

Machine Unlearning: A Comprehensive Survey

WEIQI WANG, University of Technology Sydney, Australia

ZHIYI TIAN, University of Technology Sydney, Australia

CHENHAN ZHANG, University of Technology Sydney, Australia

SHUI YU, University of Technology Sydney, Australia

As the right to be forgotten has been legislated worldwide, numerous studies have sought to design unlearning methods to protect users' privacy when they want to remove their data from machine learning service platforms. This has given rise to the concept of "machine unlearning" — a field dedicated to removing the influence of specified samples from trained models. This survey aims to systematically classify a wide range of machine unlearning studies, discussing their differences, connections, and open problems. We categorize current unlearning methods into four key areas: centralized unlearning, federated unlearning, unlearning verification, and privacy and security issues in unlearning. Given that centralized unlearning is the primary focus, we further divide its discussion into two parts: first, at a higher level, we classify centralized unlearning into exact unlearning and approximate unlearning; second, we provide a detailed introduction to the techniques used in these methods. Next, we introduce federated unlearning, another emerging area that has garnered significant attention. Following the discussion on unlearning methods, we review studies on unlearning verification and audit, which assess the effectiveness of existing unlearning algorithms. Additionally, we consider the privacy and security issues essential in machine unlearning and compile the related literature. Finally, we discuss the challenges of various unlearning scenarios and highlight potential research directions.

CCS Concepts: • Computing methodologies → Machine learning; Artificial intelligence; • Security and privacy; • Theory of computation;

Additional Key Words and Phrases: Federated Unlearning; Machine Unlearning; Unlearning Verification

1 INTRODUCTION

Over the past decade, enormously increased data and fast hardware improvement have driven machine learning developments quickly. Machine learning (ML) algorithms and artificial intelligence (AI) are embedded into day-to-day applications and wearable devices [1]. It continuously collects increasing amounts of user information, including private data such as driving trajectories, medical records, and online shopping histories [2, 3]. On the one hand, such an enormous amount of data helps to further advance ML and AI development. On the other hand, however, it poses a threat to users' privacy and creates a significant need for robust data management to ensure information security and privacy in ML [4].

Machine unlearning has drawn growing research attention as the recent legislation of the "Right to be Forgotten" in many countries. Notable instances include the GDPR in the European Union [5], the PIPEDA privacy legislation in Canada [6], and the California Consumer Privacy Act in the United States [7]. According to these laws, companies must take reasonable measures to guarantee that personal data is deleted upon request.

Authors' addresses: Weiqi Wang, Weiqi.Wang-2@student.uts.edu.au, University of Technology Sydney, Australia; Zhiyi Tian, University of Technology Sydney, Australia, zhiyi.tian@student.uts.edu.au; Chenhan Zhang, University of Technology Sydney, Australia, chenhan.zhang@student.uts.edu.au; Shui Yu, University of Technology Sydney, Australia, shui.yu@uts.edu.au.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Association for Computing Machinery.

XXXX-XXXX/2024/7-ART \$15.00

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

It indicates that individual users have the right to request companies to remove their private data, which was previously collected for ML model training. The deletion is not only erasing their data from a database, but it also needs to delete the influence of the specified samples from trained models. The process of data removal from models was first conceptualized as machine unlearning [8]. Specifically, suppose a user (Alice) wants to exercise her right [5] when quitting a ML application, then the trained model of such application must "unlearn" her data. Such a process includes two steps: first, a subset of the dataset previously used for ML model training is requested to be deleted; second, the ML model provider erases the contribution of these data from the trained models. A naive data-erasing method is retraining a new ML model from scratch [9]. However, the computation and storage costs of retraining are expensive, especially in complex learning tasks.

Many researchers have tried to find efficient and effective methods to implement unlearning rather than naive retraining, and there are several common challenges, which are summarized as follows. (1) *Stochasticity of training*: A huge amount of randomness exists in the training process in machine learning, especially in complicated models' training periods such as CNNs [10] and DNNs [11]. This randomness makes the training results non-deterministic [9] and raises challenges for machine unlearning to estimate the impact of the typical erased samples. (2) *Incrementality of training*: The training process in machine learning is incremental, meaning that the model update from one data point influences the contribution of subsequent data points fed into the model. Deciding a way to effectively remove the contributions of the to-be-erased samples from the trained model is challenging for machine unlearning [12]. (3) *Catastrophe of unlearning*: Nguyen et al. [13] indicated that an unlearned model typically has worse model utility than the model retrained from scratch. The degradation would be exponential, especially when a method tries to delete a huge amount of data samples. They referred to such sharp degradation as catastrophic unlearning [13]. Although several studies mitigate model utility degradation by bounding the loss function or restricting the unlearning update threshold, eliminating catastrophic unlearning remains an open problem. Recently, many studies have put efforts into solving these three main challenges and proposed many novel mechanisms that promote the progress of machine unlearning.

This survey aims to classify and systematize machine unlearning methods based on the research problems and objectives in the unlearning process and to review their differences and connections, as well as their advantages and disadvantages. We first categorize unlearning approaches into four main areas: centralized unlearning, federated unlearning, unlearning verification, and privacy and security issues in machine unlearning, as shown in Fig. 1. Since many studies focus on centralized unlearning, we dedicate two sections to introducing corresponding studies in this area. We classify these unlearning methods into: *Exact unlearning* [8, 9] and *Approximate unlearning* [13, 44], where exact unlearning extends the idea of naive unlearning to design efficient retraining scheme with less computation cost. Approximate unlearning tries to unlearn a model approximate to the model retrained using the remaining dataset. We also provide a detailed introduction to the popular techniques used to implement these two types of unlearning. In addition to centralized unlearning methods, we introduce federated (distributed) unlearning studies [52, 56], which are gaining increasing importance. After discussing unlearning implementation methods, verifying the unlearning effect is equally crucial and has garnered significant research attention [67]. Therefore, we subsequently review related research on unlearning verification. Finally, we consider privacy and security issues in machine unlearning essential, organizing and reviewing the related publications on privacy and security threats, defenses, and unlearning applications.

There are only a few surveys on machine unlearning because it is a relatively new research domain, and most of them are preprints on Arxiv, focusing on certain unlearning scenarios. For an introduction to machine unlearning, including discussions on exact and approximate unlearning problems and their solutions through recently proposed methods, see [92]. For information on provable machine unlearning for linear models, including algorithm introductions and experimental analysis, refer to [93]. For an overview of federated unlearning, see [94, 95], and for graph unlearning, see [96]. While Nguyen et al. [97] summarized the general unlearning framework and added the unlearning verification part to it, they focused primarily on introducing problem

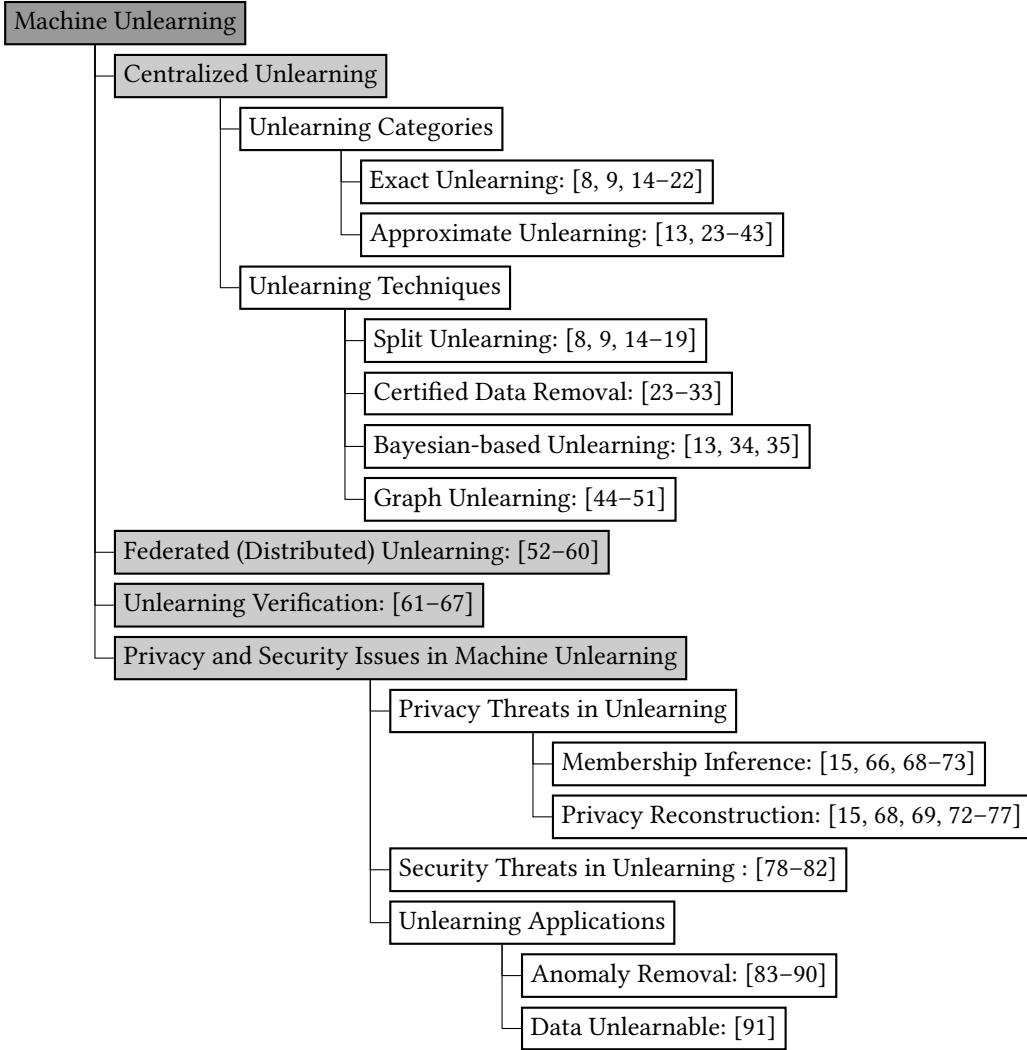


Fig. 1. Our taxonomy for machine unlearning. The introduction order will also follow this figure. We classify the current unlearning literature into four main scenarios: centralized unlearning, federated unlearning, unlearning verification, and privacy and security issues in machine unlearning.

formulations and technical definitions. However, none of these surveys reviewed the privacy and security issues in machine unlearning, nor did they comprehensively discuss the differences and connections between various unlearning scenarios.

