Trade-offs between Privacy-Preserving and Explainable Machine Learning in Healthcare

Seminar Paper

by

Tobias Budig,

Selina Herrmann,

Alexander Dietz

Institute of Applied Informatics and Formal Description Methods (AIFB)

KIT Department of Economics and Management

Advisor:	M.Sc. Konstantin Pandl
Supervisor:	Prof. Dr. Ali Sunyaev
Submitted:	September 2, 2020

Abstract

Explainability and privacy are two key concerns when training a machine learning model, especially in critical information infrastructure, such as the healthcare sector. So far, researchers are uncertain of possible trade-offs and their impact. We have conducted a systematic literature review to identify the current state of research and possible trade-offs between the explainability and privacy of machine learning models. We present possible ways of implementing explainability methods in a privacy-preserving setting with federated learning focused on the analysis of medical images. Our results show that only a few researchers have discussed possible trade-offs between explainable and privacy-preserving machine learning. The relevant papers indicate that there is a natural trade-off. A higher level of explainability can make a model more vulnerable to attacks and therefore have a higher risk of privacy leakage. For our federated learning example, we have selected three methods (SHapley Additive exPlanations, Gradient-weighted Class Activation Mapping, and Local Interpretable Model-Agnostic Explanations) that one can theoretically implement without risking privacy. However, experiments would be necessary to confirm these ideas.

Contents

List of Abbreviations

AI artificial intelligence.

CCPA California Consumer Privacy Act.

CNN convolutional neural network.

DP differential privacy.

EXML explainable machine learning.

GDPR European General Data Protection Regulation.

Grad-CAM Gradient-weighted Class Activation Mapping.

HE homomorphic encryption.

LIME Local Interpretable Model-Agnostic Explanations.

LLM locally linear maps.

ML machine learning.

PPML privacy-preserving machine learning.

SHAP SHapley Additive exPlanations.

SMC secure multiparty computation.

TEE trusted execution environment.

1 Introduction

1.1 Motivation

Machine learning (ML) has enormous potential in healthcare and has been used in medicine since the very beginning of the field [Ahmad et al., 2018, p. 1]. However, only in recent years, the importance of ML-based solutions in healthcare has been recognized, and in a few years, it will become indispensable.

For example, ML helps immensely in the detection of chronic diseases. The careful analysis of medical data ensures that maladies are detected earlier, and patients receive better care [Chen et al., 2017, p. 1]. As the University of Pennsylvania was able to co-develop a technology, which can train an artificial intelligence (AI) so that it can identify brain tumors in x-ray images using the privacy-preserving method federated learning [Intel, 2020].

The increasing computing capacity, the use of electronic health records in hospitals, and the availability of data cause this area to evolve rapidly [McCradden et al., 2020, p. 1]. Nevertheless, big data also entails ethical concerns about responsibility, trust, and accountability, among others. In the implementation of ML, public views are fundamental. On the one hand, to encourage companies to invest in AI, on the other hand, to support educational initiatives that encourage trust and support among the population. These ethnic controversies and the speed with which this technology is advancing and developing can affect public confidence.

An important issue is to protect the privacy of an individual. Therefore, privacy-preserving machine learning (PPML) tries to ensure that one cannot exploit the used training data. Especially in the healthcare sector as [McCradden et al., 2020] showed in her study about "Ethical concerns around use of artificial intelligence in health care research [...]", this is essential, as sensitive data is involved, hence trust is inevitable. While some people are open to new technologies, many cautious users are skeptical about them and wait to see how they develop [HSBC, 2017, p. 3]. Advertisers and designers make every effort that their product inspires confidence and spend vast amounts of money to ensure this. A survey by HSBC also shows the distrust towards technology in medicine. Only 14% state that they would trust a humanoid robot programmed by leading surgeons to perform open-heart surgery on them, compared to already 9% who would trust a family member led by a surgeon [p. 4-5].

An explanation is essential to build trust. However, there is no mathematical definition of explainability [Molnar, 2019, chapter 2]. A non-mathematical definition is that explainability "is the degree to which a human can understand the cause of a decision." It is easier for one to understand why predictions or decisions have been made, the better the explanation of the ML model is. However, there is the problem that many ML methods are hard to explain, as they work like a black box [Ahmad et al., 2018, p. 1]. Explainable machine learning (EXML) allows the user to understand, question, and even improve the ML system. Furthermore, the explainability and privacy of ML models is also part of ongoing political discussions, as we have seen with regulations such as European General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA). These regulations include statements about how data can be used for automated decision and some even argue for a "right for explanation" [Goodman and Flaxman, 2017] in the GDPR. Hence, the topic is not only part of current research but also essential for practical implementations. For the given practice problem, it is of interest if there are trade-offs between EXML and PPML and how these are relevant to the field of healthcare.

