer learning can be problematic for healthcare applications due to the requirement of sufficiently large data for first training and good quality data annotated by expert clinicians such as radiologists for domain-specific training.

2) Domain Adaptation: Domain adaptation is the method of learning a DL model by considering a shift between the training (often called as source domain) and test (often called as target domain) data distributions, i.e., source domain and target domain distributions are different. Domain adaptation is a special case of transfer learning that can be particularly useful for medical image analysis tasks such as MRI segmentation [136], [193], chest X-ray classification [194], and multi-class Alzheimer disease classification [195], etc. Different facets of domain adaptation have been proposed in the literature and can be broadly categorized as supervised, unsupervised, semi-supervised, and self-supervised domain adaptation methods which are described below. Please note that the definition of domain adaptation is ambiguous since it may refer to labeled data being available in the source or target domains and the definitions provided below for each method are mostly used in the literature [196].

(a) *Supervised Domain Adaptation:* This method is similar to a supervised learning strategy with the only difference of different distributions for source domain and target domain data. Supervised domain adaptation is particularly useful when a labeled data is available for the target domain and generally, the source domain also has labeled data.

(b) *Unsupervised Domain Adaptation:* In unsupervised domain adaptation, source domain data is labeled and target domain data is unlabeled. An unsupervised domain adaptation method using reverse flow and adversarial training for generating synthetic medical images is presented in [197]. In addition, the authors used self-regularization for preserving clinically-relevant features.

(c) *Semi-supervised Domain Adaptation:* In semi-supervised domain adaptation, labeled source data and partial labeled target domain.

(d) *Self-supervised Domain Adaptation:* Self-supervised domain adaptation methods aims at learning visual models without manual labeling by training generic models using auxiliary relatively simple tasks (known as pretext tasks). The supervision is provided by modifying the original

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

visual content (e.g., a set of images) according to known transformations (e.g., rotation) and then the model is trained to predict such transformations that serve as labels for the pretext tasks [198].

E. Towards Responsible ML

In this section, we provide different methods for ensuring responsible ML and we start by enlisting general responsible AI practices.

1) General Responsible AI Practices: The following are some recommended AI practices to ensure effective and reliable AI systems.[2]

- *Consider human-centered design approach*: To have a large impact on the system being developed, it is important to consider the characteristics of the users for true recommendations.
- *Evaluate training and monitoring using suitable metrics*: Instead of using multiple metrics for evaluation of model training, ensure that the metric is appropriate for the context and goals of the systems and consider users' feedback in terms of surveys.
- *Examine your raw data*: The biases and abnormalities in the datasets (e.g., missing values, class imbalance, and incorrect labels) are directly reflected by the learned ML models. To ensure the efficacy of the learning process, careful examination of the raw dataset is necessary while respecting the privacy concerns.
- *Understand limitations of the model and dataset*: It is crucial to understand the capability and limitations of the ML model and dataset, e.g., a model trained for detecting correlations cannot be used for inferences.
- *Repetitive Testing*: Once developed, ML systems should be tested again and again to ensure that they are working as intended. Rigorous tests should be performed to understand how the individual components of the ML system interact with each other. Other similar tests include testing for input drifts, using gold standard datasets, incorporating a larger sample base, and using quality checking mechanisms.
- *Continuous Monitoring and Updating*: To ensure the efficient performance of the ML systems deployed in real-time settings, continued monitoring and updating are required to identify and fix various issues encountered in realistic settings.

2) Responsible ML for Healthcare: ML/DL techniques have a great potential for clinical applications (e.g., radiologist-level pneumonia detection [11] and dermatologist-level classification of skin cancer [13], etc.) but their limited adoption in actual clinical settings indicates that these methods are not yet optimal and not ready for clinical deployment. In a recent study [199], Wiens et al. have provided a roadmap towards safe, meaningful, and responsible ML for healthcare and argued that ML deployment in any field should be carried out by an interdisciplinary team that may include different stakeholders from multi disciplines, i.e., knowledge experts, decision-makers, and users. Examples for an interdisciplinary team having different stakeholders in the healthcare ecosystem are presented in Table V. In addition, the authors also identified critical steps to be followed/considered when designing, testing, and deploying ML solutions for healthcare applications that include: (1) choosing the right problems; (2) developing a useful solution; (3) considering ethical implications; (4) rigorously evaluating the model; (5) thoughtfully reporting results; (6) deploying responsibly; and (7) making it to market.

TABLE V: Examples for interdisciplinary teams having different stakeholders from multiple domains. (Adopted from [199])

Stakeholder Category	Examples
Knowledge experts	Clinical experts, e.g., radiologist and dermatologists.
	Health information and technology experts
	ML researchers, e.g., ML engineers and data scientists.
	Implementation experts
Decision-makers	Institutional leadership
	Hospital administrators
	State and federal government
	Regulatory agencies
Users	Physicians
	Nurses
	Laboratory technicians
	Patients
	Caretakers, e.g., friends and family.

F. Tools and Libraries for Secure and Private ML

The main strength of ensuring secure ML relies on the development of security tools and algorithms. To ensure the security and privacy of ML models and data, various tools and libraries have been released so far. For example, *TensorFlow Federated*,[3] which is an open-source framework for distributed ML/DL that enables training of a global shared model in a federated environment without sharing clients' local data. CrypTen[4] is a framework for secure and privacy-preserving ML built on PyTorch that provides secure computing techniques for ML/DL model training and inference using encrypted data and PyTorch-DP[5]–a framework of PyTorch for training DL models with differential privacy. Similarly, *OpenMined*[6]–an open-source community offers various tools and libraries for building privacy-preserving ML models which are briefly described below.

- *PySyft*[7] is python library for encrypted and privacy preserving ML. It extends PyTorch, TensorFlow, and Keras and supports differential privacy, federated learning, multi-party computation, and homomorphic encryption.
- *PyGrid*[8] is a platform built on PySyft that provides a peer-to-peer network to collectively train ML models.
- *SyferText*[9] is a privacy preserving framework for NLP tasks.