Compared with existing surveys on machine unlearning, we performed a systematic literature review (SLR) following the guidelines presented by [98] and as exemplified in [99]. The contributions of this paper include:

- We systematically catalog machine unlearning studies according to their scenarios, which can be briefly classified into centralized unlearning, distributed unlearning, unlearning verification, and privacy and security issues in machine unlearning.

- For the centralized unlearning scenario, which draws the most attention, we divide related studies into exact and approximate unlearning and illustrate the connections, pros, and cons among these works.
- We further explore the privacy and security threats that target machine unlearning and discuss the applications of unlearning in defending against traditional security and privacy issues.
- We discuss different challenges between various machine unlearning areas, from centralized unlearning to federated unlearning, unlearning verification, and privacy and security issues in unlearning. We list the related open questions of each scenario and present potential research directions to solve them.

The survey's remaining sections are arranged as follows. We show a systematic literature review model of the survey in Section 2. The basic knowledge of machine unlearning and related technical tools are summarized in Section 3. After introducing the background, we introduce the survey following the taxonomy order in Fig. 1. The primary content of centralized unlearning is presented in Section 4, which includes an introduction from two main unlearning categories and corresponding in-depth techniques. We collect and introduce the federated unlearning scenarios in Section 5. Section 6 introduces the unlearning evaluation and verification methods. The privacy and security issues in machine unlearning are divided into two sections. Section 7 introduces the privacy and security threats accompanied by machine unlearning. Section 8 discusses machine unlearning applications, which are mainly applied in dealing with security issues. In Section 9, we discuss the challenges of current unlearning methods in a general way. After that, we enumerate the unaddressed issues for further research. At last, in Section 10, we provide a summary of the survey.

2 THE SLR METHOD OF THE SURVEY

We performed a systematic literature review (SLR) of the extant research on machine unlearning. The main goal was to compile a large body of research on machine unlearning and categorize them from various perspectives to facilitate analysis. Therefore, we employ the SLR method because it is especially appropriate for finding pertinent literature on a certain research topic [98]. The main process of the SLR method for the survey is presented as follows.

We first designed a search string according to the review protocol [98], identified appropriate digital databases, and defined the data extraction strategy. We focused on the keywords “unlearning” and “machine unlearning” in the search string, formulated as “*unlearning OR machine unlearning*”. We used this search string in IEEE Xplore, ACM Digital Library, Scopus, and the Web of Science to find relevant papers. Additionally, we conducted a search on Arxiv to identify further relevant literature. Consequently, a total of 972 papers were retrieved (the search was conducted on 7 July 2024). We then limited the publication years to those after 2020 and ensured that the keyword “unlearning” appeared in the title and abstract. Moreover, after filtering out duplicates and papers with fewer than six pages, 261 papers remained. Referring to the Google Scholar top publications and the China Computer Federation (CCF) recommendations lists, we focused on reviewing 103 papers from top venues. After including 33 references for related techniques, we reviewed a total of 136 references.

During the systematic literature review process, the data extraction strategy is defined as follows. Besides the basic information of papers (title, author, publication venue, year), we first categorize papers into the four main classes as Fig. 1 according to the research problems. Then, we extract the techniques utilized, the evaluating metrics, and the datasets employed in these papers, which will all be introduced in the following context.

3 BACKGROUND

3.1 Machine Unlearning

There are several urgent demands driving machine unlearning research. The foremost is the demand for privacy preservation. With the “right to be forgotten” being legislated globally [5, 6], machine unlearning ensures that users can request the erasure of their data from trained ML models. In addition to privacy concerns, other factors

Table 1. Notations in Machine Unlearning

Symbols	Description	Symbols	Description
\mathcal{Z}	data items space	\mathcal{H}	model parameters space
$D = (X, Y)$	the training dataset with inputs X and labels Y	$\mathcal{A}(\cdot)$	the ML algorithm
$D_e = (X_e, Y_e)$	the unlearned (erased) dataset	$\mathcal{U}(\cdot)$	the unlearning algorithm
$D_r = D \setminus D_e$	the remaining dataset	M	the model with parameters θ
D_{probe}	the probing data used in attacking	$M_{D \setminus D_e}$	the unlearned model

are promoting the development of machine unlearning. One significant factor is model utility. In the real world, vast amounts of data are generated daily, necessitating prompt updates to model services, as outdated data can negatively impact model performance [3, 100, 101]. An effective unlearning mechanism is crucial for mitigating the adverse effects of outdated and incorrect data on model utility. Another critical factor is security. Adversarial attacks and data poisoning [4] can easily compromise deep learning models. Therefore, detecting and removing adversarial and poisoned data is essential to ensure model security. Once an attack type is identified, the model must erase the influence of these adversarial data using the unlearning mechanism.

To better understand how unlearning mechanisms work, we first introduce the unlearning problem and process following the machine unlearning framework demonstrated in Fig. 2, and we summarize the notations in Table 1. Let \mathcal{Z} denote as a space of data items; the particular (full) training dataset is $D \in \mathcal{Z}$. A learning process can be demonstrated as Step 0 in Fig. 2, i.e., training a model M using an algorithm \mathcal{A} on the training dataset D , denoted as $M = \mathcal{A}(D)$, where model M is in a hypothesis space \mathcal{H} .

The unlearning process begins at an unlearning request when a user wants to erase his specified data D_e from the trained model, Step 1 in Fig. 2. The requested unlearning data D_e can be data samples [9], classes [33], or graph nodes [45]. Then, in Step 2, the server removes the contribution of D_e using a machine unlearning algorithm \mathcal{U} . The unlearned model can be described as $M_{D \setminus D_e} = \mathcal{U}(M, D, D_e)$. When dealing with an unlearning request, we hope that the unlearned model $\mathcal{U}(M, D, D_e)$ is the same as the retrained model $\mathcal{A}(D \setminus D_e)$ (i.e., $\mathcal{U}(M, D, D_e) \simeq \mathcal{A}(D \setminus D_e)$). Most unlearning studies ended at this step. However, we do need an effective evaluation metric to assess the unlearning effect, as in Step 3 in Fig. 2. Therefore, we add the verification step as the last step in the unlearning process, and we will introduce the related literature later.

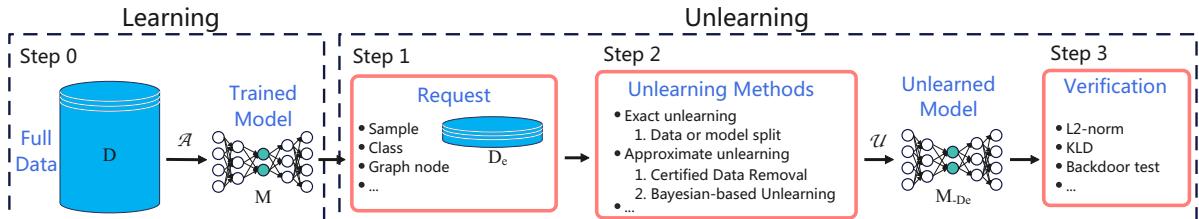


Fig. 2. Machine Unlearning Process

3.2 Machine Unlearning Evaluation Metrics

To compare the two models before and after unlearning, we need to define an evaluating metric $d(\cdot)$ between $\mathcal{A}(D \setminus D_e)$ and $\mathcal{U}(M, D, D_e)$. To this end, we briefly introduce several common evaluation distance metrics:

L_2 -Norm. In [23], the authors propose utilizing the Euclidean distance to evaluate the parameters of the retrained model and the unlearned model. Let θ represent the model parameters learned by the algorithm $\mathcal{A}(\cdot)$. The

L_2 -norm measures the distance between $\theta_{\mathcal{A}(D \setminus D_e)}$ and $\theta_{\mathcal{U}(M, D, D_e)}$, where $\theta_{\mathcal{A}(D \setminus D_e)}$ are the model parameters retrained from scratch, and $\theta_{\mathcal{U}(M, D, D_e)}$ are the model parameters resulting from the unlearning algorithm $\mathcal{U}(\cdot)$. **Kullback–Leibler divergence (KLD).** KLD is commonly used to measure the divergence between two probability distributions, often assessing the distance between retrained and unlearned models. In Bayes-based or Markov chain Monte Carlo-based unlearning methods [13], researchers utilize KLD [102] to optimize approximate models, employing it to measure the distance between two probability distributions. Recent unlearning studies have also used KLD to estimate the unlearning effect by comparing the distributions of retrained and unlearned models [13].

Evaluation Metric based on Privacy Leakage. Since membership inference attacks [68] can decide whether a sample was utilized for training a model, recently, some works have leveraged this property to verify if unlearning mechanisms remove the specific data. Some studies [61] even proposed to backdoor the unlearning samples for initial model training and then attack the unlearned model. If the unlearned model is still backdoored, this proves that the unlearning algorithm cannot unlearn samples effectively. Conversely, if the backdoor trigger cannot attack the unlearned model, it proves that the unlearning algorithm is effective. Similar methods were also used in [63, 64] to evaluate the unlearning effect.

3.3 Majority of Tools Used in Unlearning

Differential Privacy (DP). Differential privacy is a popular benchmark for privacy protection in the Statistic [103]. In a DP model, a trusted analyzer collects users' raw data and then executes a private method to guarantee differential privacy. The DP protection ensures the indistinguishability for any two outputs of neighboring datasets, where neighboring datasets mean the dataset only differs by replacing one user's data, denoted as $X \asymp X'$. A (ϵ, δ) -differential privacy algorithm $\mathcal{M} : \mathbb{X}^n \rightarrow \mathbb{Z}$ means that for every neighboring dataset pair $X \asymp X' \in \mathbb{X}^n$ and every subset $S \subset \mathbb{Z}$ has that $\mathcal{M}(X) \in S$ and $\mathcal{M}(X') \in S$ are ϵ -indistinguishable and δ -approximate. The degree of privacy protection rises with decreasing ϵ . When $\epsilon = 0$, it implies that the outputted probability distribution of mechanism \mathcal{M} cannot represent any meaningful information. A general DP mechanism based on adding Laplace noise was presented and theoretically analyzed in [103].

