(IJIASE) 2024, Vol. No. 10, Jan-Dec **e-ISSN: 2454-9258, p-ISSN: 2454-809X**

Securing Healthcare Systems: Addressing Challenges in Protecting Big Data, AI and ERP Systems

Sunil Kumar Sehrawat
Technical Analyst, Bausch Health Companies

[1]*Date of Receiving: 23 Nov 2023;*
Date of Acceptance: 28 December 2023;
Date of Publication: 15 January 2024

ABSTRACT

The paper briefly introduces big data and its role in healthcare applications. It highlights how big data architecture and techniques are continuously aiding in managing the rapid growth of data in the healthcare industry. Initially, an empirical study was conducted to analyze the impact of big data in the healthcare sector, revealing that significant advancements have been made. However, envisioning the influence of machine learning and big data on healthcare, a complex and evolving field, remains a challenge that engages the audience's curiosity and interest.

It was noted that many researchers implementing machine learning and big data analytics for disease diagnosis still need to address data privacy and security sufficiently. The specific challenges in this context include [specific challenges such as data breaches, unauthorized access, or data misuse]. To address this gap, the paper proposes a novel design for an innovative and secure healthcare information system that leverages machine learning and advanced security mechanisms to manage big data in the medical industry. The innovation involves incorporating an optimal storage and data security layer to ensure data security and privacy. Techniques such as masking encryption, activity monitoring, granular access control, dynamic data encryption, and endpoint validation have been integrated. The proposed hybrid four-layer healthcare model is presented as a more effective system for disease diagnosis using big data.

INTRODUCTION

We live in the age of algorithms, where machine learning (ML) and deep learning (DL) systems have revolutionized numerous industries such as manufacturing, transportation, and governance. Over recent years, DL has achieved state-of-the-art performance in various fields, including computer vision, text analytics, and speech processing. The extensive deployment of ML/DL algorithms across domains, such as social media, has made this technology integral to our daily lives. ML/DL is beginning to significantly impact healthcare—a field traditionally resistant to large-scale technological disruptions.

ML/DL techniques have recently demonstrated outstanding results in diverse tasks, including recognizing body organs from medical images, classifying interstitial lung diseases, detecting lung nodules, reconstructing medical photos, and segmenting brain tumours. It is anticipated that intelligent software will soon assist radiologists and physicians in patient examinations, and ML will transform medical research and practice. Clinical medicine has emerged as an exciting application area for ML/DL models, which have already achieved human-level performance in clinical pathology, radiology, ophthalmology, and dermatology. Some studies even report that DL models outperform human physicians on average. For instance, in 2018, the U.S. Food and Drug Administration (FDA) approved an intelligent diagnostic system to detect diabetes-related eye problems from medical images without human intervention.

Advancements in related technologies such as cloud/edge computing, mobile communication, and big data technology further enhance the potential of ML models for healthcare applications. These combined technologies enable ML/DL to produce highly accurate predictive outcomes and facilitate human-centred intelligent solutions.

[1] *How to cite the article:* Sehrawat S.K (2024); Securing Healthcare Systems: Addressing Challenges in Protecting Big Data, AI and ERP Systems; *International Journal of Innovations in Applied Sciences and Engineering*; Vol 10, 1-16

(IJIASE) 2024, Vol. No. 10, Jan-Dec e-ISSN: 2454-9258, p-ISSN: 2454-809X

Additionally, they offer benefits such as allowing remote healthcare services for rural and low-income areas, potentially revitalizing the healthcare industry.

Despite the impressive performance of DL algorithms, recent studies have raised concerns about the security and robustness of ML models. Szegedy et al. first demonstrated that DL models are vulnerable to carefully crafted adversarial examples. Various types of data and model poisoning attacks have been proposed against DL systems, with numerous defences also suggested in the literature. However, the robustness of these defences could be better, with many failing against specific attacks. The realization that DL models are neither secure nor robust significantly hinders their practical deployment in life-critical applications like predictive healthcare. Researchers have already demonstrated the threat of adversarial attacks on ML-based medical systems.

Ensuring the integrity and security of DL models and health data is crucial for the widespread adoption of ML/DL in the industry. This paper focuses on two key terms—security and robustness—particularly in the context of ML/DL models. Security involves possible threats or attacks on an ML/DL system that influence its behaviour or outcome, while robustness refers to the system's ability to withstand such attacks. Security is analyzed along two dimensions: attacks attempting to control the system or achieve a desired outcome and privacy attacks aimed at learning about the training data. Robustness is analyzed based on the system's ability to survive attacks and its resistance to privacy attacks. Robustness is a relative term, as the effectiveness of a system varies according to the nature of the attack.

This paper presents a comprehensive survey of existing literature on the security and robustness of ML/DL models used in healthcare systems, focusing on the dimensions above. The aim is to provide an in-depth survey of various security challenges associated with applying ML/DL in healthcare and to propose a taxonomy of potential solutions to these issues. Additionally, the paper discusses general challenges and sources of vulnerabilities hindering the safe and robust application of ML/DL in healthcare. It presents potential solutions to address security, privacy, and robustness challenges.

In summary, the specific contributions of this paper are:

1. An overview of diverse literature on applications of ML/DL techniques in four major healthcare tasks: prognosis, diagnosis, treatment, and clinical workflow.

2. A formulation of the ML pipeline for data-driven healthcare applications, identifying different sources of vulnerabilities at each stage that pose security and robustness challenges.

3. An overview of various security and robustness challenges associated with adopting ML/DL models for healthcare applications.

