Systematic Literature Review of Privacy-Preserving Deep Learning in Medical Image Analysis

Arumoy Shome
Vrije Universiteit Amsterdam, The Netherlands
a2.shome@student.vu.nl

Saba Amiri
Universiteit van Amsterdam, The Netherlands
s.amiri@uva.nl

Adam S. Z. Belloum
Universiteit van Amsterdam, The Netherlands
a.s.z.belloum@uva.nl

ABSTRACT

The success of Deep Neural Networks with image classification prompted researchers to explore the applications of Deep Learning in Medical Imaging and Medical Image Analysis (MIA). Deep Neural Networks have sufficiently demonstrated their capabilities of performing MIA tasks tirelessly and with fewer errors as opposed to their human counterpart. However the challenge of training neural networks using sensitive medical data, without violating the privacy of patients remains an active field of research. Many solutions exist to address this concern, however a systematic review and analysis of these techniques is yet to be conducted. This paper attempts to conduct the first systematic review of privacy-preserving techniques to train deep learning models. Emphasis is especially put on the performance and privacy analysis of the techniques. In addition, the communication and runtime costs, the ability of the solutions to scale, tolerance to faults and the level of security against threats and attacks are also studied.

KEYWORDS

Systematic Literature Review, Deep Learning, Privacy Preserving, Medical Imaging, Medical Image Analysis, Federated Learning, Distributed Learning

1 INTRODUCTION

Medical imaging is a vital aspect of medical science, performed for critical surgeries and early detection and prevention of chronic illnesses. Using various technologies such as Magnetic Resonance Imaging (MRI), Computer Tomography (CT) and Positron Emission Tomography (PET), the internal structure of the human body is created to study or identify any abnormalities, that may exist. The study of medical images or Medical Image Analysis (MIA) is primarily conducted by radiologists since it requires years of training and practise. MIA of a single patient involves manual inspection of over 100 medical images and comprises of repetitive tasks (such as identification of a Region of Interest (ROI) and image segmentation) which can cause cognitive fatigue. MIA is thus a laborious task limited by the skills and experience of radiologists [14]. It is also a field with very low tolerance for human error since one mistake may result in incorrect diagnosis of a patient and potentially lead to loss of human life. While humans may not always perform repetitive tasks to the best of their abilities, machines will perform them tirelessly and consistently [9]. The remarkable success of deep neural networks in image classification, popularised by the ImageNet challenge, accelerated its adoption within the field of MIA [14, 18, 21]. Deep Learning (DL) models have had tremendous success, not only within MIA but also within other fields of medical science. Although this is an active field of research, several challenges still remain [20]. One such challenge is to learn whilst preserving the privacy of the patients.

DL models require large quantities of data and powerful machines to be able to perform large volumes of computation. This may not always be possible, for instance at smaller medical institutes or for rare diseases [6]. Larger medical institutes who meet these requirements are able to train DL models locally. These models however are biased since the data is not a good representation of the general population [6, 12, 30]. This is because the institute perhaps specialises in a certain disease or only observes specific medical problems in their patients due to their geographic location. DL models thus need to be trained using data collected from several institutes. The traditional, client-server architecture for training raises several privacy concerns. Centralised training requires the data from all participating institutes to be collected in a single server. However once the data is uploaded to the server, the institutes lose their data ownership and governance rights [15]. Furthermore, there is no guarantee that the data is transmitted and stored securely, making it vulnerable to attacks. Efforts have been made to create centralised repositories such as eICU Collaborative Research Database and The Cancer Genome Atlas which contain anonymised medical data. But data privacy and protection laws such as the GDPR in Europe and HIPAA in The United States pose a large overhead in the creation of such data repositories [1]. The data is required to be anonymised such that it cannot be traced back to the original patient. While this is a step towards preserving the privacy of patients, the anonymisation and de-identification process negatively affects the utility of the data of future research. Anonymisation also does not provide any guarantees against re-identification as anonymised data contains unique statistical fingerprints which can be exploited using linkage attacks [3, 16, 23]. Distributed learning addresses the data ownership and governance concerns of centalised training by moving the model to the data. While several distributed learning methodologies exist, they require the aid of additional data privacy protection techniques in order to learn without violating the privacy of patients [1–3, 6, 12, 28, 30].

Privacy-preserving deep learning is an active field of research and has made several advances in the last decade. While many systems and methodologies exist, a systematic review of the literature has not yet been attempted. The goal of this study is to identify existing DL systems used in the field of MIA which account for patient privacy. The study is conducted by first developing a set of research questions (RQs) as presented in Section 2. Keywords are extracted from the RQs which are used to construct search queries. A set of inclusion and exclusion criterion are applied to

the shortlisted papers to form the final list of papers included in this study. Each paper is summarised and critically analysed in Section 4 following which a comparative analysis is conducted in Section 5. The study concludes by stating some of its limitations and directions for future work in Section 6. Being the first of its kind, this study wishes to aid current and future researchers to get acquainted with the state-of-the-art solutions that exist in the field of privacy-preserving deep learning. As such, prior knowledge of Deep Learning and Medical Image Analysis amongst the readers is assumed.

2 STUDY DESIGN

The over arching goal of this systematic review is to identify the existing privacy-preserving techniques which can be used to train Deep Learning (DL) models using medical data. Additionally, this study seeks to analyse these privacy-preserving techniques in an attempt to identify the ones which provide good performance, are scalable and robust, and provide maximum protection against security threats that violate the patient's privacy. To aid in this goal, the following research questions are formulated.

RQ1. What are the existing systems, platforms or techniques that facilitate deep learning on medical images that do not invade the patient's privacy?
RQ2. What are the top performing deep learning models being used to perform tasks in medical imaging?
RQ3. What are the specific attacks that can compromise the security of these systems?
RQ4. What is the amount of data leakage in these systems?

The keywords extracted from the research questions were used to construct the following search queries. *Google Scholar* and *Scopus* were queried to identify papers relevant for this study.

(1) *(deep|machine) & learning & ("medical imaging"|"medical image analysis")*
(2) *(deep|machine) & learning & ("privacy preserving"|private) & medical & (imag*|data)*
(3) *(deep|machine|federated|distributed) & learning & "privacy preserving" & medical & (imag*|data)*
(4) *(deep|machine|federated|distributed) & learning & (attack|flaw|threat) & (security|privacy)*

The obtained papers were then cross-checked with the inclusion (IC) and exclusion criterion (EC) listed below. All papers satisfying one or more inclusion criterion whilst satisfying all exclusion criterion were shortlisted via a preliminary survey of the paper. Finally, the shortlisted papers were given a secondary survey to determine if they were relevant for the study and if so, were included in the review.

IC1. The paper presents a deep learning system, platform or technique for training using medical data.
IC2. The paper presents a deep learning system, platform of technique which can be used to train using distributed data.
IC3. The paper evaluates the performance of one or more deep learning models trained using the presented technique.
IC4. The paper reports the dataset used for training along with the accuracy or error of the model(s).
IC5. The paper analyses the runtime performance and communication overhead of the proposed technique.
IC6. The paper conducts privacy and security analysis of the proposed technique.
IC7. The paper identifies vulnerabilities or describes techniques to exploit the proposed technique.
EC1. The paper is published or expected to be published in a peer-reviewed journal.
EC2. The paper is written in English.

The shortlisted papers were reviewed to extract the following data attributes for further analysis.

(1) *Key*: A unique, bibtex key to identify each paper.
(2) *Author*: Authors of the paper.
(3) *Title*: Title of the paper.
(4) *Year*: Year of publication.
(5) *Journal*: Name of the journal/conference the paper was published in.
(6) *Dataset*: Name of the dataset(s) used to evaluate the performance of the proposed system.
(7) *Model*: Neural Network or ML model(s) used for training.
(8) *Performance*: Performance (accuracy or error) of the models trained using the proposed technique.
(9) *Communication*: Communication overhead of the proposed technique, if any.
(10) *Scalability*: If relevant, the number of participants the proposed technique can accommodate.
(11) *Reliability*: If relevant, how reliable is the proposed technique during a system failure?
(12) *Runtime*: Runtime overhead of the proposed technique, if any.
(13) *Privacy*: Quantity of data leakage of the proposed technique, if any.
(14) *Security*: Threats and attacks the proposed technique is vulnerable to, if any.