[2]https://ai.google/responsibilities/responsible-ai-practices/

[3]https://www.tensorflow.org/federated
[4]https://github.com/facebookresearch/CrypTen
[5]https://github.com/facebookresearch/pytorch-dp
[6]https://www.openmined.org/
[7]https://github.com/OpenMined/PySyft
[8]https://github.com/OpenMined/PyGrid
[9]https://github.com/OpenMined/SyferText

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3013489, IEEE Reviews in Biomedical Engineering

V. Open Research Issues

In this section, various open research issues related to the domain of secure, robust, and private ML for healthcare that require further research attention are presented.

A. Interpretable ML

Although the advancement in ML/DL research has provided significant performance improvements over the previous state of the art methods in terms of performance metrics such as accuracy, precision, recall, and f1-measure, these advancements have made the learning process of modern models very complex and are usually deployed as a black-box. These black-box methods fail at providing rational or insights as well as at explaining their learning behavior and thought process for making predictions [200]. The aforementioned problem is termed as the interpretability problem of ML in the literature, which is defined as the ability to describe the internal processes of an ML system in a human-understandable manner.

Moreover, interpretability of ML/DL techniques is required to ensure algorithmic fairness, robustness, and generalization based on potentially dispersed data collected from a heterogeneous population. This can eventually help in the smooth deployment and functionality of ML/DL systems in realistic settings. For a critical application like healthcare, the ML/DL model is expected to be highly accurate and understandable at the same time. Moreover, it has been argued that clinical integration of AI models will require interpretability [201]. To perform an interpretation of ML models, questions about the fairness of model's predictions, transparency, and accountability are considered and interpretation is performed using explanation methods for justifying predictions of the model using visual, textual, or features information. For instance, Bach et al. presented a pixel-wise explanation method that uses layer-wise relevance propagation for interpreting the predictions of non-linear classifiers [202]. Similarly, for interpretation of classifiers' predictions, Ribeiro et al. presented a framework named LIME and proposed two methods for interpretability, i.e., learning a local model around the predictions and representing predictions and their explanations in a non-redundant way using a submodular optimization approach. In [203], the use of reinforcement learning (RL) is proposed to build interpretable decision support systems for heart patients and it learns what is interpretable to each user by their interactions. One yet common method for interpreting/explaining deep models, in particular, CNN is the use of saliency maps [204], [205]. These methods are particularly focused on general applications, however, more research that is specifically focused on the interpretation of ML/DL systems used in healthcare applications is required.

B. Machine Learning on the Edge

The advancements in ML research have revolutionized traditional healthcare (as discussed in earlier sections). Healthcare services will increasingly adopt the utilization of IoT devices and wearable sensors in the future, particularly with the evolution of smart cities and portable medical devices, e.g., portable MRI scanner. With such proliferation, there is a pressing need for pushing ML models training and inference on edge devices. This introduces unique challenges such as limited hardware and processing capabilities, etc. Moreover, this is crucial for portal medical devices that are utilized for patients in critical care as they cannot be moved to fixed medical equipment in the hospital. The research on enabling ML on edge devices (a.k.a fog) is in the early stages of development and requires further attention from the research community. The development of this field will enable to monitor patients in a critical situation and eventually enable continuous behavioral monitoring for improving individuals' life-style and timely detection of diseases.

C. Handling Dataset Annotation

To increase the performance of ML/DL models, one natural strategy is to acquire more labeled training data. This requires that radiologists and medical experts spend their valuable time manually annotating medical data, e.g., medical images, signals, and reports. Another important aspect is devising true validation sets that will evaluate the performance of the ML/DL models and expose the limitations of these models. Therefore, manual annotation of samples into respective categories is time consuming, costly, and a tidy process. Automatic approaches should be developed to address this issue and one such technique is active learning which can be used to annotate unlabelled data samples.

Data from multiple sources should be considered when performing annotation for specific clinical applications because single-source data might lack precise structured labels [115]. The integration of multiple source data is an important application of ML in healthcare [206], which is known as phenotyping [207]. NLP techniques and recurrent deep models can be used for extracting and integrating rich information from unstructured clinical notes to augment the capacity of data annotators.

D. Distributed Data Management and ML

In healthcare settings, the data is generated in a distributed fashion, i.e., across different departments within a hospital and even across different hospitals. This necessitates the efficient management and sharing of distributed data for clinical analysis purposes, particularly using ML/DL models. In general, for developing ML/DL models, it is assumed that complete training and validation datasets are centrally available and easily accessible. Therefore, there is an increasing demand to develop methods for distributed data management and ML.

E. Fair and Accountable ML

The literature on analyzing the security and robustness of ML/DL approaches reveals that the outcomes of these models lack fairness and accountability [163]. Whereas ensuring the fairness and accountability of predictions in life-critical applications like healthcare are of paramount importance, the *fairness* property ensures that the ML model should not favor certain cases over others. Such discrimination mainly

arises due to biases in the training data. On the other hand, *accountability* property is concerned with the interpretation of the predictions. Fairness and accountability will assist in developing models robust to biases and imperfections such as past clinical practices

F. Model-Driven ML

Although ML, AI, and big data are immensely useful tools for healthcare, these tools are not panacea and it is important to be aware of the associated caveats and pitfalls [200]. Failing to realize this, one can easily fall prey to the dangerous dogma that data once available in abundance must and will speak for itself and can handle hypothesis generation as well—which in clinical terms would mean that data mining is sufficient and independent of the need of clinical interpretation, external validation, and understanding of data's provenance [208]. To avoid the various problems that can arise from improper use of ML in healthcare, it is important to combine data-driven methods with hypothesis-driven or model-based methods (based on subject matter knowledge) and to bring scientific rigor in these studies. Properly designed experiments are also necessary for deriving causal explanations. Avenues for developing secure and robust ML solutions for healthcare that are scientifically robust and rigorous requires further attention from the community.

