Article published online: 2023-12-26

Security and Privacy in Machine Learning for Health Systems: Strategies and Challenges

Erikson J. de Aguiar, Caetano Traina Jr., Agma J. M. Traina
Institute of Mathematics and Computer Science, University of São Paulo, Brazil

Summary

Objectives: Machine learning (ML) is a powerful asset to support physicians in decision-making procedures, providing timely answers. However, ML for health systems can suffer from security attacks and privacy violations. This paper investigates studies of security and privacy in ML for health.

Methods: We examine attacks, defenses, and privacy-preserving strategies, discussing their challenges. We conducted the following research protocol: starting a manual search, defining the search string, removing duplicated papers, filtering papers by title and abstract, then their full texts, and analyzing their contributions, including strategies and challenges. Finally, we collected and discussed 40 papers on attacks, defense, and privacy.

Results: Our findings identified the most employed strategies for each domain. We found trends in attacks, including universal adversarial perturbation (UAPs), generative adversarial network (GAN)-based attacks, and DeepFakes to generate malicious examples. Trends in defense are adversarial training, GAN-based strategies, and out-of-distribution (OOD) to identify and mitigate adversarial examples (AE). We found privacy-preserving strategies such as federated learning (FL), differential privacy, and combinations of strategies to enhance the FL. Challenges in privacy comprehend the development of attacks that bypass fine-tuning, defenses to calibrate models to improve their robustness, and privacy methods to enhance the FL strategy.

Conclusions: In conclusion, it is critical to explore security and privacy in ML for health, because it has grown risks and open vulnerabilities. Our study presents strategies and challenges to guide research to investigate issues about security and privacy in ML applied to health systems.

Keywords

Adversarial attacks, machine learning, medical images, privacy

Yearb Med Inform 2023:269-81
http://dx.doi.org/10.1055/s-0043-1768731

1 Introduction

In recent years, data produced by medical systems has grown exponentially. The processing and knowledge extraction from these data contributed to the development of the so-called big data [1]. Medical systems produce complex data from sensors, imaging, or genomics data, among others. Medical complex data are essential for decision-making and generating new knowledge. Large amounts of medical images are collected to support physicians in the diagnosis process and to help identify disease patterns. Decision-making strategies are based on classical machine learning (ML) or deep learning (DL). Physicians can integrate ML techniques to analyze and assist decision-making, considering the recommendations of the models to enhance the diagnosis precision [1, 2].

Although ML can improve physicians' decision-making, ML methods applied to health systems can also suffer attacks [2, 3, 4]. Attacks on ML methods correspond to a study field called adversarial attacks (AA), which builds methods to train and test models in adversarial environments [3, 4]. ML methods are susceptible to attacks, such as poisoning the training data (data poisoning), bypassing the test data (evasion attack), invalidating the model, and exploiting backdoors [2]. For instance, Figure 1 illustrates an example of AA on optical coherence tomography (OCT) images using the projected gradient descent (PGD) [5] attack.

Unfortunately, DL models for health systems are vulnerable to AA and suffer from privacy risks. According to [4], systems that handle sensitive health data need to be designed to consider privacy risks. Concerning privacy, many governments have defined regulations to formalize organizations' data handling since growing data leakage resulted in decreasing systems confidence. The European Union proposed the general data protection regulation (GDPR) to establish rules and rights to manage sensitive data. Furthermore, in 1996 the United States proposed a regulation to handle medical data called the health insurance portability and accountability act (HIPAA). We explain all acronyms in Table 1.

This paper investigates security and privacy in ML for health, considering three perspectives: the most common attacks, suitable defenses, and privacy-preserving strategies. Finally, we highlight in this survey the following main contributions: (i) current taxonomies for security and privacy in ML for health systems; (ii) trends in attacks, defenses, and privacy-preserving strategies during the last years (2018-2022); (iii) challenges from developing attacks, defenses to detect and mitigate attacks, as well as to employ privacy methods in ML; (iv) tools and databases most applied to run experiments in security and privacy in ML for health; and (v) a summary of most relevant studies that cover strategies for attacks, defense, and privacy.

2 Background

This section addresses essential concepts about security and privacy in ML. They motivate studying security and privacy in the health environment when handling sensitive information.

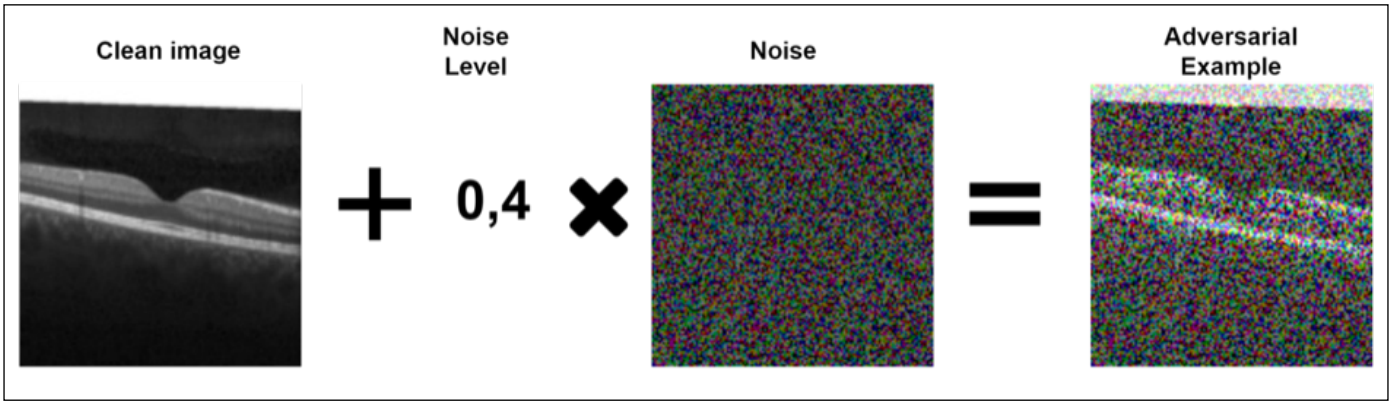

Fig. 1 Example of adversarial attack using projected gradient descent (PGD) on optical coherence tomography (OCT) image.

Table 1 Explanations of acronyms.

Explanation	Acronym	Explanation	Acronym
Adversarial Attack	AA	Generative Adversarial Network	GAN
Adversarial Example	AE	Homomorphic Encryption	HE
Carlini & Wagner	C&W	Jacobian Saliency Map Attack	JSMA
Computed Tomography	CT	Machine Learning	ML
Deep Learning	DL	Magnetic Resonance Imaging	MRI
Differential Privacy	DP	MultiParty Computation	MPC
Electrocardiograms	ECG	Optical Coherence Tomography	OCT
Electronic Health Records	EHR	Out-Of-Distribution	OOD
Fast Gradient Sign Method	FGSM	Picture Archiving and Communication Systems	PACS
Federated Learning	FL	Projected Gradient Descent	PGD
General Data Protection Regulation	GDPR	Universal Adversarial Perturbation	UAP

2.1 Security in Machine Learning

ML methods are susceptible to adversarial attacks (AA). AA can exploit vulnerabilities in ML models and data [2, 4]. Adversarial example (AE) is formally defined in Equation 1, which minimizes the distance between AE and the original example using the Euclidean distance. Equation 1 represents an AE as , the original example as , the noise level as , the class label as , and the loss function of the ML algorithm . An AE aims to induce a visual perception like the original example, fooling the ML model during the test or training phases. The AA seeks to maximize the loss of the ML algorithm, mainly used for DL methods [1, 3]. According to [3, 4], the security in ML for health can involve attacks and defense methods.

$$Min \parallel x_{adv} - x_0 \parallel + \varepsilon \times L_f(x_{adv}, y) \qquad (1)$$

AA for health cover features such as capabilities, system violations, knowledge, perturbation metrics, and classification or segmentation tasks [3, 4]. The *objective of the attack* can be poisoning or evasion. Poisoning attacks affect the training set, and evasion attacks affect the test set. *System violations* define which features of the system the attacker attempts to bypass. *System violations* target *integrity*, *availability*, and *privacy*. *Attacker knowledge* defines the permission level. The permissions are: (i) limited (black-box), which only explores the interface to access the model and test it; partial (gray-box), which explores a specific part of the system, such as the parameters; open (white-box) that targets several structures of the models, such as the hyperparameters and database. The *perturbation metrics* are used to craft examples and generate AE. Usually, these metrics are based on distances such as , , and . Examples of AA are: the fast gradient sign method (FGSM) [6], projected gradient descent (PGD) [5], One Pixel [7], jacobian saliency map attack (JSMA) [8], DeepFool [9], carlini & wagner (C&W) [10], and universal adversarial perturbations (UAPs) [11]. Besides, attacks can be against pre-processing algorithms, such as Image Scaling [12, 13]. Finally, the defenses to mitigate attacks are adversarial training [5], feature squeezing [14], defensive distillation [15], and generative adversarial network (GAN)-based (*e.g.*, Magnet) [16].