Bayesian Variational Inference. In machine learning, the Bayesian variational inference is used to approximate difficult-to-compute probability densities via optimization [104, 105]. We revisit the variational inference framework that learns approximate posterior model parameters θ using Bayesian Theory in this part. Suppose a prior belief $p(\theta)$ of an unidentified model and a complete data trainset D , an approximate posterior belief $q(\theta|D) \sim p(\theta|D)$ can be optimized by minimizing the KLD [102], $\text{KL}[q(\theta|D)||p(\theta|D)]$. KLD measures how one probability distribution $q(\theta|D)$ differs from another probability distribution $p(\theta|D)$. However, it is intractable to compute the KLD exactly or minimize the KLD directly. Instead, the evidence lower bound (ELBO) [104] was proposed to be maximized, which is equivalent to minimize KLD between the two probability distributions. ELBO follows directly from $\log(p(D))$ subtracting $\text{KL}[q(\theta|D)||p(\theta|D)]$, where $\log(p(D))$ is independent of $q(\theta|D)$. The ELBO is a lower bound of $\log(p(D))$ as $\text{KL}[q(\theta|D)||p(D|\theta)] \geq 0$. In general training situations, ELBO is maximized using stochastic gradient ascent (SGA) [104]. The primary process is approximating the expectation $\mathbb{E}_{q(\theta|D)}[\log(p(D|\theta)) + \log(p(\theta)/q(\theta|D))]$ with stochastic sampling in each iteration of SGA. We can use a simple distribution (e.g., the exponential family) to approximate computational ease posterior belief $q(\theta|D)$.

Privacy Leakage Attacks. Privacy leakage occurs in both unlearning verification and privacy threats in two parts of unlearning. In unlearning verification, researchers tried to use privacy leakage attacks to verify whether the specific data is unlearned. Regarding the privacy and security issues in unlearning, researchers have tried to design effective inference attacks tailored to machine unlearning. The basic attack of privacy leakage in a machine learning setting is membership inference, which determines if a sample was employed in the model updating process or not. When an attacker fully knows a sample, knowing which model was trained on it will

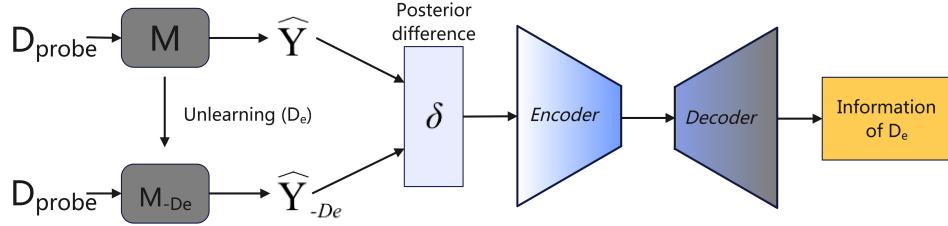


Fig. 3. Privacy Leakage: a Privacy Reconstruction Process

leak information about the model. A generic membership inference process was introduced in [70]. Shokri et al. first trained the shadow models to approach the target ML models. Then, they observed and stored the different outputs of the shadow models based on different inputs, in or not, in the trainset. They used these stored outputs as samples to train the membership inference attack model.

Model inversion [106], or privacy reconstruction [107] is another privacy threat in general machine learning. Model inversion aims to infer some lacking attributes of input features based on the interaction with the trained ML model. Salem et al. [107] proposed a reconstruction attack target recovering specific data samples used in the model updating by different model outputs before and after updating. Later, inferring the private information of updating data in conventional machine learning is transferred to inferring the privacy of the erased samples in machine unlearning. In reconstruction attacks, the adversary first collects the different outputs using his probing data D_{probe} , including the original outputs \hat{Y}_M before unlearning, and the outputs \hat{Y}_{M-De} after unlearning. Then, he constructs the attack model based on the posterior difference $\delta = \hat{Y}_{M-De} - \hat{Y}_M$. The attack model contains an encoder and decoder, which has a similar structure as VAEs [104], and the main process is shown in Fig. 3.

4 CENTRALIZED MACHINE UNLEARNING

In this section, we classify existing centralized unlearning methods by their inherent mechanism and designed purposes and present the corresponding detailed techniques.

4.1 Unlearning Solution Categories

From the former introduction, we know that naive retraining is the most effective manner to realize machine unlearning. However, it is inefficient because it requires storing the entire original dataset and retraining the model from scratch, which consumes significant storage and computational resources, especially in deep learning scenarios. Therefore, researchers tried to design effective and efficient unlearning mechanisms. Exact unlearning was proposed to reduce the computation cost by splitting training sub-models based on pre-divided data sub-sets [9]. It also divides the stochasticity and the incrementality into sub-models. It can unlearn an exact model by ensembling the consisting submodels, but it still needs to store all the split data subsets for fast retraining. Another approach is approximate unlearning [26]. It can reduce both storage and computation costs because it unlearns only based on the erased data. However, it is difficult for approximate unlearning methods to control the accuracy degradation due to the challenges in estimating stochasticity and incrementality during the training process. Most of them bounded their estimation to avoid dramatic accuracy reduction.

4.1.1 Exact Unlearning. Exact unlearning could also be called fast retraining, whose basic idea is derived from naive retraining from scratch. Following the background introduction of unlearning, we know the learning and unlearning algorithm, $\mathcal{A}(D)$ and \mathcal{U} , based on the trainset D and erased dataset $D_e \subseteq D$, respectively. If $\mathcal{U}(\cdot)$ is implemented as naive retraining, the equality between $\mathcal{A}(D \setminus D_e) \in \mathcal{H}$ and $\mathcal{U}(M, D, D_e) \in \mathcal{H}$ is absolutely

guaranteed. However, naive retraining involves high computation and storage costs, especially for deep learning models and complex datasets [63]. Unlike naive retraining, which relies on the whole remaining dataset, exact unlearning tries to retrain a sub-model only using a subset of the remaining dataset to reduce calculation cost. A general operation of exact unlearning is that they first divide the dataset into several small sub-sets. Then, they transform the learning process by ensembling the sub-models trained with each sub-set as the final model [8, 9]. So that when an unlearning request comes, they are just required to retrain the sub-model corresponding to the sub-set containing the erased data. They then ensemble the retrained sub-model and other sub-models as the unlearned model.

As we introduced above, exact unlearning aims to mitigate the computation cost when retraining a new model by transforming the original learning algorithms into an ensembling form. It divides the stochasticity and incrementality into several sub-models to reduce their influence. However, to some extent, they sacrificed the storage cost because they needed to store the whole training dataset in a divided form. In [8], Cao and Yang transformed the traditional ML algorithms into a summation form. They are only required to update several summations when an unlearning requirement comes, ensuring the method runs faster than retraining from scratch. SISA [9] is a representative exact unlearning algorithm, which splits the full training dataset into shards and trained models separately in each shard. For unlearning, they simply need to retrain the shard that includes the erased data. Study [24] proposed a framework that precisely models the impact of individual training sample on the model concerning various performance criteria and removes the impact of samples that are required to be removed. Golatkar et al. [14] proposed an unlearning method on deep networks, splitting the trained model into two parts. The core part based on the data will not be deleted, and the unlearning part with the erased data will be unlearned with parameters bound. These methods are efficient in computation, but they sacrifice the storage space to store the intermediate training parameters of different slices and the related training sub-sets.

Besides the high storage cost, another major issue with exact unlearning is that it is only suitable for scenarios where the unlearning request involves removing a few samples with low frequency. Suppose an unlearning request needs to remove many data samples (usually, they are not in the same previous divided sub-set). In that case, exact unlearning must retrain all these related sub-models or even all the sub-models in the worst situation. At the moment, exact unlearning is no longer computation efficient, and the whole training dataset and intermediate parameters still need to be stored.

4.1.2 Approximate Unlearning. Unlike exact unlearning, which only aims to reduce the retraining computation cost, approximate unlearning tries to directly unlearn based on the trained model and the erased data sample, which saves the computation and storage costs together. Approximate unlearning studies aim to unlearn a model that approaches the model trained on the remaining dataset, i.e., the unlearned model $\mathcal{U}(M, D, D_e)$ should match the retrained model $\mathcal{A}(D \setminus D_e)$. Since exact unlearning is implemented by retraining from the remaining dataset or sub-sets, they can almost guarantee equality before and after unlearning. However, since approximate unlearning tries to directly delete the influence of the unlearned samples from trained models, the core problem lies in precisely estimating and removing this contribution, which includes both stochasticity and incrementality.

The text description of the changes between two different distribution spaces before and after removing the specific data is not intuitive. Fig. 4 shows illustrated changes when adding a new point or removing a point in a classifying model. When an influential point appears, it usually pushes the line to move forward than the original classifying line to identify it, as shown in Fig. 4 (b). When this influential point is requested to be removed, the unlearning mechanism must recover the model to the original one that has not been trained by this specific point, as shown in Fig. 4 (c). However, when only unlearning a non-influential point, which may have almost non-influence on the model, the unlearned model may not change compared to the original trained model in this situation, as shown in Fig. 4 (d).

