PRIVACY-PRESERVING MACHINE LEARNING FOR HEALTHCARE: OPEN CHALLENGES AND FUTURE PERSPECTIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine Learning (ML) has recently shown tremendous success in modeling various healthcare prediction tasks, ranging from disease diagnosis and prognosis to patient treatment. Due to the sensitive nature of medical data, privacy must be considered along the entire ML pipeline, from model training to inference. In this paper, we conduct a review of recent literature concerning Privacy-Preserving Machine Learning (PPML) for healthcare. We primarily focus on privacy-preserving training and inference-as-a-service, and perform a comprehensive review of existing trends, identify challenges, and discuss opportunities for future research directions. The aim of this review is to guide the development of private and efficient ML models in healthcare, with the prospects of translating research efforts into real-world settings.

1 INTRODUCTION

Machine Learning (ML) and Deep Learning (DL) have shown great promise in many domains, leveraging the use of large datasets. Some notable contributions include *AlphaFold* (Jumper et al., 2021) for the prediction of protein structures and *Transformers* (Vaswani et al., 2017) for natural language processing. Healthcare is one of the domains in which ML is expected to provide substantial improvements in the delivery of patient care worldwide (WHO, 2021). Given the rapid growth in the number of models over the last couple of years (Ravì et al., 2016; Miotto et al., 2018; Kaul et al., 2022; Javaid et al., 2022), healthcare applications deserve special consideration considering the sensitive nature of the data that is required to train the models and the safety-critical nature of medical decision-making.

In this regard, real-world implementation of such models is still hampered by ethical and legal constraints. Legal frameworks have been developed and enforced to guarantee the transparency and privacy of ML-based healthcare solutions, such as the *Health Insurance Portability and Accountability Act (HIPAA)* in the United States (Gostin et al., 2009) and the *General Data Protection Regulation (GDPR)* in Europe (Voigt & Von dem Bussche, 2017). Therefore, there is a crucial need for Privacy-Preserving Machine Learning (PPML) in healthcare to enable the implementation of trustworthy systems in the future. The main goal of this review is to provide a comprehensive overview of state-of-the-art PPML in healthcare and encourage the development of new methodologies that tackle specific challenges relevant to the nature of the domain.

Motivation. There exist several related literature reviews that focus on a specific subset of PPML for healthcare. Several highlight recent advancements in federated learning (Xu et al., 2021; Ali et al., 2022; Joshi et al., 2022; Nguyen et al., 2022), cryptographic techniques (Zalonis et al., 2022), or security aspects of ML models, such as adversarial attacks (Liu et al., 2021). Existing review articles cover a wide range of applications related to health and input data modalities, ranging from IoT sensors to medical images (Qayyum et al., 2020). Compared to existing work, our review has three main contributions with the intent of bridging between research pertaining to ML for healthcare and cybersecurity. First, we distinguish between PPML for training and inference, i.e., *ML-as-a-service*. Second, we focus on state-of-the-art (SOTA) literature published in the last three years, considering the high proliferation of ML in healthcare and recent methodological advancements in ML and DL (e.g., network architectures, model pre-training, etc.). Third, we consider studies that

develop or apply methodologies using two popular modalities based on publicly available datasets and state-of-the-art in ML for healthcare, namely, medical images and data extracted from Electronic Health Records (EHR) (Kaul et al., 2022). Although we acknowledge the importance of security for ML models, it is out of the scope of this paper since we primarily focus on privacy.

To this end, we review papers that meet the following inclusion criteria:

1. We include recently published work i.e., publication year ≥ 2020,
2. We include articles that focus on the application or development of PPML either for model training and/or inference, including but not restricted to homomorphic encryption, differential privacy, federated learning, and multi-party secure computation.
3. We include articles that consider clinical tasks involving medical images and/or EHR data.

In Section 2, we provide background knowledge about concepts and terminology concerning PPML. In Section 3, we provide an overview of the state-of-the-art pertaining to PPML for training (Section 3.1) and for inference (Section 3.2). Later in Section 4, we discuss open challenges and derive future directions. Finally, we provide concluding remarks in Section 5.

2 Privacy-preserving machine learning: background & terminology

2.1 Federated Learning

Since medical data is highly sensitive, data sharing is difficult, and subject to ethical restrictions and legal constraints if at all possible. Federated learning (FL) (McMahan et al., 2017) aims to overcome the challenges of data sharing by enabling collaborative training, which does not require that the involved parties share their training data. Therefore, the data remains private to each local node within the FL network, such that only the model updates are shared and integrated in a centralized model.

Federated averaging (McMahan et al., 2017) is the most common form of FL. In this setting, a centralized server is connected to N entities, which have their own training data. The central server orchestrates the collaborative training process as follows: (1) the initial model is distributed amongst all entities, (2) each entity performs a training iteration on their local model using their own training data, typically one epoch, and shares its resulting model parameters with the central server, (3) the server averages the model parameters shared by all entities and distributes the resulting (averaged) model amongst all entities, and (4) steps (2) and (3) are repeated sequentially until a performance threshold or a specific number of training iterations is achieved. FL has proven to be very efficient in training models with strong performance, while avoiding the need for data sharing (McMahan et al., 2017). However, FL might be vulnerable to privacy issues such as reconstruction attacks (Liu et al., 2022), thus requiring that it is combined with other privacy-preserving methods to ensure robust privacy guarantees (Nguyen et al., 2022).

