nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-021-00390-3

Check for updates

Adversarial interference and its mitigations in privacy-preserving collaborative machine learning

Dmitrii Usynin[1,2,3], Alexander Ziller[2,3,4], Marcus Makowski[2], Rickmer Braren[2], Daniel Rueckert[1,4], Ben Glocker[1], Georgios Kaissis[1,2,3,4,6] ✉ and Jonathan Passerat-Palmbach[1,3,5,6]

Despite the rapid increase of data available to train machine-learning algorithms in many domains, several applications suffer from a paucity of representative and diverse data. The medical and financial sectors are, for example, constrained by legal, ethical, regulatory and privacy concerns preventing data sharing between institutions. Collaborative learning systems, such as federated learning, are designed to circumvent such restrictions and provide a privacy-preserving alternative by eschewing data sharing and relying instead on the distributed remote execution of algorithms. However, such systems are susceptible to malicious adversarial interference attempting to undermine their utility or divulge confidential information. Here we present an overview and analysis of current adversarial attacks and their mitigations in the context of collaborative machine learning. We discuss the applicability of attack vectors to specific learning contexts and attempt to formulate a generic foundation for adversarial influence and mitigation mechanisms. We moreover show that a number of context-specific learning conditions are exploited in similar fashion across all settings. Lastly, we provide a focused perspective on open challenges and promising areas of future research in the field.

Machine learning (ML), a rapidly developing subdomain of artificial intelligence, utilizes large quantities of data to train high-performance algorithms for tasks such as image analysis or language modelling. Such datasets are not always obtainable due to legal (regulatory frameworks such as the European General Data Protection Regulation (GDPR)[1] or the United States Health Insurance Portability and Accountability Act (HIPAA)[2]), ethical (sharing sensitive data without explicit informed consent) or practical (expensive data transfer) considerations.

Whenever individual data owners' data lack the required quality or quantity to successfully conduct a (single actor) ML task, the distributed training of ML models over confederations of data owners can offer regulation-compliant solutions. Such protocols allow the joint model to benefit from a richer, more diverse data pool, allowing data owners to enjoy better model generalization and utility, and can be financially incentivized through reimbursement for allowing others to use their data while preserving data ownership and governance schemes. Moreover, algorithmic services provided by third parties such as cloud providers over the network (Machine Learning as a Service) can allow the utilization of computational resources or models that would otherwise be locally unavailable.

Scenarios in which more than one data owner collaborates to train a common model are termed collaborative machine learning (CML)[3]. Whenever data provided by multiple data owners are centrally aggregated at a single site, the process is referred to as centralized CML, which, historically, has been the paradigm of choice for large-scale model training. However, due to privacy and governance concerns, a new paradigm has recently emerged in which data do not leave the site but instead models or their parameter updates are transmitted, referred to as decentralized CML[4], on which we focus here. Among its most commonly used implementations is federated learning (FL)[5], which allows to perform a CML training procedure by relying on the remote execution of ML algorithms and sharing of model parameters between sites. Besides FL, other CML implementations exist, such as peer-to-peer learning[6] or split learning[7]. The former method allows model training similar to FL, but obviates the aggregation server by sharing model updates with neighbouring clients instead. The latter involves training the model up to a specific layer (cut layer) locally and then sharing the activations with the next client or the aggregation server, where subsequent computations are performed.

A number of requirements are common to all CML protocols and summarized as trustworthy artificial intelligence[4,8]: integrity, verifiability and privacy preservation are critical for sensitive applications such as healthcare, where the correctness of results is paramount and the disclosure of confidential data must be avoided at all costs. One must, however, consider the possibility that not all protocol participants intend to cooperate honestly, or that they may attempt to interfere by subverting or invalidating the learning process to obtain knowledge that participants have not consented to sharing. We refer to such entities as adversaries.

In this Perspective, we present an overview of adversarial attacks against CML systems noting two different goals: utility compromise and privacy violation. We present an overview of mitigations against these attacks alongside their limitations. We conclude by discussing open challenges and future research directions in the field.

Overview and definitions

CML naturally lends itself to the study of adversarial interference as it harbours a number of risks not present in single-actor ML. As in all multi-actor environments, it is impossible to guarantee that all participants will both adhere to the learning protocol and refrain from trying to reveal other parties' confidential information. This is exacerbated by the fact that adversaries have ample opportunity

[1]Department of Computing, Imperial College London, London, UK. [2]Department of Diagnostic and Interventional Radiology, Technical University of Munich, Munich, Germany. [3]OpenMined, Oxford, UK. [4]Institute for Artificial Intelligence in Medicine and Healthcare, Technical University of Munich, Munich, Germany. [5]ConsenSys Health, New York, NY, USA. [6]These authors contributed equally: Georgios Kaissis, Jonathan Passerat-Palmbach. ✉e-mail: g.kaissis@tum.de

Table 1 | Overview of attacks against CML systems

	Adversarial goal	Minimal model access	Adversarial presence	References
Privacy-centred attacks				
Model inversion	Reconstruction of training data	Black box	Out of the network	12,43,45–48
Membership inference	Inferring presence of an individual in the training set	Black box	Out of the network	15,17,30,32,34,35,40,52
Attribute inference	Inferring sensitive value of a record	Black box	Out of the network	13,16,39,59,84
Model extraction	Retrieval of the target model	Black box	Out of the network	50,51
Side-channel attack	Obtaining information about the training protocol	Black box	In the network	27,89,90
Utility-centred attacks				
Model poisoning	Degrading utility of the target model	White box	In the network	18,57,91,92
Back-door insertion	Running an auxiliary learning task	White box	In the network	23,24,55,93
Evasion attack	Misclassification of data at inference time	Black box	Out of the network	26,92

to remain concealed while executing such attacks. Moreover, a number of adversarial entry points exist that are specific to CML (for example, eavesdropping on model updates to perform model inversion attacks), and not well defined in single-actor ML, where the model/data are directly available. Thus, although a subset of the attacks described below are applicable to any ML scenario, CML offers the broadest attack surface due to the flexibility conferred on the adversary. We outline the exact collaborative settings at risk and define the assumptions required for adversarial exploitation below.

Moreover, we distinguish between the following types of adversary based on their location in the system: we define an adversary contributing to the learning process as an in-the-network attacker and an adversary who only interacts with the final model as an out-of-the-network attacker. The two types have different capabilities in relation to the learning protocol: an attacker able to obtain complete access to the collaboratively trained models (including all model (hyper-)parameters) is said to possess white-box access, whereas an attacker has black-box knowledge when they only have access to the model in an inference setting and are limited to requesting predictions on new data supplied by themselves.