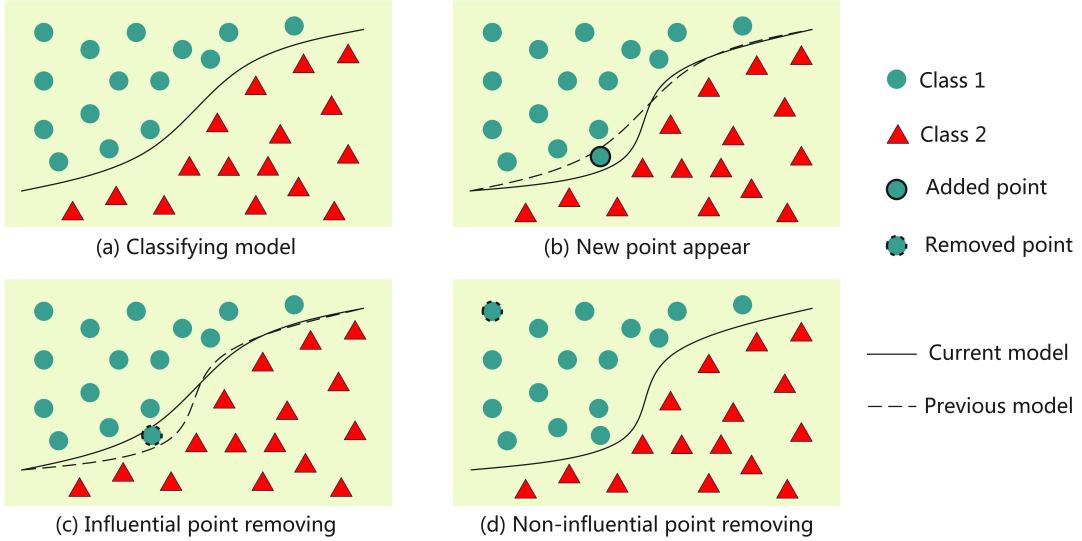


Fig. 4. The model changes when adding a new point or removing a point. (a) A normally trained classifying model classifies classes 1 and 2. (b) When a new point appears, the model is trained based on it, and the classifying line is pushed to classify it. (c) When we need to remove an influential point, we should recover the contribution of this data point on the model. (d) When we remove a Non-influential point, the model may not need to change a lot.

Many methods were proposed to implement approximate unlearning efficiently and effectively. The popular solutions are certified-removal [25] and Bayes-based mechanisms [13], which are introduced in technical detail in Section 4.2. Here, we briefly summarize current studies about approximate unlearning. A certified-removal mechanism [25] was an approximate unlearning method similar to differential privacy. Guo et al. [25] defined the ϵ -approximate unlearning, which makes sure that the model after and before unlearning must be ϵ -indistinguishable as the definition in DP. The difference between ϵ -approximate unlearning and ϵ -DP is that the mechanism \mathcal{A} on differential privacy is needed never to memorize the data in the first place, which is impossible in machine unlearning. The machine learning model does not learn anything from the training dataset if \mathcal{A} is differentially private [9]. A similar solution to [25] is introduced in [26], which erases the influence of the specified samples from the gradients during unlearning. To mitigate accuracy degradation, model parameter updates are constrained by a differential unlearning definition.

Different from indistinguishable unlearning, other methods used the Bayes theorem to unlearn an approximate posterior. In [13], Nguyen et al. employed Bayesian inference to unlearn an approximate model using the erased data. Moreover, Fu et al. [35] developed a similar unlearning method, which uses Markov chain Monte Carlo (MCMC) methods. In [34], the authors also explained the effectiveness of the approximate unlearning method from the perspective of MCMC. Nevertheless, since those techniques are approximately unlearning the contribution of all the erased data, including the inputs and labels, they inevitably decrease the model accuracy to some extent after unlearning.

Main Challenges of Approximate Unlearning. As we discussed at the beginning, there are three main challenges in machine unlearning. In centralized scenarios, these studies mainly aim at solving the basic machine unlearning problem, which unavoidably faces three challenges: stochasticity of training, incrementality of training, and catastrophe of unlearning. The exact unlearning methods extend the retraining idea, which avoids facing these challenges but consumes lots of storage costs. The approximate unlearning methods face these challenges

Table 2. Popular Unlearning Techniques

Unlearning Literature	Taxonomy	Requests Type	Techniques	Realization Method	Year
SISA [9]	Exact unlearning	Samples	Split Unlearning	Data and Model Partition	2021
Amnesiac unl. [15]	Exact unlearning	Samples	Split Unlearning	Partially retraining	2021
GraphEraser [45]	Exact unlearning	Graph nodes	Split Unlearning	Data Partition	2022
RecEraser [19]	Exact unlearning	Samples	Split Unlearning	Balanced Data Partition	2022
ARCANE [18]	Exact unlearning	Samples	Split Unlearning	Partition by Class	2022
HedgeCut [28]	Exact unlearning	Samples	Split Unlearning	Tree ensemble learning	2021
DeltaGrad [23]	Exact unlearning	Samples	Split Unlearning	L-BFGS [110]	2020
ERASER [22]	Exact unlearning	Samples	Split Unlearning	Inference Serving-Aware	2024
L-CODEC [31]	Approximate unlearning	Samples	Certified Data Removal	Markov Blanket selection	2022
PUMA [24]	Approximate unlearning	Samples	Certified Data Removal	SME	2022
Certified Removal [25]	Approximate unlearning	Samples	Certified Data Removal	LP	2019
(ϵ, δ) -unl. [27]	Approximate unlearning	Samples	Certified Data Removal	Perturbed gradient descent	2021
[33]	Approximate unlearning	Samples	Certified Data Removal	Influence Theory	2023
Graph unl. [44]	Approximate unlearning	Graph nodes	Certified Data Removal	Certified removal	2022
Gif [51]	Approximate unlearning	Graph nodes	Certified Data Removal	Influence function	2023
SUMMIT [50]	Approximate unlearning	Graph nodes	Certified Data Removal	Multi-Objective Optimization	2024
EUBO, rKL [13]	Approximate unlearning	Samples	Bayesian Unlearning	VBI	2020
MCU [34]	Approximate unlearning	Samples	Bayesian Unlearning	Monte cario-based	2022
BIF [35]	Approximate unlearning	Samples	Bayesian Unlearning	MCMC	2022

Algorithm abbreviations

VBI: Variational Bayesian Inference, SME: Store medial estimation, MCU: Monte Cario-based machine unlearning, FIM: Fisher Information Matrix, MCMC: Markov chain Monte Carlo, BIF: Bayesian inference forgetting, LP: Loss perturbation, TF-IDF: Term Frequency Inverse Document Frequency, EUBO: Evidence upper bound, rKL: reverse Kullback–Leibler

directly, and we here list the relevant work about how to estimate the contribution of erased samples to overcome the stochasticity and incrementality of training, and how to prevent unlearning catastrophe.

- To overcome the stochasticity and incrementality challenges when estimating the unlearning influence, one popular strategy is based on the first-order and second-order influence function [108], which is calculated based on the perturbation theory [109].
- The unlearning catastrophe appears commonly in approximate unlearning, and many studies try to propose some methods to solve this problem. In certified removal and Bayesian-based methods, they usually set a threshold to limit the unlearning update extent [13, 25]. In [38], Wang et al. solves this problem by adding a model utility compensation task during unlearning optimization and finding the optimal balance based on multi-objective training methods.

4.2 Detailed Techniques of Centralized Unlearning

This section presents the popular techniques used in existing unlearning methods, in both exact and approximate unlearning studies. Exact unlearning extends the idea of retraining and tries to reduce the computation cost of unlearning. Approximate unlearning was proposed to find a way to reduce computation and storage consumption together. The dominant studies are summarized in Table 2, where the primary technique used in exact unlearning is split learning. Two primary techniques used in approximate unlearning are certified data removal and Bayesian-based unlearning. The introduction of these techniques will primarily follow the sequence of the unlearning techniques presented in Fig. 1.

4.2.1 Split Unlearning. Since most exact unlearning methods attempted to partition the training dataset into multiple subsets and divide the ML model learning process, we call this kind of unlearning technique split unlearning. The main procedure of split unlearning is illustrated in Fig. 5 (b). By contrast, the process of the

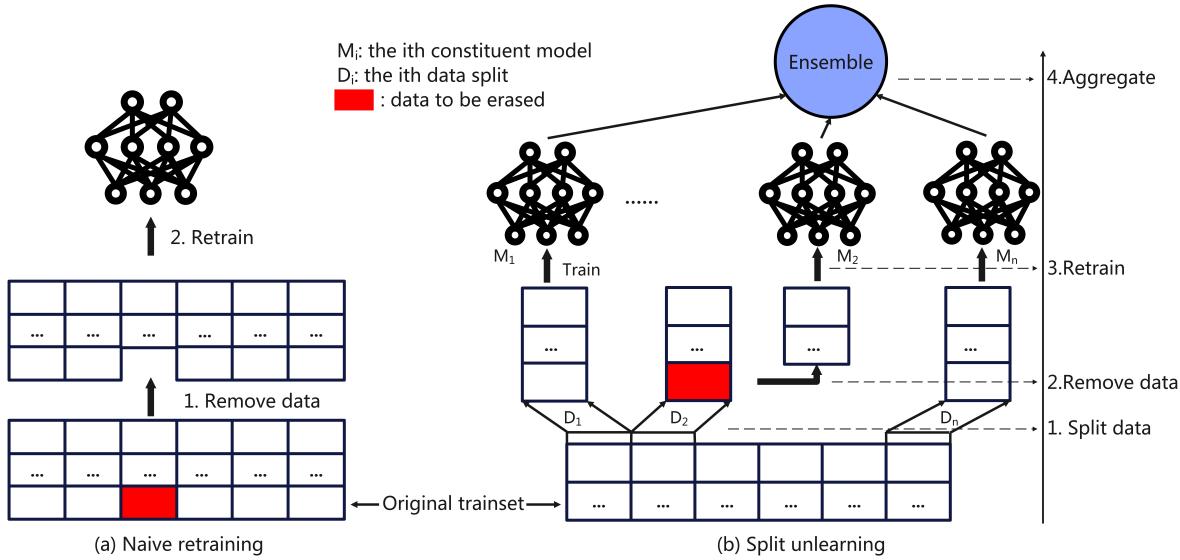


Fig. 5. (a) Naive unlearning. There are only two steps: delete the specified samples from the whole dataset and retrain a model based on the remaining dataset. (b) Split unlearning. It contains four steps: 1. split the original dataset into n shards, 2. remove the erased data from the corresponding shard, 3. retrain the sub-model of this shard, 4. ensemble all sub-models as the final model.

naive retraining method is shown in Fig. 5 (a), where we need two steps to realize it: first, delete the samples from the training dataset; second, retrain a new model using the remaining dataset. Since it needs to store the whole training dataset and retrain from the remaining dataset, it often entails significant computational and stored overhead. They proposed many exact unlearning methods to reduce the huge computation cost of naive retraining, and the majority of them are based on split learning techniques, either on data or model. As shown in Fig. 5 (b), the split unlearning technique can be summarized into four steps. Unlike naive unlearning trained based on the remaining trainset, the first phase of split unlearning is dividing the original full trainset into multiple disjoint shards. All the constituent models are trained based on each split data slice. Then, in the second phase, when the unlearning request comes, they only need to erase the requested samples from the split slice and retrain this slice's constituent model in the third phase. In the last phase, the split unlearning aggregates the retrained and other constituent models together as a new unlearned model.