2.2 Differential Privacy

Differential Privacy (DP) has its origins in statistical analysis of databases. Its main aim is to address the paradox of learning nothing about specific individuals, while learning useful information about the general population (Dwork et al., 2014). In the FL context, it is usually incorporated in the form of additive noise to model updates, either artificially or using a differentiable private optimizer, prior to transferring the updates from the entities to the central server (Abadi et al., 2016). The amount of artificial noise added is directly proportional to the degree of privacy desired (i.e., privacy budget) (Zhang et al., 2021c). DP can successfully make privacy attacks fail, such as reconstruction attacks, as the added noise hinders the inference of actual knowledge about the training data by the attacker. However, adding too much noise (i.e., high privacy budget) can hamper learning and negatively impact the model accuracy (Chilukoti et al., 2022).

2.3 HOMOMORPHIC ENCRYPTION

In mathematics, the term *homomorphic* refers to the transformation of a given set into another while preserving the relation between the elements in both sets. Thus, Homomorphic Encryption (HE) refers to the conversion of plaintext into ciphertext while preserving the structure of the data. Consequently, specific operations applied to the ciphertext will provide the same results as if they were applied to the plaintext but without compromising the encryption (Acar et al., 2018). That is, the plaintext data is never accessed nor decrypted as the operations are directly applied to the encrypted data. The result of the transformations on the ciphertext can only be decrypted back to plaintext by the encryption key owner.

Despite the benefit of provable privacy guarantees, the range of operations available in HE is restricted to addition and multiplication i.e., *fully* homomorphic encryption. This limits the set and number of transformations applicable to the data and requires the use of approximations for more complex operations (e.g., HE-ReLU is the polynomial approximation of the ReLU function (Yue et al., 2021b)). This also significantly increases the computational time needed to process encrypted text compared to plaintext by several orders of magnitude (Popescu et al., 2021).

2.4 MULTI-PARTY SECURE COMPUTATION

Secure Multi-Party Computation (SMPC) (Goldreich, 1998) provides a framework in which two or more parties jointly compute a public function with their data while keeping the inputs private and hidden from other parties using cryptographic protocols. Most protocols used for SMPC with more than two parties are based on Secret Sharing (SS). In SS, a portion of the secret input is shared among a number of other parties. Most ML methods use Shamir's SS and additive SS (Singh & Shukla, 2021b). Although these methods are considered information-theoretic secure cryptosystems, recent studies show that leakage of global data properties can occur in some scenarios (Zhang et al., 2021a).

3 OVERVIEW OF STATE-OF-THE-ART

Following the inclusion criteria described in Section 1, we summarize existing work on PPML for healthcare based on whether the work focuses on model training (Table 1) or model inference (Table 2). For each study (row) we describe several attributes. *Use case* provides a succinct summary of the objective of the study. *Model* reports the ML or DL architecture that was employed to model the task. *Dataset* summarizes the dataset that was used for model training and evaluation. Additionally, we use the * symbol to indicate the use of a private dataset. *ML task* describes the nature of the prediction task (e.g., binary or multi-class classification). *Input modality* reports the nature of the model's input data, which could either be *I* for medical images or *E* for EHR data. In the *Validation* column, we report whether the trained model was internally and/or externally evaluated, with ✔indicating the use of internal validation i.e., test set from the same distribution of the training data, and ✔✔indicating the assessment of the generalization of the model on an external test dataset. Lastly, *Metrics* lists the evaluation metrics used to describe the performance of the proposed model.

3.1 PRIVACY-PRESERVING TRAINING FOR HEALTHCARE

As observed in Table 1, the most commonly used privacy-preserving approach for model training is FL, either independently or in combination with DP. DP is added to increase the privacy of the FL training updates i.e., adding noise to the shared weights, thus making the system more robust to privacy threats, such as reconstruction attacks by an external actor intercepting the communication channel or an *honest-but-curious* central server (Nguyen et al., 2022).

The second most commonly investigated approach for private training is HE, which leverages encryption schemes to provide privacy with provable mathematical guarantees. However, as described in the previous section, training ML models on encrypted data increases the computational resources needed and the processing overhead by several orders of magnitude (Zhang et al., 2021b). It also adds noise to the training process due to the approximations of activation functions, especially in large models.

Table 1: **Summary of PPML in healthcare for model training.** We summarize studies that focus on developing PPML in the context of model training. We group them based on the methodology considered, i.e. federated learning, homomorphic encryption, and differential privacy.