1.2 Objectives

Privacy and explainability are crucial when using ML in healthcare. There is currently only limited literature discussing both topics and especially their trade-offs. Therefore, our main objective with this work is to examine possible trade-offs between EXML and PPML. An overview would benefit future research on these topics. To answer this question, we want to discuss the findings of the relevant literature and present the effects of EXML on privacy exploitation risk, how PPML can hamper this risk and what are overall outcomes of PPML and EXML on a model's performance. Thus, giving an overview of the latest research on these topics. As healthcare is one of the fields, we want to assess the relevance of EXML and PPML for healthcare. To do so, we examine if different EXML methods can be implemented in a PPML setting. In this context, we have chosen the example of federated learning for Image Analysis as the implementation of EXML would be highly beneficial in this application.

2 Background

2.1 Privacy-Preserving Machine Learning

As mentioned in our introduction, one of the key challenges of ML in healthcare is keeping the patient's data private and secure. For medical research purposes, it is highly beneficial to share data and collect larger data sets to train better ML models. Therefore, helping medical personnel make more precise and correct decisions. However, medical records are some of the most sensitive data, and keeping this data private is inevitable. Due to this need, researchers have developed privacy-preserving methods. We will introduce the most crucial ones that are or will be relevant to healthcare applications, such as differential privacy (DP), trusted execution environment (TEE), cryptographic approaches, and federated learning.

2.1.1 Differential Privacy

DP is a privacy-preserving method that is based on a probabilistic model [Dwork and Roth, 2014]. Given the situation of a database where each row represents the information about an individual, the goal is to ensure that a single change to the database does not affect the results by much. Therefore, the input of an individual is privacy secured. Adding random noise, mostly LaPlace or Gaussian distributed, to the data accomplishes this.

Definition 1. (Differential Privacy) A randomized mechanism $\mathcal{M} : D \to R$ with domain D and range R satisfies (ε, δ)-differential privacy if for any two adjacent inputs $x, y \in D$ and for any subset of outputs of S $\subseteq R$ it holds that

$$\Pr[\mathcal{M}(x) \in S] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in S] + \delta,$$

where δ represents the privacy budget [Dwork and Roth, 2014, p. 17-18].

In general, a smaller (ε, δ)-value provides higher privacy. This definition provides multiple advantages such as "composability, group privacy, and robustness to auxiliary information" [Abadi et al., 2016, p. 2].

2.1.2 Trusted Execution Environment

A TEE is defined as "a secure, integrity-protected processing environment, consisting of processing, memory, and storage capabilities" [Asokan et al., 2014, p. 1]. TEEs are already in use in many of our mobile devices as a way to secure two-factor authentication or IoT applications [Ekberg et al., 2013, p. 1].

In the context of ML TEEs can provide a solution for protecting the privacy of training data as the training of a model can be run within the TEE. A user sends encrypted data to the enclave, which then decrypts the data and trains the model [Narra et al., 2019, p. 1]. This approach is also interesting for healthcare applications as a clinic could send sensitive encrypted data to a research site to support the training of a ML model, for example, on image classifications of brain tumors. By doing so, a large number of clinics can train the model as long as the clinic trusts the TEE. Current work also proposes the use of TEE in a federated learning setting [Mo and Haddadi, 2019, p. 1]. This approach should hamper the information leakage as the model is trained within a TEE on the client-side, and one only encrypted data is sent to the central server.

2.1.3 Cryptographic Approaches

Two of the most relevant cryptographic approaches are secure multiparty computation (SMC) and homomorphic encryption (HE). SMC is a method that "enables computation on sensitive data from multiple sources while maintaining privacy" [Chen et al., 2019a, p. 1]. It guarantees that only each computation reveals the outcome to other users. The theoretical approach was introduced in 1982 by A. C. Yao [Yao, 1982], but due to the high need for computation power, it took until 2004 for a first general notable implementation [Malkhi et al., 2004]. One of the most used methods is Shamir's secret sharing [Shamir, 1979].

HE requires algorithms to allow certain operations to be carried out on ciphertexts [Aslett et al., 2015, p. 2]. By doing so, there is no difference between operating on the ciphertext or the original message. The user encrypts the data with a public key, and only certain individuals with a private secret key can decrypt the data at any point in time. In healthcare, HE can be used to share data between research institutes or store sensitive data securely in a cloud [Raisaro et al., 2018]. Furthermore, there are also first implementations of HE in other PPML methods, like federated learning, to ensure a higher level of privacy [Xu et al., 2019]. Current restraints in HE include the high computational cost [Aslett et al., 2015, p. 9-10], the inability to perform division operations, and the large data size of the ciphertext.

2.1.4 Federated Learning

By definition, federated learning is the collective training of one model by multiple clients orchestrated by one central server. It provides the advantage of decentralized training data and a high level of privacy by design. Especially in the context of the GDPR federated learning is currently a promising technique to ensure PPML.