3 RELATED WORK

This section presents prior work and concepts that are relevant this this review. Related literature such as other literature studies and systematic reviews are presented first. An overview of data privacy techniques are presented next, followed by security threats and attacks which can be used to exploit deep neural networks.

3.1 Related Literature

The paper by Zerka et al. (2020) is the only other systematic review that was identified to be related to this review. The Zerka study reviewed 6 papers that presented distributed machine learning techniques which do not violate the privacy of patients. The Zerka study emphasises on the ethics and data governance of distributed machine learning. In contrast, the goal of this study is to analyse privacy-preserving techniques specifically designed for deep learning models. The emphasis of this study is on performance, runtime and security aspect of the techniques [40]. Other literature studies by Greenspan et al. (2016), Shen et al. (2017), Litjens et al. (2017) and Suzuki et al. (2017) present an overview of deep learning in the field of medical image analysis. However, since these studies were conducted during the period in which deep learning was gaining its

momentum, the focus is towards the novelty of deep learning rather than on its application. The studies are not systematic reviews thus do not perform any analysis of the models which were presented. Finally, these studies do not analyse the privacy aspect of deep learning. [11, 20, 29, 32].

3.2 Data privacy techniques

Prior to the advent of deep learning and to-date, anonymisation or de-identification of data remains the most popular method to preserve the privacy of patients when sharing medical data. While no standardised methods for de-identification exist and different policies propose different requirements, three primary approaches are identified: 1. de-identification or the removal of patient identifiers 2. pseudonymisation or replacement of patient identifiers with unique pseudonyms and 3. anonymisation which entails de-identification followed by removal of further information in order to minimize the probability of re-identification. Anonymisation is still popular due to its simplicity and the fact that it is built into existing medical image analysis tools. Anonymisation however does not guarantee complete privacy as existing literature has demonstrated that it is possible to reconstruct the original data by combining the anonymised data with other public datasets, commonly referred to as linkage attacks [19, 33].

Rather than altering the data as done in de-identification and pseudonymization, Differential Privacy (DP) preserves the privacy of patients by injecting noise. This allows for statistical analysis of the dataset without compromising the sensitive details of an individual patient. Differentially private data is resistant to linkage attacks however perturbing the data leads to its degradation which may lead to poor model performance and quality of research and analysis [3, 13, 30, 36].

Encryption is the gold standard for secure communication and data transfer, originating from the field of cryptography. The state-of-the-art encryption schemes cannot be cracked using a brute force attack. They can be applied to the models or the data alike, making them an ideal choice when dealing with sensitive data. Homomorphic Encryption (HE) is an encryption scheme which allows certain computations (such as addition, subtraction and multiplication) to be conducted directly on encrypted data. HE enables models to be trained directly on unperturbed, encrypted data, however it does so with additional computational overhead [13, 16, 36, 39].

3.3 Security threats and attacks

In classical software development, computers strictly follow a specific set of programmed instructions. In contrast, deep learning algorithms develop their own rules based on a substantial amount of data provided to them. This behaviour often leads to neural networks being interpreted as a black box, preventing users from understanding its inner workings. This black box condition makes neural networks a potential target for exploitation, thus identification of its security threats and vulnerabilities must be prioritised. Existing research has presently identified several key threats, which are summarised below.

Adversarial examples are inputs that are often indistinguishable from typical inputs, yet contain intentional feature changes that leads to incorrect classification [34]. Several studies emphasise adversarial attacks to be an intrinsic flaw of deep learning [4, 22, 24, 34]. Adversarial attacks are significant because they question the robustness of the deep learning models and the essence of what it truly learnt. They are especially crucial in terms of computer vision applications such as self driving cars, facial recognition systems and medical imaging [22]. Unlike adversarial attacks that target problems within the neural networks, data poisoning exploits the heavy reliance of the models on data. The aim is to modify the training data such that the model can learn malicious intent and manifest that as its predictions. Open and public datasets that are used for training are especially open to data poisoning [8, 37].

In addition to attacks which target the data, the DL model itself can be exploited. By observing the gradients and parameters of a trained network, parts of the dataset can be obtained. Model inversion, membership inference and reconstruction attacks often utilise this technique to obtain the training data or infer if a public dataset was used to train the model. Combined with linkage and tracking attacks, the presence of an individual in the dataset and their sensitive information can be obtained [13, 27, 30, 31].

4 RESULTS

This section presents the discoveries made while conducting this systematic review. The privacy-preserving deep learning techniques can be broadly classified into 3 categories namely: *centralised learning*, *distributed learning* and *synthetic data generation*. The review additionally found several solutions that specialise in medical image analysis tasks using neural networks. These solutions and techniques are summarised in this section along with their critical assessment and comparisons made where possible.

4.1 Asynchronous Distributed Learning

Shokri et al. (2015) present the very first privacy preserving system for collaborative deep learning on sensitive data, commonly known now as Federated Learning (FL). The system enables multiple participants in possession of sensitive data, to train their models locally. The model parameters are then selectively shared with a centralised server where they are aggregated to form a global set of parameters. The local models are able to download the most frequently updated global parameters resulting in robust local models. Since the data no longer requires to be pooled at a central server, the system enables institutes to partake in collaborative research with full rights and control over their data. Compared to centralised training, the data leakage is drastically reduced since now there is only indirect leakage in the form of model parameters. The system is tested by performing image classification and regression tasks on the MNIST and SVHN datasets. Although the system is able to attain accuracy similar to centralised training, it should have also been tested with sensitive medical data. The authors propose a technique to further reduce the data leakage using Differential Privacy (DP). A comparative analysis is done to identify the trade-off between privacy and accuracy. The accuracy of the models fall with increased privacy. The analysis shows that the secure models are able to achieve their past accuracy when the number of participants and the quantity of shared parameters is increased. Despite the thorough privacy analysis conducted, the authors fail to conduct fault tolerance and

Table 1: Summary of privacy preserving techniques

Key	Dataset(s)	Model(s)	Performance	Communication	Scalability	Reliability	Runtime	Privacy	Security
Asynchronous distributed learning									
shokri 2015 privacy	MNIST and SVHN	MLP and CNN	Accuracy of models similar to centalized training	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Indirect data leakage in the form of model gradients	Model-inversion and adversarial attacks
amir2017 scalable	Parkinson's telemonitoring skin segmentation	ML models for classification and regression	Better performance for classification tasks using proposed aggregation scheme	Not evaluated	Scales upto 100 nodes	Stress tested for fault tolerance and liveliness	Not evaluated	Not evaluated	Adversarial attacks
aono2017 privacy	MNIST and SVHN	MLP	Accuracy of model similar to unencrypted gradients	Communication overhead doubles with encrypted gradients	Not evaluated	Not evaluated	Not evaluated	No gradients leaked to parameter server	Adversarial attacks
vizitiu 2019 towards	MNIST and X-ray of coronary angiography	CNN	Accuracy of model marginally lower to plaintext training	Not evaluated	Not evaluated	Not evaluated	Cyphertext training takes 30 times longer	No data leakage	Adversarial attacks
jeon2019 privacy	CIFAR-10 and CIFAR-100	VGG and ResNet	Models achieve higher accuracy compared to LSGD	Lower communication costs compared to LSGD	Not evaluated	Not evaluated	Not evaluated	Gradient of a single layer leaked to parameter server	Secure [1]
phuong 2019 privacy	MNIST, CIFAR-10, CIFAR-100 and UCI	MLP, CNN and ResNet	Models achieve accuracy and F1 score similar to centralized training	Not evaluated	Not evaluated	Not evaluated	Not evaluated	No data leakage	Adversarial attacks

reliability tests for the system. Additionally, all analysis were conducted using the accuracy of the models. Without a description of the data distribution, it is difficult to validate the legitimacy of the reported scores [30].