2.2 Privacy in Machine Learning

Organizations have been concerned about privacy due to the growing data leakage and establishing of privacy regulations, such as GDPR [17-19]. Privacy violations are increasing and require mitigation. ML models can suffer data leakage, resulting in privacy disasters for organizations. According to

[18,19], challenges to privacy in ML include developing robust defenses to mitigate attacks, such as membership inference or re-identification. Threat models, attacks, defenses, and features categorize privacy in ML. **Threat models** can be Linkage [20], Inference [21], and Re-identification [22]. **Attacks** are Feature Estimation [23], Membership Inference [24], Model Memorization [25], Extraction [26], and DeepFakes [27]. Attack features are knowledge of the attacker and attack objective. The **attacker's knowledge** are black-box, gray-box, and white-box. Moreover, the **attack objective** targets models or training data [19].

Privacy-preserving strategies are obfuscation, cryptography, and aggregation [19]. *Obfuscation* methods hide sensitive variables using perturbations that seek to mitigate privacy risks, such as differential privacy (DP) [28] and GAN-based ones. *Cryptographic* methods use algorithms to hide user identities, using homomorphic encryption (HE) [29] and multiparty computation (MPC) [30]. These methods encrypt sensitive information, enabling complex operations on the encrypted data [19]. The *aggregation* methods work on collaborative training, including federated learning (FL) [31]. FL creates clean models and sends them to the organizations that handle sensitive data. These organizations train models on sensitive data without making it public and send the trained model to a server that aggregates the models on a general model [17-19].

3 Materials and Methods

We applied a methodology of software engineering proposed by [32] to conduct this research on security and privacy in ML for health. We investigate papers from 2018 to 2022. This section describes the method applied to search and select the relevant papers. We carried out the methodology encompassing the six steps, as follows: (i) define the research questions; (ii) select the databases; (iii) select the proper keywords; (iv) define the search string; (v) define inclusion and exclusion criteria; (vi) perform data extraction. The main purpose of this research is to identify strategies and issues of security and privacy in ML for health.

We **define our research question** to guide this work. First, we did an initial search to raise relevant papers and authors from the literature based on papers [2] and [4]. Afterward, we did a manual search to analyze papers that cited [2] and [4]. Also, we selected papers by analyzing abstracts and titles to collect important topics of security and privacy in ML for health. Thus, we collected candidate papers and analyzed their discussions, including or excluding papers if following the main topic (security and privacy in ML). Finally, referring to the papers collected, we defined research questions that guided the selection of the set of studies: (i) what state-of-art attack the study applied? (ii) has it employed defense to mitigate the attack? (iii) which features of defense contribute to mitigate the attacks? (iv) has it applied privacy-preserving ML techniques? (v) what metrics were applied to quantify attacks and defenses in machine learning for health?

The **databases selected** were the most used ones in computer science for health research, following the study of [33], such as ACM Digital Library, IEEE Explore, PubMed, Web@Science, and ScienceDirect. The percentage of papers found in each database are: EI Compendex (25.64%), ACM Digital Library (24.44%), IEEE Explore (1.13%), PubMed (3.98%), Web@Science (2.03%), and ScienceDirect (42.78%). Based on the research questions and topics, we *selected the keywords* most commonly used in the candidate papers initially collected. We used the Mendeley platform[1] to identify common keywords. The keywords selected were adversarial machine learning, privacy, security, deep learning, medical systems, medical image, and healthcare systems.

The *search string* was drawn to cover variants of topics related to deep learning, machine learning, adversarial attacks, privacy, and medical systems. We identified relevant topics dependent on the manual search and fine-tuned terms based on papers [2] and [4], as well as the most cited authors. Our search string was defined in the review process based on the initial search conducted by the following steps:

1. Manual search based on papers cited and keywords extracted from [2].
2. We select the most used databases for Computer Science, such as ACM digital library, IEEE Explore, IE Compedex, Web@Science, PubMed, and Science Direct. These databases are collected according to [32] and validated on papers [2-4, 18, 19] that are reviews related to security or privacy in ML.
3. We extracted keywords from papers [2-4, 18, 19] and fine-tuned keywords using the Mendeley platform that stored papers from manual searches. The keywords are adversarial machine learning, privacy, security, deep learning, medical, medical image, and healthcare.
4. Having to define keywords, we composed the search string, placed in the box as follows:

> („deep learning" OR „machine learning" OR „artificial intelligence") AND („medical" OR „healthcare") AND („medical image" OR „medical imaging") AND („adversarial attacks" OR „adversarial perturbations" OR defenses or privacy)

Therefore, after searching papers in the database, we refined the relevant papers, and we applied a *selection criteria* to include or exclude primary studies. The *inclusion criteria* are:

- The study addresses any topic about adversarial attacks or defenses of machine learning in the medical field;
- The study addresses any topic about privacy concerns in machine learning applied to the medical field;
- The study includes strategies of attack or defense in machine learning applied to the medical field;
- The paper is applied to complex data, such as medical images;
- The study is a research paper;
- The study is available;
- The study was written in English.

Also, we defined the following *exclusion criteria*:

- The study is not related to machine learning security or privacy in the medical field;
- The study does not discuss strategies or problems of adversarial attacks, defenses, or privacy applied to the medical field;

[1] https://www.mendeley.com/search/

- The paper is not aimed to complex data;
- The study is gray literature, i.e., tutorials, electronic books, pre-print, tutorials, or technical reports;
- The paper is not available;
- The study is not written in English.

Finally, we collected the papers based on the search string and stored them in the Mendeley platform. These papers are fine-tuned by removing duplicates, considering their title and abstract, and analyzing the full papers. The *data extraction* comprehends relevant information from studies, such as title, authors, interest topics, strategies applied, and challenges. To complete the data extraction, we defined two taxonomies for security and privacy in ML for health, respectively. Figure 3 describes the security taxonomy inspired in [3]. We built the taxonomy of security following this specification: (i) we select the main topic of adversarial attack in health systems; (ii) we specify and group aspects analyzed in the literature, such as features, category, defenses, and health task; (iii) we classify strategies following [2] that defines which features are systems' violations, the goal, and knowledge, as well as the categories following attacks method based on gradient, optimization, and preprocessing; (iv) finally, we select strategies and papers collected from the literature that address this strategy. Figure 4 shows the taxonomy of privacy inspired in [19] that collects the following aspects from literature: (i) the main topic; (ii) the group of aspects analyzed, such as threat model, attacks, defenses, and features; (iii) we classify strategies following features and defenses, for instance, according to [19] defenses are obfuscation, cryptography, and aggregation; (iv) the strategies selected correspond to papers collected from the literature that address these strategies for health task.

4 Result

This section presents our findings about security and privacy in ML for health, based on the selected literature works from 2018 to 2022. Figure 2 shows a pipeline to collect the papers. We followed three stages: (i) search strategy, (ii) selection and exclusion criteria, and (iii) data extraction.

Based on Figure 2, we describe the following stages. *Stage 1, the search strategy*, defined the initial search following a manual search to collect primary studies, such as relevant surveys and reviews. Such reviews are [2] and [4] used to guide research questions outlined in Section 3. Also, based on the manual search, we selected databases regarding computer science and medicine related to [2] and [4]. The main topics collected from the manual search assisted in defining the keywords: adversarial machine learning, privacy, security, deep learning, medical systems, medical image, and healthcare systems. We built a search string based on keywords and constraining them for the period between 2018 to 2022, English language, and if it is a research paper. Stage 1 returned 1,330 primary studies that will be fed to stage 2. Stage 2 *filters studies* following the selection and exclusion criteria (see Section 3). We removed 246 duplicated papers, reducing from 1,330 to 1,084 papers. Thereafter, we filtered papers by title and abstract, removing 880 from 1,084 to 204 papers. Finally, we filtered from 204 to 40 papers by analyzing the full text. Stage 3 does the *data extraction*, considering the main topics about adversarial attacks, defenses, and privacy results. In summary, we identify titles, authors, challenges, and strategies applied to the main topics posed.