Since its introduction in 2016, [Konečný et al., 2016], a wide range of applications use federated learning. One of the most prominent is the use in the Gboard mobile keyboard by Google. Each user downloads the current model, trains it with their local data, and then sends the updated model back to the server, which averages all inputs to compute the latest model. By doing so, the user can get recommendations for any input on their smartphone based on the data of millions of other users without ever revealing their private data to any other user or server [Yang et al., 2018, p. 8]. Federated learning currently faces also challenges in the implementation and security [Kairouz et al., 2019], which are also relevant to the field of healthcare. First, the central server can be an exploitable weakness as it has to handle all the clients' inputs. Furthermore, it requires a certain level of trust by the clients as some centralized authority has to decide on what to train and how to train the model. There is a risk of information leakage from the server to the client or at the server. An adversary could use this information to reconstruct a client's data based on the model and the gradient update.

Therefore, it is advisable to implement further privacy-preserving techniques such as DP, TEE, or cryptographic approaches in the design of a federated learning system to hamper such attacks. So far, it seems that DP and cryptographic approaches are the only efficient techniques [Huang et al., 2020]. Due to the promising results concerning privacy, researchers also try to implement applications in various fields of healthcare, such as dentistry [Schwendicke et al., 2020], wearables [Chen et al., 2019b], and image analysis [Li et al., 2019].

2.2 Explainable Machine Learning

In this section, we want to give a brief overview of EXML techniques with a focus on the explanation of images predictions.

EXML refers to an ML system that "produces details or reasons to make its functioning clear and easy to understand" [Arrieta et al., 2020]. To reach this, different approaches are developed depending on the class of ML model. Here we differentiate between transparent and complex ML models, which consider the inherent complexity. For complex models, we need to generate an explanation after training. These so-called "post-hoc" techniques can be classified into two dimensions. First, the scope of the explanation, global or local, describes whether the aim of an explanation is only for one prediction or the overall behavior. Second, explainable algorithms can be, on the one hand, model-specific if they only work with some kind of model architecture. On the other hand, model-agnostic algorithms are independent of the model properties and can, therefore, be applied to all kinds of ML models. [Molnar, 2019, 2].

2.2.1 Transparent and complex Machine Learning models

Simple ML models, like linear regression or decision trees, can be interpreted without further techniques. Here, humans can interpret the weights of the regression to understand the model's decisions. Their advantage is providing out-of-the-box explainability, but shallow models are not as powerful as complex ones in nonlinear domains.

On the other hand, we need to use post-hoc methods for complex ML models like neural networks to achieve explainability. In this case, we can not interpret the neural network's weights directly due to the huge number of parameters and the inner complexity. [Molnar, 2019, 2.2] Therefore, there is a trade-off between the intrinsic model explainability and the model accuracy [Arrieta et al., 2020, p. 100].

2.2.2 Local and global explanations of Machine Learning models

ML explanations can have a different scope. On the one hand, local explainability describe which input features have a big impact on one specific predicted result. Common methods are Local Interpretable Model-Agnostic Explanations (LIME), Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP). They are useful to justify an end-user [Doshi-Velez and Kim, 2017, p. 7]. On the other hand, global explanations describe the behavior of the whole model. Methods, such as ProtoDash [Gurumoorthy et al., 2017], try to investigate and describe the patterns learned by the model. Incorrect learnings (e.g. biases) can be discovered here [Doshi-Velez and Kim, 2017, p. 7]. However, local explanations can also help to identify biases in ML predictions [Selvaraju et al., 2017].

2.2.3 Model-agnostic and model-specific explainability

Model-agnostic post-hoc algorithms only need the input object and the result of an ML model to provide explainability for this model. Therefore, they do not depend on specific model structure or architecture. There are two main classes of model-agnostic post-hoc methods.

First, explanation by simplification, where we try to approximate parts of the complex ML model with simpler ones. Here, LIME is the most known approach. It builds local linear models trained by the original model input-output pairs to explain individual predictions. One can apply it to tabular, image, or text data. LIME for image explanations will be discussed later in section 5.3.2.

The second class, feature relevance explanations, tries to rank or measure the impact of features on the predicted result. SHAP is currently the most popular approach in this field. It uses each possible combination of input features to get a marginal contribution to the result [Lundberg and Lee, 2017]. SHAP can be used for every neural network like a convolutional neural network (CNN).

In contrast, model-specific techniques can use internal parameters like architecture or optimization algorithm. They can benefit from direct access to the weights of a model. For Instance, Grad-CAM, the most common model-specific technique for CNNs, utilizes the gradient of the last convolutional layer [Selvaraju et al., 2017].

3 Methodology

3.1 Data Collection

Our approach for this paper can be separated into two different methods. For section 2 and section 5.2 we used a non-systematic literature review. The trade-offs between PPML and EXML were analyzed by a systematic literature review. To do so, we focused our search on the following scientific databases: IEEE Xplore, ArXiv, ACM Digital Library, AIS Electronic Library, Science Direct, and Scopus.

We believe that these databases are the most relevant for the discussed topics and cover a wide range of journals and conference publications. We used the following search string: TIKEAB((explain* OR interpret*) AND "machine learning" AND (privacy OR federated OR trust*)).