The first split unlearning is proposed by Cao and Yang [8]. They split the original learning algorithms into a summation form. In a regular machine learning form, the model directly learns from the training dataset. However, in the summation form, they first train a small number of constituent models, which learn from several parts of the full trainset and then aggregate these intermediate models as the final learning model. So that when unlearning, they only need to retrain the constituent model that contains the information of erased data. It can efficiently speed up retraining time and reduce computation costs. In [8], the authors indicated that support vector machines, naive Bayes classifiers, k-means clustering, and many ML algorithms could be implemented in a summation form to reduce the retraining cost. The statistical query (SQ) learning [111] guarantees the summation form. Although algorithms in the Probably Approximately Correct (PAC) setting can transform to the SQ learning setting, many complex models, such as DNNs, cannot be efficiently converted to SQ learning.

Then, Bourtoule et al. [9] and Yan et al. [18] proposed advantaged methods unlearn samples suitable on deep neural networks. The primary idea of [9, 18] is also similar to the process shown in Fig. 5 (b). In [9], Bourtoule et al.

named their unlearning method the SISA training approach. SISA can be implemented on deep neural networks, training multiple sub-neural networks based on divided sub-datasets. When the unlearning request comes, SISA retrains the model of the shard, which contains the information about the erased samples. SISA is effective and efficient as it aggregates all sub-models final prediction results rather than aggregates all these models. Unlike the original split unlearning dividing the dataset and transforming learning algorithms to summation form, Yan et al. proposed ARCANE [18], which transforms conventional ML into ensembling multiple one-class classification tasks. When many unlearning requests come, it can reduce retraining costs, which was not considered in previous work.

Chen et al. [19] extended exact unlearning methods to recommendation tasks and proposed RecEraser, which has similar architecture as split unlearning in Fig. 5 (b). RecEraser is tailored to recommendation systems, which can efficiently implement unlearning. Specifically, they designed three data division schemes to partition recommendation data into balanced pieces and created an adaptive aggregation algorithm utilizing an attention mechanism. They conducted the experiments on representative real-world datasets, which are usually employed to assess the effectiveness and efficiency of recommendation models.

Besides the above popular ML models, Schelter et al. proposed HedgeCut [28], which implemented machine unlearning on tree-based ML models in a split unlearning similar form. Tree-based learning algorithms are developed by recursively partitioning the training dataset, locally optimizing a metric such as Gini gain [112]. HedgeCut focuses on implementing fast retraining for these methods. Furthermore, they evaluated their method on five publicly available datasets on both accuracy and running time.

Another method that is similar to split unlearning is Amnesiac Unlearning [15]. The intuitive idea of Amnesiac Unlearning is to store the parameters of training batches and then subtract them when unlearning requests appear. In particular, it first trains the learning model by adding the total gradients $\sum_{e=1}^E \sum_{b=1}^B \nabla_{\theta_{e,b}}$ to the initial model parameters θ_{initial} , where E is the training epochs, and B is the data batches. In the model training process, they kept a list called SB , which records the batches holding the private data. This list could be formed as an index of batches for each training example, an index of batches for each category or any other information expected. When the unlearning request comes, a model using Amnesiac unlearning needs only to remove the updates from each batch $sb \in SB$ from the learned model θ_M . As Graves et al. [15] stated, using Amnesiac unlearning effectively and efficiently removes the contribution of the erased samples that could be detected through state-of-the-art privacy inference attacks and does not degrade the accuracy of the model in any other way.

4.2.2 Certified Data Removal. Certified data removal unlearning methods usually define their unlearning algorithms as ϵ -indistinguishable unlearning, which is similar to the differential privacy definition [113]. An example is presented in Figure 6. Most of them use the Hessian matrix [114] to evaluate the contribution of erased data samples for unlearning subtraction. After estimating the impact of the erased data samples, they unlearn by subtracting these impacts with an updating bound from the unlearning model.

In [25], Guo et al. proposed a certified data removal method, which assumes removing the last training sample, (x_n, y_n) . Specifically, they defined a removal mechanism that approximately minimizes $\mathcal{L}(\theta; D')$ with $D' = D \setminus (x_n, y_n)$. The loss gradient at sample (x_n, y_n) can be denoted as $\Delta = \lambda \nabla \ell(\theta^T \cdot x_n, y_n)$ and the Hessian of $L(\cdot; D')$ at θ by $H_\theta = \nabla^2 L(\theta; D')$. Then, they

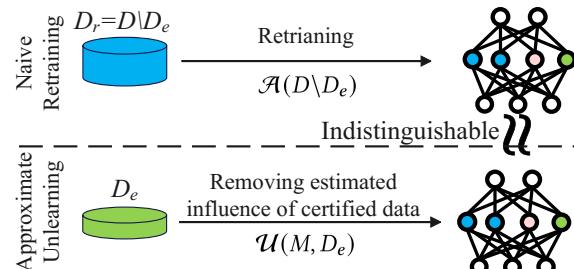


Fig. 6. The approximate unlearning by certified data removal. The unlearning algorithm \mathcal{U} includes estimating the influence of specified data and removing the estimation from trained models. The unlearned model is expected to approach the retrained model.

applied a one-step Newton update to the model parameters impact of the erased point (x_n, y_n) on the model θ . Under their observation, they found that directly removing the Hessian contribution from the gradient will reveal the private information of the erased data. They used the loss perturbation technique [115] to hide this information. It used a random linear term to perturb the empirical risk and ensure that the outputs of their method $\mathcal{U}(D, D_e, \mathcal{A})$ is ϵ -indistinguishable between the retrained model $\mathcal{A}(D \setminus D_e)$. In [31] and [63], they designed unlearning algorithms following the certified data removal definition in [25]. Mehta et al. [31] unlearned via their proposed efficient Hessians, L-FOCI [31]. Thudi et al. [63] used membership inference as a verification error to adjust the unlearning process on stochastic gradient descent (SGD) optimization.

Another similar unlearning method is PUMA. In [24], Wu et al. proposed a new data removal method through gradient re-weighting called PUMA, which also used the Hessian Vector Product (HVP) term. They first estimated and recorded individual contributions of (x_i, y_i) , where the estimation is limited to less than one dot product between the pre-cached HVP term and individual gradient. When the unlearning request comes, they subtract the estimate of the erased samples to revise the model.

Ginart et al. [29] extended certified data removal to k-means clustering algorithms. They formulated the unlearning problem of efficiently removing personal data information from trained clustering models. They offered two different deletions for k-means clustering, quantized k-mean and divide-and-conquer k-means. In their work, both algorithms have theoretical guarantees and strong empirical results.

To retrain SGD-based models fast, DeltaGrad was proposed by Wu et al. [23] to unlearn small changes of data inspired by the idea of "differentiating the optimization path" concerning the training dataset and Quasi-Newton methods. They theoretically proved that their algorithm could approximate the right optimization path rapidly for the strongly convex objective. DeltaGrad starts with a "burn-in" period of first iterations, where it computes the full gradients precisely. After that, it only calculates the complete gradients for every first iteration. For other iterating rounds, it operates the L-BGFS algorithm [110] to compute Quasi-Hessians approximating the true Hessians, keeping a set of updates at some prior iterations.

For a deeper understanding of certified machine unlearning, Sekhari et al. [26] further given a strict separation between ϵ -indistinguishable unlearning and differential privacy. Different from [25], in order to utilize tools of differential privacy (DP) for ML, the most straightforward manner is to forget the special dataset of erasure demands D_e and create an unlearning mechanism \mathcal{U} that solely relies on the learned algorithm $\mathcal{A}(D)$. In particular, the unlearning method is of the form $\mathcal{U}(D_e, \mathcal{A}(D)) = \mathcal{U}(\mathcal{A}(D))$ and makes sure the true unlearned model $\mathcal{U}(\mathcal{A}(D))$ is ϵ -indistinguishable to $\mathcal{U}(\mathcal{A}(D \setminus D_e))$. Notice the difference between [25] and [26]. In the definition of [25], their ϵ -indistinguishable unlearning is between $\mathcal{U}(\mathcal{A}(D))$ and $\mathcal{A}(D \setminus D_e)$, but here is between $\mathcal{U}(\mathcal{A}(D))$ and $\mathcal{U}(\mathcal{A}(D \setminus D_e))$. Such a pair of algorithms in [26] would be differential private for D , where the neighboring datasets mean that for two datasets with an edit distance of m samples. The guarantee of DP unlearning is more powerful than the model distribution undistinguishable unlearning in [25], and therefore, it suffices to satisfy it. Based on the definition of [26], they pointed out that any DP algorithm automatically unlearns any m data samples if they are private for datasets with the distance m . Therefore, they derivate the bound on deletion capacity from the standard performance guarantees for DP learning. Furthermore, they determine that the existing unlearning algorithms can delete up to $\frac{n}{d^{1/4}}$ samples meanwhile still maintaining the performance guarantee w.r.t. the test loss, where n is the size of the original trainset, and d is the dimension of trainset inputs.