Rather than sharing the gradients of the local models with the parameter server, as proposed by Shokri et al. (2015), Phuong et al. (2019) put forth the idea of sharing the weights of the local models. The authors propose two systems: 1. Server-aided Network Topology (SNT) and 2. Fully-connected Network Topology (FNT). The SNT system mimics the traditional FL system proposed by Shokri et al. (2015) with additional security against an "honest-but-curious" parameter server by encrypting the local weights before transmission. Assuming there are L number of institutes, the SNT system comprises of L training or participant nodes and a parameter server. Each participant connects with the parameter server using a separate TLS connection, thus a total of L connections are required. The FNT system does not contain a parameter server and instead all participants are connected to each other using $L(L-1)/2$ connections. This is similar to the Cyclic Weight Transfer (CWT) system proposed by Chang et al. (2018) where a single model is trained by each participant multiple times. Thorough performance and runtime analysis of the proposed systems are conducted. A MLP is used for text (various UCI datasets) and CNN and ResNet are used for image (MNIST, CIFAR-10 and CIFAR-100) datasets to perform classification tasks. The analysis shows that the models trained using the proposed systems are able to achieve an accuracy and F1 score similar to that of centralised training. The authors recommend utilising the SNT system when the number of participants is $L \geq 20$ and the FNT system otherwise. It is however unclear how this threshold was identified since the experiments were conducted

using only 5 participants. It is also unclear which system (SNT or FNT) has been used for the experiments and how they compare against each other. Although runtime analysis has been done, the results of the proposed systems are not compared to traditional FL thus the feasibility and motivation for sharing weights instead of the gradients is unclear. Although the distribution of the data is described, the data is randomly and equally split amongst the participants which is not representative of what is observed in reality. The experiments were also conducted using non-medical datasets [26].

Federated Learning (FL) addresses the privacy concerns raised regarding centralised databases when dealing with medical data. Since the data remains on-site, FL is able to alleviate all concerns regarding data governance and ownership. However, as pointed out by the Shokri et al. (2015) and Phuong et al. (2019), FL is not inherently privacy preserving. The extent to which parameters are shared with the central server has a direct effect on the privacy resulting in a trade-off between privacy and accuracy of the models. Additionally, FL systems are prone to model-inversion attacks where even a small quantity of leaked gradients can be used to obtain the training data [2, 30]. Aono et al. (2017) propose an improvement to the original FL system by using Homomorphic Encryption (HE) to encrypt the model gradients. The HE encryption scheme not only enables learning over encrypted data, but also enables all gradients to be shared with the central server, thus boosting the accuracy. The improved system is tested by training a MLP to perform classification on the MNIST and SVHN datasets. The results are promising as the new system does not leak any gradients to the parameter server. Although the new system is able to match the accuracy of models trained without encrypted gradients, a communication overhead between participants and parameter server is observed. Thorough computational analysis conducted by the authors show that the communication overhead posed by the additional encryption-decryption phase can be reduced by multithreading the procedure. However, only accuracy of the models was using to evaluate the model and distribution of the data was not reported. Authors propose two encryption schemes: Paillier-based and LWE-based but only conducted analysis with LWE-based encryption. Further analysis, preferably using medical data, needs to be conducted to also validate the Paillier-based encryption scheme. The overall outcomes of the improved solution are positive since it shifts the original privacy-accuracy trade-off to a privacy-efficiency trade-off. The authors propose to address this trade-off by utilising more compute power and dedicated software and hardware in the future [2].

While Aono et al. (2017) opted to encrypt only the gradients of the models, Vizitiu et al. (2019) encrypted the entire data using Homomorphic Encryption (HE). The authors employ the MORE encryption scheme which allows limited - albeit sufficient - operations to be performed on encrypted data directly. The model is thus able to learn from encrypted data directly, without the requirement of an encryption key. A CNN is trained on unencrypted (plaintext) and encrypted (cyphertext) data using the MNIST digits dataset and another containing X-ray coronary angiography. Analysis of the models indicate that training using cyphertext provides a fully secure and private model whose accuracy is only marginally lower compared to plaintext training. There is however a significant runtime overhead as the model trained using cyphertext takes 30 times longer to train. The authors provide a complete security analysis of MORE and Fully Homomorphic Encryption (FHE) schemes. Although FHE is identified as the more secure encryption scheme, it also is computationally more demanding. The MORE encryption scheme is more simplistic since with enough pairs of encrypted-unencrypted data, an attacker may be able to compute the secret key. The MORE encryption scheme is however able to provide "good enough" security at cheaper computations making it more suitable for DL. HE addresses limitations of both DP and FL. Since the data is encrypted, model inversion attacks are no longer possible and no noise needs to be injected thus preserving the data quality. The authors make an assumption that the data can be centralised on account of it being encrypted. This assumption may not always hold due to restrictions posed by data privacy laws. Institutes may still raise concerns regarding data ownership or may not wish to share the data with an external server. Further analysis is thus required to compare the runtime and performance when the data is distributed. This analysis will also help understand the pros and cons of encrypting the data versus the gradients of the model [36].

Secure multi-party computing (SMPC) is a technique used to perform secure computations over encrypted, unperturbed data. The data is split across multiple participating servers such that no single participant can access all the data. Participants perform computations on their share of the data and the results are aggregated to obtain the final result. Private multi-party ML (PMPML) builds upon SMPC for running machine learning algorithms on large scale, distributed data [25]. As with FL, PMPML brings data governance back to its owners. However preceding PMPML systems adopt a peer-to-peer communication approach where all peers know each other. This is a fatal security flaw as a malicious peer can infer data present at all other peers by forwarding a fabricated model. InsuLearn, a distributed learning platform for classification and regression of medical imaging data is presented by Amir et al. (2017). The models are trained at each institute independently and later combined at secure coordinator nodes. The system was deployed in a distributed fashion and stress tested for fault tolerance, liveliness and scalability. The system scales well up to 100 nodes and is able to continue aggregating models to a reasonable extent as the rate of node failure is increased. InsuLearn addresses the limitations of existing PMPML systems with a novel aggregation technique which is tolerant to malicious peers. The aggregation technique is tested using 5 ML algorithms and the results are compared to a naive aggregation scheme and models trained on a centralised dataset. InsuLearn performs better than naive aggregation for classification tasks and slightly worse than insecure warehouse technique for regression tasks. The system was tested by performing skin segmentation and regression on Parkinson's telemonitoring data. The testing was done however using ML models such as Support Vector Machines and Random Forest, thus it remains to be seen if the aggregation technique can be extended to neural networks. The solution also hinges upon the assumption that the centralised aggregation node is secure. The authors however fail to propose protocols to achieve this security [1].

A communication overhead is posed by Federated Learning (FL) where all participants communicate with the parameter server

twice: once to upload their local gradients and later to download the aggregated gradients from the parameter server. This communication scheme does not scale well with increase in the number of participants and depth of models. Computation costs also increase at a larger scale since all participating models need to be trained simultaneously and managing a large number of participants becomes complicated. Jeon et al. (2019) propose an alternative solution with the goal of reducing the communication bandwidth and maintaining computational efficiency. Instead of training full local networks, only the first layer (L^1) is trained locally. The rest of the network (layers L^2 to L^k) are kept on the central server which is trained using the weights of each participant. Once the training of the central model is completed, the output of L^k is transmitted to each participant where the gradients are calculated. The gradients are transmitted back to the central server where backpropagation is conducted until L^2. The gradient of L^2 is transmitted to all participants to finally calculate the loss. The communication overhead is analysed for image classification using CIFAR-10 and CIFAR-100. Two models, ResNet and VGG are used. The models are trained using the proposed method and Large Scale Stochastic Gradient Descent (LSGD) [7]. The communication costs of the new system are noted to be significantly lower while achieving significantly higher accuracy compared to LSGD. Compared to FL, participants communicate with the central server twice as much in the new system. However, since the participants only communicate the parameters of a single layer as opposed to an entire network, the quantity of data communicated is expected to reduce. Empirical evidence remains to be collected such that the communication overhead for FL and the new system are analysed. The authors also fail to compare the performance of their system to centralised training thus it is difficult to understand how the accuracy of models is affected by the new distributed protocol. As with FL, the new system is not inherently privacy preserving since the central server and participants can still be breached. The distributed nature of the network may act as protection against model-inversion attacks however no guarantees can be made if the attacker obtains access to both the participant and the central server. The computation burden in the proposed system is shifted to the central server. This is a clear advantage since the performance of the central server can be improved by upgrading its hardware, thus removing the bottleneck posed by the heterogeneous hardware of participating institutes observed in prior FL systems [12].