Reference	Use case	Model	Dataset	ML task	Input modality	Validation	Metrics
FEDERATED LEARNING							
Dou et al. (2021)	COVID19 image interpretation	RetinaNet	Lung CTs*	Object detection	I	✔✔	mAP Specifity Recall AUROC
Field et al. (2022)	Decision support system	Logistic regression	Australian radio-oncology clinical centers*	Risk prediction	I+E	✔	AUROC C-index
Dayan et al. (2021b)	FL benchmark in healthcare	NN, LSTM, CNN	MIMIC-III PhysioNet ECG	Mortality prediction Multi-class	I+E	✔	AUROC AUPRC F1-Score
Yang et al. (2022)	FL benchmark in healthcare	Transformer, EfficientNetB0, ResNet-NC-SE	eICU, ISIC19, HAM10000, PhysioNet ECG	Mortality, LOS, Discharge time, Acuity prediction	I+E	✔	AUROC AUPRC
Sadilek et al. (2021)	FL benchmark in healthcare	Logistic regression, NN, DNN, Generalized linear model	UCI Heart failure MIMIC-III Malignancy in SARS-CoV-2	Risk prediction	E	✔	AUROC
Loftus et al. (2022)	COVID19 detection	DenseNet	5 institute COVID19 X-ray*	Binary	I	✔✔	AUROC AUPRC
Wolff et al. (2022)	Coronary artery calcification prediction	Random forest	CAC risk factors*	Risk prediction	E	✔	Sensitivity Specificity
Wang & Zhou (2022)	Gene expression cancer inference	-	iDASH 2020	Multi-class	E	✔	Accuracy AUC Recall Precision F1-score
Islam et al. (2022a)	Diabetic kidney disease risk prediction	Logistic regression, MLP	Cerner's Health Facts EMR	Risk prediction	E	✔	F1-Score
Deist et al. (2020)	Post-treatment 2-year survival	Logistic regression	8 international healthcare centers*	Survival prediction	E	✔	RMSE Accuracy AUROC
HOMOMORPHIC ENCRYPTION							
Boulila et al. (2022)	COVID19 detection	MobileNetV2	COVID19 X-ray	Multi-class	I	✔	Accuracy Precision Recall F1-score
Ma et al. (2020)	Heart & thyroid disease classification	XGBoost	UCI Heart Disease Kaggle Hypothyroid	Binary	E	✔	Accuracy
Paul et al. (2021)	ICU patient outcome	LSTM	MIMIC-III	Binary	E	✔	Recall Precision AUROC
Chen et al. (2022)	Dermatology diagnostics	SVM	UCI Dermatology	Multi-class	E	✔	Accuracy
Zhang et al. (2022)	Skin cancer classification	CNN	HAM10000	Multi-class	I	✔	Accuracy Overhead
DIFFERENTIAL PRIVACY							
Zhang et al. (2021b)	Thoracic pathology detection	DenseNet-121	CheXpert	Multi-class	I+E	✔	AUROC Accuracy
Chilukoti et al. (2022)	COVID19 detection	EfficientNetB2	COVID19 X-ray	Binary	I	✔	Accuracy
Suriyakumar et al. (2021)	Clinical prediction tasks	CNN DenseNet-121, Logistic regression, GRU-D	NIH Chest X-ray CIFAR MIMIC-III	Multi-class Binary	I+E	✔	AUROC

The third most common approach is standalone DP, which is less computationally demanding and provides strong privacy guarantees. However, the increase in privacy guarantees is negatively cor-

Table 1 Continued: **Continued summary of PPML in healthcare for model training.** We summarize here studies that use a combination of federated learning and differential privacy, blockchain, Secure Multi-Party Computation (SMPC), image encryption, and image modification.

Reference	Use case	Model	Dataset	ML task	Input modality	Validation	Metrics
FEDERATED LEARNING + DIFFERENTIAL PRIVACY							
Islam et al. (2022b)	Cardiomyopathy risk prediction	Random forest Naive Bayes	iDash '21 Breast Cancer TCGA	Risk prediction	E	✔	AUROC
Kerkouche et al. (2021)	In-hospital mortality prediction	Custom FCNN	USA Premier Healthcare*	Prediction	E	✔	AUROC Overhead
Dayan et al. (2021a)	COVID19 oxygen requirement prediction	ResNet34 DeepCrossNet	20 international institutes*	Risk prediction	I+E	✔✔	AUROC Recall Specificity
BLOCKCHAIN							
Zerka et al. (2020)	Distributed training	ResNet18	NSCLC-Radiomics	Binary	I	✔✔	AUROC
Warnat-Herresthal et al. (2020)	Disease classification	DNN	Blood transcriptomes*	Binary	E	✔	Accuracy
SMPC							
Hong et al. (2020)	Tumor samples classification	Logistic regression	iDASH 2019	Binary	E	✔	Accuracy Overhead
IMAGE ENCRYPTION							
Huang et al. (2022)	Brain tumor COVID19	DenseNet-121, XceptionNet	MRI Brain Tumor, COVID19 X-ray	Multi-class	I	✔	F1-score
IMAGE MODIFICATION							
Montenegro et al. (2021)	Glaucoma recognition	VGAN-based CNN	Warsaw-BioBase Disease-Iris v2.1	Binary	I	✔	F1-Score Accuracy

related with model accuracy, as it is associated with an increase in the quantity of noise applied. Therefore, the trade-off between privacy (i.e., privacy budget) and model accuracy is a relevant factor to take into account for the inclusion of DP in any ML solution. There are other PPML approaches for model training that have been evaluated in related work, including the addition of a blockchain ledger to avoid the centralization of training (i.e., fully distributed learning), image modification to increase data privacy in the context of model explainability, and SMPC as an alternative encryption scheme to HE.