4. A taxonomy of different solutions to ensure secure and robust application of ML/DL techniques in healthcare.

5. Highlighting 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.

APPLICATIONS OF HEALTHCARE USING MACHINE LEARNING

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

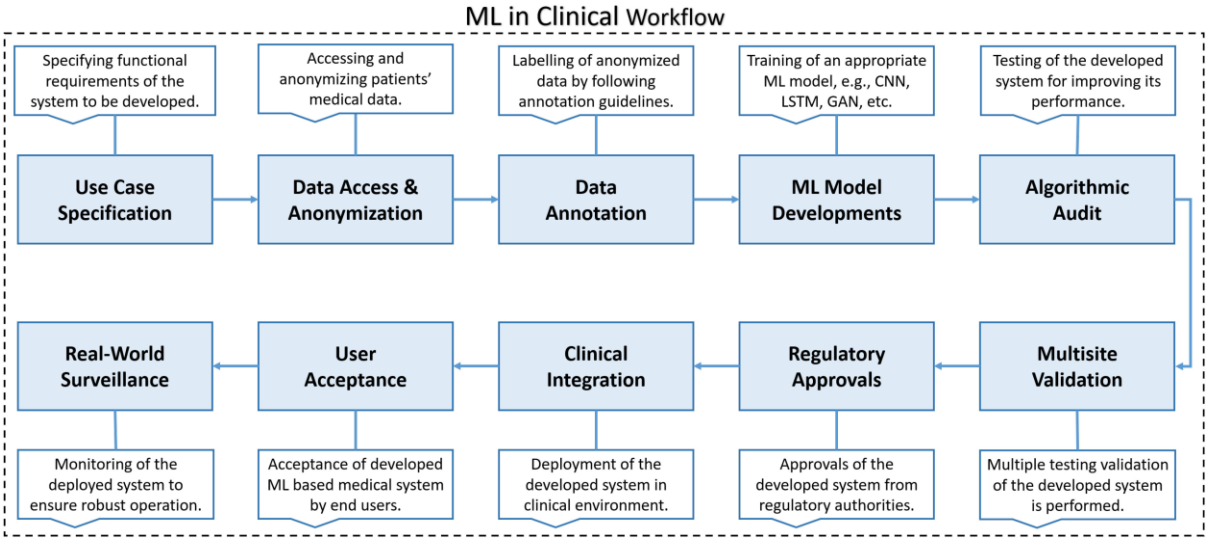

Fig. 1. The illustration of major phases for development of machine learning (ML) based healthcare systems.

A. ML in Healthcare: The Big Picture

Fig. 1 presents a roadmap of the primary phases involved in developing an ML-based healthcare system, underscoring the immense potential of machine learning in revolutionizing healthcare. The major types of ML/DL applicable to healthcare are briefly described below, offering a glimpse into the transformative power of these technologies.

1. Unsupervised Learning: Techniques that utilize unlabelled data fall under unsupervised learning methods. Common examples include clustering data points based on similarity metrics and dimensionality reduction, which projects high-dimensional data into lower-dimensional subspaces (also known as feature selection). Additionally, unsupervised learning can be applied for anomaly detection, such as clustering anomalies [1]. Classic examples in healthcare include predicting heart diseases using clustering and predicting hepatitis using principal component analysis (PCA), a dimensionality reduction technique.

2. Supervised Learning: This Method, which maps the relationship between inputs and outputs using labelled training data, is a key component of ML in healthcare. If the output is discrete, the task is called classification; if it is continuous, it is called regression. Real-world examples of its use in healthcare include classifying various types of lung diseases (nodules) and recognizing different body organs from medical images . In some cases, ML methods are neither purely supervised nor unsupervised, using a combination of labelled and unlabelled data, known as semi-supervised learning. For instance, a semi-supervised learning model could be used to predict the progression of a disease based on a combination of labelled patient data and unlabelled environmental data.

3. Semi-Supervised Learning: This method is particularly beneficial in healthcare, where obtaining sufficient labelled data is often challenging. It is used when both labelled and unlabelled samples are available for training, typically with a small amount of labelled data and a large amount of unlabelled data. Different aspects of semi-supervised learning using various techniques have been proposed in the literature, such as semi-supervised clustering for healthcare data and a semi-supervised approach for activity recognition using sensor data. Examples also include applying semi-supervised learning to medical image segmentation.

4. Reinforcement Learning: Methods that learn a policy function from observations, actions, and rewards over time fall into reinforcement learning (RL). RL has significant potential to transform healthcare applications. Recently, it has been used for context-aware symptom checking for disease diagnosis[2]. The potential of RL in healthcare is further highlighted by its success in the Go game, where a computer using RL combined with supervised and unsupervised learning defeated a human champion.

B. Applications of ML in Healthcare

Healthcare providers grapple with vast amounts of heterogeneous data daily, posing a challenge for traditional methods. However, ML/DL methods offer a practical and effective solution to analyze and process this information, providing actionable insights. Fig. 2 illustrates how various data sources, such as genomics, medical records, social media, and environmental data, can enrich healthcare data. The primary applications of ML/DL in healthcare, including prognosis, diagnosis, treatment, and clinical workflow, are detailed below, demonstrating the tangible benefits of these technologies.

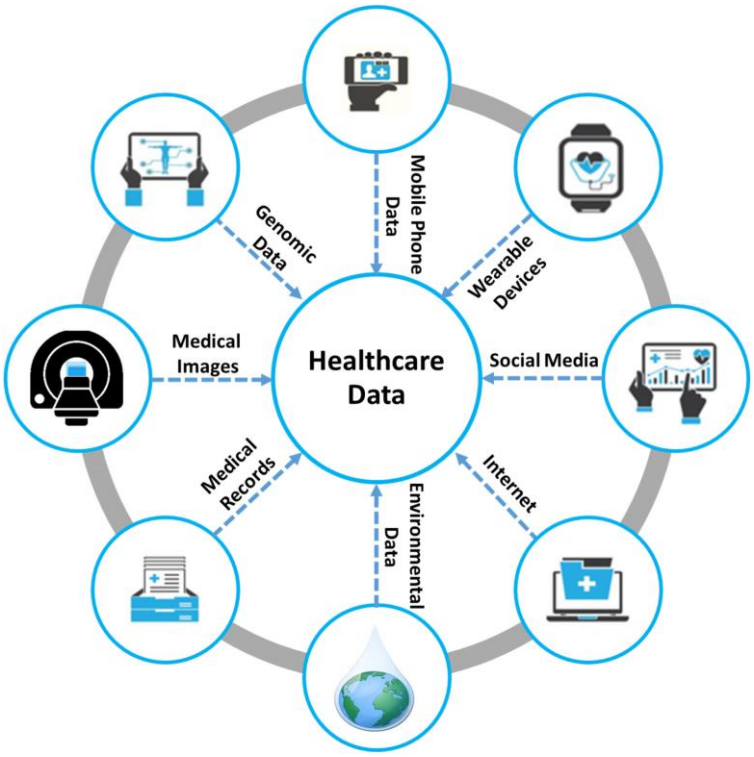

Fig. 2. An example of the diverse sources that go into creating healthcare data.