4.2 Synchronous Distributed Learning

Distributed training overcomes the limitations of centralised training at the cost of added complexity. The complexity stems from the parallel training of several models and the heterogeneity in the network and hardware capabilities of participating institutes. Chang et al. (2018) hypothesise that synchronous training techniques can perform just as well as centralised training. In order to validate this hypothesis, the authors perform analysis of 3 synchronous learning techniques: 1. Model ensembling (ME) 2. Single weight transfer (SWT) and 3. Cyclic weight transfer (CWT). ResNet, a CNN is trained using the learning techniques mentioned above and their performance is compared to centralised training. 4 institutions are simulated and the models are trained to perform image classification on 3 distinct datasets: retinal fundus images, mammography and ImageNet. In ME, 4 models are trained separately at each site and their output is averaged. In SWT, a single model is trained at the first site until an acceptable learning accuracy is achieved and then transferred to the other institutes once for fine-tuning. Finally in CWT, the models are trained at all institutes several times for specific number of epochs until an acceptable learning accuracy is achieved. All proposed techniques perform better than a model trained at a single institute which shows the benefits of collaborative training. Thorough analysis conducted by the authors dictate that SWT performs better than ME. CWT performs the best as its performance is comparable to that of centralised training. Although the model is trained at each institute several times under the CWT technique, empirical evidence suggests that no overfitting occurred. The authors also analysed the effect of transfer frequency on model accuracy. The outcomes indicate that the model performs better with a high frequency of transfer with an added cost of longer training time. The robustness of CWT was analysed next by testing with and without an institute containing heterogeneous data. Heterogeneity in the data was introduced by using images with non-standardised quality, resolution and size. Additionally, the quantity of data per patient was also not equal in order to simulate realistic data. The scalability of CWT was analysed by training the model using 6000 patient samples randomly distributed across 20 institutes. The authors ensured that the model would not be able to perform better than random classification when trained at a single institute. The results indicate that CWT is highly robust and scalable as it is able to achieve accuracy similar to centralised training when all 20 institutes were used for training. Furthermore, it is able to maintain this performance as the variability in the data is increased. Although the authors try to emulate variability of realistic data, the emulation is lacking since the data was sampled from the same dataset. This does not accurately represent the domain difference observed in reality and further testing using data derived from unique patient populations is proposed. The analysis was conducted using CNNs to perform binary classification. Further experimentation is required to see the effects of CWT for multi-label classification, performed using other network architectures such as autoencoders, GANs and RNNs. Finally, the authors fail to conduct any runtime performance of the synchronous learning techniques. Specifically, it would have been nice to see analysis on runtime performance of CWT and FL as the system is scaled. This analysis would have aided in determining the threshold after which FL becomes feasible since the network and hardware capabilities of the participating institutes are non-uniform and thus unpredictable [6].

Sheller et al. (2018) address the limitations of Chang et al. (2018) by conducting an analysis of Single Weight Transfer (SWT), Cyclic Weight Transfer (CWT) and Federated Learning (FL) for image classification using medical data. Compared to FL, the bandwidth requirements in SWT are lesser since each participant transmits the model once and receives it twice, once for training and the other to receive the final model. A common problem in SWT is the drop in performance with the increase in number of participants due to catastrophic forgetting [17]. CWT addresses this issue to an extent by varying the number of epochs of training at each site.

Table 2: Summary of privacy preserving techniques (cont.)

Key	Dataset(s)	Model(s)	Performance	Communication	Scalability	Reliability	Runtime	Privacy	Security
Synchronous distributed learning									
chang 2018 distributed	Proprietary Retinal fundus and mammography and ImageNet	ResNet	Accuracy of model trained using CWT similar to centralised training	Not evaluated	Scalable upto 20 participants	Not evaluated	Not evaluated	Indirect data leakage in the form of model parameters	Model-inversion and adversarial attacks
sheller 2018 multi	BraTS	U-Net	Accuracy of model trained using FL performs best	SWT consumes lower bandwidth than FL and CWT	CWT not scalable to many participants	Not evaluated	Not evaluated	Indirect data leakage in the form of model parameters	Model-inversion and adversarial attacks
beaulieu 2018 privacy	eICU and TCGA	CNN	AUROC of model trained using CWT and DP improves as number of participants increase	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Indirect data leakage in the form of model parameters	Adversarial attacks
Centralised Learning									
wu2019 p3sgd	Proprietary pathology images	ResNet, AlexNet, VGG and MobileNet	Slight decrease in accuracy when regularized with P3SGD	NA	NA	NA	Not evaluated	NA	Adversarial attacks
Synthetic data generation									
torkzadehmahani 2019dp	MNIST	Logistic Regression and MLP	AUCROC of models trained using synthetic data similar to real data	NA	NA	NA	NA	NA	Model-inversion and adversarial attacks

U-Net, a CNN specialising in image segmentation is trained using the BraTS brain tumour segmentation dataset, using FL, SWT and CWT. The training is performed using both a skewed distribution of data amongst institutes (representing realistic data) and a simulated distribution where each institute is roughly given images for the same number of patients. FL outperforms SWT and CWT in both data distributions with CWT performing second best due to some amount of catastrophic forgetting. Empirical evidence provided by the authors suggest that SWT and CWT do not scale well to a large number of institutions with small amounts of data. Authors additionally show that CWT is infeasible as running validation tests add communication costs above FL [28].

Beaulieu et al. (2018) apply Differential Privacy (DP) to Cyclic Weight Transfer (CWT) in an attempt to address the privacy concerns of CWT. The authors test their solution on the eICU Collaborative Research Database containing de-identified demographic data of intensive care patients and the Cancer Genome Atlas dataset (TCGA) which contains data from breast invasive carcinoma patients. The model is trained to perform image classification using 4 techniques: 1. centralised 2. centralised with DP 3. CWT and 4. CWT with DP. The performance of the model trained using technique 4 is slightly lower than technique 3. The performance of the model trained using technique 3 is better than technique 2 but not equal to technique 1. The performance of model trained using CWT and DP is observed to improve as the number of participants is increased. Although the proposed solution addresses the privacy concerns of CWT, the testing distribution is felt to be unrealistic since each institute is given equal number of examples. The authors wish to address this limitation in the future by using secondary models. The secondary or local models are trained only at their respective institutes in an attempt to capture local trends and biases. Although the authors claim that the communication costs are significantly reduced compared to FL, no runtime analysis is done to support such claims. The performance analysis is also done only using the AUROC metric without any comments on the

distribution of the data. Further analysis is required to compare and contrast CWT with FL when both are put under DP. As with Chang et al. (2018), the analysis was only conducted with very few number of participants. This makes it unclear how the proposed solution scales or how many participants are required before the performance of the model plateaus. Finally, it is unclear how the results of the secondary models will be integrated with the global model and how the final model will be tested. The authors also fail to conduct any privacy tests in order to determine the quantity of data leakage and how it compares to FL [3].