Most of the reviewed studies use a single source of input data i.e., image or EHR and only one medical dataset. Although some studies train their models on several datasets, including popular computer vision benchmarks, the vast majority restrict their evaluation to one input modality from the same dataset. This limits the generalization of the results and neglects the potential improvement in predictive performance that could result from combining different data sources in multi-modal learning settings (Ramachandram & Taylor, 2017). Furthermore, most studies perform internal validation, such that the test sets are from the same distribution as the training dataset. This is generally a common challenge in healthcare applications considering distribution shifts across different hospitals, for example due to differences in patient demographics. Finally, most existing work focuses on convolutional neural networks to handle computer vision tasks. However, validation schemes and metrics reported are not consistent, making the comparison among them very difficult.

3.2 PRIVACY-PRESERVING INFERENCE FOR HEALTHCARE

We now focus on the literature employing PPML methods for inference, as summarized in Table 2. We frame PPML for inference as providing private *machine-learning-as-a-service* (MLaaS) or *inference-as-a-service* (IaaS) (Lins et al., 2021). In this scenario, a model with strong performance is controlled by a single party (i.e., model owner), and other external parties (i.e., clients) would like the model to perform inference on their own data. The external parties can share data samples with the model owner and their predictions are sent back. Due to legal and/or ethical constraints related to privacy, clients cannot disclose their data with the model owner, thus requiring the use of PPML to maintain the privacy of the data they wish to share.

Table 2: **Summary of PPML in healthcare for model inference.** We summarize studies that focus on developing PPML in the context of model inference. We group them based on the methodology considered, i.e. homomorphic encryption, combination of federated learning and Secure Multi-Party Computation (SMPC), differential privacy and SMPC, federated learning with blockchain and SMPC, and finally federated learning with differential privacy and homomorphic encryption.

Reference	Use case	Model	Dataset	ML task	Input modality	Validation	Metrics
HOMOMORPHIC ENCRYPTION							
Yue et al. (2021a)	Breast and cervical cancer classification	Convolutional LSTM	Cervigram Image, BreaKHis	Multi-class Binary	I	✔	AUROC
T'Jonck et al. (2022)	Breast cancer EHR classification	NN, SVM	UCI IRIS, UCI Breast Cancer	Multi-class Binary	E	✔	Accuracy Privacy budget Overhead
Sarkar et al. (2022)	Cancer inference via gene expression	SVM, Logistic regression, DNN	iDASH 2020	Multi-class	E	✔✔	Accuracy AUROC
Vizitiu et al. (2020)	Coronary angiography image classification	CNN	Whole-body circulation*, X-ray coronary angiography*	Multi-class Regression Binary	I	✔	Accuracy overhead
FEDERATED LEARNING + SMPC							
Ziller et al. (2020)+	Pediatric chest x-ray classification	ResNet18	Optical Coherence Tomography, Chest X-Ray	Multi-class	I	✔✔	AUROC Latency
Kaissis et al. (2021)+	Pediatric chest x-ray classification	ResNet18	Optical Coherence Tomography, Chest X-Ray	Multi-class	I	✔✔	AUROC Latency
DIFFERENTIAL PRIVACY + SMPC							
Singh & Shukla (2021a)	Chest X-ray pneumonia classification	CNN, VGG16	Chest X-ray Pneumonia	Binary	I	✔	Accuracy
Jarin & Eshete (2021)	Accuracy vs. privacy trade-off analysis	NN	IDC, MIMIC-III	Multi-class Binary Multi-class	I	✔	Accuracy Precision Recall Privacy
FEDERATED LEARNING + BLOCKCHAIN + SMPC							
Kasyap & Tripathy (2021)	Medical image classification	CNN	MedMNIST	Multi-class	I	✔	Accuracy
FEDERATED LEARNING + DIFFERENTIAL PRIVACY + HOMOMORPHIC ENCRYPTION							
Gopalakrishnan et al. (2021)	Multi-modal medical data classification	CNN	MedMNIST	Binary Ordinal regression Multi-class	I	✔	Accuracy Privacy budget Overhead

+ Kaissis et al. (2021) is an extension of Ziller et al. (2020).

Compared to the number of studies addressing PPML for training, a relatively fewer number have explored PPML for inference. Most studies within the theme of PPML for inference, focus on the deployment of the trained model as a service and its use by third parties. The most common approach for delivering PPML IaaS is HE, which ensures with provable mathematical guarantees that neither the model owner nor any intermediate party are able to inspect the original data nor the detection result i.e., both are encrypted and can only be decrypted by the data owner. Another common approach is SMPC, which also leverages encryption schemes, being used in combination with other privacy-preserving collaborative approaches such as FL, DP and blockchain.

Similar to PPML for training, most studies here use a single source of input data (i.e., images in most cases), neglecting many other diverse medical modalities of varying characteristics. The lack of use of benchmark medical datasets and inconsistent validation schemes and metrics hinders the generalization of the proposed approaches.