To classify and match the attacks with mitigation strategies, we first present an overview and a number of definitions, which we also summarize in Table 1. Furthermore, we provide a pictorial representation of the attacks in Fig. 1. We loosely classify the majority of the attacks into privacy-centred versus utility-centred attacks, motivated by the two main goals of CML: training a model with high utility while preserving the privacy of data owners. While this classification is arbitrary, it has been previously deployed in other studies[9]. However, different categories (for example, attacks focused on the data versus the model or stratified by attacker type) remain conceivable.

Privacy-centred attacks. We first describe privacy-centred attacks, attempting to obtain information leaked unintentionally by clients during training, for which—whenever applicable—we assume an honest-but-curious adversary who follows the learning protocol without deviation while attempting to extract as much information as possible, possibly by colluding with other adversaries[10,11]. One such attack is model inversion (MInv) (Fig. 1a), in which the adversary, by studying how specific inputs affect the model's behaviour, reverse-engineers the model to obtain data that were used for model training. First proposed by Fredrikson et al.[12], these attacks allow partial or full reconstruction of the training data. Next, we describe attribute-inference attacks (Fig. 1b), which target attributes of a specific client rather than their membership in a dataset. They can allow the adversary to determine sensitive values that belong to a data record[13] or specific features of an image[14]. Last, we discuss membership-inference attacks (MIA) (Fig. 1c), which allow the adversary to determine whether a particular individual's data are part of the training dataset[15–17].

Utility-centred attacks. A different approach is used by adversaries who attempt to change the outcome of the learning protocol or to subvert the model's utility (utility-centred attacks). We generally assume an active (that is, dishonest) adversary for these attacks, who may deviate from the learning protocol to influence the final model[10]. We first consider model poisoning (Fig. 1g). A first instance of model poisoning sees the adversary perturb either the data used for training the joint model or their corresponding label, deploying a technique known as data poisoning. The second subset of model-poisoning attacks targets the training process itself by manipulating the parameters of the model directly[18] or the objective function[19]. In both forms, the attacker can severely reduce the utility of the model or introduce an unintended collateral task[20,21]. Back-door attacks (Fig. 1f) employ collateral tasks to provoke unintended (adversarial) behaviour to a specific combination of input features at inference time. Such attacks can introduce unfair targeted model bias against specific individuals[22]. Of note, back-door attacks differ from adversarial samples, as they modify the model itself, whereas the latter rely on data perturbation during training or at inference time[23,24].

Under certain circumstances, the attacker also controls the physical environment the model is deployed in, allowing them to identify specific input combinations that make an already trained model behave in an unpredictable or unwarranted manner. Such attacks are termed evasion attacks (Fig. 1h) and can be executed by an adversary modifying the data that the model performs inference on, resulting in unintended behaviour[25,26]. Such attacks are particularly risky in contexts where an adversary has the ability to modify the environment, for example, the modification of road signs to target autonomous vehicles[26].

Miscellaneous attacks. In certain scenarios, the components (weights, architecture) of the trained network are considered trade secrets. Model-extraction attacks (Fig. 1d) can steal the model for profit or as a foundation for white-box inference attacks. Moreover, we consider attacks targeting the hardware of the CML system rather than, for example, the training procedure. Such side-channel attacks (Fig. 1e) utilize data not directly related to the learning task itself, such as CPU/GPU usage or network latency, to obtain