4.3 Centralised Learning

CNNs are susceptible to overfit or memorise certain aspects of the training data [38]. This is a cause for concern when working with pathological data which contain sensitive information regarding patients. While FL is a promising solution at a macro scale, Wu et al. (2019) propose a solution at a micro scale. P3SGD is a novel regularisation technique for models being trained with pathological data. Most pathological databases contain several images obtained from the same patient. Contrary to previous work which focus on ensuring privacy at the image level, P3SGD utilises Differential Privacy (DP) to ensure privacy at the patient level which has the added benefit of reducing overfitting in CNNs. Top image classification CNNs such as ResNet, AlexNet, VGG and MobineNet are trained to detect the presence of glomerulus using real pathology images (image classification task). The models are trained using SGD and regularised using P3SGD. The performance of the models are then compared to results obtained when the models are regularised using state-of-the-art techniques such as Dropout and Batch Normalisation. A thorough analysis of the results indicate that there is a slight decrease in the performance of the models regularised using P3SGD. A model-inversion attack is additionally performed and models regularised using P3SGD are found to be resistant to such attacks. The performance analysis were done using only the accuracy of the models. Unfortunately, the authors fail to provide any information regarding the distribution of the data without which it is difficult to ascertain the legitimacy of the metrics reported. P3SGD was developed exclusively for pathological image classification and extending it to other tasks is proposed as future work. It would also be informative to analyse the performance of P3SGD when models are trained using modern optimisation algorithms such as AdaGrad and Adam [38].

4.4 Synthetic Data Generation

Preceding literature in the field of synthetic data generation focuses solely on the generation of images. This, although useful in circumstances where data is scarce, cannot be used by supervised models which require the corresponding labels as well. Torkzadehmahani et al. (2019) present the first Differentially Private Conditional GAN (DP-CGAN), a platform for secure training of GANs on sensitive data. Using Differential Privacy (DP) and the state-of-the-art Renyi Differential Privacy Accountant (RDP), the system is able to protect GANs from malicious model-inversion and membership inference attacks. The MNIST dataset was used to train the model. The synthetic data and label produced was used to train a Logistic Regression (LR) and a Multi Layer Perceptron (MLP) to classify the digits. Promising results were obtained as the classifiers trained on data generated by DP-CGAN obtain similar scores to when the models are trained using insecure data. The authors report only the AUROC metric which is not enough information to gauge the capabilities of the system. Further information such as the distribution of the data and metrics such as the accuracy, precision, recall and F1 score should have also been reported. The system addresses the lack of sensitive data by synthetically generating it. However, sensitive data is still required to facilitate the synthetic generation. If the sensitive data is already available, then the value of the solution proposed is unclear. The system was tested using the MNIST dataset which is publicly available. Although the authors propose to extend the system to a more challenging dataset such as CIFAR100, experiments with truly sensitive data such as medical imaging would have been valuable to look at [35].

4.5 Specialised Solutions

ATMOSPHERE is a container-based, federated, Infrastructure as a Service (IaaS) which can be used to develop medical-imaging specific applications. Blanquer et al. (2019) present a solution for early detection of the Rheumatic Heart Disease (RHD) built atop ATMOSPHERE. RHD when detected in its early stages is curable but leads to severe health disorders including death if left untreated. Moreover, the research in the early detection of RHD is limited since there is no standardised test to indicate positive early stage presence. CNNs were trained on videos from 4615 echo-cardio studies to classify data into three categories: 1. Definite RHD 2. Borderline RHD and 3. No RHD (normal). The federated infrastructure provided by ATMOSPHERE enables medical facilities across the globe to collaborate without compromising the privacy of the patients. The solution is container-based, thus can dynamically scale based on the demand and utilises the parallel compute power of GPUs for training. The application development process, network topology, data access layer and deployment procedure are well documented. The paper however fails to report the performance and evaluation of the model itself. Details on the dataset (such as its distribution), testing methodology and evaluation metrics used would have presented a complete picture of the capabilities of the solution presented. Limitations of the work done and possible areas of improvement would have enabled future work to be conducted. Finally, the system is not publicly available so it is difficult to gauge if the project is being actively developed or if an active user base is present [5].

Kim et al. (2019) present the first client-server system for semantic medical image segmentation with identify preserving, distributed learning. Inspired by GANs, the system utilises three networks: 1. an image encoder 2. a discriminator and 3. a medical image analysis network (such as a CNN for segmentation in this case). The image encoder is deployed at the client facilities and is used to convert the patient data into an obfuscated signal. The signal which contains enough semantic information, is sent to a centralised server for further analysis. The discriminator and the image analysis network are deployed at the server. The discriminator network is used to identify two signals originating from the same patient. Finally, the image analysis network is used to perform the medical imaging task. The results are sent back to the client

Table 3: Summary of privacy preserving techniques (cont.)

Key	Dataset(s)	Model(s)	Performance	Communication	Scalability	Reliability	Runtime	Privacy	Security
Specialised Solutions									
blanquer 2019 medical	Proprietary eco-cardio data	Not mentioned	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Indirect data leakage in the form of model gradients	Model-inversion and adversarial attacks
kim2019 privacy	Proprietary MRI images	Autoencoder and CNN	System is able to obtain good accuracy without compromising privacy	Not evaluated	Not evaluated	Not evaluated	Not evaluated	No data leakage	Model-inversion and adversarial attacks
yue2019 privacy	Proprietary colposcopy and breast tissue biopsy images	CNN with LSTM cell	System achieves higher accuracy and lower false-negatives compared to competing models	Not evaluated	Not evaluated	Not evaluated	Not evaluated	No data leakage	Adversarial attacks
gibson 2018 niftynet	NA	NA	NA	Not evaluated	Not evaluated	Not evaluated	Not evaluated	NA	NA

where the image encoder is used to decode the results. The paper strategically outlines the existing solutions and their limitations which makes the importance and value of their solution clear. The system is evaluated using two independent test sets and compared with the performance when non-encoded images are used. The results indicate that the system is able to maintain its performance without compromising the patient's privacy. Both datasets however contained MRI images of adult brains which does not highlight the generalisation capabilities of the network. Further experiments need to be conducted using datasets with a larger domain shift (for instance MRI images of infant to adult brains) to identify the gaps in the model's ability to generalise. The authors propose to also expose the system to different image modalities such as MRI and CT scans in an attempt to improve the model's generalisation abilities. Although the proposed system facilitates multi-centre learning, the performance of the network on a distributed dataset and the runtime performance of the entire system were not conducted [16].

The dynamic changes in lesions is an important field of study in medical image analysis. The data for such a task is a succession of images taken over time which capture the changes that occur in lesions, commonly known as time-series medical images. Yue et al. (2019) present HE-CLSTM, a system capable of performing analysis on encrypted time-series medical images. The system addresses two short comings of the existing body of work which are: 1. existing algorithms only support analysis over singular encrypted images rather than the entire image sequence and 2. the emphasis is put on improving the accuracy of the algorithms rather than reducing the false negatives. This is a critical limitation since a missed diagnosis may result in serious health repercussion for the patient. The system utilises Homomorphic Encryption (HE) to first encrypt the data before transferring it to a server where the analysis is conducted by a CNN equipped with LSTM cells. The system was evaluated using two independent datasets containing images from colposcopy and breast tissue biopsy respectively. The system is able to achieve a substantial improvement in 5 evaluation metrics compared to competing models. It is able to attain a high accuracy and a low false negative score by utilising a weighted unit and a sequence voting layer. Although the paper presents an empirical analysis of the model's performance, it fails to conduct any runtime evaluation of the system. The system was evaluated with a single client but it would have been helpful to see the performance of the system when multiple clients share encrypted data with the server [39].

DL models used in medical imaging are typically specialised towards a specific task or an organ. However there is a substantial overlap in the implementation and software pipeline of such models. Medical data is generally enriched with additional metadata concerning the patient which require to be removed or anonymised prior to its use. This procedure is often laborious and time consuming. After the data has been anonymised, data extraction and augmentation requires domain knowledge as they are derived from the task and organ of concern. To reduce duplication of work and facilitate an easy way to share scientific models, Gibson et al. (2018) present NiftyNet. NiftyNet is a deep learning platform which facilitates common medical tasks using deep learning. The platform provides existing solutions to common infrastructural needs for medical imaging such as data loading, data augmentation, loss functions and evaluation metrics. The platform also provides a database

containing pre-built and pre-trained deep learning models commonly used in medical imaging. NiftyNet hides away the technical complexity of building DL models thus allowing researchers to rapidly generate prototype models for tasks such as image segmentation and regression. NiftyNet is still in its infancy as the platform supports only a handful of models to serve as proof-of-concept. The platform does not contain state-of-the-art models and lacks support for image classification, registration and pathology detection. NiftyNet relies on users to provide the data and thus fails to cater to scenarios where the data is distributed across several users. The authors also fail to comment on the privacy and security aspect of the system as the emphasis is put on technical aspect [10].