4.1 Overview

This section presents an overview of results that summarize the main strategies and the taxonomy proposed. Our findings comprehend 40 papers related to 3 domains: attacks

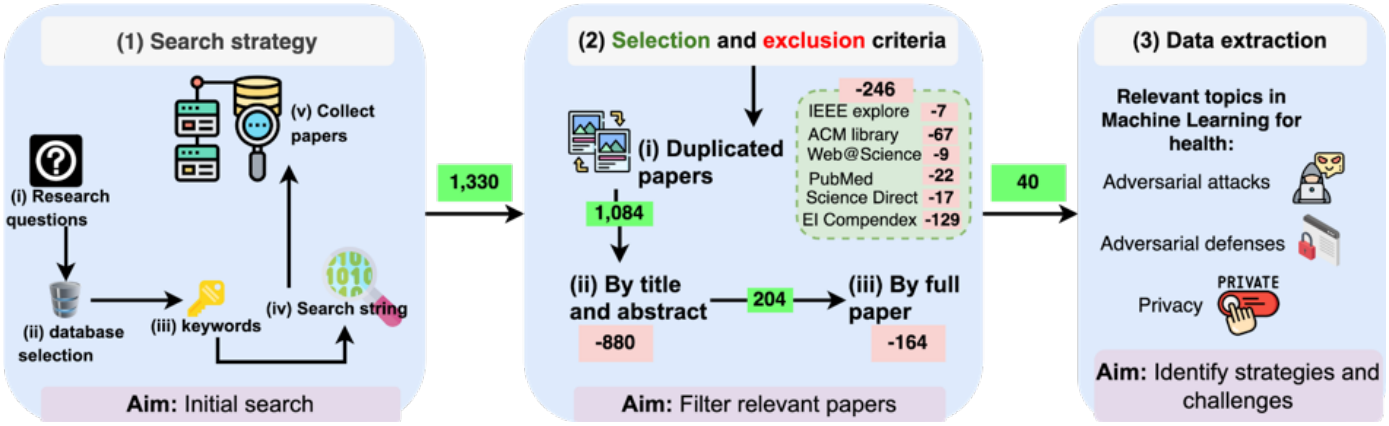

Fig. 2 Pipeline of the literature review. This review collects relevant papers from the literature from 2018 to 2022, including security and privacy in machine learning for health. The research issues focus on adversarial attacks, defenses, and privacy concerns.

with 17 papers (42.50%), defenses with 14 papers (35.00%), and privacy with 9 papers (22.50%). The main topics of the papers are strategies to attack DL classifiers tested on medical images, techniques to identify or mitigate attacks, and strategies to privacy-preserving medical images with sensitive attributes. In the literature, most attacks applied in DL for healthcare are FGSM (23.53%) [4], PGD (11.76%) [5], GAN-based (17.65%) [34], and UAPs (11.76%) [11]. Furthermore, we found that the most employed defenses are frequency domain (13.33% of the papers), GANs (26.67% of the papers), and adversarial training (20.00% of the papers) to mitigate or identify AE.

We proposed two taxonomies to summarize the main strategies found and to classify the papers collected. We were inspired by [3] and [19] to build our taxonomies and extend them to DL for healthcare. Figure 3 presents a taxonomy of security in ML for health, regarding the attack category, attacker knowledge, defense features, and defense category. Attacks are classified into categories: Gradient-based, Optimization, and pre-processing. Other significant aspects of attacks are the features that classify a system violation, the objective, and the knowledge. Defensive methods are organized as pre-processing, identification with out-of-distribution (OOD) and GANs, mitigation with frequency domain and adversarial training, as well as the Surrogate model with GANs. Our taxonomy classifies the papers as targeting the attack strategy or defense strategy.

Our results show that the most employed strategies for privacy-preserving in ML are: FL [31] with 44.44% of the papers, DP [28] with 22.22% of the papers, HE [29] with 11.11% of the papers, and MPC [30] with 11.11% of the papers (see Section 2). Moreover, DL models could be attacked for feature estimation [23], membership [24], model memorization [25], and extraction [26]. The privacy attacks are modeled by the threat model following linkage [20], inference [21], and re-identification [22]. These results are summarized in our taxonomy of privacy in ML for health that define relevant topics and papers addressing privacy-preserving strategies. Figure 4 illustrates our taxonomy, drawing papers that applied the strategy presented in green squares. The threat models are state-of-art papers, such as [20-22]. In addition, attacks are outlined in papers [23-27]. For health, the defensive methods most employed are DP [28, 68, 70], GAN-based [65, 71], MPC [67], HE [72], and FL [66, 67, 69, 72]. Also, we list privacy features in ML, such as permission level and attack objective.

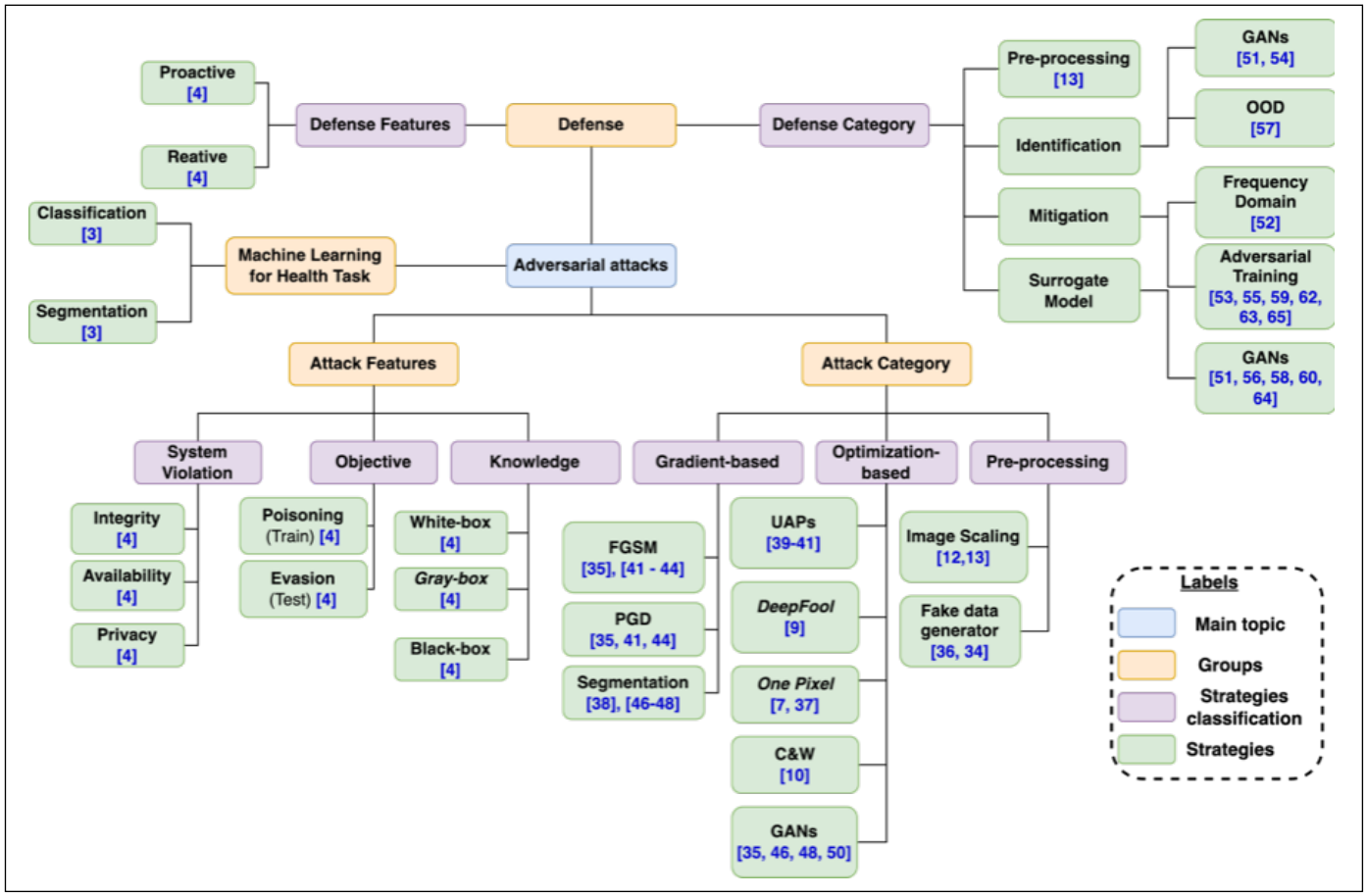

Fig. 3 Taxonomy of security in machine learning for health. The figure shows the definition of adversarial attacks (bottom part) and defensive methods (upper part).