VI. Conclusions

The use of machine learning (ML)/deep learning (DL) models for clinical applications has great potential to transform traditional healthcare service delivery. However, to ensure a secure and robust application of these models in clinical settings, different privacy and security challenges should be addressed. In this paper, we provided an overview of such challenges by formulating the ML pipeline in healthcare and by identifying different sources of vulnerabilities in it. We also discussed potential solutions to provide secure and privacy-preserving ML for security-critical applications like healthcare. Finally, we presented different open research problems that require further investigation.

Acknowledgement

The publication of this article was funded by the Qatar National Library (QNL). The statements made herein are solely the responsibility of the authors.

References

[1] S. Latif, J. Qadir, S. Farooq, and M. Imran, "How 5G wireless (and concomitant technologies) will revolutionize healthcare?" *Future Internet*, vol. 9, no. 4, p. 93, 2017.

[2] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D. N. Metaxas, and X. S. Zhou, "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1332–1343, 2016.

[3] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[4] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *International Conference on Information Processing in Medical Imaging*. Springer, 2015, pp. 588–599.

[5] J. Schlemper, J. Caballero, J. V. Hajnal, A. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for mr image reconstruction," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 647–658.

[6] J. Mehta and A. Majumdar, "Rodeo: robust de-aliasing autoencoder for real-time medical image reconstruction," *Pattern Recognition*, vol. 63, pp. 499–510, 2017.

[7] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[8] K. Bourzac, "The computer will see you now," *Nature*, vol. 502, no. 3, pp. S92–S94, 2013.

[9] L. Xing, E. A. Krupinski, and J. Cai, "Artificial intelligence will soon change the landscape of medical physics research and practice," *Medical physics*, vol. 45, no. 5, pp. 1791–1793, 2018.

[10] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[11] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[12] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[13] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[14] S. Latif, M. Asim, M. Usman, J. Qadir, and R. Rana, "Automating motion correction in multishot mri using generative adversarial networks," *Published as Workshop Paper at 32nd Conference on Neural Information Processing Systems (NIPS 2018)*, 2018.

[15] X.-W. Chen and X. Lin, "Big data deep learning: challenges and perspectives," *IEEE access*, vol. 2, pp. 514–525, 2014.

[16] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2017.

[17] K. Papangelou, K. Sechidis, J. Weatherall, and G. Brown, "Toward an understanding of adversarial examples in clinical trials," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 35–51.

[18] H. Kim, D. C. Jung, and B. W. Choi, "Exploiting the vulnerability of deep learning-based artificial intelligence models in medical imaging: Adversarial attacks," *Journal of the Korean Society of Radiology*, vol. 80, no. 2, pp. 259–273, 2019.

[19] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, 2019.

[20] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[21] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 6103–6113.

[22] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[23] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[24] A. K. Pandey, P. Pandey, K. Jaiswal, and A. K. Sen, "Datamining clustering techniques in the prediction of heart disease using attribute selection method," *heart disease*, vol. 14, pp. 16–17, 2013.

[25] K. Polat and S. Güneş, "Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system," *Applied Mathematics and computation*, vol. 189, no. 2, pp. 1282–1291, 2007.

[26] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," in *Supervised and Unsupervised Learning for Data Science*. Springer, 2020, pp. 3–21.

[27] M. N. Sohail, J. Ren, and M. Uba Muhammad, "A euclidean group assessment on semi-supervised clustering for healthcare clinical implications based on real-life data," *International journal of environmental research and public health*, vol. 16, no. 9, p. 1581, 2019.

[28] A. Zahin, R. Q. Hu *et al.*, "Sensor-based human activity recognition for smart healthcare: A semi-supervised machine learning," in *International Conference on Artificial Intelligence for Communications and Networks*. Springer, 2019, pp. 450–472.

[29] D. Mahapatra, "Semi-supervised learning and graph cuts for consensus based medical image segmentation," *Pattern recognition*, vol. 63, pp. 700–709, 2017.

[30] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert, "Semi-supervised learning for network-based cardiac mr image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 253–260.

[31] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 2, no. 4.

[32] H.-C. Kao, K.-F. Tang, and E. Y. Chang, "Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[33] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, p. 484, 2016.

[34] A. Collins and Y. Yao, "Machine learning approaches: Data integration for disease prediction and prognosis," in *Applied Computational Genomics*. Springer, 2018, pp. 137–141.

[35] P. Afshar, A. Mohammadi, and K. N. Plataniotis, "Brain tumor type classification via capsule networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3129–3133.

[36] W. Zhu, C. Liu, W. Fan, and X. Xie, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 673–681.

[37] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, p. 395, 2012.

[38] Z. Wang, A. D. Shah, A. R. Tate, S. Denaxas, J. Shawe-Taylor, and H. Hemingway, "Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning," *PLoS One*, vol. 7, no. 1, p. e30412, 2012.

[39] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A machine learning-based framework to identify type 2 diabetes through electronic health records," *International journal of medical informatics*, vol. 97, pp. 120–127, 2017.

[40] B. Nestor, M. McDermott, W. Boag, G. Berner, T. Naumann, M. C. Hughes, A. Goldenberg, and M. Ghassemi, "Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks," *arXiv preprint arXiv:1908.00690*, 2019.

[41] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, no. 11, p. 226, 2018.

[42] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing mri," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 72–82, 2008.

[43] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2016, pp. 241–246.

[44] Y. Chen, Y. Xie, Z. Zhou, F. Shi, A. G. Christodoulou, and D. Li, "Brain mri super resolution using 3d deep densely connected neural networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 739–742.

[45] K. Sirinukunwattana, S. e Ahmed Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images." *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[46] H. Wang, A. C. Roa, A. N. Basavanhally, H. L. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi, "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," *Journal of Medical Imaging*, vol. 1, no. 3, p. 034003, 2014.