3.3 Open Challenges

There is no *one-size-fits-all* PPML approach for model training or inference by design. We observe that previous work pick and choose PPML approaches based on the intended clinical use case. Currently, there is no consensus on what different "privacy models" look like in healthcare. Since the methodology depends on the use case, we also observe a clear trade-off between privacy and accuracy, based on the availability of computational resources. For instance, standalone FL is

computationally faster than HE, but it does not provide strong privacy guarantees. On the other hand, HE and DP can provide strong privacy guarantees but they add noise to the model both for private training and private inference resulting in less accurate solutions. For HE, this is especially critical for model training where successive layers of approximations are needed to perform operations that are not supported, such as *softmax*, or that are computationally inefficient, such as max pooling. In general, encryption-based options are provably secure but computationally inefficient, since they increase the processing overhead of training using cyphertext compared to plaintext data.

Additionally, the availability of computational resources is a decisive factor in choosing a particular PPML methodology. For instance, in the HE scenario, sending data over a communication channel does not require infrastructure for model training but still requires handling the encryption/decryption process appropriately. In the FL context, it requires that the entity has allocated resources for model training.

The centralization of model training in FL poses an additional security threat. Relying on a single central server entails a single point of failure that is highly susceptible to security attacks such as Denial-of-Service. Although blockchain has been proposed to achieve fully distributed training and mitigate this threat, it increases the complexity of the information technology infrastructure significantly, requiring dedicated resources for the implementation of the distributed ledger and modeling framework.

Most existing work use a single dataset and do not conduct external validation, thus arising concerns about the generalization of the results. We observe that existing work focus on a limited set of medical datasets. Additionally, some work only evaluate their solutions on computer vision benchmark datasets (e.g., *MNIST* or *CIFAR-10*) inferring that good performance on these datasets will provide similar results on medical image data (Festag & Spreckelsen, 2020; Onesimu & Karthikeyan, 2020). However, this assumption is not empirically supported by work that uses both medical and non-medical datasets (Suriyakumar et al., 2021; Zhang et al., 2021b; Gopalakrishnan et al., 2021; Jarin & Eshete, 2021; Vizitiu et al., 2020) and must, therefore, be avoided.

MLaaS for healthcare has not been explored thoroughly. As demonstrated by the limited literature on this topic, we observe that the literature is highly skewed towards PPML for training. Considering disparities in technical capabilities and expertise, information technology resources, and availability of data across medical institutions, the case in which an entity does not have enough resources to perform model training independently is highly likely. Thus, the usage of third-party models as inference systems that can run on proprietary data is a prominent scenario that has not been thoroughly explored and should be considered in future research. MLaaS can provide access to models with strong performance, enabling full preservation of data privacy using PPML methods. This makes it a more efficient solution for small-scale or low-resource medical entities to access and leverage third-party knowledge.

4 Future research directions

4.1 Comprehensive evaluation on diverse medical datasets

For the sake of comparison and generalization of results, studies should complement their internal dataset evaluation with additional extensive evaluation on benchmark medical datasets. This is due to the fact that most of the existing work use a single dataset and do not perform external validation. The number of studies that use external datasets for validation is marginal. Only 7 out of the 36 studies considered validated their results with an independent test set. This hampers model generalization and hinders performance comparison among approaches built for the same medical task. For benchmarking, we suggest *MedMNIST* (Yang et al., 2023), which contains curated datasets for different medical tasks and modalities. Therefore, similar to *MNIST* or *CIFAR-10* for computer vision models, this medical dataset could be employed as a common benchmark for medical applications.

4.2 Multi-modal models

Current advances in ML for healthcare are moving towards multi-modal learning, where several sources of information are combined to improve performance (Ramachandram & Taylor, 2017; Soenksen et al., 2022). This approach not only tends to provide better performance but also ensures

a comprehensive understanding of the different physiological variables involved in studying and modeling the development of human biology and pathology. As observed in Sections 3.1 and 3.2, most work is restricted to a single modality. To develop robust and strong ML models, the use of different data sources to develop multi-modal systems is paramount. Notwithstanding that, the use of more clinical data entails more privacy concerns (e.g., individuals may be identified using correlated data) and requires more training resources due to increased model complexity. Therefore, additional privacy and computational constraints must be considered in the design of these algorithms.

4.3 Machine learning as a service (MLaaS)

The deployment of PPML within MLaaS is a very promising opportunity to access strong proprietary models by less resourceful institutions. Indeed, one of the main objectives of ML in healthcare is to develop efficient and scalable solutions that improve healthcare delivery. In addition to lack of resources, the deployment of these systems in medical settings can also be highly challenging (Kreuzberger et al., 2022; Wiesenfeld et al., 2022). The development of MLaaS is significantly less investigated than PPML for model training. Therefore, further research on this topic is required to provide secure, private and efficient data sharing between third-party model providers and client institutions. Reducing obstacles for clinical institutions to access powerful inference systems could lead to a major improvement in healthcare delivery across regions, bypassing physical barriers. It can also lead to an increase in the confidence and widespread adoption of ML in healthcare. It is important to note that the success of MLaaS is dependent on improvements in model generalizability and fairness in external datasets.