1. Applications of ML in Prognosis: Prognosis involves predicting the expected progression of a disease in clinical practice. It includes identifying symptoms and signs of a specific disease and determining whether they will worsen, improve, or remain stable over time. Prognosis also involves identifying potential associated health problems and complications, the ability to perform routine activities and the likelihood of survival. Multi -modal patient data (e.g., phenotypic, genomic, proteomic, pathology tests, medical images) can empower ML models to facilitate prognosis, diagnosis, and treatment in clinical settings. ML models have been developed to identify and classify different types of cancers, such as brain tumours and lung nodules. Recent translational research efforts aim to leverage ML for disease prognosis, predicting symptoms, risks, survivability [3], and recurrence and contributing to personalized medicine. However, personalized medicine is still in its early stages, requiring extensive development in bioinformatics, robust validation strategies, and demonstrably practical ML applications to achieve significant translational impact. The future of ML in healthcare holds promise, with potential advancements in areas such as real-time disease monitoring, personalized treatment plans, and improved patient outcomes.

2. Applications of Machine Learning in Treatment:

a) Image Interpretation:

Medical images are essential in routine clinical practice, where expert physicians and radiologists analyze and interpret these images, writing detailed radiology reports for each examined organ. However, generating these reports can be challenging for less experienced radiologists or healthcare providers in rural areas, where healthcare

services might need to improve. The report-writing process is often tedious and time-consuming for seasoned professionals, especially with a high volume of patients. Researchers have explored natural language processing (NLP) and machine learning (ML) techniques to address these issues[3,4]. For instance, a method using NLP for annotating clinical radiology reports has been proposed.

A multi-task ML framework for automatic tagging and describing medical images has also been developed. An end-to-end architecture combining CNN and RNN for thorax disease classification and reporting in chest X-rays has also been presented. Moreover, a novel multi-modal model utilizing CNN and LSTM networks has been created for automatic report generation.

b) ML in Real-Time Health Monitoring:

Real-time monitoring of critical patients is crucial in treatment processes. Continuous health monitoring through wearable devices, IoT sensors, and smartphones is becoming increasingly popular. Typically, health data collected via wearable devices and smartphones is transmitted to the cloud for analysis using ML/DL techniques, with outcomes sent back to the device for appropriate action[7]. A similar system architecture integrating mobile and cloud for heart rate monitoring using PPG signals has been developed. A review of ML techniques for human activity recognition and remote patient monitoring using wearable devices is also available. However, sharing health data with the cloud raises privacy and security concerns, which are discussed in the subsequent section.

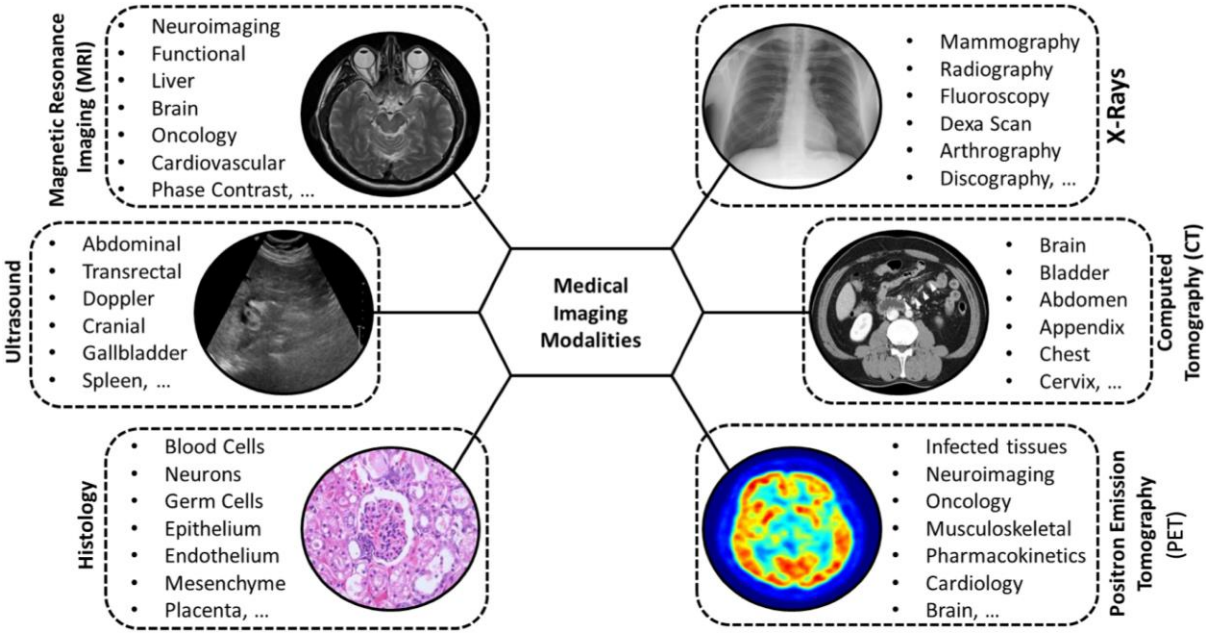

Fig. 3. A classification of frequently utilised medical imaging techniques

APPLICATIONS OF ML IN CLINICAL WORKFLOWS

a) Disease Prediction and Diagnosis:

The early prediction and diagnosis of diseases using medical data is a promising application of ML. Studies have demonstrated the potential of predictive healthcare for timely treatment. For example, cardiovascular risk prediction using various ML algorithms has shown improved prediction efficacy. Additionally, surveys on ML techniques for detecting and diagnosing diseases like diabetes, dengue, hepatitis, heart disease, and liver conditions have been conducted. The potential of ML-based methods for cancer prediction and prognosis has also been highlighted.