This search string requires the publication to have at least one of the search terms in the fields of explainability and privacy, respectively. Furthermore, it requires the papers to discuss ML. For the database ArXiv, we included only papers published since 2018 as papers on ArXiv are not peer-reviewed. We conducted the search on June 3rd, 2020 and had 526 results shown in Table 1 grouped by their database.

Database	**Number of Results**
IEEE Xplore	100
ArXiv	150
ACM Digital Library	189
AIS Electronic Library	5
Science Direct	21
Scopus	61

Table 1: results of literature review

3.2 Data Analysis

As the following step, we analyzed our results and looked for papers that discuss PPML and EXML. The majority of the found papers only discussed one of the topics and were, therefore, not applicable for a literature review on the trade-offs. We were able to identify three papers that discussed trade-offs. Section 4.1 examines these papers. It seems like the topic is part of ongoing research and can be a topic for further papers. Shokri et al. [Shokri et al., 2019] and Harder et al. [Harder et al., 2020] have stated to be the first ones to discuss possible trade-offs. Besides, these papers were recently published.

Backward search from our results has shown that ML models are vulnerable to Membership Inference Attacks. In other words, the attacker can exploit these black-box models only by accessing input and output queries to reconstruct members of the training set of the model [Shokri et al., 2017], [Truex et al., 2019], [Song et al., 2019]. Even worse, [Oh et al., 2019] showed, that it is possible to reverse-engineer some parts of Neural Networks like architecture or hyperparameters.

These results and the mentioned points in the papers raise three questions

1. Does Explainable Machine Learning increase privacy exploit probability?
2. Which Privacy-Preserving Machine Learning techniques save Machine Learning models from exploitation?
3. What impact do Explainable and Privacy-Preserving Machine Learning have on the model's accuracy?

which will be discussed in section 4.2.

4 Trade-offs

After conducting the systematic literature review as described in 3.2, we got three relevant papers. Two of it - [Harder et al., 2020] and [Shokri et al., 2019] - discuss the privacy-explainable trade-off directly. Furthermore, [Arrieta et al., 2020] focus on EXML in general but also investigate the privacy-explainable question from data-fusion perspective. To start, we first present the relevant key findings of the three relevant papers for further discussion.

4.1 Relevant Papers

4.1.1 Privacy risks of Explaining Machine Learning Models

[Shokri et al., 2019] discuss the possibilities for an adversary to use a model explanation to infer sensitive data of the training's set. In their work, they focus on Membership Inference Attacks and Reconstruction Attacks. The authors conducted experiments using gradient-based attribution methods or record-based influence measures. The used data sets included two sets with binary features and up to circa 200,000 records. Another two sets with mixed Features and up to circa 100,000 records. Furthermore, the authors used CIFAR-100, a benchmark data set for image classification. They used "fully connected multi-layer networks with tanh activations" [p.4] for training the datasets with binary and mixed features and a convolutional neural network for the CIFAR-100 image dataset.

When running a Membership Inference Attack, the adversary tries to determine the training set's data point. On the other hand, a Reconstruction Attack tries to get as many data of the training set as possible, basically reconstructing the training set.

They show that an adversary can exploit record and feature-based explanations. Furthermore, she can get the training set membership information of a data point. The only way to reduce information leakage is by adding noise to the data, using DP. In addition, they were able to conduct Reconstruction Attacks to extract parts of a training set given a record-based explanation. In this context, they also state that minorities in the training data have a high risk of being revealed by this attack.
The authors believe to be the first to discuss the trade-offs between explainability and privacy of a ML model.

4.1.2 Interpretable and Differential Private Predictions

[Harder et al., 2020] main question is if it is possible to have an explainable model without data lost cause of privacy protection. They claim to be the first ones to enquire about this question. Their approach is using locally linear maps (LLM) on a family of simple models, which should ensure privacy and do not expose the whole model, only the LLM.

LLM act as an approximation of differentiable functions, i.e., a collection of piecewise linear functions. If sufficient linear maps are available, local models compared to complex model counterparts have a relatively low loss inaccuracy. The more complex the data, the higher the loss of prediction accuracy. For using LLM, explainability and privacy are necessary.

The authors explain two ways to gain explainability. One relies on inherently explainable models. The other is post-processing schemes, for example, using gradient-based attributions. To provide privacy, they use DP, which lead them to the conflict that adding a large among noise ensures a high level of privacy, but is disadvantageous for predicting accuracy. They also state that high dimensional parameters imply high security lost. That is why they propose using a relatively small network, which can be partly trained and guarantees privacy using LLM. Their contributions are proposing a novel family of explainable models, providing explanations for 'local' and 'global' DP on classification and suggesting using random projections to deal better with privacy and accuracy trade-offs [p. 2].

After explaining their method using LLM, they eventuate the trade-off between privacy, explainability, and accuracy with experiments. They realize that it is tough to assure all three assertions. For example, reducing the level of privacy and removing random projections implies better explainability results. Meanwhile, private training benefits form increasing the dimensionality of random projections. Also, raising the number of LLM, the accuracy of private models decreases because the privacy budget is distributed over more parameters.