The aforementioned methods are trying to address the basic unlearning problem from a certified data removal perspective. Neel et al. [27] extended the definition of unlearning to include updates, which encompass both "add" $(D \cup \{z\})$ and "delete" $(D \setminus \{z\})$, where $z = (x, y)$ and $z \in \mathcal{Z}$ is a data point. They follow the definition in [25] and define similar (ϵ, δ) -indistinguishability. Furthermore, they extend (ϵ, δ) -indistinguishability to both "add" and "delete" updates, which can also be denoted as (ϵ, δ) -publishing.

4.2.3 Bayesian-based Unlearning. Different from certified data removal that unlearns samples by subtracting corresponding Hessian matrix estimation from trained models, Bayesian-based unlearning tries to unlearn an approximate posterior as the model is trained by employing the remaining dataset. The exact Bayesian unlearning posterior can be derived from the Bayesian rule as $p(\theta|D_r) = p(\theta|D) p(D_e|D_r)/p(D_e|\theta)$, where θ is the posterior (i.e., model parameters). The erased dataset and the remaining dataset are two independent subsets of the full training dataset. If the model parameters θ are discrete-valued, $p(\theta|D_r)$ can be directly obtained from the Bayesian rule [13]. Additionally, employing a conjugate prior simplifies the unlearning process.

Nevertheless, it is challenging to get the exact posterior in practice, not to mention the unlearning posterior. In [13], Nguyen et al. tried doing likewise at the beginning. They defined the loss function using the KLD between the approximate predictive distribution $q_u(y|D_r)$ and the exact predictive distribution $p(y|D_r)$. They bounded this loss function by the KLD between posterior beliefs $q_u(\theta|D_r)$ and $p(\theta|D_r)$ and further proposed evidence upper bound (EUBO) as the loss function to unlearn the approximate unlearning posterior. To avoid the overestimation of using KL divergence to optimize the posterior, they introduced an adjusted likelihood to control the unlearning extent.

In [34], the authors also studied the problem of "unlearning" particular erased subset samples from a trained model with better efficiency than retraining a new model from scratch. Toward this purpose, Nguyen et al. [34] proposed an MCMC-based machine unlearning method deriving from the Bayesian rule. They experimentally proved that MCMC-based unlearning could effectively and efficiently unlearn the erased subsets of the whole training dataset from a prepared model.

Fu et al. [35] converted the MCMC unlearning problem into an explicit optimization problem. Then they proposed ϵ -knowledge removal, which was a little similar to certified removal methods, but they defined that KLD between unlearned and retrained models must be less than ϵ . To quantify the explicit ϵ -knowledge removal, they proposed a knowledge removal estimator to assess the difference between the original and unlearned distributions. As they indicated, though their algorithm cannot wholly remove the learned knowledge from the already trained distribution, their method can still help the unlearned model approach its local minimum.

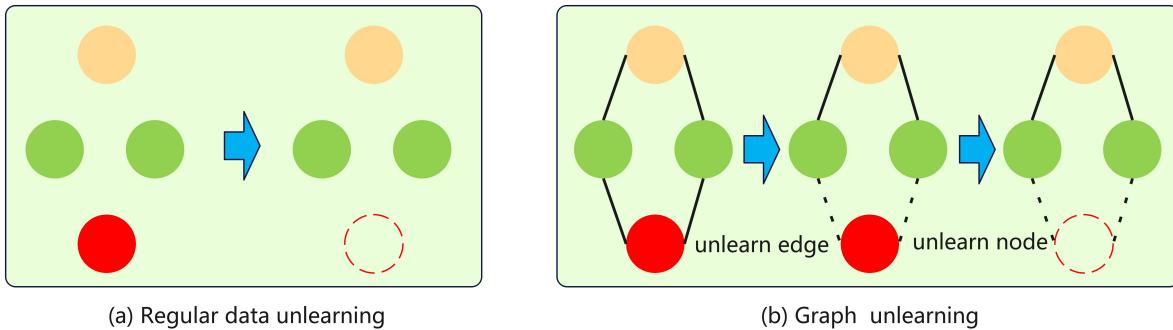


Fig. 7. (a) Regular data unlearning, which only unlearns the data sample. (b) Graph unlearning. Since the graph data contains both node features and connective edge information, we may need to unlearn edge information or both edges and nodes in graph unlearning.

4.2.4 Graph Unlearning. We introduce graph unlearning as a representative kind of irregular data unlearning. In [44, 45, 47, 50], researchers extend regular data machine unlearning to a graph data scenario. Graph structure data are more complex than standard structured data because graph data include not only the feature information of nodes but also the connectivity information of different nodes, shown in Fig. 7. Therefore, Chien et al. [44]

proposed node unlearning, edge unlearning, and both node and edge unlearning for simple graph convolutions (SGC). Besides the different information unlearned in a graph learning problem, they found another challenge associated with feature mixing during propagation, which needs to be addressed to establish provable performance guarantees. They gave the theoretical analysis for certified unlearning of GNNs by illustrating the underlying investigation on their generalized PageRank (GPR) extensions and the example of SGC.

Chen et al. [45] found that applying SISA [9] unlearning methods to graph data learning will severely harm the graph-structured information, resulting in model utility degradation. Therefore, they proposed a method called GraphEraser to implement unlearning tailored to graph data. Similar to SISA, they first cut off some connecting edges to split the total graph into some sub-graphs. Then, they trained the constituent graph models on these sub-graphs and ensembled them for the final prediction task. To realize graph unlearning efficiently, they proposed two graph partition algorithms and corresponding aggregation methods based on them.

In [47], Cong and Mahdavi filled in the gap between regularly structured data unlearning and graph data unlearning by studying the unlearning problem on the linear-GNN. To remove the knowledge of a specified node, they design a projection-based unlearning approach, PROJECTOR, that projects the weights of the pre-trained model onto a subspace irrelevant to the deleted node features. PROJECTOR could overcome the challenges caused by node dependency and is guaranteed to unlearn the deleted node features from the pre-trained model.

5 FEDERATED UNLEARNING

We have introduced popular centralized machine unlearning in the former section. Their main purpose is to reduce the calculation and storage costs and the accuracy degradation caused by unlearning. Recently, we noticed that some researchers paid attention to realizing unlearning in other scenarios, such as in federated learning (FL) scenarios. In this section, we will introduce the literature about federated unlearning.

FL was initially introduced to protect the privacy of participating clients during the machine learning training process in distributed settings. All participants will only upload their locally trained model parameters instead of their sensitive local data to the FL server during model training processes. Therefore, in a federated learning scenario, limited access to the dataset will become a unique challenge when implementing unlearning.

Since the local data cannot be uploaded to the federated learning (FL) server side, most federated unlearning methods try to erase a certain client's contribution from the trained model by storing and estimating the contribution of uploaded parameters. In this situation, they can implement federated unlearning without interacting with the client, shown as the server-side federated unlearning in Fig. 8 (a). The two representative methods are [56, 60]. Liu et al. [56] proposed “FedEraser” to sanitize the impact of a FL client on the global FL model. In particular, during FL training process, the FL-Server maintains the updates of the clients at each routine iteration and the index of the related round to calibrate the retrained updates. Based on these operations, they reconstructed the unlearned FL model instead of retraining a new model from scratch. However, FedEraser can only unlearn one client's data, which means it must unlearn all the contributions of this specific client's data. It is unsuitable for a client who wants to unlearn a small piece of his data. Study [60] tried to erase a client's influence from the FL model by removing the historical updates from the global model. They implemented federated unlearning by using knowledge distillation to restore the contribution of clients' models, which does not need to rely on clients' participation and any data restriction.

Wang et al. [52] explored the problem of how to selectively unlearn a category from a trained CNN classification model in FL. The biggest challenge is that the local data utilized for the FL global model training cannot be accessed globally. Therefore, they explored the inner influence of each channel and observed that various channels have distinct impacts on different categories. They proposed a method that does not require accessing the data used for training and retraining from scratch to cleanly scrub the information of particular categories from the global FL model. A method called TF-IDF was introduced to quantize the class discrimination of channels. Then,

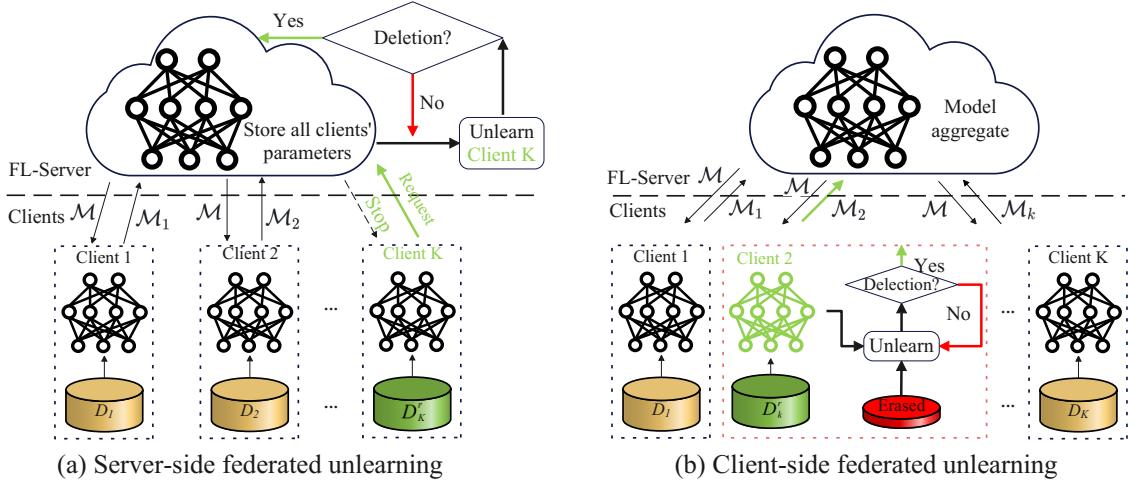


Fig. 8. Comparison between (a) server-side federated unlearning and (b) client-side federated unlearning

they pruned the channels with high TF-IDF scores on the erased classifications to unlearn them. Evaluations of CIFAR10 and CIFAR100 demonstrate the unlearning effect of their method.