(IJIASE) 2024, Vol. No. 10, Jan-Dec **e-ISSN: 2454-9258, p-ISSN: 2454-809X**

b) ML in Computer-Aided Detection or Diagnosis:

Computer-aided detection (CADe) or diagnosis (CADx) systems are designed to assist radiologists by automatically interpreting medical images. These systems leverage ML/DL, traditional computer vision, and image processing techniques. IBM's Watson is a notable example of a CADx system that integrates various ML techniques. Additionally, automated detection tasks in medical imaging, such as fatty liver detection in ultrasound kurtosis imaging, exemplify CADe/CADx systems [5].

c) Clinical Reinforcement Learning:

Reinforcement learning (RL) aims to develop policy functions for precise decision-making in uncertain environments to maximize accumulated rewards. In clinical medicine, RL can optimize diagnosis and treatment for patients with varying characteristics. Different RL techniques (e.g., Q-value iteration, tabular Q-learning, fitted Q-iteration, deep Q-learning) for sepsis treatment in ICUs using real-world medical data have been evaluated. These studies found that even simpler Q-learning methods can effectively learn policies for sepsis treatment, performing comparably to more complex procedures like deep Q-learning.

d) ML for Clinical Time-Series Data:

Modelling clinical time-series data is a critical task in clinical workflows. Applications include predicting clinical interventions in ICUs using CNN and LSTM, mortality prediction in traumatic brain injury (TBI) patients, and estimating mean arterial blood pressure (ABP) and intracranial pressure (ICP) for cerebrovascular autoregulation in TBI patients. Recent studies have utilized attention models for ICU forecasting tasks, integrating clinical notes with multivariate and time-series data. Additionally, ML techniques have been investigated to predict unexpected respiratory decompensation[9].

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 [28]. 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 [19]. In [18], 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 clinical settings, healthcare professionals spend significant time on documentation tasks, such as creating clinical notes, discharge summaries, and radiology reports. According to Dr Simon Wallace, clinicians dedicate approximately 50% of their time to these tasks, leading to increased demotivation due to high clinical workloads, administrative responsibilities, and a lack of leisure time. This often results in more time spent on documentation than direct patient interactions[10].

Clinical speech and audio processing technologies hold immense promise in addressing these challenges. They offer speech interfaces for hands-free services, automatic transcription of patient conversations, and the synthesis of clinical notes. The implementation of these tools in clinical environments can bring about a transformative shift. For patients, speech can become a new modality for assessing their condition. Clinicians can achieve greater efficiency and time savings, and the healthcare industry can benefit from enhanced productivity and cost reductions. This potential should inspire optimism and hope for a more streamlined future in clinical documentation.

Research has shown that speech processing has the potential to revolutionize healthcare. It can be used to identify speech-related disorders, such as vocal hyperfunction, and conditions that manifest through speech, like dementia. For instance, linguistic features have been successfully used to identify Alzheimer's disease. This demonstrates the immense value of speech-processing technologies in healthcare[11].

(IJIASE) 2024, Vol. No. 10, Jan-Dec e-ISSN: 2454-9258, p-ISSN: 2454-809X

SECURE, PRIVATE, AND ROBUST ML FOR HEALTHCARE: CHALLENGES

In this section, we delve into the security and robustness challenges that machine learning (ML) and deep learning (DL) models face in healthcare settings. We will discuss various sources of vulnerabilities throughout the ML pipeline, including data collection, data annotation, and model deployment. Understanding these challenges is crucial for healthcare professionals and researchers involved in clinical decision support and machine learning in healthcare settings.

Sources of Vulnerabilities in the ML Pipeline

ML applications in healthcare face numerous privacy and security challenges. As illustrated in Fig. 4, we will discuss these challenges and identify potential sources of vulnerabilities at different stages of the ML pipeline. It's crucial to emphasize that these discussions are aimed at ensuring the safety and security of patient data, providing reassurance about the robustness of our approach[12].

Vulnerabilities in Data Collection

Training ML/DL models for clinical decision support necessitates collecting large amounts of data, such as electronic health records (EHRs), medical images, and radiology reports. This process is often time-consuming and requires significant human effort. Despite careful data collection practices aimed at ensuring effective diagnoses, several vulnerabilities can impact the functionality of ML/DL systems:

- Instrumental and Environmental Noise: Data may contain artefacts due to instrumental and environmental disturbances. For example, multishot MRI, a modality used to acquire high-resolution medical images, is highly sensitive to motion. Even slight movements by the subject, such as head movement or respiration, can introduce artefacts that increase the risk of misdiagnosis[13].

Unqualified Personnel: Healthcare environments are interdisciplinary, often requiring more qualified personnel to develop and maintain ML/DL systems. Effective data-driven healthcare requires workers with strong statistical and computational backgrounds, such as engineers and data scientists. However, hospitals typically rely on physician researchers, who may need more computational expertise.

Vulnerabilities Due to Data Annotation

Most ML/DL applications in healthcare are supervised learning tasks requiring extensive labelled training data. Labelling data samples, known as data annotation, should ideally be conducted by experienced clinicians to create domain-enriched datasets essential for developing effective ML/DL models. However, several issues arise:

- Ambiguous Ground Truth: Ground truth can be ambiguous in medical datasets. Even expert clinicians may disagree on diagnostic tasks. This issue is exacerbated by malicious users who may manipulate data to complicate diagnoses and evade detection, even by human reviewers.