[47] Y. Yu, H. Lin, J. Meng, X. Wei, H. Guo, and Z. Zhao, "Deep transfer learning for modality classification of medical images," *Information*, vol. 8, no. 3, p. 91, 2017.

[48] J. Antony, K. McGuinness, N. E. O'Connor, and K. Moran, "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 1195–1200.

[49] E. Kim, M. Corte-Real, and Z. Baloch, "A deep semantic mobile application for thyroid cytopathology," in *Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations*, vol. 9789. International Society for Optics and Photonics, 2016, p. 97890A.

[50] M. F. Stollenga, W. Byeon, M. Liwicki, and J. Schmidhuber, "Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation," in *Advances in neural information processing systems*, 2015, pp. 2998–3006.

[51] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[52] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.

[53] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *Journal of digital imaging*, pp. 1–15, 2019.

[54] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, "Low-dose ct with a residual encoder-decoder convolutional neural network," *IEEE transactions on medical imaging*, vol. 36, no. 12, pp. 2524–2535, 2017.

[55] M. Usman, S. Latif, M. Asim, and J. Qadir, "Motion corrected multishot mri reconstruction using generative networks with sensitivity encoding," *arXiv preprint arXiv:1902.07430*, 2019.

[56] F. E.-Z. A. El-Gamal, M. Elmogy, and A. Atwan, "Current trends in medical image registration and fusion," *Egyptian Informatics Journal*, vol. 17, no. 1, pp. 99–124, 2016.

[57] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *Ieee Access*, vol. 6, pp. 9375–9389, 2017.

[58] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration–a deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, 2017.

[59] S. Miao, Z. J. Wang, and R. Liao, "A cnn regression approach for real-time 2d/3d registration," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1352–1363, 2016.

[60] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[61] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid, "Medical image retrieval using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 8–20, 2017.

[62] J. Zech, M. Pain, J. Titano, M. Badgeley, J. Schefflein, A. Su, A. Costa, J. Bederson, J. Lehar, and E. K. Oermann, "Natural language–based machine learning models for the annotation of clinical radiology reports," *Radiology*, vol. 287, no. 2, pp. 570–580, 2018.

[63] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[64] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9049–9058.

[65] Y. Xue, T. Xu, L. R. Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 457–466.

[66] V. Jindal, "Integrating mobile and cloud for ppg signal selection to monitor heart rate during intensive physical exercise," in *Proceedings of the International Conference on Mobile Software Engineering and Systems*. ACM, 2016, pp. 36–37.

[67] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31 314–31 338, 2015.

[68] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PloS one*, vol. 12, no. 4, p. e0174944, 2017.

[69] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 01, p. 1, 2017.

[70] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[71] H.-Y. Ma, Z. Zhou, S. Wu, Y.-L. Wan, and P.-H. Tsui, "A computer-aided diagnosis scheme for detection of fatty liver in vivo based on ultrasound kurtosis imaging," *Journal of medical systems*, vol. 40, no. 1, p. 33, 2016.

[72] Z. Zhang *et al.*, "Reinforcement learning in clinical medicine: a method to optimize dynamic treatment regime over time," *Annals of translational medicine*, vol. 7, no. 14, 2019.

[73] A. Raghu, "Reinforcement learning for sepsis treatment: Baselines and analysis," 2019.

[74] H. Suresh, "Clinical event prediction and understanding with deep neural networks," Ph.D. dissertation, Massachusetts Institute of Technology, 2017.

[75] C.-S. Rau, P.-J. Kuo, P.-C. Chien, C.-Y. Huang, H.-Y. Hsieh, and C.-H. Hsieh, "Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models," *PloS one*, vol. 13, no. 11, p. e0207192, 2018.

[76] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[77] O. Ren, A. E. Johnson, E. P. Lehman, M. Komorowski, J. Aboab, F. Tang, Z. Shahn, D. Sow, R. Mark, and L.-w. Lehman, "Predicting and understanding unexpected respiratory decompensation in critical care using sparse and heterogeneous clinical data," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2018, pp. 144–151.

[78] A. K. Jha, "The promise of electronic records: around the corner or down the road?" *Jama*, vol. 306, no. 8, pp. 880–881, 2011.

[79] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than english: opportunities and challenges," *Journal of biomedical semantics*, vol. 9, no. 1, p. 12, 2018.

[80] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, and H. Xu, "Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines," *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 331–336, 2017.

[81] D. S. Wallace, "The role of speech recognition in clinical documentation," *Nuance Communications*, 2018, access on: 14-Dec-2019. [Online]. Available: https://www.hisa.org.au/slides/hic18/wed/SimonWallace.pdf

[82] M. Ghassemi, J. H. Van Stan, D. D. Mehta, M. Zañartu, H. A. Cheyne II, R. E. Hillman, and J. V. Guttag, "Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: Initial results for vocal fold nodules," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1668–1675, 2014.

[83] C. Pou-Prom and F. Rudzicz, "Learning multiview embeddings for assessing dementia," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2812–2817.

[84] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[85] J. B. Andre, B. W. Bresnahan, M. Mossa-Basha, M. N. Hoff, C. P. Smith, Y. Anzai, and W. A. Cohen, "Toward quantifying the prevalence, severity, and cost associated with patient motion during clinical mr examinations," *Journal of the American College of Radiology*, vol. 12, no. 7, pp. 689–695, 2015.

[86] A. K. Manrai, G. Bhatia, J. Strymish, I. S. Kohane, and S. H. Jain, "Medicine's uncomfortable relationship with math: calculating positive predictive value," *JAMA internal medicine*, vol. 174, no. 6, pp. 991–993, 2014.

[87] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1721–1730.