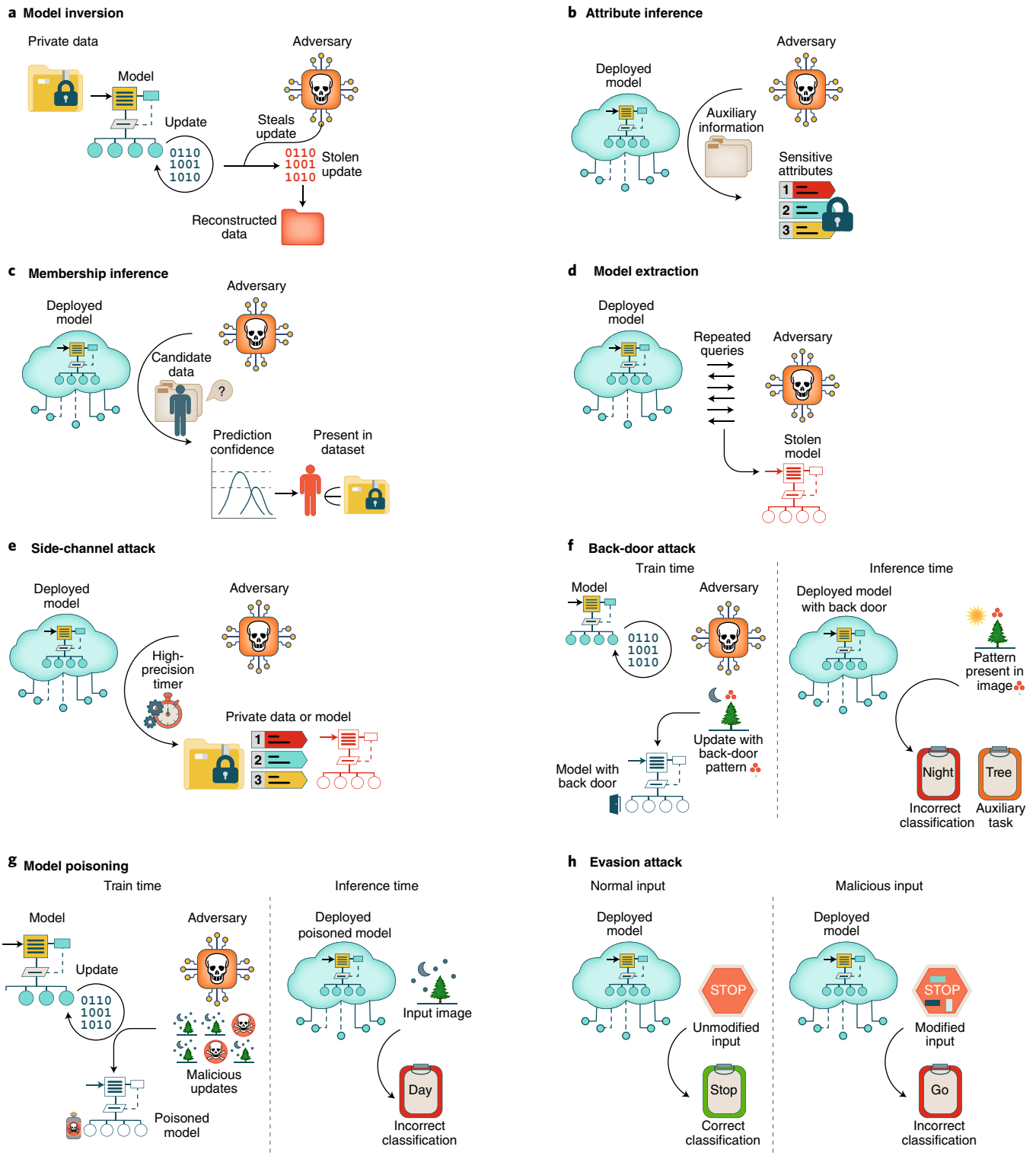

Fig. 1 | Overview of the attacks. a, Model inversion attack: the adversary steals a model update and reverse-engineers private data from it. **b**, Attribute inference attack: the adversary utilizes auxiliary information in conjunction with observing the model's output to reconstruct sensitive attributes of individuals in the dataset. **c**, Membership inference attack: the adversary probes the model with candidate data to infer based on model confidence whether an individual was used to train the model. **d**, Model extraction: the adversary aims repeated queries at the model with the aim of inferring the model structure or parameters. **e**, Side-channel attack: the adversary monitors the latency of a model's execution to infer sensitive data. **f**, Back-door attack: during training, the adversary covertly inserts a specific pattern in the training data. During inference time, the pattern triggers an incorrect model behaviour or allows execution of an auxiliary task. **g**, Model poisoning: during training, the adversary inputs maliciously crafted inputs to influence the model's training. This leads to incorrect model behaviour at inference time. **h**, Evasion attack: the adversary controls the model's deployment environment and manipulates it to cause incorrect model behaviour.

information about the learning procedure that later can be exploited for other attacks[27,28]. We discuss an auxiliary line of malicious interference (model replacement) in the Supplementary Information.

Attacks

Privacy-centred attacks. *Membership-inference attacks.* We begin by analysing MIA, targeting membership disclosure membership of a given data record. Most MIA implementations rely on the fact that predictive models produce different results when performing inference on previously seen data compared with data they have not interacted with. Thus, models produce higher-confidence predictions on data they have been trained on, which is exploited by the adversary[17,29]. If inference confidence is found to lie above a specific threshold, the model is assumed to have learnt how the data point changes the prediction outcome and it is concluded that it was used for training. This simple approach has been used by a number of studies as a baseline to obtain results much more accurate than random guessing[17,30]. Alternative implementations of MIA rely on prediction entropy[31] or observation of model outputs in reaction to input perturbations[32]. Such variants were shown to be more effective in contexts where simple empirical mitigations were deployed[33]. A number of works have proposed that model overfitting is an essential component to an MIA's success[17,34,35]. However, modern attack strategies theoretically demonstrate that overfitting is not a requirement for an attack's success[32,36] but merely allows the attacker to perform inference with higher confidence[37]. We present further information and references on the effects of overfitting and model generalization in the Supplementary Information.

Attribute-inference attacks. Attribute-inference (also called feature reconstruction) attacks aim to recover characteristics of the training data learnt by the model. Examples in computer vision include the reconstruction of attributes such as facial appearance or race[14] and, in genetics applications, of sequence information[38] used to infer specific diseases[39]. These attacks have a flexible threat model and can be executed both under white-box model access[13] and black-box model access[40]. White-box access typically yields a higher attack accuracy, being able to extract complex features from layers of the model that have learnt the most information about the training set[17,41,42]. In addition, white-box adversaries can benefit from access to unencrypted model updates, specifically gradient information, which can either be used as input into generative adversarial networks (GANs) to generate data with the same features[41], or into attacker-controlled classification models, which determine the presence of a specific sensitive feature in a record[40,42].

Model-inversion attacks. In addition to membership or attribute inference, training data can be reconstructed either wholly or partially via MInv. For attribute inference, we note that the attacker must possess auxiliary information about the victim (for example, personal information such as age or race) to exploit the relationship between these attributes and the sensitive features they are attempting to infer (for example, credit rating or disease diagnosis). In contrast, MInv concentrates on the relationship between the model input and model output, trying to reverse-engineer the internal representations produced by the model to reveal the training data and makes no such assumptions[12,43].

MInv has a flexible threat model and relies on diverse exploitation methods. For instance, GANs can be used for reconstruction[44,45] by utilizing model updates obtained during training. Ideally, GANs produce samples that are identical to the training data. However, in practice, GANs are typically only able to produce images that are similar to the training data, but otherwise share few features with it[45]. Beyond GANs, attackers have successfully utilized gradient updates, designing optimization algorithms that mimic the honest update by perturbing attacker-controlled data samples[46–48]. These attacks pose a particular threat in CML, as they only require access to an unencrypted model update to reconstruct sensitive training data in full[49].

Model-extraction attacks. We conclude this section by mentioning model extraction, assuming that the model itself is a secret that must not be disclosed to any participant. Such attacks attempt to steal (hyper-)parameters of the joint model, which include weights, architecture and so on[50,51]. Before the model is stolen, these attacks operate under very tight threat models with black-box access, in which adversaries can only make predictions on the parameters of the model by querying it with data that they control and observing the changes in the prediction vector. However, once the model has been successfully extracted, all above-mentioned attacks that require white-box access can be performed[13,52].

Utility-centred attacks. In addition to preserving privacy, CML must be robust to adversarial influence aiming at undermining the utility of the resulting model. The healthcare domain, for instance, has seen a number of case studies in which malicious inputs can endanger patient safety, such as the miscalculation of a recommended dose of anticoagulant medication[12,53] leading to overdosing, or the misprediction of early stage breast cancer[18] and thyroid disease[54].

Model-poisoning attacks. A large number of utility-centred attacks are executed during model training. The adversaries normally belong to the training consortium and have full control over the data that they contribute to the protocol, as well as the corresponding labels. We first consider the case of input data perturbations, starting with the random perturbation of image pixels. Random perturbations are neither optimized nor efficient in terms of the number of pixels they have to alter or the computational power necessary to guarantee a successful attack[55]. Yet ref. [56] demonstrated that even the random perturbation of a single pixel can be enough to achieve high rates of misclassification by the target model. However, in cases in which models are complex, or when images have high resolution, more care is required when crafting such adversarial samples. Successful approaches address both the issue of computational complexity and the guarantees of attack success. They rely on a number of image-generation strategies such as the fast gradient sign method[57] or projected gradient descent[58]. A technique relying on generation of both adversarial images and labels is called bilateral adversarial attack[59]. It is able to circumvent all defence mechanisms discussed below with the exception of multi-step adversarial training[58], highlighting the fact that generic defences against adversarial influence represent an important open problem in CML.