- Improper Annotation: Proper guidelines, privacy, and legal considerations should inform the annotation of data samples for life-critical healthcare applications. Many widely used healthcare datasets are annotated for coarse-grained labels, whereas ML/DL models need fine-grained and hidden strata within clinical environments for real-life utility. Inadequate labelling can lead to various efficiency challenges.

Efficiency Challenges

The quality of healthcare data used to build ML/DL models is impacted by several efficiency challenges, including issues arising from improper data collection and annotation. These challenges are not to be underestimated, as they can significantly affect the performance and reliability of ML/DL systems in clinical settings. By acknowledging these challenges, we demonstrate our understanding of the complexities involved, fostering a sense of awareness and preparedness among healthcare professionals[14].

(IJIASE) 2024, Vol. No. 10, Jan-Dec **e-ISSN: 2454-9258, p-ISSN: 2454-809X**

a) Limited and Imbalanced Datasets: The datasets used to train ML/DL models often require expansion. This is a crucial issue in healthcare, where the demand for extensive datasets impedes the practical use of DL methods. Many severe health conditions are rare, affecting only a few in thousands or millions of patients, which poses a challenge in training and optimizing ML/DL algorithms for such tasks. Addressing this issue is vital to ensure the effectiveness of ML/DL in healthcare.

b) Data Augmentation: Data augmentation is a widely used technique to address the scarcity of large-scale medical datasets. This involves applying various methods (such as cropping, flipping, rotation, and translation) to diversify and expand the training data. Transformation techniques like Gaussian augmentation are also employed. However, it's important to note that data augmentation can compromise the robustness of ML/DL systems, as the augmented data's distribution may diverge from the original training data's true, often unknown distribution. Studies indicate that Gaussian data augmentation does not enhance models' robustness against iterative attacks, underscoring the need for careful implementation.

c) Class Imbalance and Bias: Class imbalance is a significant challenge in supervised ML/DL, where the distribution of samples among classes is uneven. Training models on imbalanced datasets can lead to biased outcomes, disproportionately affecting certain categories. In healthcare, such biases can have serious implications, underscoring the necessity for the development of effective mitigation strategies. The impact of class imbalance and bias in supervised ML/DL is substantial, directly influencing the outcomes of these models[15].

d) Sparsity: Data sparsity, characterized by missing values, is joint in real-world datasets due to various factors (e.g., unmeasured or unreported samples). Missing data can significantly degrade the performance of ML/DL models.

3) Model Training vulnerabilities include improper or incomplete training, privacy breaches, and model poisoning or theft. Improper training occurs when models are trained with incorrect parameters (e.g., learning rate, epochs, batch size). ML/DL models are also highly susceptible to security and privacy threats such as adversarial attacks and model and data poisoning. These vulnerabilities hinder the effective deployment of ML/DL systems in security-critical (e.g., digital forensics, biometrics) and life-critical (e.g., self-driving cars, healthcare) applications. Ensuring the security and integrity of ML/DL systems is crucial for these applications. Various security threats to ML/DL systems are discussed in the next section.

4) Vulnerabilities in the Deployment Phase: Deploying ML/DL techniques in clinical environments involves human-centric decisions, making robustness, fairness, and accountability critical. Key vulnerabilities during deployment include:

Distribution Shifts: Distribution shifts are standard in realistic healthcare settings. For example, DL models trained on images from one imaging centre may need to improve when deployed on images from another centre. Similarly, ML models developed with historical patient data may be less effective when applied to new patients, raising concerns about predictive accuracy. These differences can also be exploited to create adversarial examples [16].

- Incomplete Data: Real-world healthcare data like EHRs often contain missing observations or variables. Ignoring missing values during analysis can lead to inaccuracies, as their relationships with observed or unobserved data are still being determined. Using incomplete data for training ML/DL models can result in false positives (misdiagnosing a healthy person) and false negatives (failing to diagnose a patient). Both scenarios can have severe consequences, emphasizing the need for complete and accurate healthcare data to ensure reliable predictions [17-25].

5) Vulnerabilities in the Testing Phase

Vulnerabilities in the testing phase concern interpreting results from ML/DL systems, including misinterpretation, false positives, and false negatives. As discussed earlier, these outcomes often arise from incomplete or inefficient model training or incomplete data fed during inference. Ultimately, the true potential of ML in healthcare lies not merely in its mechanical application but in the careful and informed application of analytical methods [26].

(IJIASE) 2024, Vol. No. 10, Jan-Dec e-ISSN: 2454-9258, p-ISSN: 2454-809X

THE SECURITY OF ML: AN OVERVIEW

This section provides an overview of ML security, particularly from the healthcare perspective, and highlights various security challenges.

1) Security Threats

Security threats to ML systems can be broadly categorized into influence attacks, security violations, and attack specificity [27]. A taxonomy of these security threats is depicted in Fig. 2.

a) Influence

Influence attacks can be of two types:

1. Causative Attacks: Attempt to control the training data.

2. Exploratory Attacks: Exploit misclassification of the ML model without interfering with the model training.

b) Security Violations

Security violations concern the availability and integrity of services and can be categorized into three types:

1. Integrity Attacks: Aim to increase the false-negative rate of the deployed model when given harmful inputs.

2. Availability Attacks: Aim to increase the false-positive rate of the classifier in response to benign inputs.

3. Privacy Violation Attacks: Concern about exposing sensitive and confidential information from the training data, the trained model, or both.

c) Attack Specificity

The specificity of an attack can be defined in two ways:

1. Targeted Attacks: Intended for a specific input sample or a group of samples.

2. Indiscriminate Attacks: Cause the ML model to fail indiscriminately.

The first axis in the taxonomy defines adversaries' capabilities, such as their ability to modify the training process by injecting poisoned data or not (i.e., accessing training data). If the attacker cannot access the training data, they can perform an exploratory attack. For instance, in a disease classification problem, the adversary can exploit query-response pairs to induce misclassification.