[88] X. A. Li, A. Tai, D. W. Arthur, T. A. Buchholz, S. Macdonald, L. B. Marks, J. M. Moran, L. J. Pierce, R. Rabinovitch, A. Taghian *et al.*, "Variability of target and normal structure delineation for breast cancer radiotherapy: an rtog multi-institutional and multiobserver study," *International Journal of Radiation Oncology* Biology* Physics*, vol. 73, no. 3, pp. 944–951, 2009.

[89] F. Xia and M. Yetisgen-Yildiz, "Clinical corpus annotation: challenges and strategies," in *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*, 2012.

[90] S. T. M. Ataky, J. de Matos, A. d. S. Britto Jr, L. E. Oliveira, and A. L. Koerich, "Data augmentation for histopathological images based on gaussian-laplacian pyramid blending," *IEEE International Joint Conference on Neural Networks (IJCNN 2020), Glasgow, UK*, 2020.

[91] J. F. R. Rochac, L. Liang, N. Zhang, and T. Oladunni, "A gaussian data augmentation technique on highly dimensional, limited labeled data for multiclass classification using deep learning," in *2019 Tenth International Conference on Intelligent Control and Information Processing (ICICIP)*. IEEE, 2019, pp. 145–151.

[92] N. Carlini and D. Wagner, "Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples," *arXiv preprint arXiv:1711.08478*, 2017.

[93] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *29th International Conference on Machine Learning*, 2012, pp. 1807–1814.

[94] S. Alfeld, X. Zhu, and P. Barford, "Data poisoning attacks against autoregressive models," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[95] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," *arXiv preprint arXiv:1611.03814*, 2016.

[96] T. J. Pollard, I. Chen, J. Wiens, S. Horng, D. Wong, M. Ghassemi, H. Mattie, E. Lindmeer, and T. Panch, "Turning the crank for machine learning: ease, at what expense?" *The Lancet Digital Health*, vol. 1, no. 5, pp. e198–e199, 2019.

[97] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1322–1333.

[98] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[99] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[100] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, 2017, pp. 506–519.

[101] M. Usama, J. Qadir, A. Al-Fuqaha, and M. Hamdi, "The adversarial machine learning conundrum: Can the insecurity of ml become the achilles' heel of cognitive networks?" *arXiv preprint arXiv:1906.00679*, 2019.

[102] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[103] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, "Systematic poisoning attacks on and defenses for machine learning in healthcare," *IEEE journal of biomedical and health informatics*, vol. 19, no. 6, pp. 1893–1905, 2014.

[104] S. G. Finlayson, H. W. Chung, I. S. Kohane, and A. L. Beam, "Adversarial attacks against medical deep learning systems," *arXiv preprint arXiv:1804.05296*, 2018.

[105] N. Karimian, M. Tehranipoor, D. Woodard, and D. Forte, "Unlock your heart: Next generation biometric in resource-constrained healthcare systems and iot," *IEEE Access*, vol. 7, pp. 49 135–49 149, 2019.

[106] U. Kumar, E. Tripathi, S. P. Tripathi, and K. K. Gupta, "Deep learning for healthcare biometrics," in *Design and Implementation of Healthcare Biometric Systems*. IGI Global, 2019, pp. 73–108.

[107] S.-K. Kim, C. Y. Yeun, E. Damiani, and N.-W. Lo, "A machine learning framework for biometric authentication using electrocardiogram," *IEEE Access*, vol. 7, pp. 94 858–94 868, 2019.

[108] L. Wieclaw, Y. Khoma, P. Fałat, D. Sabodashko, and V. Herasymenko, "Biometrie identification from raw ecg signal using deep learning techniques," in *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 1. IEEE, 2017, pp. 129–133.

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

[110] J. Chaudhry, "Securing healthcare data using biometric authentication," *Security and Privacy in Communication Networks*, p. 123, 2018.

[111] A. Mohsin, A. Zaidan, B. Zaidan, S. A. bin Ariffin, O. Albahri, A. Albahri, M. Alsalem, K. Mohammed, and M. Hashim, "Real-time medical systems based on human biometric steganography: A systematic review," *Journal of medical systems*, vol. 42, no. 12, p. 245, 2018.

[112] Y. Sun, F. P.-W. Lo, and B. Lo, "Security and privacy for the internet of medical things enabled healthcare systems: A survey," *IEEE Access*, vol. 7, pp. 183 339–183 355, 2019.

[113] J. Zhang and E. Bareinboim, "Fairness in decision-making—the causal explanation formula," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[114] P. Schulam and S. Saria, "Reliable decision support using counterfactual models," in *Advances in Neural Information Processing Systems*, 2017, pp. 1697–1708.

[115] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, and R. Ranganath, "Opportunities in machine learning for healthcare," *arXiv preprint arXiv:1806.00388*, 2018.

[116] E. Begoli, T. Bhattacharya, and D. Kusnezov, "The need for uncertainty quantification in machine-assisted medical decision making," *Nature Machine Intelligence*, vol. 1, no. 1, p. 20, 2019.

[117] A. Khademi, S. Lee, D. Foley, and V. Honavar, "Fairness in algorithmic decision making: An excursion through the lens of causality," in *The World Wide Web Conference*. ACM, 2019, pp. 2907–2914.

[118] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Advances in Neural Information Processing Systems*, 2017, pp. 656–666.

[119] B. David, R. Dowsley, R. Katti, and A. C. Nascimento, "Efficient unconditionally secure comparison and privacy preserving machine learning classification protocols," in *International Conference on Provable Security*. Springer, 2015, pp. 354–367.

[120] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018, pp. 19–35.

[121] M. Liu, H. Jiang, J. Chen, A. Badokhon, X. Wei, and M.-C. Huang, "A collaborative privacy-preserving deep learning system in distributed mobile environment," in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2016, pp. 192–197.

[122] D. Malathi, R. Logesh, V. Subramaniyaswamy, V. Vijayakumar, and A. K. Sangaiah, "Hybrid reasoning-based privacy-aware disease prediction support system," *Computers & Electrical Engineering*, vol. 73, pp. 114–127, 2019.