4.4 Integration of SOTA and advances in deep learning

Future work should also investigate the integration of recent advances in DL and ML models in healthcare, considering that most of the current PPML work focuses on convolutional neural networks. For instance, the *Transformer* architecture and its variants (Dosovitskiy et al., 2020), which are considered the current SOTA for many computer vision or natural language processing tasks, are only adopted by Yang et al. (2022) in the current related literature. Adopting SOTA architectures can take advantage of the latest advances in research, both in terms of optimizing hardware and software, to maintain performance improvements in clinical prediction tasks.

4.5 Global and local explainability

Transparency and model explainability are essential for trustworthy artificial intelligence (OECD, 2023b). However, PPML methods, such as data encryption or noise addition, hinder global model and local prediction explainability. The collision between two key principles for trustworthy artificial intelligence, secure and PPML OECD (2023a) and explainability, highlights an important research problem that is currently under-investigated. Only Montenegro et al. (2021) attempt to address this problem, which should encourage future work in this research direction.

5 Conclusion

In this paper, we introduce and summarize recent literature concerning PPML for model training and inference in the healthcare domain. We highlight trends, challenges and promising future research directions. In conclusion, we recognize the lack of consensus when it comes to defining the requirements of privacy-preserving frameworks in healthcare. This requires collaboration between machine learning scientists, healthcare practitioners, and privacy and security experts. From the perspective of advancing ML approaches, we encourage researchers to perform comprehensive evaluation of proposed algorithms on diverse medical datasets to increase generalization, to investigate the constraints of PPML in multi-modal learning settings, to further consider the promise of MLaaS in healthcare as a catalyst for improved healthcare delivery, and to adopt state-of-the-art advances in deep learning architectures to enhance model performance. Our suggestions aim to address research gaps and guide future research in PPML to facilitate the future adoption of trustworthy and private ML for healthcare.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, 51(4):1–35, 2018.

Mansoor Ali, Faisal Naeem, Muhammad Tariq, and Geroges Kaddoum. Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey. *IEEE journal of biomedical and health informatics*, 2022.

Wadii Boulila, Adel Ammar, Bilel Benjdira, and Anis Koubaa. Securing the classification of covid-19 in chest x-ray images: a privacy-preserving deep learning approach. In *2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, pp. 220–225. IEEE, 2022.

Yange Chen, Qinyu Mao, Baocang Wang, Pu Duan, Benyu Zhang, and Zhiyong Hong. Privacy-preserving multi-class support vector machine model on medical diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3342–3353, 2022.

Vijay Srinivas Tida Sai Venkatesh Chilukoti, Sonya Hsu, and Xiali Hei. Privacy-preserving deep learning model for covid-19 disease detection. *arXiv preprint arXiv:2209.04445*, 2022.

Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27(10): 1735–1743, 2021a.

Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27(10): 1735–1743, 2021b.

Timo M Deist, Frank JWM Dankers, Priyanka Ojha, M Scott Marshall, Tomas Janssen, Corinne Faivre-Finn, Carlotta Masciocchi, Vincenzo Valentini, Jiazhou Wang, Jiayan Chen, et al. Distributed learning on 20 000+ lung cancer patients–the personal health train. *Radiotherapy and Oncology*, 144:189–200, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Qi Dou, Tiffany Y So, Meirui Jiang, Quande Liu, Varut Vardhanabhuti, Georgios Kaissis, Zeju Li, Weixin Si, Heather HC Lee, Kevin Yu, et al. Federated deep learning for detecting covid-19 lung abnormalities in ct: a privacy-preserving multinational validation study. *NPJ digital medicine*, 4 (1):60, 2021.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Sven Festag and Cord Spreckelsen. Privacy-preserving deep learning for the detection of protected health information in real-world data: Comparative evaluation. *JMIR Formative Research*, 4(5): e14064, 2020.

Matthew Field, David I Thwaites, Martin Carolan, Geoff P Delaney, Joerg Lehmann, Jonathan Sykes, Shalini Vinod, and Lois Holloway. Infrastructure platform for privacy-preserving distributed machine learning development of computer-assisted theragnostics in cancer. *Journal of Biomedical Informatics*, 134:104181, 2022.

Oded Goldreich. Secure multi-party computation. *Manuscript. Preliminary version*, 78(110), 1998.

Aparna Gopalakrishnan, Narayan P Kulkarni, Chethan Raghavendra, Raghavendra Manjappa, Prasad B Honnavalli, and Sivaraman Eswaran. Primed: Private federated training and encrypted inference on medical images in healthcare. *Available at SSRN 4196696*, 2021.

Lawrence O Gostin, Laura A Levit, Sharyl J Nass, et al. Beyond the hipaa privacy rule: enhancing privacy, improving health through research. 2009.

Cheng Hong, Zhicong Huang, Wen-jie Lu, Hunter Qu, Li Ma, Morten Dahl, and Jason Mancuso. Privacy-preserving collaborative machine learning on genomic data using tensorflow. In *Proceedings of the ACM Turing Celebration Conference-China*, pp. 39–44, 2020.

Qi-Xian Huang, Wai Leong Yap, Min-Yi Chiu, and Hung-Min Sun. Privacy-preserving deep learning with learnable image encryption on medical images. *IEEE Access*, 10:66345–66355, 2022.