The second dimension of attacks concerns the type of security violations an adversary can execute, such as learning about users' privacy in training data or increasing the classifier's false-negative or false-positive rate. Each type of security violation poses severe problems for healthcare applications, where preserving user privacy is crucial, and models with minimal uncertainty are highly desirable.

The third dimension describes the adversary's specific objectives, whether they aim for a targeted attack, such as forcing a classifier to misclassify a specific input (e.g., bypassing a disease detection system by misclassifying the input as benign), or an indiscriminate attack that causes widespread failure of the classifier.

2) Adversarial Machine Learning (ML)

Adversarial attacks have emerged from efforts to identify vulnerabilities in ML/DL models during training and inference, posing significant security threats to these systems [20], [28]–[10]. The primary goal of an adversary in such attacks is to generate adversarial examples by adding small, carefully crafted perturbations to actual input samples to compromise the integrity of the ML/DL system. Generally, there are two types of adversarial attacks:

1. Type 1: Description of the first type.

2. Type 2: Description of the second type.

a) Poisoning Attacks

Poisoning attacks are adversarial strategies targeting the model's training phase. By manipulating the training data, these attacks aim to mislead the learning process of machine learning (ML) or deep learning (DL) models.

b) Evasion Attacks

Evasion attacks occur during the inference phase. In these attacks, the attacker manipulates test data to compromise the integrity of the ML/DL model and introduces harmful inputs to mislead the model [20].

3) Adversarial Attacks in Healthcare Applications

In the healthcare sector, poisoning attacks present a significant concern. The direct manipulation of existing training data, a common strategy in these attacks, is often challenging or unfeasible. However, the addition of new samples to the training dataset can be relatively easy, posing significant risks to the applicability of ML/DL systems. This underscores the urgency of detecting and mitigating poisoning attacks to ensure robust ML/DL applications in healthcare. For instance, systematic poisoning attacks on six conventional ML models for hypothyroid diagnosis have demonstrated the potential to prevent accurate diagnosis, a real-world implication that cannot be ignored.

Adversarial attacks in healthcare settings present a unique and urgent threat. Unlike adversarial examples in other domains, healthcare may face unintentional adversarial patients, leading to severe ethical issues. Studies have shown that patients with identical predictive features can experience significantly different treatment outcomes, highlighting the potential harm of these attacks. Recent research has demonstrated white box and black box adversarial attacks against clinical applications such as fundoscopy, dermoscopy, and chest X-ray analysis. Additionally, potential incentives for adversarial attacks in clinical trials are likely to increase with the growing use of ML in computer-aided diagnosis and decision support systems, further underscoring the need for immediate action[21].

Adversarial ML poses a significant dilemma for the security and privacy of ML/DL models in healthcare biometrics applications, leading to severe unintended consequences. Biometrics, like palm vein readers, fingerprint scanners, ECG, iris scanners, and face recognition, offer advantages such as fraud detection and medical records and facilities protection. However, these systems, which often rely on ML/DL techniques, are vulnerable to security and privacy attacks. For instance, an adversary can bypass a face recognition system deployed to restrict unauthorized access in secure areas.

ML FOR HEALTHCARE: CHALLENGES

1) Safety Challenges

Ensuring the safety of ML/DL systems for patients is a complex challenge that requires a continuous focus throughout the ML/DL lifecycle. While these systems often perform excellently in controlled lab environments, this does not guarantee safety in real-world applications. Clinicians often deal with common health conditions but must also diagnose rare, subtle, and hidden conditions. Enabling ML/DL systems to perform well on such outliers and edge cases is crucial for patient safety, a safety challenge that cannot be overlooked.

(IJIASE) 2024, Vol. No. 10, Jan-Dec **e-ISSN: 2454-9258, p-ISSN: 2454-809X**

2) Privacy Challenges

Privacy is a significant concern in data-driven healthcare, which focuses on using patient data by ML/DL systems to make predictions. Patients expect healthcare providers to safeguard confidential information, such as age, sex, date of birth, and health data. Privacy threats can involve the unauthorized disclosure of information or the malicious use of data[22].

The nature of the collected data, its creation environment, and patient demographics influence privacy. Mitigating privacy breaches requires anonymizing data to prevent re-identification of individuals and ensure secure data transfers within healthcare facilities. Privacy concerns also arise with the adoption of ML/DL in biometric healthcare systems, whether offline (e.g., face or fingerprint recognition to secure medical facilities) or online (e.g., real-time medical systems and IoT device authentication). Therefore, performing worst-case robustness tests is crucial to ensure the security and privacy of these systems.

3) Ethical Dilemmas

In the realm of user-centric machine learning applications, such as healthcare, the significance of upholding ethical standards in data usage cannot be overstated. It is not just about gathering data to construct machine learning models, but also about proactively understanding the demographics and sociological aspects of the targeted user base. Moreover, recognizing the potential risks to a patient's well-being and dignity posed by data collection is a crucial aspect of ethical data usage. Neglecting these ethical considerations can lead to adverse outcomes when implementing machine learning in practical contexts. Therefore, a comprehensive understanding of AI systems, especially in uncertain and complex scenarios, is essential to ensure their fair and ethical operation[23].

4) Grappling with Causality

In healthcare, grasping causality holds significant importance as many critical healthcare issues demand causal reasoning, such as hypothetical scenarios ("What if?"). For instance, understanding the implications of prescribing treatment A versus treatment B requires causal analysis. Traditional learning algorithms often need to improve such queries, necessitating a causal modelling approach. Deep learning models, while powerful, operate as black boxes devoid of explicit causal connections, relying on patterns and correlations. While this lack of causality may not hinder predictive accuracy, it raises urgent concerns about the interpretability of outcomes. Furthermore, leveraging causal reasoning can enhance fairness in decision-making by accurately estimating the causal effects of variables on target outputs.