Different from unlearning a whole client's influence and unlearning a class, Liu et al. [55, 57] investigated how to unlearn data samples in FL, shown as the client-side federated unlearning in Fig. 8 (b). In [57], they first defined a federated unlearning problem and proposed a fast retraining method to withdraw the influence of data from the FL model. Then, they proposed an efficient federated unlearning method following the Quasi-Newton methods and the first-order Taylor approximate method [57]. They utilized the practical Fisher Information Matrix to model the Hessian matrix at a low cost. When unlearning, they subtract the estimation matrix from gradients with a bound. In [55], they implement federated unlearning based on Bayesian inference, and they propose a parameters self-sharing method to reduce the model utility degradation.

6 UNLEARNING EVALUATION AND VERIFICATION

6.1 The Unlearning Verification and Auditing Methods

Recent studies paid a huge amount of attention to unlearning problem-solving; however, verifying the unlearning effect is also an important problem in machine unlearning. In Section 3, we have introduced the L2-norm [23], KLD [13], and privacy leakage as the unlearning verification metrics. The common metrics also include accuracy in assessing the performance of the unlearned model and running time in evaluating the unlearning efficiency. We have listed recent evaluation and verification metrics in Table 3, where the popular experimental datasets are furthermore presented. Since accuracy, running time, L2-norm, and KLD are fixed evaluation metrics, we do not discuss them in detail here. In this section, we introduce some new evaluation metrics that recent studies tailored to unlearning. Most of them are based on privacy leakage.

In [118], Liu and Tsaftaris indicated that the data auditing problem in machine unlearning is becoming popular. They used a calibration dataset and the Kolmogorov-Smirnov (KS) [119] distance to confirm if an erased dataset is removed. Their challenge was that the method would fail when the calibration dataset was out of high quality, or the query samples were similar to the training dataset. The EMA [66] method was proposed to overcome these limitations when auditing data erasure from a trained model. EMA is a 2-step data removal auditing method that ensembles several metrics and statistical tools. In the first step of EMA, the auditor uses various metrics

Table 3. Evaluation and Verification Metrics

Evaluation Metrics	Description	Usage	Literature
Accuracy	Model accuracy on erased datasets and remaining datasets	To evaluate the predictive accuracy of the unlearned model	[9, 25], ...
Running time	The training time of unlearning process	To evaluate the unlearning efficiency	[8, 25], ...
L2-norm	The parameters differences between the retrained and unlearned models $\ \theta_1 - \theta_2\ $	To evaluate the indistinguishability between two models	[23]
KL-Divergence	The KLD between the distribution of the unlearned and retrained model: $KL(\mathcal{A}(D_r) \mathcal{U}(D_e, \mathcal{A}(D)))$	To evaluate the indistinguishability between model parameters	[13]
JS-Divergence	The distance between the predictions of retrained and unlearned model: $JS(\mathcal{A}, \mathcal{U}) = 0.5 \cdot KL(\mathcal{A} Q) + 0.5 \cdot KL(\mathcal{U} Q)$	To evaluate the indistinguishability between model outputs	[116]
Membership inference	Recall (#detected objects / #erased objects)	To verify if the erased sample is unlearned by the model	[15]
Epistemic uncertainty	$\text{efficacy}(\theta; D) = \begin{cases} \frac{1}{tr(I(\theta; D))}, & \text{if } tr(I(\theta; D)) > 0 \\ \infty, & \text{otherwise} \end{cases}$	To evaluate how much information the model exposes	[117]
EMA	Ensembled Membership Auditing: Ensemble multiple membership metrics and utilizes Kolmogorov-Smirnov (KS) statistical tools to obtain a final auditing score	To verify if a unlearned model memorizes a query dataset	[118]
MIB	Membership Inference via Backdooring: Achieve membership inference for the backdoored data by querying a certain number of black-box queries	To verify if the model unlearns the backdoored data	[61]
Forgetting rate (FR)	$FR = \frac{AF - BF}{BT}$, where AF, BF and BT are defined below	To measure the rate of samples that are modified from member to non-member after unlearning	[64]

$tr(I(\theta; D))$ is the trace of $I(\theta; D)$, and $I(\theta; D)$ is the Fisher Information matrix [117].

to infer if the model remembers the samples of the query dataset. EMA aggregates a final auditing score in the second step by ensembling many membership metrics and statistical tools. They verified whether a trained model forgets query data by the final auditing score. They conducted the experiments using standard datasets, including MNIST, SVHN, and the Chest X-ray dataset.

Inspired by backdoor attacks in ML, Hu et al. [61] proposed Membership Inference via Backdooring (MIB). MIB leverages the property of backdoor attacks that backdoor triggers will misadvise the trained model to predict the backdoored sample to other wrong classes. The main idea of MIB is that the user proactively adds the trigger to her data when publishing them online so that she can implement the backdoor attacks to determine if the model has been trained using her dataset. MIB evaluates the membership inference for the triggered data by calculating the results of a certain number of black-box queries to the targetted model. Sequentially, they observed the results between the targeted model and clean models to infer whether the model was trained based on the backdoored samples.

A similar membership-inference-based method was proposed in [64]. Ma et al. [64] verified the effect of unlearning methods by the proposed forgetting rate (FR) metric. The evaluation metrics were defined using the observation of membership inference. Suppose D_e is the erased dataset; the FR of an unlearning method is denoted as $FR = \frac{AF - BF}{BT}$. In their FR definition, BF and AF are the samples in D_e that are predicted as false by a membership inference attack before and after machine unlearning operations. BT is the sample size in

D_e , which is indicated to be correct by a membership inference attack before machine unlearning. According to the definition, FR presents an instinctive evaluation of how many data points are altered from member to non-member by unlearning. If an unlearning method achieves that $AF > BF$ on the condition that $BT > 0$, it means this method is effective. By contrast, the unlearning will be meaningless. They also proposed an unlearning method called Forsaken and evaluated Forsaken's performance on eight standard benchmark datasets. Forsaken's main idea is to generate masked gradients continuously, apply them to the model's neurons, and model them to forget the contribution of the specific samples. To implement this process, they introduced a mask gradient generator. The proposed mask gradient generator could adjust the volume to prevent unlearning catastrophes by observing the updated model.

Sommer et al. [65] introduced "Athena", which leverages the property of backdooring techniques to verify the effectiveness of unlearning. Athena effectively and confidently certifies whether the data is deleted from an unlearning method. Thus, it provides a basis for quantitatively inferring unlearning. In their backdoor-based verification scheme, they first backdoor users' data and then test the backdoor success probability to infer if the data is unlearned.

6.2 The Employed Datasets

We collect the commonly employed datasets in machine unlearning studies and present the detail introduction of them in Table 4. There are four main types of data: Image, Tabular, Text and Graph. Most of the unlearning studies use image datasets and train classification models based on these image datasets. For tabular datasets, most of them are used in recommendation systems. The unlearning studies that investigate how to unlearn a recommendation model will use these tabular datasets. Graph data is employed for node classification and link prediction tasks, which is usually used in graph unlearning studies. For convenience to find the related studies, we link the corresponding unlearning studies at the last column in Table 4.

7 PRIVACY AND SECURITY THREATS ON MACHINE UNLEARNING

Although unlearning methods were initially proposed to safeguard users' privacy, many researchers have noticed that unlearning brings new privacy threats simultaneously. Similar to machine learning scenarios, there are two main kinds of privacy threats in unlearning. They are membership inference attacks and private information reconstruction attacks. In this section, we detail the two prevalent unlearning privacy threats.

7.1 Privacy Threats on Machine Unlearning

7.1.1 Membership Inference Attacks in Unlearning. Chen et al. [68] first pointed out that when a model is unlearned, the discrepancy in the outputs from the model before and after unlearning leaks privacy of erased data. Then, they proposed the corresponding membership inference attack pipeline in unlearning, which includes three phases: posterior generation, feature construction, and membership inference.

- (1) **Posteriors Generation.** Suppose that the attacker can access two versions of the trained model, the model $\theta_{\mathcal{A}}$ before unlearning and the model $\theta_{\mathcal{U}}$ after unlearning. Assume a target data point e , the attacker queries $\theta_{\mathcal{A}}$ and $\theta_{\mathcal{U}}$, and has the corresponding posteriors, $p(\theta_{\mathcal{A}})$ and $p(\theta_{\mathcal{U}})$, which also called as confidence values in [70].
- (2) **Feature Construction.** After achieving the two posteriors $p(\theta_{\mathcal{A}})$ and $p(\theta_{\mathcal{U}})$, the attacker sums them to make the inference feature vector F . Common methods exist to construct the feature vector shown in [68].
- (3) **Inference.** After the attacker finishes training the attack model based on the created features F , he inputs the collected feature to the inference model to predict if the specific sample e occurs in the erased dataset of unlearning models.