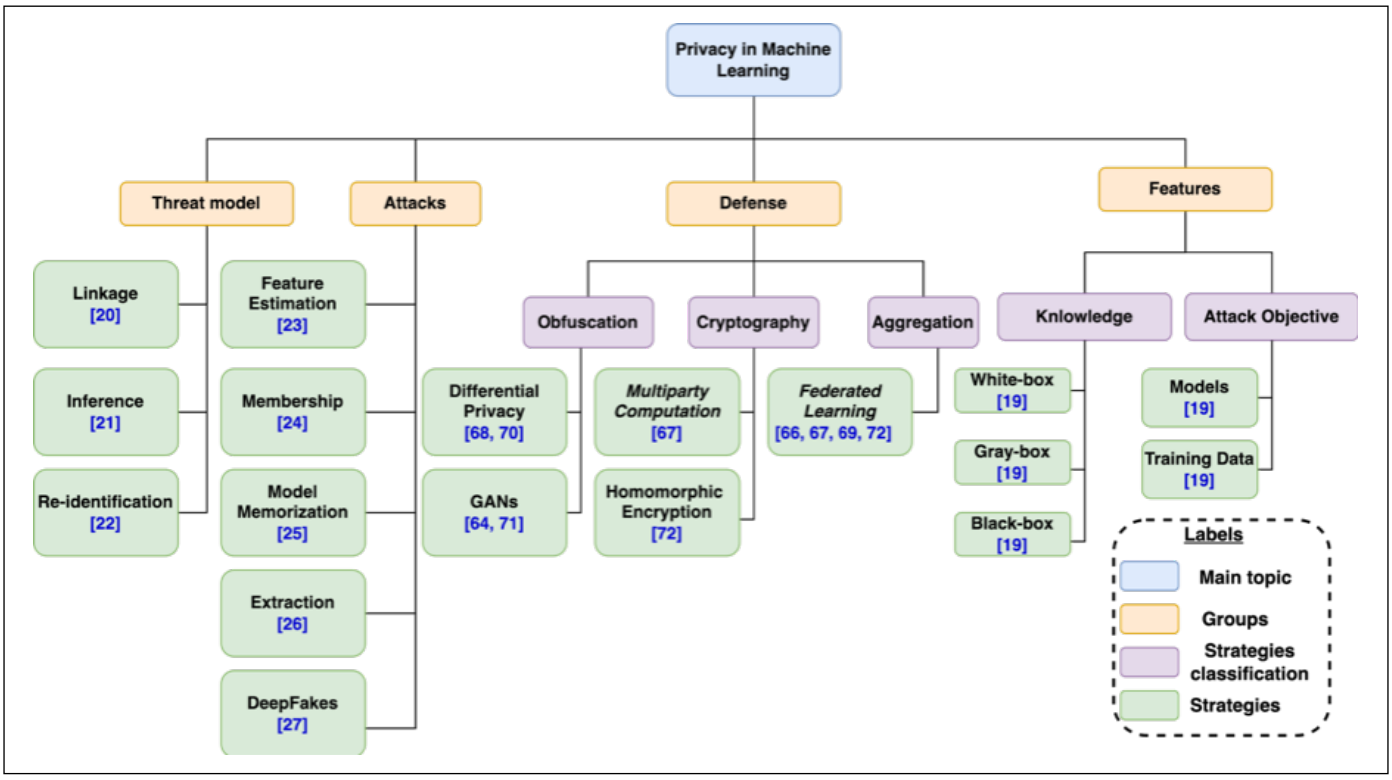

Fig. 4 Taxonomy of privacy in machine learning for health. The figure shows the definition of mitigation methods (center part), privacy attacks (left part), and features of attacks (right part).

In terms of medical datasets, papers in the literature are usually collected from public (e.g., Kaggle[2]) or private datasets, selecting different categories of medical images, such as X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Dermoscopy, Fundoscopy, and OCT. Most images analyzed correspond to brain, chest, skin, and eye, as well as COVID-19 images. Figure 5(a) shows the most employed medical datasets, including X-ray, CT, MRI, Dermoscopy, Fundoscopy, and EHRs. These datasets are exploited to generate attacks or to build defenses and privacy-preserving strategies. X-ray images are widely addressed to outline attack and defense strategies. Note that the papers collected from the literature are detailed in Tables 2, 3, and 4.

The papers collected from the literature described target databases, attack methods (see Table 2), defense methods (see Table 3), and privacy-preserving strategies (see Table 4). The next section describes the highlighted strategies applied to the attack, defense, and privacy-preserving ML models in health.

4.2 Highlighted Strategies of Security in Machine Learning for Health

Security strategies in ML for health applications must be aware of attacks and defenses for ML models. We summarized the literature collected from attacks in Table 2 and defenses in Table 3.

Papers have applied attacks such as FGSM, PGD, One Pixel, and UAPs. Furthermore, the authors propose strategies to attack the segmentation or classification task. Such papers [38, 46, 48] investigated attacks to fool the segmentation task using UNet[3] to generate perturbed masks. In the classification task, papers [35] and [41-44] employed the FGSM attack, [35, 41, 44] the PGD attack, [39, 40] the UAP attack, [37] the One Pixel attack, and [46, 48, 60] GANs-based attack. As far as DeepFake attacks are concerned, which generate fake data, e.g., inserting a malign tumor into a medical image that is supposed to be benign. These papers collect medical databases, including diverse categories. Figure 5(b) illustrates the categories of the most employed data by studies collected from the literature. The authors mostly applied Chest X-rays images due to the COVID-19 pandemic. Also, they have employed images of Dermoscopy, Fundoscopy, and OCT, as well as EHRs. Finally, MRI and CT images are applied less frequently than X-rays.

Regarding defenses in adversarial environments for health, papers explore strategies based on pre-processing, identification, mitigation, and surrogate models. Trending strategies are identifying attacks with GANs [51, 52] and OOD [57]. In addition, [53, 55, 59-64] develop strategies to mitigate AA us-

[2] https://www.kaggle.com/

[3] https://pytorch.org/hub/mateuszbuda_brain-segmentation-pytorch_unet/

Table 2 Summary of paper about attacks in machine learning for health.

ID	Paper	Objective
1	[35]	Investigating the advances of adversarial attacks in health. The authors applied the FGSM and PGD attacks to fool DL models and reduce their performance. The attack is tested on Chest X-ray images.
2	[36]	Create fake images with GAN to fool DL models. The authors inject an anomaly in clean X-ray images. They aim to need clarification on the model when it classifies.
3	[37]	Reduce performance of DL for classification dermoscopy images with Nevos Melanocíticos using the One-pixel attack.
4	[38]	Explore attacks on segmentation using the Multi-Scale attack to generate malicious masks on Dermatological lesion images.
5	[39]	The authors applied natural perturbation to reduce DL performance. Attacks with natural features can contribute to hidden perturbation on the clean image. The tests were performed on Fundoscopy, Chest X-Ray, and Dermoscopy images.
6	[40]	Exploring Universal Adversarial Perturbation (UAPs) applied to Chest X-ray, Melanoma, and OCT images. This paper aims to demonstrate that UAPs can reduce the performance of CNNs
7	[41]	Investigating the effect of natural images of ImageNet with attacks. The proposal addresses adversarial attacks using UAPs against images from OCT, Melanoma, and X-ray.
8	[42]	Explore the FGSM attacks on Chest X-ray images of COVID-19. The authors compare the attacked images in front of models trained on clean images to reduce their performance.
9	[34]	Generate Malicious 3D images that contain beginning or malign features when the X-Ray machine transfers the image to the database.
10	[43]	The authors studied the effects of batch normalization to produce vulnerable networks. The experiments are carried out using FGSM and PGD attacks on X-ray medical images.
11	[44]	The authors proposed an attack on COVID-19 X-ray images. They intend to test white-box and black-box attacks, as well as compare the no-sign attack to FGSM and PGD.
12	[45]	Proposed a black-box attack in Medical images of Magnetic Resonance (MRI), CT Scans, and X-rays using the Watermark technique. This technique embedded a watermark in the image by Krawtchouk moments to produce adversarial images that fool DL models.
13	[46]	This paper addressed a method to attack DL models for image segmentation. The authors create a technique based on differential evolution, which optimizes the space of attack aimed at better than gradient-based attack. The datasets attacked were Glaucoma, Lung, Melanoma, and ultrasound.
14	[47]	The vulnerability of DL models during the segmentation of medical images is evaluated. The authors applied the Adaptive Segmentation Mask Attack (ASMA) attack that generates adversarial masks to segment dermoscopy and glaucoma images.
15	[48]	Proposed an approach to attack segmentation models for medical images. The authors employ a Variational Autoencoder to generate adversarial examples of CT images.
16	[49]	Exploring vulnerabilities of CNNs NasNet-Large and Inception-ResNet-v2 on Chest X-ray images.
17	[50]	A GAN to generate fake X-ray images of mammogram breast cancer. The authors create a DeepFake strategy to insert cancerous tissues into the image.

ing adversarial training and applying GANs. GANs methods could generate synthetic AE to teach models the features of an AE during adversarial training. The work in [49] aims to identify anomalies in the frequency domain and mitigate AA in medical images. Besides, we found that the surrogate models can reduce attack effects, and GANs strategies tend to be most applied in this context, such as [54], [59-63]. Figure 5(c) summarizes the number of studies that proposed defensive strategies, corresponding attacks, and defenses. To evaluate defenses, we raised attacks FGSM, PGD, One Pixel, C&W, segmentation, and GANs. FGSM and segmentation attacks are most used to test the strengths of defenses, such as identification, GANs, surrogate model, frequency domain, and adversarial training. Furthermore, papers mostly build defenses based on surrogate models with GANs.

4.3 Highlighted Strategies of Privacy in Machine Learning for Health

We analyzed trends in privacy in ML for health (detailed in Table 4), collecting privacy-preserving strategies, such as FL, DP, HE, and MPC, as well as other defenses against re-identification attacks. Note that [68, 70] and [64, 71] present obfuscation with DP and GANs, respectively. Other strategies are the cryptographic one with MPC in [67] and HE in [72]. Besides, the most employed strategy is the FL addressed by [66, 67, 69, 72] to protect privacy in datasets of X-rays images, Electronic Health Records (EHR), and Electrocardiograms (EGG). The works in [66, 67, 68, 72] carry out strategies to protect DL models trained on X-ray images. Another important issue is tackled in [65], which investigates protections against re-identification in Picture Archiving and Communication Systems (PACS), and [70] that mitigates leakage in EHRs data. On the other hand, [27] generates DeepFake images in EGGs.