5) Standardizing Data Exchange

A deep understanding of healthcare tasks often necessitates the integration of diverse patient data beyond clinical imaging, such as electronic medical records (EMRs). However, the lack of adherence to data exchange standards in healthcare IT systems poses a significant obstacle to effective data integration and exchange across specialities and organizations. Therefore, implementing standardized data exchange protocols is not just a recommendation, but a necessity for leveraging multi-modal data to enhance algorithmic understanding and improve clinical decision-making. With the widespread adoption of data exchange standards, the efficacy of machine learning and deep learning systems in healthcare can be maintained[24].

6) Addressing Distribution Shifts

Data distribution shifts pose significant challenges, particularly in clinical settings where training and testing data distributions may diverge due to various factors like medical institutions and devices. Due to these distribution shifts, machine learning models trained on public databases often fail to perform adequately in real-world clinical environments. Addressing this challenge requires moving beyond the assumption of similar data distributions during model training, especially given the life-critical nature of clinical applications. Ensuring machine learning techniques' smooth and safe operation in clinical settings demands strategies to mitigate the impact of distribution shifts.

(IJIASE) 2024, Vol. No. 10, Jan-Dec e-ISSN: 2454-9258, p-ISSN: 2454-809X

ENSURING SECURE, PRIVATE, AND ROBUST ML FOR HEALTHCARE: SOLUTIONS

In this section, we provide an overview of proposed methods to ensure the security, privacy, and robustness of machine learning in healthcare applications. In Table II, we summarize articles focused on "Secure and privacy-preserving ML for healthcare" and describe various approaches for achieving secure, private, and robust machine learning. Additionally, we present a taxonomy of commonly used approaches in Figure 6, followed by individual descriptions of each approach.

A. Preserving Privacy in ML

Privacy preservation is paramount in healthcare, where user-centric applications involve collecting sensitive personal data. ML model training and inference must not compromise user privacy. Data anonymization is commonly used to mitigate privacy risks but may not always suffice. Various cryptographic approaches, including homomorphic encryption, garbled circuits, and secret sharing, offer robust solutions for preserving privacy while enabling effective machine learning on sensitive data. These techniques ensure that computations on encrypted data do not compromise user privacy, thus facilitating secure and privacy-preserving machine learning in healthcare.

1) Federated Learning: Google Inc. recently introduced the concept of federated learning (FL) [16], where a collaborative machine learning (ML) model is developed by utilizing data from various distributed devices. Each device trains the model using its local data and then shares only the model parameters with a central model, ensuring the privacy of raw data. An FL-based decentralized scheme, employing the iterative cluster primal-dual splitting (cPDS) algorithm, is proposed for predicting hospitalization in patients with heart-related diseases using large-scale electronic health records (EHR). Additionally, various configurations of split learning deep learning (DL) models, such as simple vanilla, U-shaped, and vertically partitioned data-based configurations, are presented in under the framework named SplitNN, which ensures patient data privacy without sharing critical information with the server. Furthermore, a framework called federated autonomous deep learning (FADL) utilizing distributed EHR is introduced in [30]. A comparative analysis of different privacy-preserving techniques is provided in Table III, which can guide researchers and practitioners in selecting the most suitable technique for their specific use case[25].

B. Countermeasures Against Adversarial Attacks

Countermeasures against adversarial attacks are classified into three categories in recent literature:

(1) modifying the model,

(2) modifying the data, and

(3) adding auxiliary model(s) [13].

1) Modifying Model:

This category involves altering the parameters or features of the trained ML model. One widely used method is Defensive Distillation, initially proposed by Hinton et al. [14] and later adopted by Papernot et al. as a defence against adversarial attacks [15]. Defensive Distillation works by training a model to be more robust to adversarial attacks by smoothing its output probabilities. However, Carlini and Wagner demonstrated that their C&W attack could bypass defensive distillation [16]. Other methods include Network Verification, which involves verifying the integrity of the network's weights and biases, Gradient Regularization, which penalizes large gradients to prevent overfitting, and utilizing Generative ML Models. These methods, while diverse in their approach, all aim to enhance the robustness and security of ML models against adversarial attacks.

C. Causal Models for Healthcare

In the realm of healthcare, the importance of posing causal questions cannot be overstated, despite the challenges posed by ethical constraints on experimentation. Retrospective observational data is often used to train models for making counterfactual predictions. Two foundational approaches to causality, potential outcomes and causal graphical models, have been instrumental in this field. Various methods for providing causal inferences and

reasoning in healthcare have been presented, such as Gaussian processes-based counterfactual causal models and probabilistic graphical models for analysing causality in health conditions.

REVISED: ADDRESSING DISTRIBUTION SHIFTS

Various techniques have emerged in the literature to tackle the challenge of data distribution shift. These methods, such as transfer learning and domain adaptation, offer solutions.

1) Transfer Learning: This approach alleviates the necessity for a large-scale dataset by retraining deep learning (DL) models on application-specific datasets, typically smaller than the original dataset. The goal is to transfer the knowledge gained from one domain to another. However, transfer learning in healthcare applications faces challenges due to the need for ample initial training data and high-quality annotations by expert clinicians.

2) Domain Adaptation: Domain adaptation addresses the discrepancy between training and test data distributions. It's particularly relevant in medical image analysis tasks like MRI segmentation, chest X-ray classification, and Alzheimer's disease classification. Various methods within domain adaptation exist, including:

a) Supervised Domain Adaptation: This method is akin to supervised learning but with differing distributions between the source and target domains. It's effective when labelled data is available for both domains.

b) Unsupervised Domain Adaptation: Here, the source domain data is labelled while the target domain data is unlabelled. Techniques such as reverse flow and adversarial training, along with self-regularization, have been employed to preserve clinically relevant features.

c) Semi-supervised Domain Adaptation: This approach involves labelled source data and partially labelled target domain data.

d) Self-supervised Domain Adaptation: These methods aim to train visual models without manual labelling by utilizing auxiliary tasks, known as pretext tasks, to provide supervision [26].