Humayera Islam, Khuder Alaboud, Tanmoy Paul, Md Kamruz Zaman Rana, and Abu Mosa. A privacy-preserved transfer learning concept to predict diabetic kidney disease at out-of-network siloed sites using an in-network federated model on real-world data. In *AMIA Annual Symposium Proceedings*, volume 2022, pp. 264. American Medical Informatics Association, 2022a.

Tanzir Ul Islam, Reza Ghasemi, and Noman Mohammed. Privacy-preserving federated learning model for healthcare data. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0281–0287. IEEE, 2022b.

Ismat Jarin and Birhanu Eshete. Pricure: privacy-preserving collaborative inference in a multi-party setting. In *Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics*, pp. 25–35, 2021.

Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, 3:58–73, 2022.

Madhura Joshi, Ankit Pal, and Malaikannan Sankarasubbu. Federated learning for healthcare domain-pipeline, applications and challenges. *ACM Transactions on Computing for Healthcare*, 3(4):1–36, 2022.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Georgios Kaissis, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima Jr, Jason Mancuso, Friederike Jungmann, Marc-Matthias Steinborn, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, 3(6):473–484, 2021.

Harsh Kasyap and Somanath Tripathy. Privacy-preserving decentralized learning framework for healthcare system. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(2s):1–24, 2021.

Deeksha Kaul, Harika Raju, and BK Tripathy. Deep learning in healthcare. *Deep Learning in Data Analytics: Recent Techniques, Practices and Applications*, pp. 97–115, 2022.

Raouf Kerkouche, Gergely Acs, Claude Castelluccia, and Pierre Genevès. Privacy-preserving and bandwidth-efficient federated learning: An application to in-hospital mortality prediction. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 25–35, 2021.

Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. Machine learning operations (mlops): Overview, definition, and architecture. *arXiv preprint arXiv:2205.02302*, 2022.

Sebastian Lins, Konstantin D Pandl, Heiner Teigeler, Scott Thiebes, Calvin Bayer, and Ali Sunyaev. Artificial intelligence as a service: Classification and research directions. *Business & Information Systems Engineering*, 63:441–456, 2021.

Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

Pengrui Liu, Xiangrui Xu, and Wei Wang. Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives. *Cybersecurity*, 5(1):1–19, 2022.

Tyler J Loftus, Matthew M Ruppert, Benjamin Shickel, Tezcan Ozrazgat-Baslanti, Jeremy A Balch, Philip A Efron, Gilbert R Upchurch Jr, Parisa Rashidi, Christopher Tignanelli, Jiang Bian, et al. Federated learning for preserving data privacy in collaborative healthcare research. *Digital Health*, 8:20552076221134455, 2022.

Zhuoran Ma, Jianfeng Ma, Yinbin Miao, Ximeng Liu, Kim-Kwang Raymond Choo, Ruikang Yang, and Xiangyu Wang. Lightweight privacy-preserving medical diagnosis in edge computing. *IEEE Transactions on Services Computing*, 15(3):1606–1618, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Helena Montenegro, Wilson Silva, and Jaime S Cardoso. Privacy-preserving generative adversarial network for case-based explainability in medical image analysis. *IEEE Access*, 9:148037–148047, 2021.

Dinh C Nguyen, Quoc-Viet Pham, Pubudu N Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(3):1–37, 2022.

OECD. Robustness, security and safety (principle 1.4), 2023a. URL `https://oecd.ai/en/dashboards/ai-principles/P8`.

OECD. Transparency and explainability (principle 1.3), 2023b. URL `https://oecd.ai/en/dashboards/ai-principles/P7`.

J A Onesimu and J Karthikeyan. An efficient privacy-preserving deep learning scheme for medical image analysis. *Journal of Information Technology Management*, 12(Special Issue: The Importance of Human Computer Interaction: Challenges, Methods and Applications.):50–67, 2020.

Jestine Paul, Meenatchi Sundaram Muthu Selva Annamalai, William Ming, Ahmad Al Badawi, Bharadwaj Veeravalli, and Khin Mi Mi Aung. Privacy-preserving collective learning with homomorphic encryption. *IEEE Access*, 9:132084–132096, 2021.

Andreea Bianca Popescu, Ioana Antonia Taca, Cosmin Ioan Nita, Anamaria Vizitiu, Robert Demeter, Constantin Suciu, and Lucian Mihai Itu. Privacy preserving classification of eeg data using machine learning and homomorphic encryption. *Applied Sciences*, 11(16):7360, 2021.

Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14:156–180, 2020.

Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2016.

Adam Sadilek, Luyang Liu, Dung Nguyen, Methun Kamruzzaman, Stylianos Serghiou, Benjamin Rader, Alex Ingerman, Stefan Mellem, Peter Kairouz, Elaine O Nsoesie, et al. Privacy-first health research with federated learning. *NPJ digital medicine*, 4(1):132, 2021.

Esha Sarkar, Eduardo Chielle, Gamze Gursoy, Leo Chen, Mark Gerstein, and Michail Maniatakos. Scalable privacy-preserving cancer type prediction with homomorphic encryption. *arXiv preprint arXiv:2204.05496*, 2022.