Back-door attacks. In contrast to classical adversarial data generation for utility destruction, back-door insertion attacks undermine utility by introducing malicious collateral tasks to model training, such as the targeted misclassification of specific victims' data[23]. In certain cases, they allow the adversary to exploit features that already exist in the data as triggers for this malicious behaviour. Thus, even without adversarial data perturbation, the attack is effective against victims with data containing these respective triggers[24]. Back-door insertion is facilitated by an increase in model complexity, as overparametrized models are able to learn the triggering features even if the corresponding labels during training are noisy[55,60]. Attempts to mitigate such attacks by reducing model complexity usually lead to an undesirable decrease in model utility.

Evasion attacks. Of note, the same techniques for performing adversarial sample attacks during training are also valid for inference time attacks, such as the evasion attacks described above. These attacks are, as a complication, highly dependent on the ability of an adversary to control the environment the model is deployed in. One of the most commonly discussed literature examples is the

Table 2 | Overview of attack mitigations

	Overview	Advantages	Disadvantages	References
Privacy-centred defences				
Homomorphic encryption	Training is performed on encrypted data, only decrypting the final output	- Formal guarantees	- Computational overhead	94–96
		- Secure by design	- Function approximations reduce model utility	
		- Several schemes available	- Susceptible to utility and membership attacks	
Secure multi-party computation	Compute shared function without revealing inputs to other clients	- Formal guarantees	- Communication overhead	97–99
		- Secure by design	- Susceptible to utility and membership attacks	
		- Many implementations available		
Trusted execution environments	Run (part of) training on secure enclave	- Formal guarantees	- Additional hardware requirements	100–102
		- Secure by design	- GPU training nascent	
			- Susceptible to side-channel attacks	
Knowledge distillation	Transfer of knowledge from a public model to the private one	- Prevents overfitting	- Requires publicly available dataset	62,64,65
		- Scalable	- No formal guarantees	
		- No computation overhead		
Split learning	Model is trained locally up to a cut layer, the rest is trained on the other host	- Scalable	- Susceptible to reconstruction attacks	7,73
		- Reduced communication overhead compared with FL	- Susceptible to utility and membership attacks	
Utility-centred defences				
Data analysis	Analyse data from other clients and perform pre-processing if required	- Empirically effective	- Violates privacy	75–77
			- Difficult to execute under data protection regulations	
Update analysis	Analyse updates from other clients and determine whether they should be aggregated	- Flexible metrics for updates	- Violates privacy	57,74,76
		- Empirically effective	- Not effective against back-door attacks	
Robust aggregation	Replace update averaging with an aggregation based on utility and/or data analysis	- Allows more efficient training	- No formal guarantees	70,91,103
		- Empirically effective	- Can reduce model utility	
			- Some updates are discarded thus wasting computation resources	
			- Susceptible to privacy attacks	
Adversarial training	Train the model on adversarially crafted samples in addition to regular ones	- Empirically effective	- No formal guarantees	58,75,104,105
		- Easy to implement	- Has a small impact on model utility	
		- No computation overhead	- Susceptible to privacy attacks	

Continued

Table 2 | Overview of attack mitigations (continued)

	Overview	Advantages	Disadvantages	References
Shared defences				
Regularization	Modification of model training aimed at increasing model generalization	- Large number of implementations	- No formal guarantees	33,72
		- Prevents overfitting	- Very limited impact on most attacks	
		- Easy to include in training		
		- Improves the performance of the model		
Model pruning	Discard specific neurons/units of the model based on a pre-defined strategy	- Large number of implementations	- No formal guarantees	68
		- Prevents overfitting	- Can reduce model utility	
		- Easy to include in training		
Differential privacy	Targeted perturbation of certain stages of the protocol to make the algorithm approximately invariant to addition/exclusion of data points	- Formal guarantees	- Reduced model utility	105–109
		- Implicit regularization	- Increased training time	
		- Improved robustness	- Disputed effectiveness against utility attacks	
		- Scalable to many parties		

modification of traffic signs, resulting in unexpected behaviour from self-driving cars[26].

So far we have assumed a single adversary; however, extended cases exist, in which more than one (potentially colluding and/or otherwise coordinating) adversary is present in a system, termed Byzantine attacks, which we describe in the Supplementary Information.

Attack mitigations

Privacy-centred defences. Advances in the field of privacy-preserving ML in recent years have yielded theoretical frameworks and software libraries both for centralized and decentralized approaches in which participants do not share their raw data directly. For an overview of the techniques of differential privacy (DP), secure multi-party computation (SMPC) and homomorphic encryption (HE), we refer to our previous work[4] and have provided a brief introduction to key terms in the Supplementary Information.

Such techniques often provide participants with provable privacy guarantees, often at the expense of model utility. Alternative mitigation strategies have been introduced that impact utility less, but only provide empirical privacy guarantees. We discuss these techniques in this section.

A number of defences have been proposed in literature that rely on the principles of security through obscurity. Such methods include the addition of noise to the confidence vector at inference stage or the return of discrete labels instead of full prediction vectors. We note that these methods were proven to be ineffective against any privacy-oriented attack and hence we do not consider them here[15,32,33,61]. On the other hand, security techniques such as inference query analysis, which prohibit access to the model based on some specific heuristic, have proven effective against model-extraction attacks[50].

Distillation. As mentioned above, overfitting is often regarded as a contributor to an attack's confidence, and several methods attempt to reduce it in CML. Distillation allows the knowledge of an ensemble of models to be transferred into a single, typically simpler, model[62]. While originally this method was designed to reduce overfitting and improve performance in the context of CML, it is can also defend against model poisoning[56,63].

Newer frameworks allow combining distillation with differential privacy, such as private aggregation of teacher ensembles (PATE)[64], which utilizes a publicly available dataset to transfer knowledge from models trained locally to a central model under a DP mechanism. This setting has found usage in medical imaging for its privacy guarantees and scalability[65–67], but suffers when data are unevenly distributed or there is a lack of publicly available data for the task.

Model pruning. Pruning entails the removal of individual units (neurons) from the network according to a predefined strategy[68–71]. By their removal, their exploitability is limited, as disabled units can no longer be targeted by adversaries during, for example, a model-poisoning attempt[68]. Pruning is a hybrid defence strategy that can be used to mitigate both privacy-centred and utility-centred attacks. However, newer work shows that sophisticated attack mechanisms developed on the basis of shadow training[17] can circumvent this defence in many scenarios[32,34,69]. Furthermore, model introspection can be problematic in privacy-critical scenarios, as it can itself reveal sensitive features of the training data.