TOWARDS RESPONSIBLE ML

Ensuring responsible machine learning (ML) involves several practices:

1) Human-Centred Design: User characteristics are essential for impactful system development.

2) Evaluation Metrics: Use appropriate metrics aligned with system goals and gather user feedback through surveys.

3) Raw Data Examination: Thoroughly analyse datasets for biases, abnormalities, and privacy concerns.

4) Model and Dataset Limitations: Understand the ML model's and dataset's capabilities and constraints.

5) Repetitive Testing: Continuously test ML systems to ensure proper functioning, considering factors like input drifts and quality checks.

6) Continuous Monitoring and Updating: Regularly monitor and update deployed ML systems to address issues encountered in real-world settings.

RESPONSIBLE ML FOR HEALTHCARE

While ML and DL techniques hold promise for clinical applications, their limited adoption indicates the need for further refinement. A multidisciplinary approach involving stakeholders from various fields is crucial for safe and meaningful deployment in healthcare settings. Critical steps include problem selection, solution development, ethical considerations, rigorous evaluation, transparent reporting, responsible deployment, and market readiness.

Tools and Libraries for Secure and Private ML

Securing ML models and data requires the development of specialized tools and algorithms. Several frameworks and libraries have been introduced for this purpose:

1) TensorFlow Federated: Enables distributed ML training without sharing local client data.

2) CrypTen: Facilitates secure and privacy-preserving ML using encrypted data within PyTorch.

3) PyTorch-DP: A framework for training DL models with differential privacy.

4) OpenMined: Offers tools and libraries for building privacy-preserving ML models, including PySyft, PyGrid, and SyferText.

CONCLUSION

Machine Learning (ML) and Deep Learning (DL) models have emerged as powerful tools with transformative potential in healthcare. These models promise to improve diagnostic accuracy, treatment effectiveness, and patient outcomes. However, alongside their benefits, they also bring forth significant challenges, particularly concerning security and privacy.

One primary concern is the vulnerability of sensitive medical data to unauthorized access, manipulation, or breaches. Patient confidentiality and data integrity are paramount in healthcare, making it imperative to safeguard against security threats and privacy breaches. Moreover, the potential biases embedded within datasets used to train these models can lead to unfair treatment or inaccurate predictions, further exacerbating ethical and privacy concerns[28].

Addressing these challenges requires a multifaceted approach grounded in responsible practices, interdisciplinary collaboration, and the adoption of secure ML tools. Responsible practices entail thoroughly examining datasets for biases and anomalies, transparent reporting of model performance and limitations, and continuous monitoring and updating of deployed systems. Interdisciplinary collaboration involving healthcare professionals, data scientists, ethicists, and policymakers fosters a holistic understanding of complex issues. It ensures that solutions are aligned with ethical principles and regulatory requirements.

Furthermore, adopting secure ML tools and frameworks is essential for mitigating security risks and preserving patient privacy. Techniques such as differential privacy, federated learning, and homomorphic encryption offer avenues for conducting ML/DL operations while protecting the confidentiality and integrity of sensitive data. Open-source initiatives and community-driven efforts provide access to tools and libraries designed explicitly for privacy-preserving ML, enabling researchers and practitioners to develop secure and ethically sound solutions.

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 et al., “Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition,” IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1332–1343, May 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 Trans. Med. Imag., vol. 35, no. 5, pp. 1207–1216, May 2016.

[4] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, “Multi-scale convolutional neural networks for lung nodule classification,” in Proc. Int. Conf. Inf. Process. Med. Imag., 2015, pp. 588–599.

[5] J. Schlemper, J. Caballero, J. V. Hajnal, A. Price, and D. Rueckert, “A deep cascade of convolutional neural networks forMR image reconstruction,” in Proc. Int. Conf. Inf. Process. Med. Imag., 2017, pp. 647–658.

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

[7] M. Havaei et al., "Brain tumor segmentation with deep neural networks," Med. Image Anal., 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," Med. Phys., vol. 45, no. 5, pp. 1791–1793, 2018.

[10] B. E. Bejnordi et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," J. Amer. Med. Assoc., vol. 318, no. 22, pp. 2199–2210, 2017.

[11] P. Rajpurkar et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," 2017, arXiv:1711.05225.

[12] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," J. Amer. Med. Assoc., vol. 316, no. 22, pp. 2402–2410, 2016.

[13] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.

[14] S. Latif,M.Asim, M. Usman, J. Qadir, and R. Rana, "Automatingmotion correction in multishot MRI using generative adversarial networks," in Proc. 32nd Conf. Neural Inf. Process. Syst., 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 Bioinformat., 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 Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases, 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," J. Korean Soc. Radiol., 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 Trans. Neural Netw. Learn. Syst., vol. 30, no. 9, pp. 2805–2824, Sep. 2019.

[20] C. Szegedy et al., "Intriguing properties of neural networks," 2013, arXiv:1312.6199.

[21] A. Shafahi et al., "Poison frogs! Targeted clean-label poisoning attacks on neural networks," in Proc. Adv. Neural Inf. Process. Syst., 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, andV.Kumar, "Anomaly detection:Asurvey," ACM Comput. Surv., vol. 41, no. 3, pp. 1–58, 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. Güne¸s, "Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system," Appl. Math. Comput., 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. Berlin, Germany: Springer, 2020, pp. 3–21.

[27] M. N. Sohail, J. Ren, andM. U.Muhammad, "A Euclidean group assessment on semi-supervised clustering for healthcare clinical implications based on real-life data," Int. J. Environ. Res. 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 Proc. Int. Conf. Artif. Intell. Commun. Netw., 2019, pp. 450–472.

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

[30] W. Bai et al., "Semi-supervised learning for network-based cardiac MR image segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, 2017, pp. 253–260.