[123] H. Takabi, E. Hesamifard, and M. Ghasemi, "Privacy preserving multi-party machine learning with homomorphic encryption," in *29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.

[124] M. Kim, Y. Song, S. Wang, Y. Xia, and X. Jiang, "Secure logistic regression based on homomorphic encryption: Design and evaluation," *JMIR medical informatics*, vol. 6, no. 2, p. e19, 2018.

[125] D. Bogdanov, L. Kamm, S. Laur, and V. Sokk, "Implementation and evaluation of an algorithm for cryptographically private principal component analysis on genomic data," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 5, pp. 1427–1432, 2018.

[126] M. Min, X. Wan, L. Xiao, Y. Chen, M. Xia, D. Wu, and H. Dai, "Learning-based privacy-aware offloading for healthcare iot with energy harvesting," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4307–4316, 2018.

[127] B. K. Beaulieu-Jones, W. Yuan, S. G. Finlayson, and Z. S. Wu, "Privacy-preserving distributed deep learning for clinical data," *Machine Learning for Health (ML4H) Workshop at NeurIPS*, 2018.

[128] H. Zhu, X. Liu, R. Lu, and H. Li, "Efficient and privacy-preserving online medical prediagnosis framework using nonlinear svm," *IEEE journal of biomedical and health informatics*, vol. 21, no. 3, pp. 838–850, 2016.

[129] O. Choudhury, A. Gkoulalas-Divanis, T. Salonidis, I. Sylla, Y. Park, G. Hsu, and A. Das, "Differential privacy-enabled federated learning for sensitive health data," *arXiv preprint arXiv:1910.02578*, 2019.

[130] D. Liu, T. Miller, R. Sayeed, and K. Mandl, "Fadl: Federated-autonomous deep learning for distributed electronic health record," *Machine Learning for Health (ML4H) Workshop at NeurIPS*, 2018.

[131] L. Faes, S. K. Wagner, D. J. Fu, X. Liu, E. Korot, J. R. Ledsam, T. Back, R. Chopra, N. Pontikos, C. Kern *et al.*, "Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study," *The Lancet Digital Health*, vol. 1, no. 5, pp. e232–e242, 2019.

[132] N. u. . h. O'Reilly, "Challenges to AI in healthcare accessed online: 16 oct 2019."

[133] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" in *Advances in Neural Information Processing Systems*, 2018, pp. 3539–3550.

[134] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "Practical guidance on artificial intelligence for health-care data," *The Lancet Digital Health*, vol. 1, no. 4, pp. e157–e159, 2019.

[135] T. Panch, H. Mattie, and L. A. Celi, "The "inconvenient truth" about ai in healthcare," *Npj Digital Medicine*, vol. 2, no. 1, pp. 1–3, 2019.

[136] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," *NeuroImage*, vol. 194, pp. 1–11, 2019.

[137] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large datasets (how to break anonymity of the netflix prize dataset)," *University of Texas at Austin*, 2008.

[138] P. Mohassel and Y. Zhang, "Secureml: A system for scalable privacy-preserving machine learning," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 19–38.

[139] Y. Aono, T. Hayashi, L. Wang, S. Moriai *et al.*, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1333–1345, 2017.

[140] S. Carpov, T. H. Nguyen, R. Sirdey, G. Constantino, and F. Martinelli, "Practical privacy-preserving medical diagnosis using homomorphic encryption," in *2016 IEEE 9th International Conference on Cloud Computing (CLOUD)*. IEEE, 2016, pp. 593–599.

[141] D. Kahrobaei, A. Wood, and K. Najarian, "Homomorphic encryption for machine learning in medicine and bioinformatics," *ACM Comput. Surv.*, 2020.

[142] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 1175–1191.

[143] M. S. Hossain and G. Muhammad, "Emotion recognition using secure edge and cloud computing," *Information Sciences*, vol. 504, pp. 589–601, 2019.

[144] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine learning classification over encrypted data." in *NDSS*, vol. 4324, 2015, p. 4325.

[145] A. Gribov, K. Horan, J. Gryak, K. Najarian, V. Shpilrain, R. Soroushmehr, and D. Kahrobaei, "Medical diagnostics based on encrypted medical data," in *International Conference on Bio-inspired Information and Communication*. Springer, 2019, pp. 98–111.

[146] O. Ohrimenko, F. Schuster, C. Fournet, A. Mehta, S. Nowozin, K. Vaswani, and M. Costa, "Oblivious multi-party machine learning on trusted processors," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 619–636.

[147] F. Shaon, M. Kantarcioglu, Z. Lin, and L. Khan, "Sgx-bigmatrix: A practical encrypted data analytic framework with trusted processors," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1211–1228.

[148] R. Kunkel, D. L. Quoc, F. Gregor, S. Arnautov, P. Bhatotia, and C. Fetzer, "Tensorscone: A secure tensorflow framework using intel sgx," *arXiv preprint arXiv:1902.04413*, 2019.

[149] Z. Sun, Y. Wang, M. Shu, R. Liu, and H. Zhao, "Differential privacy for data and model publishing of medical data," *IEEE Access*, vol. 7, pp. 152 103–152 114, 2019.

[150] W. Huang, S. Zhou, Y. Liao, and H. Chen, "An efficient differential privacy logistic classification mechanism," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 620–10 626, 2019.

[151] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *International journal of medical informatics*, vol. 112, pp. 59–67, 2018.

[152] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," *Published as Workshop Paper at 32nd Conference on Neural Information Processing Systems (NIPS 2018)*, 2018.

[153] C. Dwork, "Differential privacy," *Encyclopedia of Cryptography and Security*, pp. 338–340, 2011.

[154] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 308–318.