Regularization. In addition, methods associated with regularization as a technique for overfitting prevention, such as L2-regularization, label smoothing, early stopping or MixUp[72], have been proposed as measures to increase attack resilience[17,35]. By reducing overfitting, these techniques often simultaneously improve generalization performance. This also makes them suitable for empirical mitigations of confidence-based attacks as well as train-time utility-oriented attacks. However, many of these techniques offer weak privacy enhancing attributes, and several have been circumvented[32,33,69].

Split learning. Split learning, described above, has been proposed as superior to FL[7] as it allows the consortium to reduce communication overhead, thereby providing an empirical reduction in the information shared between participants and resolving the issue of unintended information disclosure associated with FL. However, this method is no more effective than the standard FL approach

(and in fact may be more vulnerable, due to directly sharing intermediate activations), and does not prevent attacks based on generative reconstruction[43] from succeeding due to label leakage[73].

Utility-centred defences. *Data and update analysis.* Strategies that attempt to minimize the adversarial affect on the utility of the final model are based on two principles: detection and sanitation. Detection aims to determine which updates can be considered malicious. This can be achieved through comparison of model utility between updates[57,74] or by analysis of the projected distribution of data (or the data itself) for each participant[75–78]. Such techniques, although simple and effective, come at the cost of contradicting the very privacy-preserving properties that the system was designed for, and are thus complicated to reconcile with settings in which the requirements for privacy protection outweigh other CML protocol goals. Moreover, their effectiveness against, for example, model-poisoning attacks is reduced by the utilization of cryptographic techniques, as encrypted models are unable to be introspected and analysed.

Robust aggregation. Robust aggregation techniques are designed to limit the adversarial impact on the final model. One such technique is Krum[79], which selects only the update that results in the highest utility of the model, discarding the rest. Such selection could, however, result in adversarial attacks that manage to produce sufficiently good results to be selected each round and remain concealed[80]. Moreover, colluding adversaries (that is Byzantine attacks, see Supplementary Information), also reduce the effectiveness of this aggregation strategy[18]. An improved version of this algorithm (Bulyan)[81] combines two aggregation methods: trimmed mean[74] and Krum. However, these methods were only empirically evaluated and cannot provide any formal robustness guarantees, which was exploited by refs. [18,55,80], resulting in severe model-utility degradation. This has been addressed recently by the introduction of robust aggregation techniques providing certifiable defences[82].

Adversarial training. Finally, we describe an alternative use of adversarial samples, designed to allow clients to define the upper bound on a potential deviation from the projected final model through their use 'for good'. This approach, called adversarial training, involves consciously including adversarial samples in model training to render the model more robust to actual adversaries, mainly by reducing overfitting. While such models are empirically resilient against model poisoning and most advanced adversarial attacks so far[20,83], adversarial training cannot certifiably guarantee to be effective against any arbitrary adversarial attack[84].

An overview of the mitigation strategies described above can be found in Table 2.

Table 3 | Overview of proposed defences based on the attack type

	Proposed defenses	References
Privacy-centred attacks		
Model inversion	DP, HE, SMPC, TEE, knowledge distillation, model pruning	43,45,47,48
Membership inference	DP, knowledge distillation, model pruning, regularization	15,17,30,32,34, 35,40,52
Attribute inference	DP, model pruning, regularization	13,16,39,59,84
Model extraction	DP, inference query analysis, model pruning	50,51
Side-channel attack	DP, TEE	27,89,90,102,109
Utility-centred attacks		
Model poisoning	DP (disputed), regularization, data analysis, update analysis, robust aggregation, adversarial training	18,57,92,105
Back-door insertion	DP (disputed), data analysis, regularisation, robust aggregation, model pruning	23,24,55,93
Evasion attack	Data analysis, adversarial training	26,92

TEE, trusted execution environments.

Open challenges and future directions

In this Perspective, we surveyed the body of current literature on attacks against collaborative ML systems. We provide a concise summary of these attacks as a reference for practitioners in Table 3. We conclude by describing open challenges and promising future research directions.

Multiple attack mitigation. We identify only a small number of works on the topic of mitigation of multiple attack types simultaneously[69,85]. We believe this to be due to the conflicting nature of mitigation strategies: a key component of utility-centred mitigation is the analysis of data, model updates or the model itself as described above, which risks disclosing sensitive information and violating the privacy properties of the CML protocol. Correspondingly, mitigations against privacy leakage impede utility protection. This tension field can undermine the trustworthiness of the overall system: if the federation cannot be guaranteed confidentiality of their data and reliability of the joint model simultaneously, clients are unlikely to participate. We thus believe improved multiple attack-mitigation techniques not relying on privacy- or utility-critical methods to represent a promising direction for future work.

Beyond information hiding. In terms of attack targets, many published works assume that clients are facing an in-the-network adversary and hence most advanced defence mechanisms are aimed at reducing the disclosure of information between the client and the rest of the federation, including the central server. Counter-productively, this assumption directly serves adversaries wishing to remain concealed during their attacks who can benefit from such privacy-preserving mechanisms. An example of this is encrypted update aggregation, which can conceal poisoned model updates. Consequently, we see an open research frontier regarding mechanisms that rely less on information hiding and more on the preservation of appropriate information flow (contextual integrity[86]) while ascertaining and verifying correctness, privacy and utility (structured transparency[87]), thus precluding or reducing the overall effectiveness of any attack. We have summarized individual techniques above and in our previous work[4], and anticipate future work to synthesize them and evaluate their theoretical guarantees and ramifications.

Privacy–utility trade-offs. Moreover, we note that in current published literature, certain mitigation strategies are observed more frequently than others, particularly those that prioritize defences against utility-centred attacks based on update analysis[53,54,88]. We assume the reason for this to be a strong incentive to focus on accuracy, robustness and utility of the trained models. Such mitigations

may thus protect well against utility-centred attacks to the detriment of privacy guarantees. Beyond the necessary societal, political, moral and ethical discussion on the consequences of this trade-off, we also view the development of high-utility privacy-preserving protocols and privacy-centred defences as a crucial research direction.

Generic susceptibility and defences. Studies have evaluated the role of individual CML components in the context of generic attacks[21,63,69]. They conclude that overfitting contributes to both privacy-centred and utility-centred attacks, show that DP and HE provide rigorous privacy/secrecy guarantees at the cost of reduced utility, identify white-box model access as the most intrusive yet most common scenario in CML, and deduce that the statistical data distribution substantially impacts the fidelity of utility-centred attacks. However, while the general setup of current CML settings is similar, certain attacks and adaptations are only possible in certain learning contexts. Considering the rapid evolution of CML protocols and the concurrent development of new attacks, we thus find that further investigation is required to identify defence mechanisms that are context agnostic and generally applicable to a variety of learning scenarios.