5 DISCUSSION

This study identified 15 papers that presented techniques to train Deep Learning (DL) models using sensitive data, without violating the privacy concerns of the patients. The techniques can be broadly classified into 3 categories: 1. Synthetic data generation 2. Centralised training and 3. Distributed Training. While majority of the papers presented a distributed training solution, the study was able to identify two papers which presented a solution using synthetic data and centralised training respectively. An overview of privacy-preserving deep learning is presented in Figure 1.

The synthetic data generation technique is felt to be a viable solution to address the lack of data at remote medical institutes or for rare diseases. Since the synthetic data has a similar distribution to the original data, model-inversion and membership inference attacks can still be carried out on the DL network to reveal the original data. Additional data privacy techniques such as Differential Privacy (DP) or Homomorphic Encryption (HE) can be used to secure the models. However training conducted using synthetic data does not fix the bias in the models, only amplifies it. This technique is thus felt to be appropriate for testing proof-of-concept models using local data only.

In the occasion that the model does not require to be trained using data across multiple institutes or in regions where the privacy laws are relaxed, a centralised training approach can be adopted. However, this is a rare possibility as bulk of the cases require a distributed approach. Two approaches of training, namely, synchronous and asynchronous are identified amongst the distributed training techniques. Cyclic Weight Training (CWT) is identified as the most suitable synchronous technique in which a model is trained at each institute several times, in a pre-determined or random order. In contrast, asynchronous training techniques such as Federated Learning (FL) and Secure Multi-Party Machine Learning (SMPML) train multiple models in parallel and utilise a parameter server to aggregate the results. Occam's Razor dictates that CWT, being the simpler technique, be chosen where possible. However, due to the large communication overhead it poses, it becomes infeasible when training across many institutes. The requirement for further analysis is felt to determine the exact threshold beyond which the added complexity of distributed training is warranted. Additional research is also required to determine how well FL systems scale since addition of every new participant or increase in the depth of the neural networks, add to the communication and computation costs.

None of the techniques mentioned so far are inherently privacy-preserving and require the aid of data privacy techniques. Differential Privacy (DP) and Homomorphic Encryption (HE) are identified as the two most common privacy techniques used across all training protocols. DP introduces a trade-off between the performance of the model and the privacy of the patients by probabilistically injecting noise into the dataset. While the presence of more noise ensures higher privacy, it also negatively affects the performance of the model. Currently, the DP limit for a given dataset is set by the medical institutes who own it, ideally the limit should be set by the patients themselves. Consider a scenario where a patient P first visits an institute A for MRI scans and later another institute B for a second opinion such that both institutes are in possession of P's data. Individually, the institutes may set a DP limit of 0.1 for their patients before sharing their data. However a total of 0.2% of P's data is then shared which might not be acceptable to P. In contrast, HE originating from the field of cryptography, enables DL models to operate directly over unperturbed, encrypted data. HE is able to lift the trade-off noted in DP for a higher runtime cost. HE is felt to be the most secure data privacy technique and its use is recommended when privacy is of utmost importance and runtime is not a concern.

Compared to centralised training, the data leakage in distributed training is drastically reduced through indirect leakage of model parameters only. However, distributed training techniques are vulnerable to model-inversion and inference attacks which can exploit a small quantity of leaked gradients in order to reconstruct the dataset. DP and HE can be employed in an attempt to safeguard the privacy of the patients however the DL models themselves remain vulnerable to adversarial and data poisoning attacks. These attacks may not breach the privacy of the patients but can raise concerns regarding the robustness of the models. Another alarming outcome is that the attacks may change the predictions of the model which may lead to the incorrect diagnosis thereby putting the patient's life at risk. While model parameters and data secured with HE are resistant to brute force attacks, they can however be de-crypted if the attacker is in possession of sufficient quantity of plaintext and its corresponding cyphertext.

The field of privacy-preserving deep learning is somewhat paradoxical in nature. While the goal is to develop systems and techniques to privately train DL models using medical data, such data is not publicly available. The existing research is thus restrained to experiments conducted using centralised data repositories which do not accurately depict the domain shift observed in truly distributed data. The research is also limited primarily to Multi Layer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs), trained using Stochastic Gradient Descent (SGD), to perform classification tasks. The need for further research is felt to train other types of neural networks such as Generative Adversarial Networks (GANs), Recurrent Neural Networks (RNNs) and Autoencoders using modern optimisation algorithms such as AdaGrad, Adam and RMSProp for tasks besides image classification. Taking a step back, if a distributed learning system using one of the techniques proposed above were to be constructed, the ownership of such a system is unclear at this point. For instance, who would be responsible for it construction and maintenance? How will the project be financed?

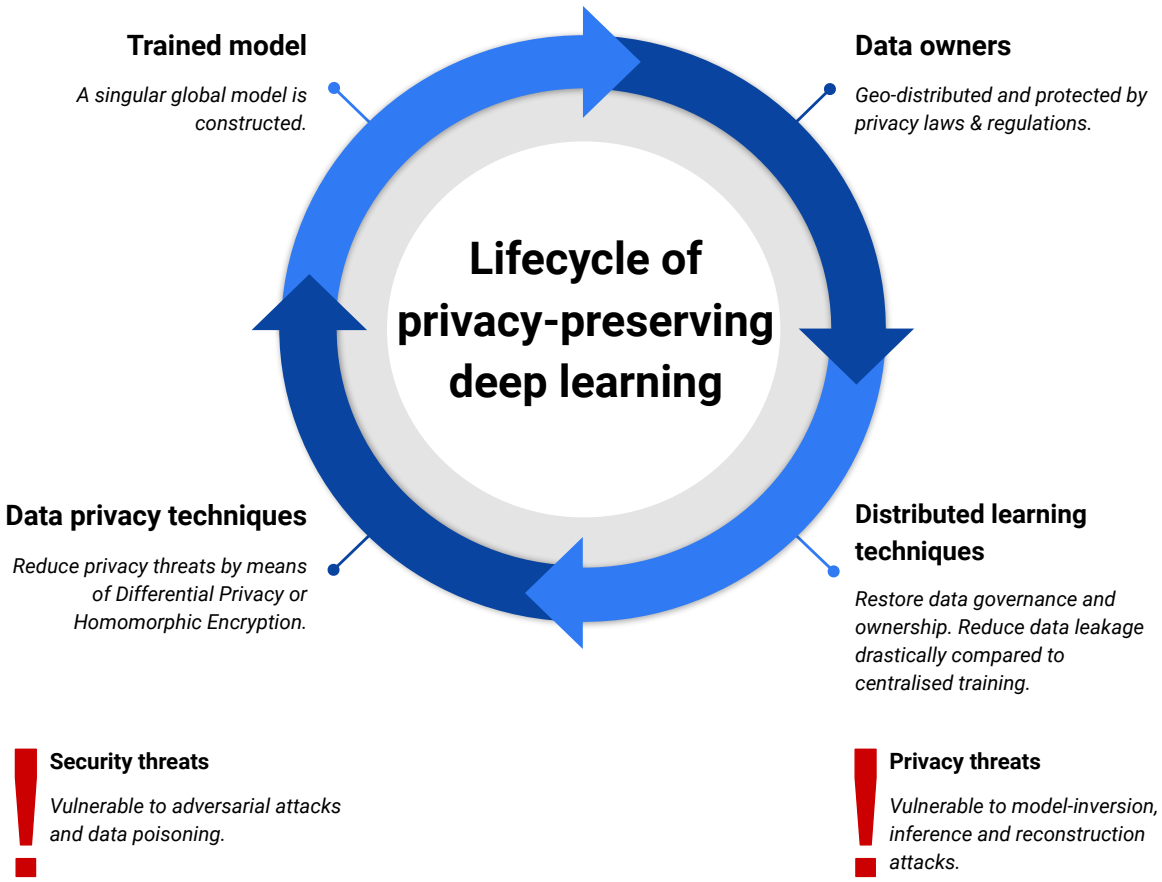

Figure 1: Summary of privacy-preserving deep learning

And where will the parameter server for the asynchronous techniques be located geographically? Construction of such a system not only requires a joint initiative from all participating institutes but also warrants the development of legislative policies to ensure a fair and conflict free operation of the system. Small medical institutes, in remote geographical locations will benefit the most from such a technological advancement. However, since they are unable to contribute due to lack of data and infrastructure, concerns are raised regarding the accessibility of the system to small medical institutes.