In conclusion, it is possible to use LLM to provide explainability and privacy. However, the data set, which they use, is simple and relatively small, so there is still the question: what are the limit of complexity for the LLM models? They also state that "several open questions for future research remain" [p. 7].

4.1.3 Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI

The literature review "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI" by [Arrieta et al., 2020] gives an overview of the current EXML research. Moreover, the authors provide a global taxonomy of EXML by classify current explainable techniques. In the beginning, they state that responsible AI needs explainable AI as well as privacy-preserving AI. Therefore, the authors investigated not only current EXML techniques but also their privacy impact.

Despite the explainability of ML models enable third parties to investigate the model, many papers about EXML do not cover the privacy topic.

A general privacy concern of ML models are attacks like membership inference attacks or approaches to reverse engineer the model's parameter, even with black-box models where third persons only have access to the input features and output predictions. EXML can, therefore, be used to increase the attacker's success probability.

Furthermore, the authors explore the field of responsible ML for data fusion. They state that EXML can compromise privacy for data fusion on different levels (data, model, Knowledge) [Arrieta et al., 2020, p. 107].

As a result, the authors conclude that further research is needed to ensure explainability as well as privacy for ML in general and data fusion.

4.2 Results

In the next subsections, we describe the contributions of the reviewed literature to the questions motivated in section 3.2.

4.2.1 Does Explainable Machine Learning increase privacy exploit probability?

In general, ML models are vulnerable to privacy leaks through membership inference attacks [Shokri et al., 2017] or reverse-engineering attacks [Oh et al., 2019]. [Shokri et al., 2019, p. 1] showed that EXML techniques can increase the privacy exploit probability because some EXML techniques like "gradient-based methods can leak a significant amount of information, much beyond what is leaked through the predicted labels." Moreover, they showed that record-based influence measures - a technique that explains the result by outputting the most critical point from the training set - is even worse. An adversary could reconstruct nearly 99% [Shokri et al., 2019, p. 11] of the used example. Experiments by [Shokri et al., 2019, p. 9] show, that especially minorities are vulnerable to these attacks. Here, the authors demonstrated for a diabetic hospital dataset that data rows of minorities like children or African-American people could be regenerated. Outliers are more likely to memorize by the model during training is the explanation for this behavior by [Shokri et al., 2019, p. 9].

In contrast, new, designed explainability techniques can be privacy-preserving. [Harder et al., 2020] describes a novel approach with LLM to guarantee privacy by making the gradient differential private. Nevertheless, in this paper, they have specific constraints like a relatively simple dataset and the lack of interaction with a complex counterpart. [Harder et al., 2020, p. 7]. Therefore, we cannot interpolate these results for other cases.

Furthermore, [Arrieta et al., 2020] states that common EXML techniques like LIME or SHAP are not investigated towards privacy concerns yet.

4.2.2 Which Privacy-Preserving Machine Learning techniques save Machine Learning models from exploitation?

To tackle the privacy question raised by model explanations, both - [Shokri et al., 2019] and [Harder et al., 2020] - propose to use noise towards the gradients to guarantee data protection. There are two techniques offered. First, differential private training should be immune to gradient-based attacks due to "gradient-based explanations only interact with the model, and not with the underlying training set." [Shokri et al., 2019, p. 7]. Especially smoothed gradients are resistant to the attacks performed by [Shokri et al., 2019, p. 7]. The average of the gradients in the surrounding area to the original point, together with adding Gaussian noise, seem to be a privacy-preserving approach, as the authors state.

Second, [Harder et al., 2020] developed an explainable method to provide privacy-preserving local and global explanations. The aim is to provide a set of LLMs to approximate the neural network's predictions. The authors realize this by adding weighted linear functions where the differential private stochastic gradient descent computes the linear coefficients. That means, adding noise to the gradient to make it differential private.[Harder et al., 2020, p. 4].

In addition, the authors used random projection (Johnson-Lindenstrauss transformation) to reduce the dimensions of the privatized parameters to increase accuracy [Harder et al., 2020, p. 4].

4.2.3 What impact do Explainable and Privacy-Preserving Machine Learning have on the model's accuracy?

[Harder et al., 2020, p. 7] states that there is a natural trade-off between privacy and accuracy. He also describes a triangle trade-off between privacy, explainability, and accuracy of an ML model. All three goals can not be achieved simultaneously. If one focuses on two of the three trades, the third one will be decreasing.

Also, DP comes with the cost of accuracy lost, which one could tackle by using a bigger dataset as [Shokri et al., 2019, p. 7] states. [Arrieta et al., 2020, p. 101] summarizes that there is a "need for further research toward the development of XAI Tools capable of explaining ML models while keeping the model's confidentiality in mind."

5 Implementation of Explainable Machine Learning in Federated Learning for Image Analysis

In the following section, we like to present a highly relevant example of the use of PPML in healthcare and how one can implement EXML methods in this scenario.