[155] M. McDermott, S. Wang, N. Marinsek, R. Ranganath, M. Ghassemi, and L. Foschini, "Reproducibility in machine learning for health," *Presented at the Internation Conference on Learning Representative (ICLR) 2019 Reproducibility in Machine Learning Workshop*, 2019.

[156] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, "Scalable private learning with pate," *International Conference on Learning Representations (ICLR)*, 2018.

[157] Y.-X. Wang, B. Balle, and S. Kasiviswanathan, "Subsampled r\'enyi differential privacy and analytical moments accountant," *arXiv preprint arXiv:1808.00087*, 2018.

[158] H. B. McMahan, G. Andrew, U. Erlingsson, S. Chien, I. Mironov, N. Papernot, and P. Kairouz, "A general approach to adding differential privacy to iterative training procedures," *NeurIPS 2018 workshop on Privacy Preserving Machine Learning*, 2018.

[159] N. Phan, X. Wu, H. Hu, and D. Dou, "Adaptive laplace mechanism: Differential privacy preservation in deep learning," in *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2017, pp. 385–394.

[160] F. McSherry and K. Talwar, "Mechanism design via differential privacy." in *FOCS*, vol. 7, 2007, pp. 94–103.

[161] C. Dwork and F. D. McSherry, "Exponential noise distribution to optimize database privacy and output utility," Jul. 14 2009, uS Patent 7,562,071.

[162] H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," *Proceedings of the 20 th International Conference on Artificial Intelligence and Statistics (AISTATS) JMLR: WCP volume 54*, 2017.

[163] A. Qayyum, M. Usama, J. Qadir, and A. Al-Fuqaha, "Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward," *arXiv preprint arXiv:1905.12762*, 2019.

[164] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Deep Learning Workshop, NIPS*, 2014.

[165] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 582–597.

[166] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, 2017, pp. 3–14.

[167] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient SMT solver for verifying deep neural networks," in *International Conference on Computer Aided Verification*. Springer, 2017, pp. 97–117.

[168] A. S. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[169] J. Bradshaw, A. G. d. G. Matthews, and Z. Ghahramani, "Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks," *arXiv preprint arXiv:1707.02476*, 2017.

[170] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks meet interpretability: Attribute-steered detection of adversarial samples," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 7717–7728.

[171] N. Carlini, "Is ami (attacks meet interpretability) robust to adversarial examples?" *arXiv preprint arXiv:1902.02322*, 2019.

[172] L. Nguyen, S. Wang, and A. Sinha, "A learning and masking approach to secure learning," in *International Conference on Decision and Game Theory for Security*. Springer, 2018, pp. 453–464.

[173] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári, "Learning with a strong adversary," *arXiv preprint arXiv:1511.03034*, 2015.

[174] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[175] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *Published as a Workshop Paper at International Conference on Learning Representative (ICLR)*, 2015.

[176] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*, 2018. [Online]. Available: http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018_03A-4_Xu_paper.pdf

[177] W. He, J. Wei, X. Chen, N. Carlini, and D. Song, "Adversarial example defense: Ensembles of weak defenses are not strong," in *11th USENIX Workshop on Offensive Technologies (WOOT)'17)*, 2017.

[178] J. Gao, B. Wang, Z. Lin, W. Xu, and Y. Qi, "Deepcloak: Masking deep neural network models for robustness against adversarial samples," *arXiv preprint arXiv:1702.06763*, 2017.

[179] S. Garg, V. Sharan, B. Zhang, and G. Valiant, "A spectral view of adversarially robust features," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 10 159–10 169.

[180] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=rJUYGxbCW

[181] G. Jin, S. Shen, D. Zhang, F. Dai, and Y. Zhang, "APE-GAN: adversarial perturbation elimination with GAN," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3842–3846.

[182] J. Lu, T. Issaranon, and D. Forsyth, "Safetynet: Detecting and rejecting adversarial examples robustly," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 446–454.

[183] D. Gopinath, G. Katz, C. S. Pasareanu, and C. Barrett, "Deepsafe: A data-driven approach for checking adversarial robustness in neural networks," *arXiv preprint arXiv:1710.00486*, 2017.

[184] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," *International Conference on Learning Representations (ICLR)*, 2017.

[185] F. Tramer, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *International Conference on Learning Representations (ICLR)*, 2018.

[186] G. K. Santhanam and P. Grnarova, "Defending against adversarial attacks by leveraging an entire GAN," *arXiv preprint arXiv:1805.10652*, 2018.

[187] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in *International Conference on Learning Representations (ICLR)*, 2018.

[188] L. Schott, J. Rauber, M. Bethge, and W. Brendel, "Towards the first adversarially robust neural network model on mnist," *In Seventh International Conference on Learning Representations (ICLR 2019)*, pp. 1—17, 2019.

[189] P. Schulam and S. Saria, "What-if reasoning with counterfactual gaussian processes," *History*, vol. 100, p. 120, 2017.

[190] R. C. Sato and G. T. K. Sato, "Probabilistic graphic models applied to identification of diseases," *Einstein (São Paulo)*, vol. 13, no. 2, pp. 330–333, 2015.

[191] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. 10, 2019.

[192] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[193] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.-E. de Leeuw, C. M. Tempany, B. van Ginneken *et al.*, "Transfer learning for domain adaptation in mri: Application in brain lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 516–524.

[194] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1038–1042.

[195] C. Wachinger, M. Reuter, A. D. N. Initiative *et al.*, "Domain adaptation for alzheimer's disease diagnostics," *Neuroimage*, vol. 139, pp. 470–479, 2016.

[196] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *arXiv preprint arXiv:1812.02849*, 2019.

[197] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2572–2581, 2018.

[198] J. Xu, L. Xiao, and A. M. López, "Self-supervised domain adaptation for computer vision tasks," *IEEE Access*, vol. 7, pp. 156 694–156 706, 2019.