6 CONCLUSION

Medical Image Analysis is an indispensable aspect of medical science that is highly sensitive to human error. Artificial Intelligence such as Deep Neural Networks have had tremendous success in the field and have helped reduce the burden from their human counterparts. Neural networks are however "data hungry" and require vast amounts of sensitive, medical data to learn. Besides the growing privacy concerns raised by training DL models using private medical data, it is also difficult to obtain such an immense quantity from a single institute. This paper conducted a systematic review to identify and analyse the existing solutions that can be used to train DL models in a private manner. In order to accomplish such a task, research questions were first developed to help identify relevant papers to be included in the review. These papers were then held against certain inclusion-exclusion criterion to obtain a final list of papers. The selected papers were then reviewed in full and data extracted from the papers were analysed. In light of the results obtained through this review, the research questions are revisited below.

RQ1. *What are the existing systems, platforms or techniques that facilitate deep learning on medical images that do not invade the patient's privacy?*
This review identified several privacy-preserving systems and solutions which can be used to train DL models. The

identified techniques were broadly classified as per three guiding principles, namely: *centralised learning*, *distributed learning* and *synthetic data generation*. While interesting techniques were discovered under synthetic data generation and centralised learning, the most prominent techniques approached the problem in a distributed fashion. All solutions utilised additional data privacy techniques to ensure privacy and security of the patient's data. *Differential Privacy (DP)* and *Homomorphic Encryption (HE)* were the most prominent with the former being computationally simpler and the later being more secure.

RQ2. *What are the top performing deep learning models being used to perform tasks in medical imaging?*
Heterogeneity in the training techniques, the neural networks trained, the datasets used and the model performance metrics reported, posed a challenge in the identification of the top performing DL model. The nature of medical image analysis is such that it requires the DL models to be trained with highly specialised datasets to perform a single task for a specific organ. The need for additional research is felt to identify the top performing DL architectures. This effort is to be conducted such that the DL models, the tasks they perform and the evaluation metrics used are standardised. One place to start could be to analyse the results of *Kaggle* competitions for medical imaging tasks.

RQ3. *What are the specific attacks that can compromise the security of these systems?*
All solutions, without the aid of data privacy techniques are vulnerable to model-inversion and model inference attacks. The solutions are tolerant to such attacks by using data privacy techniques such as DP and HE. Additional measures are required to protect the models themselves from adversarial and data poisoning attacks.

RQ4. *What is the amount of data leakage in these systems?*
Distributed learning techniques drastically reduce the data leakage compared to centralised learning. Since the data remains on-site and only select parameters are shared with the parameter server, there is still indirect data leakage. These leaked model parameters - no matter how small - can be exploited by attacks to reconstruct fragments of the original dataset, thus compromising the privacy of the patients. The data leakage can be minimised by sharing less model parameters, using a lower DP limit or by encrypting the parameters and data.

In this era of "big data', where many are becoming conscious of how their data is being used, privacy-preserving deep learning is gaining momentum. Being the first of its kind, this systematic review hopes to be valuable to researchers working at the intersection of medical science and artificial intelligence. With the existing solutions discussed, attention is now drawn to two avenues of research which remain unexplored.

By and large, Medical Imaging data such as MRI and CT scans, are volumetric in nature. The state-of-the-art solutions require the three dimensional data be segmented into a two dimensional format, resulting in loss of valuable information. Geometric Deep Learning (GDL) is a new field of research, focused on developing deep learning models which are able to train using volumetric data. As such, the application of GDL to MIA remains open for further investigation.

The last decade has ushered in tremendous progress in the field of deep learning. The field was re-energised after the discovery of Convolusional Neural Networks that revolutionised computer vision and enabled neural networks to surpass humans in pattern recognition and image classification tasks. Since then, deep learning has been widely adopted in fields such as robotics and natural language processing, and has received recognition for solving difficult problems. The exact inner workings of neural networks are still however unclear, thereby raising serious ethical concerns regarding their application in medical science. Hybrid intelligence is a new field of research which seeks to enhance the human intellect using artificial intelligence, rather than replacing it. Additionally, Bayesian Deep Learning is another exciting field of research which is tasked with the creation of DL models that provide a measure of uncertainty to the predictions they make. Developments in these field may be beneficial to privacy-preserving deep learning since it will enable humans to view the outcomes of neural networks with a sense of trust.

REFERENCES

[1] Alborz Amir-Khalili, Soheil Kianzad, Rafeef Abugharbieh, and Ivan Beschastnikh. 2017. Scalable and fault tolerant platform for distributed learning on private medical data. In *International Workshop on Machine Learning in Medical Imaging*. Springer, 176–184.

[2] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. 2017. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security* 13, 5 (2017), 1333–1345.

[3] Brett K Beaulieu-Jones, William Yuan, Samuel G Finlayson, and Zhiwei Steven Wu. 2018. Privacy-preserving distributed deep learning for clinical data. *arXiv preprint arXiv:1812.01484* (2018).

[4] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 387–402.

[5] Ignacio Blanquer, Angel Alberich-Bayarri, Fabio García-Castro, George Teodoro, André Meirelles, Bruno Nascimento, Wagner Meira Jr, and Antonio Luiz P Ribeiro. 2019. Medical Imaging Processing Architecture on ATMOSPHERE Federated Platform.. In *CLOSER*. 589–594.

[6] Ken Chang, Niranjan Balachandar, Carson Lam, Darvin Yi, James Brown, Andrew Beers, Bruce Rosen, Daniel L Rubin, and Jayashree Kalpathy-Cramer. 2018. Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association* 25, 8 (2018), 945–954.

[7] Jianmin Chen, Xinghao Pan, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. 2016. Revisiting distributed synchronous SGD. *arXiv preprint arXiv:1604.00981* (2016).

[8] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526* (2017).

[9] Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. 2017. Machine learning for medical imaging. *Radiographics* 37, 2 (2017), 505–515.

[10] Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, et al. 2018. NiftyNet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine* 158 (2018), 113–122.

[11] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. 2016. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging* 35, 5 (2016), 1153–1159.

[12] Joohyung Jeon, Junhui Kim, Joongheon Kim, Kwangsoo Kim, Aziz Mohaisen, and Jong-Kook Kim. 2019. Privacy-Preserving Deep Learning Computation for Geo-Distributed Medical Big-Data Platforms. In *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks–Supplemental Volume (DSN-S)*. IEEE, 3–4.

[13] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* (2020), 1–7.