Shreyansh Singh and KK Shukla. Privacy-preserving machine learning for medical image classification. *arXiv preprint arXiv:2108.12816*, 2021a.

Shreyansh Singh and KK Shukla. Privacy-preserving machine learning for medical image classification. *arXiv preprint arXiv:2108.12816*, 2021b.

Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digital Medicine*, 5(1):149, 2022.

Vinith M Suriyakumar, Nicolas Papernot, Anna Goldenberg, and Marzyeh Ghassemi. Chasing your long tails: Differentially private prediction in health care settings. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 723–734, 2021.

Kristof T'Jonck, Chandrakanth R Kancharla, Bozheng Pang, Hans Hallez, and Jeroen Boydens. Privacy preserving classification via machine learning model inference on homomorphic encrypted medical data. In *2022 XXXI International Scientific Conference Electronics (ET)*, pp. 1–6. IEEE, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Anamaria Vizitiu, Cosmin Ioan Niţă, Andrei Puiu, Constantin Suciu, and Lucian Mihai Itu. Towards privacy-preserving deep learning based medical imaging applications. In *2019 IEEE international symposium on medical measurements and applications (MeMeA)*, pp. 1–6. IEEE, 2020.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Qingyong Wang and Yun Zhou. Fedspl: federated self-paced learning for privacy-preserving disease diagnosis. *Briefings in Bioinformatics*, 23(1):bbab498, 2022.

Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N Ahmad Aziz, et al. Swarm learning as a privacy-preserving machine learning approach for disease classification. *BioRxiv*, pp. 2020–06, 2020.

WHO. Who issues first global report on artificial intelligence (ai) in health and six guiding principles for its design and use, 2021. URL `https://www.who.int/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use`.

Batia Mishan Wiesenfeld, Yin Aphinyanaphongs, and Oded Nov. Ai model transferability in healthcare: a sociotechnical perspective. *Nature Machine Intelligence*, 4(10):807–809, 2022.

Justus Wolff, Julian Matschinske, Dietrich Baumgart, Anne Pytlik, Andreas Keck, Arunakiry Natarajan, Claudio E von Schacky, Josch K Pauling, and Jan Baumbach. Federated machine learning for a facilitated implementation of artificial intelligence in healthcare–a proof of concept study for the prediction of coronary artery calcification scores. *Journal of Integrative Bioinformatics*, 19(4), 2022.

Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5:1–19, 2021.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Seongjun Yang, Hyeonji Hwang, Daeyoung Kim, Radhika Dua, Jong-Yeup Kim, Eunho Yang, and Edward Choi. Towards the practical utility of federated learning in the medical domain. *arXiv preprint arXiv:2207.03075*, 2022.

Zijie Yue, Shuai Ding, Lei Zhao, Youtao Zhang, Zehong Cao, Mohammad Tanveer, Alireza Jolfaei, and Xi Zheng. Privacy-preserving time-series medical images analysis using a hybrid deep learning framework. *ACM Transactions on Internet Technology (TOIT)*, 21(3):1–21, 2021a.

Zijie Yue, Shuai Ding, Lei Zhao, Youtao Zhang, Zehong Cao, Mohammad Tanveer, Alireza Jolfaei, and Xi Zheng. Privacy-preserving time-series medical images analysis using a hybrid deep learning framework. *ACM Transactions on Internet Technology (TOIT)*, 21(3):1–21, 2021b.

Jasmin Zalonis, Frederik Armknecht, Björn Grohmann, and Manuel Koch. Report: State of the art solutions for privacy preserving machine learning in the medical context. *arXiv preprint arXiv:2201.11406*, 2022.

Fadila Zerka, Visara Urovi, Akshayaa Vaidyanathan, Samir Barakat, Ralph TH Leijenaar, Sean Walsh, Hanif Gabrani-Juma, Benjamin Miraglio, Henry C Woodruff, Michel Dumontier, et al. Blockchain for privacy preserving and trustworthy distributed machine learning in multicentric medical imaging (c-distrim). *Ieee Access*, 8:183939–183951, 2020.

Li Zhang, Jianbo Xu, Pandi Vijayakumar, Pradip Kumar Sharma, and Uttam Ghosh. Homomorphic encryption-based privacy-preserving federated learning in iot-enabled healthcare system. *IEEE Transactions on Network Science and Engineering*, 2022.

Wanrong Zhang, Shruti Tople, and Olga Ohrimenko. Leakage of dataset properties in multi-party machine learning. In *USENIX Security Symposium*, pp. 2687–2704, 2021a.

Xinyue Zhang, Jiahao Ding, Maoqiang Wu, Stephen TC Wong, Hien Van Nguyen, and Miao Pan. Adaptive privacy preserving deep learning algorithms for medical data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1169–1178, 2021b.

Xinyue Zhang, Jiahao Ding, Maoqiang Wu, Stephen TC Wong, Hien Van Nguyen, and Miao Pan. Adaptive privacy preserving deep learning algorithms for medical data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1169–1178, 2021c.

Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Da Lima Costa Junior, Jason Mancuso, Marcus Makowski, Daniel Rueckert, Rickmer Braren, et al. Privacy-preserving medical image analysis. *arXiv preprint arXiv:2012.06354*, 2020.