Summary and outlook. We here present an overview of attacks on CML and evaluate them under the previously proposed mitigation strategies alongside their limitations. For research areas requiring large quantities of data, such as ML, to reach their full potential, secure and reliable protocols and systems have to be designed. This is not possible without a detailed analysis of common exploitation vectors that could undermine their trustworthiness. Our work is designed to bridge the gap between ML practitioners and security researchers by providing insights on the points at which these protocols may fail. We believe the ability to provide rigorous privacy guarantees on high-utility models that are robust to adversarial influence to be the crucial foundation for the development of trustworthy collaborative learning protocols offering equitable outcomes to all parties involved.

Received: 7 January 2021; Accepted: 11 August 2021;
Published online: 17 September 2021

References

1. Radley-Gardner, O., Beale, H. & Zimmermann, R. (eds) *Fundamental Texts On European Private Law* (Hart Publishing, 2016); http://www.bloomsburycollections.com/book/fundamental-texts-on-european-private-law-1
2. *Health Insurance Portability and Accountability Act* (CDC, 2020).
3. Drainakis, G., Katsaros, K. V., Pantazopoulos, P., Sourlas, V. & Amditis, A. Federated vs. centralized machine learning under privacy-elastic users: a comparative analysis. In *2020 IEEE 19th International Symposium on Network Computing and Applications (NCA)* 1–8 (IEEE, 2020); https://doi.org/10.1109/nca51143.2020.9306745
4. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
5. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & Agüera y Arcas, B. In *Proc. 20th International Conference on Artificial Intelligence and Statistics* Vol. 54 (eds Sing, A. & Zhu, J.) 1273–1282 (PMLR, 2017)
6. Warnat-Herresthal, S. et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).
7. Vepakomma, P., Gupta, O., Swedish, T. & Raskar, R. Split learning for health: distributed deep learning without sharing raw patient data. Preprint at https://arxiv.org/abs/1812.00564 (2018).
8. Brundage, M. et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. Preprint at https://arxiv.org/abs/2004.07213 (2020).
9. Jere, M. S., Farnan, T. & Koushanfar, F. A taxonomy of attacks on federated learning. *IEEE Secur. Priv.* **19**, 20–28 (2021).
10. Evans, D., Kolesnikov, V. & Rosulek, M. A pragmatic introduction to secure multi-party computation. *Found. Trends Priv. Secur.* **2**, 70–246 (2018).
11. Riazi, M. S. & Koushanfar, F. Privacy-preserving deep learning and inference. In *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* 1–4 (IEEE, 2018).
12. Fredrikson, M. et al. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In *Proc. 23rd USENIX Security Symposium 14* 17–32 (USENIX, 2014).
13. Ganju, K., Wang, Q., Yang, W., Gunter, C. A. & Borisov, N. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proc. 2018 ACM SIGSAC Conference on Computer and Communications Security* 619-633 (ACM, 2018); https://doi.org/10.1145/3243734.3243834
14. Mansourifar, H. & Shi, W. Vulnerability of face recognition systems against composite face reconstruction attack. Preprint at http://arxiv.org/abs/2009.02286 (2020).
15. Long, Y., Bindschaedler, V. & Gunter, C. A. Towards measuring membership privacy. Preprint at http://arxiv.org/abs/1712.09136 (2017).
16. He, Y., Rahimian, S., Schiele, B. & Fritz, M. In *Computer Vision – ECCV 2020: Lecture Notes in Computer Science* Vol. 12368 (eds Vedaldi, A. et al.) 519–535 (Springer, 2020).
17. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* 3–18 (IEEE, 2017).
18. Fang, M., Cao, X., Jia, J. & Gong, N. Local model poisoning attacks to Byzantine-robust federated learning. In *Proc. 29th USENIX Security Symposium 20* 1605–1622 (USENIX, 2020).
19. Bhagoji, A. N., Chakraborty, S., Mittal, P. & Calo, S. In *International Conference on Machine Learning* 634–643 (PMLR, 2019).
20. Hayes, J. & Ohrimenko, O. *Contamination Attacks and Mitigation in Multi-Party Machine Learning* (NeurIPS, 2018).
21. Chang, H., Shejwalkar, V., Shokri, R. & Houmansadr, A. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer. Preprint at http://arxiv.org/abs/1912.11279 (2019).
22. Wenger, E., Passananti, J., Yao, Y., Zheng, H. & Zhao, B. Y. *Backdoor Attacks on Facial Recognition in the Physical World* (CVPR, 2021).
23. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D. & Shmatikov, V. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics* 2938–2948 (PMLR, 2020).
24. Bagdasaryan, E. & Shmatikov, V. *Blind Backdoors in Deep Learning Models* (USENIX Security, 2021).
25. Biggio, B. et al. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 387–402 (Springer, 2013).
26. Chernikova, A., Oprea, A., Nita-Rotaru, C. & Kim, B. Are self-driving cars secure? Evasion attacks against deep neural networks for steering angle prediction. In *2019 IEEE Security and Privacy Workshops (SPW)* 132–137 (IEEE, 2019).
27. Yan, M., Fletcher, C. W. & Torrellas, J. Cache telepathy: leveraging shared resource attacks to learn DNN architectures. In *Proc. 29th USENIX Security Symposium 20* 2003–2020 (USENIX, 2020).
28. Timon, B. Non-profiled deep learning-based side-channel attacks with sensitivity analysis. *IACR Trans. Cryptogr. Hardw. Embed. Syst.* **2019**, 107–131 (2019).
29. Leino, K. & Fredrikson, M. Stolen memories: leveraging model memorization for calibrated white-box membership inference. In *Proc. 29th USENIX Security Symposium 20* 1605–1622 (USENIX, 2020).
30. Rahman, M. A., Rahman, T., Laganière, R., Mohammed, N. & Wang, Y. Membership inference attack against differentially private deep learning model. *Trans. Data Priv.* **11**, 61–79 (2018).
31. Song, L. & Mittal, P. Systematic evaluation of privacy risks of machine learning models. In *Proc. 30th USENIX Security Symposium 21* 2615–2632 (USENIX, 2021).
32. Choo, C. A. C., Tramer, F., Carlini, N. & Papernot, N. In *International Conference on Machine Learning* 1964–1974 (PMLR, 2021).
33. Kaya, Y., Hong, S. & Dumitras, T. On the effectiveness of regularization against membership inference attacks. Preprint at https://arxiv.org/abs/2006.05336 (2020).
34. Park, Y. & Kang, M. Membership inference attacks against object detection models. Preprint at http://arxiv.org/abs/2001.04011 (2020).
35. Salem, A. et al. *Ml-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models* (NDSS, 2019).
36. Long, Y. et al. Understanding membership inferences on well-generalized learning models. Preprint at https://arxiv.org/abs/1802.04889 (2018).
37. Hayes, J., Melis, L., Danezis, G. & De Cristofaro, E. Logan: membership inference attacks against generative models. *Proc. Priv. Enhanc. Technol.* **2019**, 133–152 (2019).
38. Samani, S. S. et al. Quantifying genomic privacy via inference attack with high-order SNV correlations. In *2015 IEEE Security and Privacy Workshops* 32–40 (IEEE, 2015); https://ieeexplore.ieee.org/document/7163206/
39. Wu, M. et al. Evaluation of inference attack models for deep learning on medical data. Preprint at http://arxiv.org/abs/2011.00177 (2020).