We observed that privacy-preserving strategies are evaluated over attacks against privacy. Figure 5(d) shows re-identification, inference, and DeepFakes attacks against FL, DP, HE, and MPC. FL is the most applied privacy-preserving strategy, which privately trains minimal models to share sensitive data. Our findings corroborate that the inference attack is the most applied attempt to infer sensitive attributes from a dataset. Re-identification attacks are exploited in the health context as well. DP and HE play a role in mitigating re-identification attacks in datasets of images and EHR. In summary, the papers related to AA, defenses, and privacy concerns contribute to improving the discussion of security and privacy in ML for health.

Table 3 Summary of paper about defenses in machine learning for health.

ID	Paper	Objective
18	[51]	Identifying adversarial examples using GANs. The proposed method analyzes the frequency domain to recognize images affected by perturbations.
19	[52]	It proposed a compression method based on an image JPEG to filter the frequency domain and mitigate the effect of adversarial attacks.
20	[53]	The proposal seeks to improve model generalization by adding synthetic data with adversarial examples. These examples are essential to model and learn AE. The method proposed to enhance the detection of AE on CT Scan datasets.
21	[54]	The work identifies malicious labels on data using a GAN, improving the model's robustness.
22	[55]	The paper reduces the adversarial attacks on the Gradient Descent method. They propose the Stochastic Coordinate Descent to minimize the loss when are adversarial medical examples.
23	[56]	The authors proposed a defense method by denoising operators to mitigate perturbations on medical images. This defensive method aims to mitigate FGSM and PGD attacks on Chest X-rays images.
24	[57]	This defensive method proposes to improve the robustness of DL models by evaluating Out-Of-Distribution (OOD). The authors proposed the Mahalanobis confidence score to detect OOD and mitigate the effects of AA on healthy and infected blood smear images. Also, the authors applied detection against FGSM, BIM, DeepFool, and C&W attacks.
25	[58]	The method proposed is based on the ensemble of CNNs to mitigate adversarial attacks, such as FGSM and One Pixel. The ensemble was tested in Lung nodules CT images.
26	[59]	Proposed a defensive plan against FGSM, PGD, Basic Iterative Method (BIM), and Momentum Iterative Method (MIM) attacks on Chest X-rays images. The authors employ the Multivariate Gaussian Model to identify features that do examples malicious.
27	[60]	The authors proposed the MedRDF, a framework to defend medical diagnosis against adversarial attacks and enhance the robustness of the DL model. The framework has a voting system to select denoised images and evaluate ones for the Robust Metric. The author conducted experiments on COVID-19 images of X-rays.
28	[61]	The method proposed is based on adversarial training using images orthogonal momentous. The authors tested the technique for classification and segmentation in X-rays and histopathology images when performing the attacks PGD and FGSM.
29	[62]	Analyze DL models under attacks to compare the effects of their complexity on adversarial robustness. The authors show that in adversarial training, all model complexity carries out similar robustness. They execute tests on Chest X-ray, Dermoscopy, and OCT images.
30	[63]	This paper aims to analyze the adversarial robustness in MRI knee images when applied to CNNs to reconstruct images.
31	[64]	This paper proposed a robust method to improve DL models against the FGSM attack trained on MRI brain tumor images. The strategy proposed aims to apply adversarial training. Therefore, adding the Gaussian noise to correct the confidence of the classifier.

Table 4 Summary of paper about privacy in machine learning for health.

ID	Paper	Objective
32	[65]	Privacy-preserving in medical images. The authors proposed an approach to mitigate re-identification attacks, integrating Picture Archiving and Communication System (PACS).
33	[66]	A method for sharing private medical data between organizations using FL. Their proposal shares various minimal models that train on sensitive data. The central server aggregates minimal models into a general model, which results in a model trained on private data.
34	[67]	Develop a framework to improve the security of Federated Learning with a central aggregation server. The proposal combines Differential Privacy and Multiparty Computation to protect the aggregation server. Thus, they mitigate leakage at the server level when sharing chest X-ray COVID-19 images.
35	[68]	The authors propose a PyTorch framework to private the chest X-ray images during the gradient calculus for segmentation and classification. The method applies Gaussian Differential Privacy to improve privacy-preserving in Stochastic Gradients.
36	[69]	Exploits the poisoning attacks and proposes the MediSecFed to protect the privacy of chest X-ray images. They are sharing medical images using FL for privacy-preserving.
37	[27]	Generate DeepFake electrocardiograms (EGGs) using a network called WaveGAN. The authors generate synthetic data aimed at hidden EEG real to enhance the privacy-preserving of users.
38	[70]	Propose a method based on DP to publish EHR diabetes medical data and make available private way sensitive data. Also, the authors tested the DP on a mini-bath of CNNs to train models and prevent attacks.
39	[71]	This paper proposed a GAN to generate high-quality synthetic data. Thus, the authors became the original data private and made it valuable for training other models. The experiments were performed on EHR to generate synthetic attributes.
40	[72]	This paper proposed a privacy-preserving strategy based on FL and HE to protect the aggregation server. The tests were carried out with chest X-ray COVID-19 images to train CNNs private way.

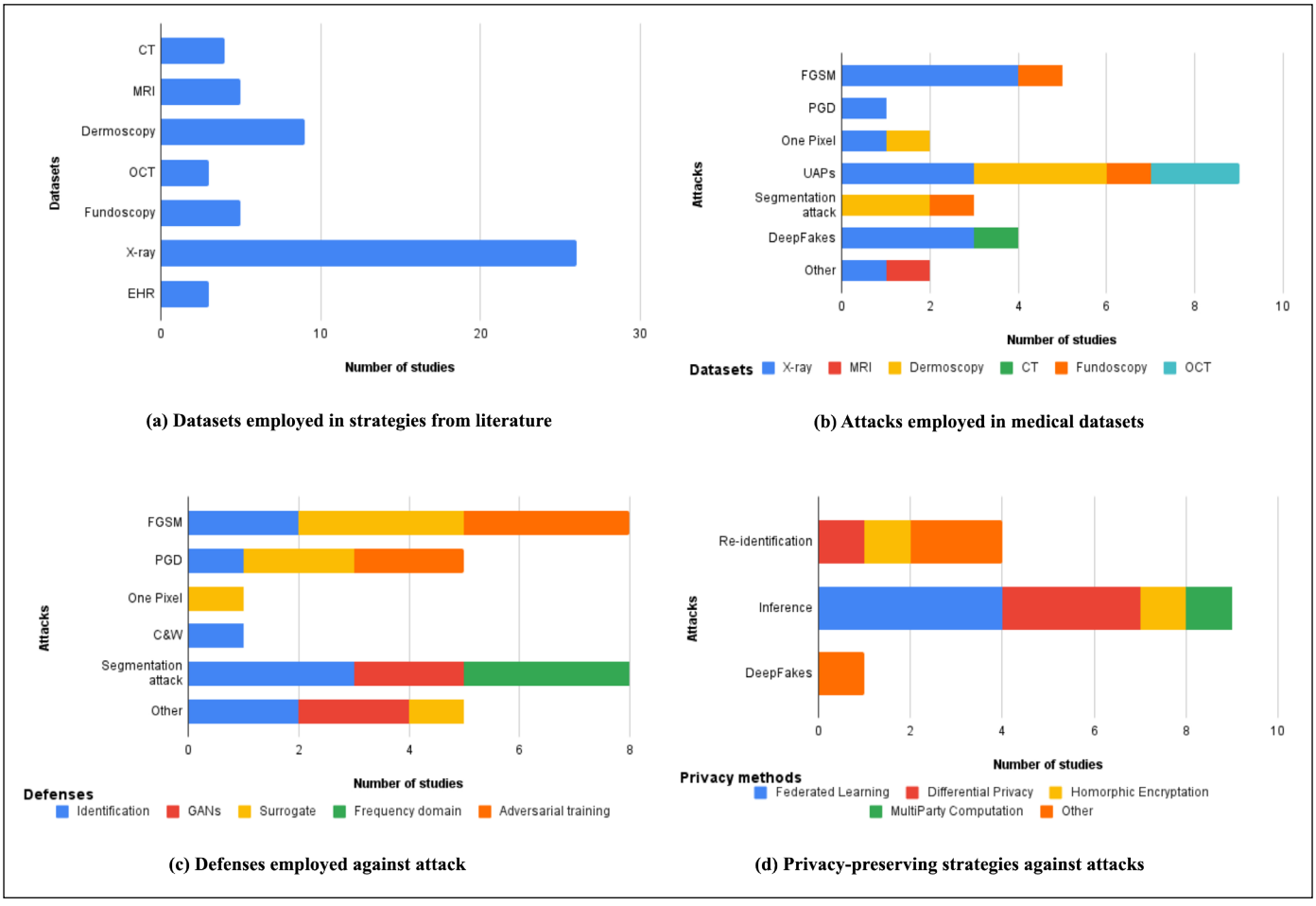

Fig. 5 literature review results related to most used medical datasets, attacks, defenses, and privacy-preserving strategies.