5.1 Convolutional Neural Networks for image analysis

CNNs are a special class of neural networks that use convolutions in place of general matrix multiplication [Goodfellow et al., 2016]. Forms of CNN are widely used in the field of visual computing and have shown great performance in object recognition and image classification. [Lawrence et al., 1997, Cireşan et al., 2011, Ciregan et al., 2012] Therefore CNN is the standard for medical imaging applications.

5.2 Federated Learning for medical image analysis

Medical images, taken by techniques like Magnet Resonance Imaging, X-ray, and ultrasound, present one of the greatest opportunities for using AI in healthcare. Clinics have millions of medical information. This data would be a good training set for a ML model. [Dash et al., 2019] However, especially in the field of diagnosis, which most image analysis is part of, it is crucial to have PPML and EXML models. Images like a brain scan are highly sensitive data, and it is in the patient's interest to have them maximally secured and not open to any other non-authorized third party. If one uses this data to train a ML model, it is even more important to ensure the privacy of the training data as multiple hospitals might use the final model. Therefore, it should not be possible to extract any of the original training data from the model.

For these reasons, we discuss the use of one of the most promising approaches in the field, federal learning for the application of medical image analysis. First implementations showed that it is possible to train a model on brain tumor segmentation data with federated learning that had a 99% accuracy of a model with shared data.[Sheller et al., 2019]

Furthermore, as we have seen in our previous sections, it is also crucial in the healthcare sector to offer the explainability of ML models to the end-user like a doctor. A decision based on images like a brain scan can have far-reaching consequences, and therefore it is inevitable to explain to the decision-maker if she is supposed to rely on the results of a ML model. Due to this fact, we present possible ways to incorporate explainability methods in a model that one trained with the federal learning approach.

5.3 Explainability methods for Image Classification

5.3.1 Gradient-weighted Class Activation Mapping

[Selvaraju et al., 2017] introduced Grad-CAM in 2017. The method is based on Class Activation Maps [Zhou et al., 2016]. Class Activation Maps highlight discriminative areas of an image. However, it is only applicable to a particular kind of CNN architecture. For this reason, the researchers developed Grad-CAM as a method that did not require any particular kind of CNN architecture. One can apply it to any already trained CNN. It uses gradient information in the last convolutional layer to highlight discriminative areas such as certain objects or classes. Due to the reason that this layer has "the best compromise between high-level semantics and detailed spatial information" [Selvaraju et al., 2017, p. 2].

5.3.2 Local Interpretable Model-Agnostic Explanations

As a model-agnostic explainable method, LIME can be applied to all ML models and therefore for CNNs, too. This method aims to approximate the decisions of the model at a specific point in a linear way. To achieve this, one generates a set of random data-points. In the case of image data, this means to divide the image into sections called super-pixels. As the second step, the model performs predictions of every sample point. Due to the local aim of the explanation, all samples are weighted in descending order of distance from the original point. In the last step, the model trains a linear classifier by the weighted random sample set and its predicted values. As a result of image data, super-pixels are turned on or off to identify the areas that contribute to the classification [Ribeiro et al., 2016].

5.3.3 SHapley Additive exPlanations

SHAP introduced in section 2.2.2 is a method to explain individual predictions. This method uses optimal Shapley Values from coalitional game theory. Shapely Values indicate how the prediction can be fairly distributed among the features. To do so, one calculates each feature's contribution to an individual prediction. [Molnar, 2019, 5.10.1].

To achieve this, the model generates a super-set of all possible combinations ("coalitions") of input features for every prediction. After that, for every of these 2^F coalitions, where F is the number of features, a model is trained. The difference between the prediction of two elements of the super-set computes the marginal contribution. Finally, weighing the marginal contributions results in a feature-wise explanation. [Lundberg and Lee, 2017] describes a technique to approximate the process by reducing the number of coalitions.

DeepSHAP is a model-specific sister algorithm of SHAP and researchers use it for CNNs. It leverages knowledge from internal weights and is optimized for Neural Networks [Lundberg and Lee, 2017, p. 7].

5.4 Result

Model-agnostic (e. g. SHAP, LIME) and model-specific (e. g. Grad-CAM) explainability approaches combined with federated learning should be possible because they only need local data even if methods require access to the model's parameters. Hence, one can perform the methods on a local and global level of federated learning. For a Shapley values explainability approach (e. g. SHAP), [Wang, 2019] proposes a method to balance model explainability and privacy in a federated learning setting. They showed that this approach enables explanations at the local guest system, without getting access to detailed information on guest data. Furthermore, explainability techniques can help detect bias in the training data as it gives a visual explanation of a decision. This is especially useful in healthcare, as there is a high risk of biased data and decisions [Gianfrancesco et al., 2018].

SHAP provides a complete explanation of all features as it is based on a well-founded theory [Molnar, 2019, 5.9.4]. The unified federated features of SHAP give useful information about the contribution of the federated features from the guest party, although the guest data does not need to be released [Wang, 2019, p. 4]. Nevertheless, the Shapley Value method has a bad performance. However, SHAP provides approximations that lead to an increase in computational performance and are, therefore, more applicable.