40. Nasr, M., Shokri, R. & Houmansadr, A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)* 739–753 (IEEE, 2019).
41. Luo, X. & Zhu, X. Exploiting defenses against GAN-based feature inference attacks in federated learning. Preprint at https://arxiv.org/abs/2004.12571 (2020).
42. Melis, L., Song, C., Cristofaro, E. D. & Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)* 691–706 (IEEE, 2019).
43. He, Z., Zhang, T. & Lee, R. B. Model inversion attacks against collaborative inference. In *Proc. 35th Annual Computer Security Applications Conference* 148–162 (ACM, 2019).
44. Hitaj, B., Ateniese, G. & Perez-Cruz, F. Deep models under the GAN: information leakage from collaborative deep learning. In *Proc. 2017 ACM SIGSAC Conference on Computer and Communications Security* 603–618 (ACM, 2017).
45. Zhang, Y. et al. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 253–261 (CVPR, 2020).
46. Zhao, B., Mopuri, K. R. & Bilen, H. iDLG: Improved Deep Leakage from Gradients. Preprint at https://arxiv.org/abs/2001.02610 (2020).
47. Zhu, L., Liu, Z. & Han, S. Deep leakage from gradients. *Adv. Neural Inf. Process. Syst.* **32**, 14747–14756 (2019).
48. Geiping, J., Bauermeister, H., Dröge, H. & Moeller, M. Inverting gradients—how easy is it to break privacy in federated learning? Preprint at https://arxiv.org/abs/2003.14053 (2020).
49. Kaissis, G. et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-021-00337-8 (2021).
50. Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A. & Papernot, N. High accuracy and high fidelity extraction of neural networks. In *Proc. 29th USENIX Security Symposium 20* (USENIX, 2020).
51. Oh, S. J., Schiele, B. & Fritz, M. in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds Samek, W. et al.) 121–144 (Springer, 2019).
52. Chen, D., Yu, N., Zhang, Y. & Fritz, M. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proc. 2020 ACM SIGSAC Conference on Computer and Communications Security* 343–362 (ACM, 2020).
53. Jagielski, M. et al. Manipulating machine learning: poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy (SP)* 19–35 (IEEE, 2018).
54. Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A. & Jha, N. K. Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE J. Biomed. Health Informatics* **19**, 1893–1905 (2014).
55. Wang, H. et al. Attack of the tails: yes, you really can backdoor federated learning. *Adv. Neural Inf. Process. Syst.* **33**, 1–15 (2020).
56. Narodytska, N. & Kasiviswanathan, S. P. *Simple Black-Box Adversarial Attacks on Deep Neural Networks* Vol. 2 (CVPR Workshops, 2017).
57. Goodfellow, I. J., Shlens, J. & Szegedy, C. *Explaining and Harnessing Adversarial Examples* (ICLR, 2014).
58. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. *Towards Deep Learning Models Resistant to Adversarial Attacks* (ICLR, 2018).
59. Wang, J. & Zhang, H. Bilateral adversarial training: towards fast training of more robust models against adversarial attacks. In *Proc. IEEE International Conference on Computer Vision* 6629–6638 (IEEE, 2019).
60. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Comms ACM* **64**, 107–115 (2016).
61. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. *Membership Inference Attacks Against Machine Learning Models* (IEEE, 2017).
62. Hinton, G., Vinyals, O. & Dean, J. *Distilling the Knowledge in a Neural Network* (NIPS, 2014).
63. Papernot, N., McDaniel, P., Sinha, A. & Wellman, M. Towards the science of security and privacy in machine learning. Preprint at https://arxiv.org/abs/1611.03814 (2016).
64. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I. & Talwar, K. *Semi-Supervised Knowledge Transfer for Deep Learning from Private Training Data* (ICLR, 2017).
65. Papernot, N. et al. *Scalable Private Learning with Pate* (ICLR, 2018).
66. Fay, D., Sjölund, J. & Oechtering, T. J. Decentralized differentially private segmentation with pate. Preprint at https://arxiv.org/abs/2004.06567 (2020).
67. Müftüoğlu, Z., Kizrak, M. A. & Yildlnm, T. Differential privacy practice on diagnosis of COVID-19 radiology imaging using EfficientNet. In *2020 International Conference on INnovations in Intelligent Systems and Applications (INISTA)* 1–6 (IEEE, 2020).
68. Dhillon, G. S. et al. Stochastic activation pruning for robust adversarial defense. Preprint at https://arxiv.org/abs/1803.01442 (2018).
69. Song, L., Shokri, R. & Mittal, P. Membership inference attacks against adversarially robust deep learning models. In *2019 IEEE Security and Privacy Workshops (SPW)* 50-56 (IEEE, 2019).
70. Xie, C., Koyejo, S. & Gupta, I. In *International Conference on Machine Learning* 10495–10503 (PMLR, 2020).
71. Bau, D. et al. Understanding the role of individual units in a deep neural network. *Proc. Natl Acad. Sci. USA* **117**, 30071–30078 (2020).
72. Fu, Y., Wang, H., Xu, K., Mi, H. & Wang, Y. Mixup based privacy preserving mixed collaboration learning. In *2019 IEEE International Conference on Service-Oriented System Engineering (SOSE)* 275–2755 (IEEE, 2019).
73. Vepakomma, P., Tonde, C. & Elgammal, A. et al. Supervised dimensionality reduction via distance correlation maximization. *Electron. J. Stat.* **12**, 960–984 (2018).
74. Yin, D., Chen, Y., Ramchandran, K. & Bartlett, P. In *International Conference on Machine Learning* 5650–5659 (PMLR, 2018).
75. Steinhardt, J., Koh, P. W. W. & Liang, P. S. Certified defenses for data poisoning attacks. *Adv. Neural Inf. Process. Syst.* **31**, 3517–3529 (2017).
76. Lee, K., Lee, K., Lee, H. & Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Adv. Neural Inf. Process. Syst.* **31**, 7167–7177 (2018).
77. Metzen, J. H., Genewein, T., Fischer, V. & Bischoff, B. *On Detecting Adversarial Perturbations* (ICLR, 2017).
78. Meng, D. & Chen, H. Magnet: a two-pronged defense against adversarial examples. In *Proc. 2017 ACM SIGSAC Conference on Computer and Communications Security* 135–147 (ACM, 2017).
79. Blanchard, P., Guerraoui, R. & Stainer, J. et al. Machine learning with adversaries: Byzantine tolerant gradient descent. *Adv. Neural Inf. Process. Syst.* **31**, 119–129 (2017).
80. Baruch, G., Baruch, M. & Goldberg, Y. A little is enough: circumventing defenses for distributed learning. *Adv. Neural Inf. Process. Syst.* **32**, 8635–8645 (2019).
81. Mhamdi, E. M. E., Guerraoui, R. & Rouault, S. In *International Conference on Machine Learning* 3521–3530 (PMLR, 2018).
82. Levine, A. & Feizi, S. *(De)randomized Smoothing for Certifiable Defense Against Patch Attacks* (NeurIPS, 2020).
83. Gilmer, J. et al. In *International Conference on Machine Learning* 2280–2289 (PMLR, 2019).
84. Pinot, R., Ettedgui, R., Rizk, G., Chevaleyre, Y. & Atif, J. In *International Conference on Machine Learning* 7717–7727 (PMLR, 2020).
85. Mejia, F. A. et al. Robust or private? adversarial training makes models more vulnerable to privacy attacks. Preprint at https://arxiv.org/abs/1906.06449 (2019).
86. Nissenbaum, H. *Privacy in Context: Technology, Policy, and the Integrity of Social Life* (Stanford Univ. Press, 2009).
87. Trask, A., Bluemke, E., Garfinkel, B., Cuervas-Mons, C. G. & Dafoe, A. Beyond privacy trade-offs with structured transparency. Preprint at https://arxiv.org/abs/2012.08347 (2020).
88. Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N. & Wachinger, C. Braintorrent: a peer-to-peer environment for decentralized federated learning. Preprint at https://arxiv.org/abs/1905.06731 (2019).
89. Wang, J., Cheng, Y., Li, Q. & Jiang, Y. Interface-based side channel attack against intel SGX. Preprint at https://arxiv.org/abs/1811.05378 (2018).
90. Liu, F., Yarom, Y., Ge, Q., Heiser, G. & Lee, R. B. Last-level cache side-channel attacks are practical. In *2015 IEEE Symposium on Security and Privacy* 605–622 (IEEE, 2015).
91. Muñoz-González, L., Co, K. T. & Lupu, E. C. Byzantine-robust federated machine learning through adaptive model averaging. Preprint at https://arxiv.org/abs/1909.05125 (2019).
92. Suciu, O., Marginean, R., Kaya, Y., Daume, H. III & Dumitras, T. When does machine learning fail? Generalized transferability for evasion and poisoning attacks. In *Proc. 27th USENIX Security Symposium 18* 1299–1316 (*USENIX*, 2018).
93. Chen, X., Liu, C., Li, B., Lu, K. & Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. Preprint at https://arxiv.org/abs/1712.05526 (2017).
94. Hesamifard, E., Takabi, H. & Ghasemi, M. Cryptodl: deep neural networks over encrypted data. Preprint at https://arxiv.org/abs/1711.05189 (2017).
95. Gilad-Bachrach, R. et al. Cryptonets: applying neural networks to encrypted data with high throughput and accuracy. In *International Conference on Machine Learning* 201–210 (PMLR, 2016).
96. Mohassel, P. & Zhang, Y. SecureML: a system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)* 19–38 (IEEE, 2017).
97. Juvekar, C., Vaikuntanathan, V. & Chandrakasan, A. GAZELLE: a low latency framework for secure neural network inference. In *Proc. 27th USENIX Security Symposium 18* 1651–1669 (USENIX, 2018).
98. Goldreich, O., Micali, S. & Wigderson, A. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali* 307–328 (ACM Books, 2019).
99. Rouhani, B. D., Riazi, M. S. & Koushanfar, F. Deepsecure: scalable provably-secure deep learning. In *Proc. 55th Annual Design Automation Conference* 2 (ACM, 2018).