4.4 Tools

Tools are established in the literature to produce attack, defense, and privacy-preserving strategies. AA can be generated using Python libraries SecML[4], Adversarial Robustness Toolbox[5] (ART), and TorchAttacks[6]. ART is practical because it implements tools to generate adversarial attacks and defenses, including attacks and defenses for privacy. SecML and TorchAttacks run AA, such as FGSM, PGD, One Pixel, and others. SecML works on TensorFlow models and TorchAttacks on Pytorch models. Moreover, the most used tools to build privacy-preserving strategies are TensorFlow or PyTorch for FL, PyDP for DP from DeepMind, Microsoft SEAL for HE, and MPyC. Finally, ART can run defenses against AA, such as Adversarial Training and Defensive Distillation.

5 Discussion

This section examines trends and challenges related to attacks, defenses in adversarial environments, and privacy concerns for health.

5.1 Trends

We presented the tendencies and directions of AA regarding defenses and privacy concerns in ML for health. Figure 6 presents a timeline of the primary studies that introduced trends of AA, including defenses and privacy-preserving. Likewise, we define a timeline from 2018 to 2022 that includes the main trends. Directions in AA for health address PGD and FGSM attacks used to generate efficient AE and to distort DL models [35]. DL models ResNet50, VGG19, VGG16, and InceptionV3 are widely employed to classify medical images. Besides, attacks target to corrupt these models and reduce their performance. The trends for

[4] https://secml.readthedocs.io/en/v0.15/
[5] https://github.com/Trusted-AI/adversarial-robustness-toolbox
[6] https://github.com/Harry24k/adversarial-attacks-pytorch

analyzing attacks started in 2018, exploring attacks vulnerabilities to corrupt NasNet and Inception trained on medical images. In 2019, the papers exploited the attacks FGSM, PGD, segmentation, and GAN-based, as well as attacks to generate DeepFakes on medical images. In 2020, papers employed attacks to build DeepFakes, and run the One Pixel attack. The attack trends in 2021 were UAPs and DeepFakes generators. In addition, trends in 2022 address UAPs, FGSM, and other strategies, such as attacks based on watermarks. Finally, the directions to develop new attacks in medical images in the next years follow the DeepFakes generator and UAPs.

Defenses against AA for health systems need to improve the model's robustness. Figure 6 illustrates the timeline regarding the trends of attacks, defenses, and privacy-preserving strategies. Our review did not find defenses against AA for the health environment from 2018 to 2019. Trends for defenses in 2020 focused on mitigating attacks using GANs, adversarial training, and detecting corrupted models. GANs strategies improve the robustness of the discriminator model to identify AE and the generator to reconstruct examples without perturbations [51, 54]. In 2021, defenses focused on identifying attacks employing GANs and OOD strategies. Directions in 2022 were towards of creating novel strategies, such as Stochastic Coordinate Descent [55], perturbation denoising [56, 60, 61], and enhanced adversarial training [61, 65]. In health environments, we need to improve model defenses at the system level due to the sensitivity of the data handled. According to our extracted data, we should develop novel defenses based on GANs and propose new strategies for health systems.

Privacy-preserving trends have led to strategies to mitigate the leakage of sensitive-health data in ML for health. In 2018, according to our review, papers mainly addressed methods to mitigate re-identification attacks in PACS and integrate ML methods into medical images. Tendencies in 2019 were protecting mini-batches of DL models and EHRs by employing DP to obfuscate the original content. In 2021, papers committed FL to share medical images, DP to protect sensitive attributes, and GANs for generating synthetic attributes based on sensitive attributes. Directions in 2022 tend to generate synthetic and sensitive data to hide the original content and combine privacy strategies to enhance FL, such as HE. Moreover, literature on health privacy tends to combine FL, DP, or MPC [67]. When handling unstructured data, such as images, privacy protection methods are needed to improve their protection.

5.2 Challenges

The scenario of AA in health systems has challenges, such as building powerful defenses to the AA, which focus on poisoning and evasion, bypassing fine-tuning methods, transferability of attacks, and attacks on real-world databases. Poisoning and evasion attacks aim to explore vulnerabilities in DL, applying UAPs, AutoAttack [73], and GANs. DL models tend to use fine-tuning strategies. When an attack affects DL models, a

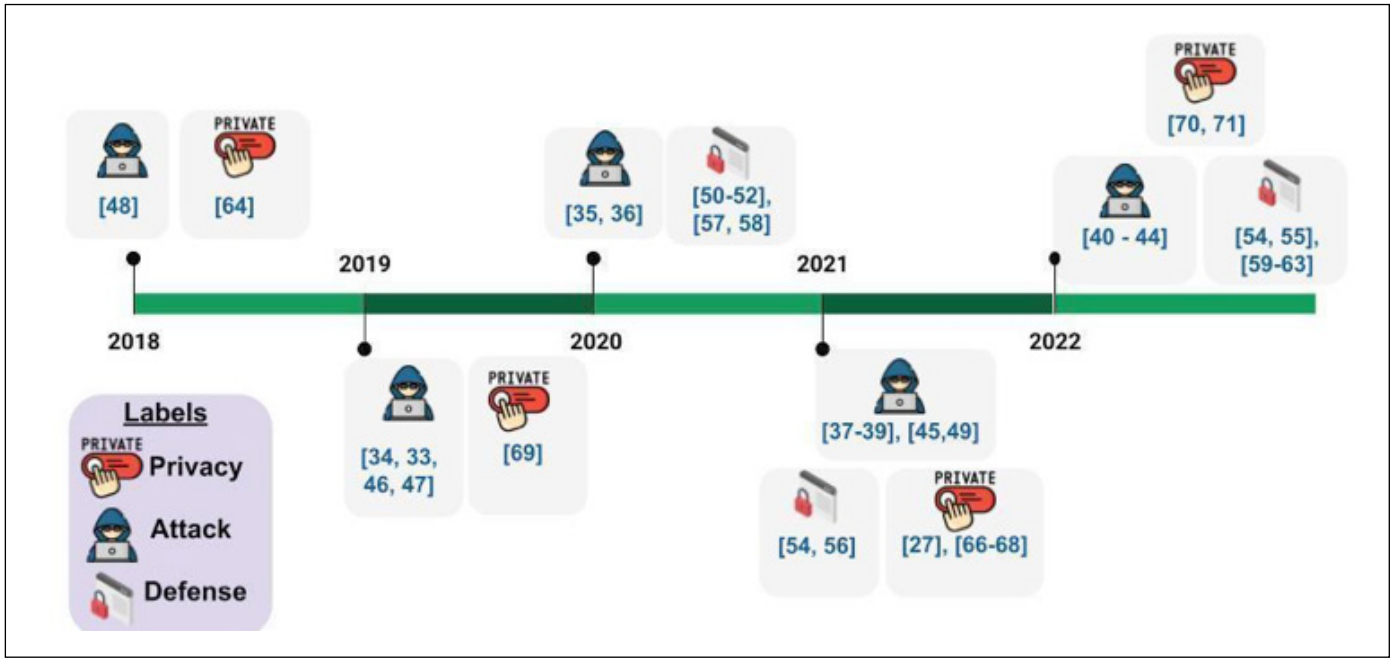

Fig. 6 Timeline of papers collected from the literature between 2018 and 2022. Each paper covers the privacy, attack, or defense domain.

challenge is to develop a method to bypass the fine-tuning strategy aimed at enhancing misclassification. Attack transferability is a relevant feature that can handle and indicate whether an attack is transferable to other domains. A challenge is treating transferability when building an attack to make it more generalizable. Developing attacks in real-world environments is arduous because the attack may have another behavior that needs fixing in the modeling phase.

Defensive methods are necessary and pose an arduous task in protecting ML for health. We collected the main challenges to creating defensive strategies using proactive and reactive approaches for applying GANs, equilibrating privacy and defense, and calibrating models. Proactive defenses identify attacks before the attack happens, and reactive defenses work after the attack happens, aimed at mitigating the ill effects. Likewise, GANs are methods for building robust defenses because they can simulate attack scenarios and generate synthetic data to emulate malicious ones. Equilibrating privacy and defenses are challenging because defenses can show more information than they should. Based on privacy concerns, papers achieve a calibrated approach as an alternative to improve the model security, because it represents a more robust approach.

We observe that privacy-preserving strategies are challenging to develop Federated Learning (FL) privacy, equilibrating privacy and accuracy scores, as well as setting the privacy budget, protecting privacy in medical images, and combining privacy methods. However, in FL, the aggregate server can suffer attacks, and its security should be improved. Besides, the privacy strategy can reduce the DL model's performance. Privacy in unstructured data is challenging because the methods proposed, such as DP, work better with tabular data. Then, we must explore the method of privacy-preserving that works in medical images. The combination of privacy techniques should be a robust strategy to improve other methods, such as combining Federated Learning (FL) with Differential Privacy (DP) or FL with MultiParty Computation (MPC). Another challenge to combine these techniques is to find a suitable method to improve the privacy budget while keeping the accuracy level.