LIME is a model-agnostic explainability method, meaning it is possible to change the underlying model. The short and human-friendly explanations are easy to understand even for a layperson or in situations with little time and where a full explanation is not required [Molnar, 2019, 5.7.4]. It provides a correct definition of distance in general, however in this case, especially for the distance between super-pixels, Molnar recommends trying to use different kernels settings. A disadvantage of LIME is that it is not stable as [Alvarez-Melis and Jaakkola, 2018] state, "On the robustness of interpretability methods showed that even close points can result into different explanations."

In federated learning, the number of members' privacy increases because the local model's parameters are merged at the central server. However, there is the problem of unequal distributed data, which could decrease the model's performance and increase the bias [Bonawitz et al., 2019, p. 10].

6 Discussion

In the previous sections, we have shown the current state of research on the trade-offs between PPML and EXML, as well as the possible implementation of EXML methods in a federated learning environment for healthcare focusing on the particular example of image analysis. We have seen that both topics are part of ongoing research and highly relevant in healthcare.

6.1 Principal findings

In general, most current ML models are vulnerable to membership inference or reverse-engineering attacks, leading to a loss of privacy or intellectual property. These results are independent from EXML methods. A key inside by our literature review is that EXML could increase the exploit probability of ML models. To tackle this issue, new EXML techniques, based on differential private gradients, are developed. The approach seems to be promising due to the properties of DP. Another approach to guarantee privacy is federated learning. It can help protect the privacy of learning data between clients and the central server and between two clients. EXML methods do not require access to the model itself (model-agnostic) or if the only access to the local model of the client (model-specific). Therefore we expect EXML techniques together with federated learning to protect privacy as [Wang, 2019] showed for Shapley Values. Furthermore, concerning our example in Section 5.2, we see EXML for image data as one of the most relevant and promising applications in healthcare. An explanation of an image is an intuitive approach that can be understood by all kinds of users.

6.2 Implications for research and practice

These results lead to the conclusion that users must handle ML models carefully. Even black-box models, which often make up application programming interfaces, can be partially reverse-engineered even without EXML [Oh et al., 2019]. Researchers should focus more on EXML methods that are privacy-preserving to provide more real-world value. By doing so, they enable medical institutions to safely useML methods. Responsible ML needs to be privacy-preserving and explainable.

6.3 Limitations and future research

Due to the few papers, we have found and the simple assumptions in the research methods, a general statement on possible trade-offs is so far not possible. A next step could be to test the common EXML algorithms as Grad-CAM, SHAP and LIME, for their influence on specific attack scenarios. DP for the gradient or training data seems to be a promising approach that researchers should continue to test. For example a possible research question could be to focus on EXML for DP and evaluate the privacy loss. In addition, federated learning seems to be another useful approach, since the data remains local, and the risk is, therefore, more in the data transfer than in the ML model itself. A significant challenge here is to carry out a meaningful fusion of the data in the central server because the populations of the nodes can be different and of different quality.

Also, we see advantages concerning the common challenges of federated learning. First, federated learning does not eliminate the risk of biased training data. However, using EXML methods can identify such bias and therefore improve the overall results and accuracy. For example, the authors of Grad-CAM show that they were able to identify gender bias in a training data set with their visual explanation technique [Selvaraju et al., 2017]. Moreover, the fairness of the federated setting is highly relevant and also linked to biased training data. Visual explanations would be able to identify the misrepresentation of minority groups. Hence, it is beneficial to implement EXML methods in federated learning for image classification in the field of healthcare.

Overall, the reviewed literature speaks not with the same voice. One group was able to reverse-engineer a black box model and could, therefore, attack privacy. The other group described a novel approach to prevent such attacks. Now it is not clear how EXML techniques influence privacy attacks. Hence, further research is needed.

References

[Abadi et al., 2016] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA. Association for Computing Machinery.

[Ahmad et al., 2018] Ahmad, M. A., Eckert, C., and Teredesai, A. (2018). Interpretable machine learning in healthcare. *The IEEE Intelligent Informatics Bulletin*, pages 1–7.

[Alvarez-Melis and Jaakkola, 2018] Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

[Arrieta et al., 2020] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

[Aslett et al., 2015] Aslett, L. J. M., Esperança, P. M., and Holmes, C. C. (2015). A review of homomorphic encryption and software tools for encrypted statistical machine learning.

[Asokan et al., 2014] Asokan, N., Ekberg, J., Kostiainen, K., Rajan, A., Rozas, C., Sadeghi, A., Schulz, S., and Wachsmann, C. (2014). Mobile trusted computing. *Proceedings of the IEEE*, 102(8):1189–1206.

[Bonawitz et al., 2019] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., McMahan, H. B., et al. (2019). Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*.

[Chen et al., 2017] Chen, M., Hao, Y., Hwang, K., Wang, L., and Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5:8869–8879.

[Chen et al., 2019a] Chen, V., Pastro, V., and Raykova, M. (2019a). Secure computation for machine learning with spdz.