100. Costan, V. & Devadas, S. Intel SGX Explained. *IACR Cryptol. ePrint Archive* **2016**, 1–118 (2016).
101. Ohrimenko, O. et al. Oblivious multi-party machine learning on trusted processors. In *Proc. 25th USENIX Security Symposium 16* 619–636 (USENIX, 2016).
102. Dessouky, G., Frassetto, T. & Sadeghi, A.-R. HybCache: hybrid side-channel-resilient caches for trusted execution environments. In *Proc. 29th USENIX Security Symposium 20* 451–468 (USENIX, 2020).
103. Sattler, F., Wiedemann, S., Müller, K.-R. & Samek, W. Robust and communication-efficient federated learning from non-iid data. *IEEE Trans. Neural Netw. Learn. Syst.* (2019).
104. Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y. & Usunier, N. In *International Conference on Machine Learning* 854–863 (PMLR, 2017).
105. Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D. & Jana, S. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)* 656–672 (IEEE, 2019).
106. Choudhury, O. et al. *Differential Privacy-Enabled Federated Learning for Sensitive Health Data* (NeurIPS, 2019).
107. Wu, B. et al. P3SGD: Patient privacy preserving SGD for regularizing deep CNNs in pathological image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2099–2108 (IEEE, 2019).
108. McMahan, H. B. et al. *A General Approach to Adding Differential Privacy to Iterative Training Procedures* (NeurIPS, 2018).
109. Xu, M., Papadimitriou, A., Feldman, A. & Haeberlen, A. Using differential privacy to efficiently mitigate side channels in distributed analytics. In *Proc. 11th European Workshop on Systems Security* 1–6 (ACM, 2018).

Acknowledgements

We thank the OpenMined community for its support. Funding: G.K. received funding from the Technical University of Munich, School of Medicine Clinician Scientist Programme (KKF), project reference H14. D.U. received funding from the Technical University of Munich/Imperial College London Joint Academy for Doctoral Studies. This research was supported by the UK Research and Innovation London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare. The funders played no role in the design of the study, the preparation of the manuscript or the decision to publish.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00390-3.

Correspondence should be addressed to Georgios Kaissis.

Peer review information *Nature Machine Intelligence* thanks Xiaoxiao Li, Tushar Semwal and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.