Finally, we highlight that the development of novel attacks, defenses, and privacy strategies have room for improvement. Each technique can contribute to another, such as exploring vulnerabilities to produce attacks leads to building novel defenses. Defensive methods can improve the robustness of DL models. Nevertheless, it can result in privacy issues. Thus, the defense method will be modeled based on gaps in defenses. In turn, privacy strategies are concerned with the performance of models because high-budget privacy levels can result in poor model accuracy.

6 Conclusion

We presented a survey on recent works from the literature and discussed health-related strategies and challenges regarding security and privacy in ML for health systems. We classified the papers into three domains: security, defenses against adversarial attacks (AA), and privacy concerns. The AA strategies cover gradient and optimization attacks, as well as defenses inspired by GANs to make adaptive strategies and generate synthetic Adversarial Examples (AE).

Regarding privacy, the strategies frequently applied are based on FL. However, each strategy comprehends issues, such as attacks that bypass fine-tuning, defenses that work reactively and proactively, and privacy based on methods for unstructured data. In summary, we highlight that security and privacy for health systems remain a strong trend for the next years. According to [17], developing ML models on sensitive data should always consider their risk and vulnerability.

Acknowledgments

We thank the support from São Paulo Research Foundation (FAPESP) grants 2016/17078-0, 2020/07200-9, and 2021/08982-3, National Council for Scientific and Technological Development (CNPq), and Coordination for Higher Education Personnel Improvement (CAPES) grant 001.

References

1. Aiello M, Cavaliere C, D'Albore A, Salvatore M. The challenges of diagnostic imaging in the era of big data. J Clin Med 2019 Mar 6;8(3):316. doi: 10.3390/jcm8030316.
2. Pitropakis N, Panaousis E, Giannetsos T, Anastasiadis E, Loukas G. A taxonomy and survey of attacks against machine learning. Comput Sci Rev 2019;34:100199. doi: 10.1016/j.cosrev.2019.100199
3. Machado GR, Silva E, Goldschmidt RR. Adversarial machine learning in image classification: A survey toward the defender's perspective. ACM Computing Surveys (CSUR) 2021 Nov 23;55(1):1-38. doi: 10.1145/3485133.
4. Biggio B, Roli F. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS '18). New York, NY, USA: Association for Computing Machinery; 2018. p. 2154–6. doi: 10.1145/3243734.3264418.
5. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083. 2017 Jun 19.
6. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. 2014 Dec 20.
7. Su J, Vargas DV, Sakurai K. J. One Pixel Attack for Fooling Deep Neural Networks. IEEE Trans Evol Comput 2019;23(5):828-41. doi: 10.1109/TEVC.2019.2890858
8. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The Limitations of Deep Learning in Adversarial Settings. 2016 IEEE European Symposium on Security and Privacy (EuroS&P), Saarbruecken, Germany; 2016. p. 372-87. doi: 10.1109/EuroSP.2016.36.
9. Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition 2016. p. 2574-82.
10. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. Towards Evaluating the Robustness of Neural Networks. 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA; 2017. p. 39-57. doi: 10.1109/SP.2017.49.
11. Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1765-73.
12. Xiao Q, Chen Y, Shen C, Chen Y, Li K. Seeing is not believing: Camouflage attacks on image scaling algorithms. 28th USENIX Security Symposium (USENIX Security 19); 2019. p. 443-60.
13. Quiring E, Klein D, Arp D, Johns M, Rieck K. Adversarial preprocessing: Understanding and preventing Image-Scaling attacks in machine learning. 29th USENIX Security Symposium (USENIX Security 20); 2020. p. 1363-80.
14. Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks.

arXiv preprint arXiv:1704.01155. 2017 Apr 4.
15. Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA; 2016. p. 582-97. doi: 10.1109/SP.2016.41.
16. Meng D, Chen H. Magnet: a two-pronged defense against adversarial examples. Proceedings of the 2017 ACM SIGSAC conference on computer and communications security; 2017 Oct 30. p. 135-47. doi: 10.1145/3133956.3134057.
17. Baracaldo N, Oprea A. Machine Learning Security and Privacy. IEEE Secur Priv 2022;20(5):11-3. doi: 10.1109/MSEC.2022.3188190.
18. Strobel M, Shokri R. Data Privacy and Trustworthy Machine Learning. IEEE Secur Priv 2022;20(5):44-9. doi: 10.1109/MSEC.2022.3178187.
19. Liu B, Ding M, Shaham S, Rahayu W, Farokhi F, Lin Z. When machine learning meets privacy: A survey and outlook. ACM Computing Surveys (CSUR) 2021 Mar 5;54(2):1-36. doi: 10.1145/3436755.
20. Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. 2008 IEEE Symposium on Security and Privacy (SP 2008), Oakland, CA, USA; 2008. p. 111-25. doi: 10.1109/SP.2008.33.
21. Krumm J. Inference attacks on location tracks. Pervasive Computing. Pervasive 2007. Lecture Notes in Computer Science, vol 4480. Berlin, Heidelberg: Springer. P. 127-43. doi: 10.1007/978-3-540-72037-9_8.
22. Henriksen-Bulmer J, Jeary S. Re-identification attacks—A systematic literature review. Int J Inf Manag 2016 Dec 1;36(6):1184-92. doi: 10.1016/j.ijinfomgt.2016.08.002.
23. Atenfiese G, Mancini LV, Spognardi A, Villani A, Vitali D, Felici G. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. International Journal of Security and Networks 2015;10(3):137-50. doi: 10.1504/IJSN.2015.071829.
24. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA; 2017. p. 3-18. doi: 10.1109/SP.2017.41.
25. Song C, Ristenpart T, Shmatikov V. Machine learning models that remember too much. Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security: 2017 Oct 30. p. 587-601. doi: 10.1145/3133956.3134077.
26. Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T. Stealing machine learning models via prediction APIs. 25th USENIX security symposium (USENIX Security 16); 2016. p. 601-18.
27. Thambawita V, Isaksen JL, Hicks SA, Ghouse J, Ahlberg G, Linneberg A, et al. DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. Sci Rep 2021 Nov 9;11(1):21896. doi: 10.1038/s41598-021-01295-2.
28. Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science 2014 Aug 10;9(3–4):211-407. doi: 10.1561/0400000042.
29. Yi X, Paulet R, Bertino E. Homomorphic Encryption. In: Homomorphic Encryption and Applications. SpringerBriefs in Computer Science. Cham: Springer; 2014. doi: 10.1007/978-3-319-12229-8_2.
30. Cramer R, Damgård I. Multiparty computation, an introduction. In: Contemporary Cryptology 2005. Advanced Courses in Mathematics - CRM Barcelona. Basel: Birkhäuser; 2005. p. 41-87. doi: 10.1007/3-7643-7394-6_2.
31. Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492. 2016 Oct 18.
32. Kitchenham B, Brereton P. A systematic review of systematic review process research in software engineering. Inf Softw Technol 2013 Dec 1;55(12):2049-75. doi: 10.1016/j.infsof.2013.07.010.
33. Brereton P, Kitchenham BA, Budgen D, Turner M, Khalil M. Lessons from applying the systematic literature review process within the software engineering domain. J Syst Softw 2007 Apr 1;80(4):571-83. doi: 10.1016/j.jss.2006.07.009.
34. Mirsky Y, Mahler T, Shelef I, Elovici Y. {CT-GAN}: Malicious Tampering of 3D Medical Imagery using Deep Learning. 28th USENIX Security Symposium (USENIX Security 19); 2019. p. 461-78.
35. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019 Mar 22;363(6433):1287-9. doi: 10.1126/science.aaw4399.
36. Mangaokar N, Pu J, Bhattacharya P, Reddy CK, Viswanath B. Jekyll: Attacking medical image diagnostics using deep generative models. 2020 IEEE European Symposium on Security and Privacy (EuroS&P), Genoa, Italy; 2020. p. 139-57. doi: 10.1109/EuroSP48549.2020.00017.
37. Allyn J, Allou N, Vidal C, Renou A, Ferdynus C. Adversarial attack on deep learning-based dermatoscopic image recognition systems: Risk of misdiagnosis due to undetectable image perturbations. Medicine (Baltimore) 2020 Dec 11;99(50):e23568. doi: 10.1097/MD.0000000000023568.
38. Shao M, Zhang G, Zuo W, Meng D. Target attack on biomedical image segmentation model based on multi-scale gradients. Information Sciences 2021 Apr 1;554:33-46. doi: 10.1016/j.ins.2020.12.013.
39. Ma X, Niu Y, Gu L, Wang Y, Zhao Y, Bailey J, et al. Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition 2021 Feb 1;110:107332. doi: 10.1016/j.patcog.2020.107332.
40. Hirano H, Minagi A, Takemoto K. Universal adversarial attacks on deep neural networks for medical image classification. BMC Med Imaging 2021 Jan 7;21(1):9. doi: 10.1186/s12880-020-00530-y.
41. Minagi A, Hirano H, Takemoto K. Natural Images Allow Universal Adversarial Attacks on Medical Image Classification Using Deep Neural Networks with Transfer Learning. J Imaging 2022 Feb 4;8(2):38. doi: 10.3390/jimaging8020038.
42. Aguiar EJ, Marcomini KD, Quirino FA, Gutierrez MA, Traina Jr C, Traina AJ. Evaluation of the impact of physical adversarial attacks on deep learning models for classifying covid cases. Proc SPIE 12033, Medical Imaging 2022: Computer-Aided Diagnosis, 120332P (4 April 2022); doi: 10.1117/12.2611199.
43. Kong F, Liu F, Xu K, Shi X. Why does batch normalization induce the model vulnerability on adversarial images? World Wide Web 2023;26:1073–91. doi: 10.1007/s11280-022-01066-7.
44. Wei C, Sun R, Li P, Wei J. Analysis of the No-sign Adversarial Attack on the COVID Chest X-ray Classification. 2022 International Conference on Image Processing and Media Computing (ICIPMC), Xi'an, China; 2022. p. 73-9. doi: 10.1109/ICIPMC55686.2022.00022.
45. Apostolidis KD, Papakostas GA. Digital Watermarking as an Adversarial Attack on Medical Image Analysis with Deep Learning. J Imaging 2022 May 30;8(6):155. doi: 10.3390/jimaging8060155.
46. Cui X, Chang S, Li C, Kong B, Tian L, Wang H, et al. DEAttack: A differential evolution based attack method for the robustness evaluation of medical image segmentation. Neurocomputing 2021 Nov 20;465:38-52. doi: 10.1016/j.neucom.2021.08.118.
47. Ozbulak U, Van Messem A, Neve WD. Impact of adversarial examples on deep learning models for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science 2019;11765:300-8. doi: 10.1007/978-3-030-32245-8_34.
48. Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D. Intelligent image synthesis to attack a segmentation CNN using adversarial learning. Simulation and Synthesis in Medical Imaging. SASHIMI 2019. Lecture Notes in Computer Science 2019;11827: 90-9. doi: 10.1007/978-3-030-32778-1_10.
49. Asgari Taghanaki S, Das A, Hamarneh G. Vulnerability analysis of chest X-ray image classification against adversarial attacks. In: Understanding and interpreting machine learning in medical image computing applications 2018 Sep 20 MLCN DLF IMIMIC 2018. Lecture Notes in Computer Science 2018;11038. doi: 10.1007/978-3-030-02628-8_10.
50. Zhou Q, Zuley M, Guo Y, Yang L, Nair B, Vargo A, et al. A machine and human reader study on AI diagnosis model safety under attacks of adversarial images. Nat Commun 2021 Dec 14;12(1):7281. doi: 10.1038/s41467-021-27577-x.
51. Park H, Bayat A, Sabokrou M, Kirschke JS, Menze BH. Robustification of Segmentation Models Against Adversarial Perturbations in Medical Imaging. In: International Workshop on Predictive Intelligence in Medicine, 2020 Oct 8. p. 46-57. Cham: Springer; 2020 .
52. Liu Q, Jiang H, Liu T, Liu Z, Li S, Wen W, et al. Defending deep learning-based biomedical image segmentation from adversarial attacks: a low-cost frequency refinement approach. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020 Oct 4. Cham: Springer; 2020: p. 342-51.
53. Liu S, Setio AA, Ghesu FC, Gibson E, Grbic S, Georgescu B, et al. No surprises: Training robust lung nodule detection for low-dose CT scans by augmenting with adversarial attacks. IEEE Trans Med Imaging 2021 Jan;40(1):335-45. doi: 10.1109/TMI.2020.3026261.
54. Qi X, Hu J, Yi Z. Missed diagnoses detection by adversarial learning. Knowl Based Syst