[Chen et al., 2019b] Chen, Y., Wang, J., Yu, C., Gao, W., and Qin, X. (2019b). Fedhealth: A federated transfer learning framework for wearable healthcare.

[Cireşan et al., 2011] Cireşan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, page 1237–1242. AAAI Press.

[Ciregan et al., 2012] Ciregan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3642–3649.

[Dash et al., 2019] Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1):54.

[Doshi-Velez and Kim, 2017] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[Dwork and Roth, 2014] Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

[Ekberg et al., 2013] Ekberg, J.-E., Kostiainen, K., and Asokan, N. (2013). Trusted execution environments on mobile devices. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, page 1497–1498, New York, NY, USA. Association for Computing Machinery.

[Gianfrancesco et al., 2018] Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

[Goodman and Flaxman, 2017] Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

[Gurumoorthy et al., 2017] Gurumoorthy, K. S., Dhurandhar, A., and Cecchi, G. (2017). Protodash: Fast interpretable prototype selection. *arXiv preprint arXiv:1707.01212*.

[Harder et al., 2020] Harder, F., Bauer, M., and Park, M. (2020). Interpretable and differentially private predictions. In *AAAI*, pages 4083–4090.

[HSBC, 2017] HSBC (2017). Trust in technology.

[Huang et al., 2020] Huang, X., Ding, Y., Jiang, Z. L., Qi, S., Wang, X., and Liao, Q. (2020). Dp-fl: a novel differentially private federated learning framework for the unbalanced data. *World Wide Web*.

[Intel, 2020] Intel (2020). Intel works with university of pennsylvania in using privacy-preserving ai to identify brain tumors.

[Kairouz et al., 2019] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. (2019). Advances and open problems in federated learning.

[Konečný et al., 2016] Konečný, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*.

[Lawrence et al., 1997] Lawrence, S., Giles, C. L., Ah Chung Tsoi, and Back, A. D. (1997). Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113.

[Li et al., 2019] Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M. J., and Feng, A. (2019). Privacy-preserving federated brain tumour segmentation.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

[Malkhi et al., 2004] Malkhi, D., Nisan, N., Pinkas, B., and Sella, Y. (2004). Fairplay - secure two-party computation system. In *USENIX Security Symposium*.

[McCradden et al., 2020] McCradden, M., Baba, A., and A. Saha A, e. a. (2020). Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: a qualitative study.

[Mo and Haddadi, 2019] Mo, F. and Haddadi, H. (2019). Efficient and private federated learning using tee.

[Molnar, 2019] Molnar, C. (2019). *Interpretable Machine Learning*. `https://christophm.github.io/interpretable-ml-book/`.

[Narra et al., 2019] Narra, K. G., Lin, Z., Wang, Y., Balasubramaniam, K., and Annavaram, M. (2019). Privacy-preserving inference in machine learning services using trusted execution environments.

[Oh et al., 2019] Oh, S. J., Schiele, B., and Fritz, M. (2019). Towards reverse-engineering black-box neural networks. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 121–144. Springer.

[Raisaro et al., 2018] Raisaro, J. L., Klann, J. G., Wagholikar, K. B., Estiri, H., Hubaux, J.-P., and Murphy, S. N. (2018). Feasibility of homomorphic encryption for sharing i2b2 aggregate-level data in the cloud. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2017:176–185. 29888067[pmid].

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

[Schwendicke et al., 2020] Schwendicke, F., Samek, W., and Krois, J. (2020). Artificial intelligence in dentistry: Chances and challenges. *Journal of Dental Research*, 99(7):769–774. PMID: 32315260.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

[Shamir, 1979] Shamir, A. (1979). How to share a secret. *Commun. ACM*, 22(11):612–613.

[Sheller et al., 2019] Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., and Bakas, S. (2019). Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., and van Walsum, T., editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 92–104, Cham. Springer International Publishing.

[Shokri et al., 2019] Shokri, R., Strobel, M., and Zick, Y. (2019). Privacy risks of explaining machine learning models. *arXiv preprint arXiv:1907.00164*.

[Shokri et al., 2017] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

[Song et al., 2019] Song, L., Shokri, R., and Mittal, P. (2019). Membership inference attacks against adversarially robust deep learning models. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE.

[Truex et al., 2019] Truex, S., Liu, L., Gursoy, M. E., Yu, L., and Wei, W. (2019). Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*.

[Wang, 2019] Wang, G. (2019). Interpret federated learning with shapley values. *arXiv preprint arXiv:1905.04519*.

[Xu et al., 2019] Xu, R., Baracaldo, N., Zhou, Y., Anwar, A., and Ludwig, H. (2019). Hybridalpha. *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security - AISec'19*.

[Yang et al., 2018] Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., Ramage, D., and Beaufays, F. (2018). Applied federated learning: Improving google keyboard query suggestions.

[Yao, 1982] Yao, A. C. (1982). Protocols for secure computations. In *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, pages 160–164.

[Zhou et al., 2016] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929.