[14] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. 2017. Deep learning applications in medical image analysis. *Ieee Access* 6 (2017), 9375–9389.
[15] Abid Khan, Umar Manzoor, Kinza Sarwar, Mansoor Ahmed, Mouzna Tahir, Adeel Anjum, Masoom Alam, Nadeem Javaid, Mohammed A Balubaid, et al. 2017. Towards preserving privacy of outsourced genomic data over the cloud. *Journal of Medical Imaging and Health Informatics* 7, 6 (2017), 1475–1482.
[16] Bach Ngoc Kim, Jose Dolz, Pierre-Marc Jodoin, and Christian Desrosiers. 2019. Privacy-Net: An Adversarial Approach for Identity-Obfuscated Segmentation of Medical Images. *arXiv preprint arXiv:1909.04087* (2019).
[17] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.
[18] June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. 2017. Deep learning in medical imaging: general overview. *Korean journal of radiology* 18, 4 (2017), 570–584.
[19] Chung-Yueh Lien, Michael Onken, Marco Eichelberg, Tsair Kao, and Andreas Hein. 2011. Open source tools for standardized privacy protection of medical images. In *Medical Imaging 2011: Advanced PACS-based Imaging Informatics and Therapeutic Applications*, Vol. 7967. International Society for Optics and Photonics, 79670M.
[20] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.
[21] Alexander Selvikvåg Lundervold and Arvid Lundervold. 2019. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik* 29, 2 (2019), 102–127.
[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).
[23] Syed Atif Moqurrab, Adeel Anjum, Umar Manzoor, Samia Nefti, Naveed Ahmad, and Saif Ur Rehman Malik. 2017. Differential Average diversity: an efficient privacy mechanism for electronic health records. *Journal of Medical Imaging and Health Informatics* 7, 6 (2017), 1177–1187.
[24] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 427–436.
[25] Olga Ohrimenko, Felix Schuster, Cédric Fournet, Aastha Mehta, Sebastian Nowozin, Kapil Vaswani, and Manuel Costa. 2016. Oblivious multi-party machine learning on trusted processors. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 619–636.
[26] Tran Thi Phuong et al. 2019. Privacy-preserving deep learning via weight transmission. *IEEE Transactions on Information Forensics and Security* 14, 11 (2019), 3003–3015.
[27] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. 2020. Updates-leak: Data set inference and reconstruction attacks in online learning. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 1291–1308.
[28] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. 2018. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop*. Springer, 92–104.
[29] Dinggang Shen, Guorong Wu, and Heung-Il Suk. 2017. Deep learning in medical image analysis. *Annual review of biomedical engineering* 19 (2017), 221–248.
[30] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1310–1321.
[31] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.
[32] Kenji Suzuki. 2017. Overview of deep learning in medical imaging. *Radiological physics and technology* 10, 3 (2017), 257–273.
[33] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.
[34] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).
[35] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. 2019. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.
[36] Anamaria Vizitiu, Cosmin Ioan Niță, Andrei Puiu, Constantin Suciu, and Lucian Mihai Itu. 2019. Towards Privacy-Preserving Deep Learning based Medical Imaging Applications. In *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 1–6.
[37] Yizhen Wang and Kamalika Chaudhuri. 2018. Data poisoning attacks against online learning. *arXiv preprint arXiv:1808.08994* (2018).
[38] Bingzhe Wu, Shiwan Zhao, Guangyu Sun, Xiaolu Zhang, Zhong Su, Caihong Zeng, and Zhihong Liu. 2019. P3sgd: Patient privacy preserving sgd for regularizing deep cnns in pathological image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2099–2108.
[39] Zijie Yue, Shuai Ding, Lei Zhao, Youtao Zhang, Zehong Cao, M Tanveer, Alireza Jolfaei, and Xi Zheng. 2019. Privacy-preserving time series medical images analysis using a hybrid deep learning framework. *ACM Transactions on Internet Technology* 37, 4 (2019), 1–22.
[40] Fadila Zerka, Samir Barakat, Sean Walsh, Marta Bogowicz, Ralph TH Leijenaar, Arthur Jochems, Benjamin Miraglio, David Townend, and Philippe Lambin. 2020. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. *JCO Clinical Cancer Informatics* 4 (2020), 184–200.

A APPENDIX

Table 4: List of selected papers

Key	Author	Title	Year	Journal
Asynchronous distributed learning				
shokri2015privacy	Shokri, Reza and Shmatikov, Vitaly	Privacy-preserving deep learning	2015	Proceedings of the 22nd ACM SIGSAC conference on computer and communications security
amir2017scalable	Amir-Khalili, Alborz and Kianzad, Soheil and Abugharbieh, Rafeef and Beschastnikh, Ivan	Scalable and fault tolerant platform for distributed learning on private medical data	2017	International Workshop on Machine Learning in Medical Imaging
aono2017privacy	Aono, Yoshinori and Hayashi, Takuya and Wang, Lihua and Moriai, Shiho and others	Privacy-preserving deep learning via additively homomorphic encryption	2017	IEEE Transactions on Information Forensics and Security
vizitiu2019towards	Vizitiu, Anamaria and Nita, Cosmin Ioan and Puiu, Andrei and Suciu, Constantin and Itu, Lucian Mihai	Towards Privacy-Preserving Deep Learning based Medical Imaging Applications	2019	2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)
jeon2019privacy	Jeon, Joohyung and Kim, Junhui and Kim, Joongheon and Kim, Kwangsoo and Mohaisen, Aziz and Kim, Jong-Kook	Privacy-Preserving Deep Learning Computation for Geo-Distributed Medical Big-Data Platforms	2019	2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks–Supplemental Volume (DSN-S)
phuong2019privacy	Phuong, Tran Thi and others	Privacy-preserving deep learning via weight transmission	2019	IEEE Transactions on Information Forensics and Security
Synchronous distributed learning				
chang2018distributed	Chang, Ken and Balachandar, Niranjan and Lam, Carson and Yi, Darvin and Brown, James and Beers, Andrew and Rosen, Bruce and Rubin, Daniel L and Kalpathy-Cramer, Jayashree	Distributed deep learning networks among institutions for medical imaging	2018	Journal of the American Medical Informatics Association
beaulieu2018privacy	Beaulieu-Jones, Brett K and Yuan, William and Finlayson, Samuel G and Wu, Zhiwei Steven	Privacy-preserving distributed deep learning for clinical data	2018	arXiv preprint arXiv:1812.01484
sheller2018multi	Sheller, Micah J and Reina, G Anthony and Edwards, Brandon and Martin, Jason and Bakas, Spyridon	Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation	2018	International MICCAI Brainlesion Workshop
Centralised learning				
wu2019p3sgd	Wu, Bingzhe and Zhao, Shiwan and Sun, Guangyu and Zhang, Xiaolu and Su, Zhong and Zeng, Caihong and Liu, Zhihong	P3sgd: Patient privacy preserving sgd for regularizing deep cnns in pathological image classification	2019	Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
Synthetic data generation				
torkzadehmahani2019dp	Torkzadehmahani, Reihaneh and Kairouz, Peter and Paten, Benedict	Dp-cgan: Differentially private synthetic data and label generation	2019	Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops

Table 5: List of selected papers (cont.)

Key	Author	Title	Year	Journal
Specialised solutions				
blanquer2019medical	Blanquer, Ignacio and Alberich-Bayarri, Angel and García-Castro, Fabio and Teodoro, George and Meirelles, André and Nascimento, Bruno and Meira Jr, Wagner and Ribeiro, Antonio Luiz P	Medical Imaging Processing Architecture on ATMOSPHERE Federated Platform.	2019	CLOSER
kim2019privacy	Kim, Bach Ngoc and Dolz, Jose and Jodoin, Pierre-Marc and Desrosiers, Christian	Privacy-Net: An Adversarial Approach for Identity-Obfuscated Segmentation of Medical Images	2019	arXiv preprint arXiv:1909.04087
yue2019privacy	Yue, Zijie and Ding, Shuai and Zhao, Lei and Zhang, Youtao and Cao, Zehong and Tanveer, M and Jolfaei, Alireza and Zheng, Xi	Privacy-preserving time series medical images analysis using a hybrid deep learning framework	2019	ACM Transactions on Internet Technology
gibson2018niftynet	Gibson, Eli and Li, Wenqi and Sudre, Carole and Fidon, Lucas and Shakir, Dzhoshkun I and Wang, Guotai and Eaton-Rosen, Zach and Gray, Robert and Doel, Tom and Hu, Yipeng and others	NiftyNet: a deep-learning platform for medical imaging	2018	Computer methods and programs in biomedicine