2021 May 23;220:106903. doi: 10.1016/j.knosys.2021.106903.
55. Yang Y, Shih FY, Roshan U. Defense Against Adversarial Attacks Based on Stochastic Descent Sign Activation Networks on Medical Images. Intern J Pattern Recognit Artif Intell 2022 Mar 15;36(03):2254005. doi: 10.1142/S0218001422540052
56. Shi X, Peng Y, Chen Q, Keenan T, Thavikulwat AT, Lee S, et al. Robust convolutional neural networks against adversarial attacks on medical images. Pattern Recognition 2022 Dec 1;132:108923. doi: 10.1016/j.patcog.2022.108923.
57. Uwimana A, Senanayake R. Out of distribution detection and adversarial attacks on deep neural networks for robust medical image analysis. arXiv preprint arXiv:2107.04882. 2021 Jul 10.
58. Paul R, Schabath M, Gillies R, Hall L, Goldgof D. Mitigating adversarial attacks on medical image understanding systems. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA; 2020. p. 1517-21. doi: 10.1109/ISBI45749.2020.9098740.
59. Li X, Zhu D. Robust detection of adversarial attacks on medical images. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA; 2020. p. 1154-8. doi: 10.1109/ISBI45749.2020.9098628.
60. Xu M, Zhang T, Zhang D. Medrdf: a robust and retrain-less diagnostic framework for medical pretrained models against adversarial attack. IEEE Trans Med Imaging 2022 Aug;41(8):2130-43. doi: 10.1109/TMI.2022.3156268.
61. Maliamanis TV, Apostolidis KD, Papakostas GA. How Resilient Are Deep Learning Models in Medical Image Analysis? The Case of the Moment-Based Adversarial Attack (Mb-AdA). Biomedicines 2022 Oct 12;10(10):2545. doi: 10.3390/biomedicines10102545.
62. Rodriguez D, Nayak T, Chen Y, Krishnan R, Huang Y. On the role of deep learning model complexity in adversarial robustness for medical images. BMC Med Inform Decis Mak 2022 Jun 20;22(Suppl 2):160. doi: 10.1186/s12911-022-01891-w.
63. Morshuis JN, Gatidis S, Hein M, Baumgartner CF. Adversarial Robustness of MR Image Reconstruction Under Realistic Perturbations. In: Haq N, Johnson P, Maier A, Qin C, Würfl T, Yoo , editors. Machine Learning for Medical Image Reconstruction. MLMIR 2022. Lecture Notes in Computer Science, vol 13587. Cham: Springer; 2022. doi: 0.1007/978-3-031-17247-2_3.
64. Gupta D, Pal B. Vulnerability Analysis and Robust Training with Additive Noise for FGSM Attack on Transfer Learning-Based Brain Tumor Detection from MRI. In: Arefin MS, Kaiser MS, Bandyopadhyay A, Ahad MAR, Ray K, editors. Proceedings of the International Conference on Big Data, IoT, and Machine Learning. Lecture Notes on Data Engineering and Communications Technologies, vol 95. Singapore: Springer; 2022. doi: 10.1007/978-981-16-6636-0_9.
65. Silva JM, Pinho E, Monteiro E, Silva JF, Costa C. Controlled searching in reversibly de-identified medical imaging archives. J Biomed Inform 2018 Jan;77:81-90. doi: 10.1016/j.jbi.2017.12.002.
66. Feki I, Ammar S, Kessentini Y, Muhammad K. Federated learning for COVID-19 screening from Chest X-ray images. Applied Soft Computing 2021 Jul 1;106:107330. doi: https://doi.org/10.1016/j.asoc.2021.107330.
67. Kaissis G, Ziller A, Passerat-Palmbach J, Ryffel T, Usynin D, Trask A, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nat Mach Intell 2021 Jun;3(6):473-84. 10.1038/s42256-021-00337-8.
68. Ziller A, Usynin D, Braren R, Makowski M, Rueckert D, Kaissis G. Medical imaging deep learning with differential privacy. Sci Rep 2021 Jun 29;11(1):13524. doi: 10.1038/s41598-021-93030-0.
69. Kumar A, Purohit V, Bharti V, Singh R, Singh SK. MediSecFed: Private and Secure Medical Image Classification in the Presence of Malicious Clients. IEEE Trans Industr Inform 2022;18(8):5648-57. doi: 10.1109/TII.2021.3138919.
70. Sun Z, Wang Y, Shu M, Liu R, Zhao H. Differential Privacy for Data and Model Publishing of Medical Data. IEEE Access 2019;7:152103-14. doi: 10.1109/ACCESS.2019.2947295.
71. Venugopal R, Shafqat N, Venugopal I, Tillbury BM, Stafford HD, Bourazeri A. Privacy preserving Generative Adversarial Networks to model Electronic Health Records. Neural Networks 2022 Sep 1;153:339-48. doi: https://doi.org/10.1016/j.neunet.2022.06.022.
72. Wibawa F, Catak FO, Kuzlu M, Sarp S, Cali U. Homomorphic Encryption and Federated Learning based Privacy-Preserving CNN Training: COVID-19 Detection Use-Case. Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference 2022 Jun 15. p. 85-90. doi: 10.1145/3528580.3532845.
73. Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. Proceedings of the 37th International Conference on Machine Learning, PMLR 2020;119:2206-16.

Correspondence to:
Erikson J. de Aguiar
E-mail: erjulioaguiar@usp.br