[199] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed *et al.*, "Do no harm: a roadmap for responsible machine learning for health care," *Nature medicine*, vol. 25, no. 9, pp. 1337–1340, 2019.

[200] S. Latif, A. Qayyum, M. Usama, J. Qadir, A. Zwitter, and M. Shahzad, "Caveat emptor: The risks of using big data for human development," *IEEE Technology and Society Magazine*, vol. 38, no. 3, pp. 82–90, 2019.

[201] X. Jia, L. Ren, and J. Cai, "Clinical implementation of ai technologies will require interpretable ai models," *Medical physics*, 2019.

[202] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[203] O. Lahav, N. Mastronarde, and M. van der Schaar, "What is interpretable? using machine learning to design interpretable decision-support systems," *Machine Learning for Health (ML4H) Workshop at NeurIPS*, 2018.

[204] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze, "Evaluating saliency map explanations for convolutional neural networks: a user study," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 2020, pp. 275–285.

[205] H. Li, Y. Tian, K. Mueller, and X. Chen, "Beyond saliency: understanding convolutional neural networks from saliency prediction on layer-wise relevance propagation," *Image and Vision Computing*, vol. 83, pp. 70–86, 2019.

[206] Y. Halpern, S. Horng, Y. Choi, and D. Sontag, "Electronic medical record phenotyping using the anchor and learn framework," *Journal of the American Medical Informatics Association*, vol. 23, no. 4, pp. 731–740, 2016.

[207] R. L. Richesson, W. E. Hammond, M. Nahm, D. Wixted, G. E. Simon, J. G. Robinson, A. E. Bauck, D. Cifelli, M. M. Smerek, J. Dickerson *et al.*, "Electronic health records based phenotyping in next-generation clinical trials: a perspective from the nih health care systems collaboratory," *Journal of the American Medical Informatics Association*, vol. 20, no. e2, pp. e226–e231, 2013.

[208] D. Belgrave, J. Henderson, A. Simpson, I. Buchan, C. Bishop, and A. Custovic, "Disaggregating asthma: big investigation versus big data," *Journal of Allergy and Clinical Immunology*, vol. 139, no. 2, pp. 400–407, 2017.

Adnan Qayyum is currently working towards Ph.D. in Computer Science at the Information Technology University (ITU) Punjab, Pakistan. His research interests include healthcare, deep/machine learning, and security of machine learning. He received the Bachelor's degree in Electrical (Computer) Engineering from COMSATS Institute of Information Technology, Wah, Pakistan, in 2014 and M.S. degree in Computer Engineering (Signal and Image Processing) from the University of Engineering and Technology, Taxila, Pakistan, in 2016.

Junaid Qadir is the director of the IHSAN—ICTD; Human Development; Systems; Big Data Analytics; Networks—Research Lab and the Chairperson of the Electrical Engineering Department at the Information Technology University (ITU) of Punjab in Lahore, Pakistan. His primary research interests are in the areas of computer systems and networking, applied machine learning, using ICT for development (ICT4D); and engineering education. He has published more than 100 peer-reviewed articles at various high-quality research venues including more than 50 impact-factor journal publications at top international research journals including IEEE Communication Magazine, IEEE Journal on Selected Areas in Communication (JSAC), IEEE Communications Surveys and Tutorials (CST), and IEEE Transactions on Mobile Computing (TMC). He was awarded the highest national teaching award in Pakistan—the higher education commission's (HEC) best university teacher award—for the year 2012-2013. He has been appointed as ACM Distinguished Speaker for a three-year term starting from 2020. He is a senior member of IEEE and ACM.

Muhammad Bilal Dr Muhammad Bilal is Associate Professor of Big Data and Artificial Intelligence (AI) at Big Data Laboratory, University of the West of England (UWE), Bristol. He holds a PhD in Big Data Analytics from UWE, Bristol. During his PhD, he developed a simulation platform for UK largest construction firm (Balfour Beatty) in which hybrid AI models (i.e. Tabular, Vision and Sequence) were operationalised in conjunction with Big Data, Scientific Visualisation, and GIS for automating non-trivial planning and execution tasks in Megaprojects. Dr Bilal has multi-disciplinary research interests that span across fields of Construction Informatics, Digital Health, Image Processing, Scientific Visualisation, AI, Computer Vision, Natural Language Processing, Geospatial Analysis Mining and Web-of-Data technologies. His inclination in these areas technologies is for solving critical real-life problems related to workers' productivity and efficiency through disruptive innovations and digital transformations. He has also led the development of several large-scale software solutions pertaining to financials and healthcare (PACS radiology). Dr Bilal has vast expertise in designing and executing collaborative research development projects. So far, he has completed RD projects of £3.7 Million at Big Data Lab in collaboration with leading UK businesses. Currently, Dr Bilal is leading Real-Time Emissions Sensing (REVIS) project (£1.79 Million) that involves IoT, GIS, Big Data, Stream Analytics and Advanced Visualisations. He has also authored more than 50 research articles at high-impact journals and international conferences.

Ala Al-Fuqaha [S'00-M'04-SM'09] received Ph.D. degree in Computer Engineering and Networking from the University of Missouri-Kansas City, Kansas City, MO, USA. He is currently a professor at the Information and Computing Technology division, college of Science and Engineering, Hamad Bin Khalifa University (HBKU). His research interests include the use of machine learning in general and deep learning in particular in support of the data-driven and self-driven management of large-scale deployments of IoT and smart city infrastructure and services, Wireless Vehicular Networks (VANETs), cooperation and spectrum access etiquette in cognitive radio networks, and management and planning of software defined networks (SDN). He is a senior member of the IEEE and an ABET Program Evaluator (PEV). He serves on editorial boards of multiple journals including IEEE Communications Letter and IEEE Network Magazine. He also served as chair, co-chair, and technical program committee member of multiple international conferences including IEEE VTC, IEEE Globecom, IEEE ICC, and IWCMC.