Table 4. The Employed Datasets in Machine Unlearning

Data Type	Name of Datasets	Feature Dimension	#. Smaples	Task Type	Employed by
Image	MNIST [120]	$28 \times 28 \times 1$	70,000	Classification	[38, 59], ...
	CIFAR10	$32 \times 32 \times 3$	60,000	Classification	[37, 38], ...
	CIFAR100	$32 \times 32 \times 3$	60,000	Classification	[37, 43], ...
	SVHN	$32 \times 32 \times 3$	99,289	Classification	[9, 39], ...
	ImageNet [121]	$224 \times 224 \times 3$	1,281,167	Classification	[37, 43], ...
	GTSRB [122]	$32 \times 32 \times 3$	51,839	Classification	[60, 84], ...
	Market-1501 [123]	$128 \times 64 \times 3$	32,668	Person re-identification	[31]
Tabular	Adult	14	48,842	Classification	[40, 68], ...
	Credit info	30	284,807	Classification and anomaly detection	[40, 87], ...
	Covtype	54	581,012		[93, 124], ...
	HIGGS [125]	28	11,000,000	Binary classification	[93, 126], ...
	YELP2018	5	1,561,406	Recommendation	[19]
	Movielens-1m	4	1,000,209	Recommendation	[19]
	Movielens-10m	4	10,000,054	Recommendation	[19]
Text	AG News [127]	3	127,600	Text classification	[23]
	RCV1 [128]	3	804,414	Text classification	[23]
Graph	Amazon Photo [129]	Features per Node: 745	Nodes: 7,650 Edges: 119,081	Node classification Link prediction	[44, 130], ...
	Cora [131]	Features per Node: 1,433	Nodes: 2,708 Edges: 5,429	Node classification Link prediction	[45, 130], ...
	Citseer [132]	Features per Node: 3,703	Nodes: 3,327 Edges: 4,732	Node classification Link prediction	[45]
	Pubmed [133]	Features per Node: 500	Nodes: 19,717 Edges: 44,338	Node classification Link prediction	[44, 45], ...
	ogbn-arxiv [134]	Features per Node: 128	Nodes: 169,343 Edges: 1,166,243	Node classification Link prediction	[44, 130], ...
	Computers [135]	Features per Node: 767	Nodes: 13,752 Edges: 245,861	Node classification Link prediction	[44, 45], ...
	CS [129]	Features per Node: 767	Nodes: 18,333 Edges: 327,476	Node classification Link prediction	[45]
	Physics [129]	Features per Node: Text	Nodes: 27,770 Edges: 352,807	Node classification Link prediction	[45]

In [68], they assumed that the attacker has admission to two ML models before and after unlearning, but it is sometimes impractical, especially in black-box learning scenarios. Lu et al. [71] further proposed a label-only membership inference method to imply if a sample is unlearned, eliminating the dependence on accessing posteriors. Their basic idea is that the same noise injection on candidate data points will show different results for the sample in or not in the training dataset. Thus, they made the adversary continuously query the original and unlearned models and add noise to modify their outputs. Observing the disturbance amplitude lets them determine whether an item is deleted.

Golatkar et al. [72] derived an upper bound to confirm the maximizing knowledge that can be extracted from a black-box model. They queried the model with a picture and obtained the related output. They used the entropy of the result probabilities to construct an effective black-box membership inference [136] attack in machine unlearning.

7.1.2 Privacy Reconstruction Attacks in Unlearning. Privacy reconstruction is another popular attack in machine unlearning. In an unlearning scenario, Gao et al. [74] proposed the deleted reconstruction attacks to recover the removed data from the outputs of the original and unlearning models. In their work, they formalized erasure inference and erasure recovering attacks. The attacker seeks to infer which sample is removed or recovers the erased sample. In particular, for the deletion inference, they formulate the objective of an erasure inference to decide if a data instance e was in or not in the erased dataset, $e \in D_e$ or $e' \notin D_e$. For the deletion reconstruction, they focused on reconstructing the erased example e . In all their reconstruction attacks, the attacker does not have any particular samples, and the purpose is to extract the features knowledge of the erased example. Specifically, the deletion reconstructions include the deleted instance reconstruction and deleted label reconstruction. As named, the deleted instance reconstruction is to extract all of the information of the erased example, and the erased label reconstruction is to infer the label of the erased point in the classification problem.

Zanella-Béguelin et al. [69] indicated that the releasing snapshot of overlapped language models would leak the privacy of the training dataset. They verified that the model updates significantly threaten the private information added to or deleted from the training dataset by many experimental results. Zanella-Béguelin et al. found five phenomenons. First, an attacker can extract particular sentences used or not in the training dataset by comparing two models. Second, analyzing more model snapshots shows more information about the updated data than considering fewer model snapshots. Third, adding or deleting other non-private data during model updates can not mitigate privacy leakage. Fourth, differential privacy can reduce privacy leakage risks and decrease trained models' accuracy. Fifth, to mitigate the privacy leakage risks while keeping the model utility, the server can limit the model parameters access or only output a subset of the results.

Many studies further utilized these privacy threats to evaluate the unlearning effect. Huang et al. [66] proposed Ensembled Membership Auditing (EMA) for auditing data erasure. They use the membership inference to assess the removing effectiveness of unlearning. Graves et al. [15] indicated that if an attacker can infer the sensitive information that was wanted to be erased, then it means that the server has not guarded the rights to be forgotten. Baumhauer et al. [76] developed linear filtration to sanitize classification models with logits prediction after class-wide deletion requests. They verified their methods by testing how well the method defends against privacy attacks.

Both [68] and [69] pointed out that differential privacy guarantees that a model does not reveal too much knowledge about any training sample. A DP-protected model can further guarantee the group's privacy by binding the impacts of a bunch of training samples. If using DP to protect the privacy of a bunch of $|D \setminus D_e|$ training examples against snapshot attacks on θ_D , $\theta_{D \setminus D_e}$, it means that $\theta_{D \setminus D_e}$ cannot be more useful than θ_D .

7.2 Security Threats on Machine Unlearning

There are several papers that address security threats targeting machine unlearning. For instance, Marchant et al. [78] examine an attacker's strategy to increase the computational cost of performing machine unlearning. Studies by Zhao et al. [79] and Hu et al. [81] investigate security vulnerabilities within unlearning systems, specifically by uploading customized malicious data update requests to negatively influence the model utility. Additionally, some studies propose backdooring or poisoning methods that exploit unlearning requests to achieve backdoor insertion, as discussed by [80] and [82]. These works collectively highlight the need for robust security measures in machine unlearning to mitigate these emerging threats.

8 MACHINE UNLEARNING APPLICATIONS

Besides the inherent demands of machine unlearning that draw much research attention, many researchers also find that machine unlearning can be applied in many other scenarios to solve related problems. We list the

Table 5. Machine Unlearning Application

Literature	Application Scenarios	Methods	Realization Methods	Evaluation Metric	Year
[91]	Data "unlearnable"	EMP	Add noise	ACC	2021
[85]	Mitigate backdoor	BAERASER	Gradient ascent method	ASR and ACC	2022
[83]	Mitigate backdoor	-	Median Absolute Deviation	L1-Norm, #FP, ACC, ASR	2019
[87]	Anomaly detection	-	Maintain memory set	#FP and #FN	2019
[86]	Repair pollution	KARMA	Cluster and unlearn	DA	2018

ASR:Attack Success Rate, ACC: Accuracy

DA:Detection Accuracy, EMP: Error-minimizing Perturbations

#FN: number of false negative, #FP: number of false positive

recent unlearning applications in Table 5. The most popular application of machine unlearning is to mitigate the anomalies, including backdoor triggers and pollution from an already-trained model.

Study [83] applied machine unlearning in detecting and mitigating backdoor attacks in DNNs. They designed two approaches to eliminate backdoored neurons from the backdoored model and repair the model to be strong against malicious pictures. Then, in [85], they offered that a backdoor model learned the poisoned decision boundary. Data points with triggers are all classified into the target class. They reversed the backdoor injection process to defend against it in machine unlearning, which is simple but effective. Their method contains two primary steps: first, they utilize a max-entropy staircase approximator to complete trigger reconstruction; second, they remove the added triggers using unlearning. They named these two key steps of BAERASER as trigger pattern recovery and trigger pattern unlearning. Via a dynamic penalty mechanism, they mitigated the sharp accuracy degradation of gradient-ascent-based machine unlearning methods.

Repairing pollution is another successful unlearning application. Cao et al. [86] proposed "KARMA" to search various subsets of original datasets and return the subset with the highest misclassifications. First, KARMA searches for possible reasons that lead to the wrong ML model classification. It clusters the misclassified samples into various domains and extracts the middle of clusters. KARMA prioritizes the search for matching examples in the original datasets using these extracted centers. Second, KARMA grows the reason discovered in the first step by discovering more training samples and creating a cluster. Third, KARMA determines if a causality cluster is polluted and calculates how many samples the cluster contains.

Machine unlearning was also applied to lifelong anomaly detection problems. Du et al. [87] proposed new approaches to solve the challenges of exploding loss and sharp accuracy degradation caused by unlearning. In their presentation, they explain their techniques, emphasizing handling false negatives. When meeting such case x_t , their purpose is, therefore, to modify the model to reduce the possibility of $p(x_t)$, or to increase the loss $\mathcal{L}(x_t)$.

Huang et al. [91] presented a method that can make samples unlearnable by injecting error-minimizing noise. This noise is intentionally synthesized to diminish the error of the samples close to zero, which can mislead the model into considering there is "nothing" to learn. They first tried the error-maximizing noise but found that it could not prevent DNN learning when used sample-wise to the training data points. Therefore, they then begin to study the opposite direction of error-maximizing noise. In particular, they proposed the error-minimizing noise to stop the model from being punished by the loss function during traditional ML model training. Therefore, it can mislead the ML model to consider that there is "nothing" to learn.

9 LESSONS LEARNT AND DISCUSSIONS

In the research domain of machine unlearning, researchers mainly face three difficult challenges, the stochasticity of training, incrementality of training, and catastrophe of unlearning, when they try to design efficient and

effective unlearning mechanisms replacing naive retraining. They tried many mechanisms to mitigate the influence of these challenges. For example, in exact unlearning, they designed split algorithms that divide the final model into several sub-models and avoid the stochasticity and incrementality of one sub-model to influence other sub-models. In approximate unlearning, they bounded the removed estimation to avoid accuracy degradation. However, it is still not easy to analyze them clearly and solve them totally. Moreover, as the research in machine unlearning in-depth, researchers extended it to new situations, such as federated learning and graph learning, and they met corresponding new challenges. Here, we discuss the differences between various unlearning scenarios and their corresponding challenges. Based on these challenges, we list some potential research directions on which we can focus.

In the common centralized unlearning scenario, retraining from scratch can achieve the best unlearning effect, but it is expensive in both computation and storage. Existing methods try to design new unlearning mechanisms to reduce the cost from the two aspects. Although they proposed many methods, they just mitigated the influence of the main challenges and still have not solved them. Existing unlearning methods were at the beginning phase, which is trying to implement effective unlearning. They proposed mechanisms, but most of them cannot guarantee the final results, so they also bound the unlearned item [25] or set an unlearning threshold [13]. Although they also introduced many metrics to verify the unlearning effect, the unlearning and verification processes are split. This means that the unlearning result is uncertain during the unlearning period before the verification is finished. Therefore, we find and list some open questions in centralized unlearning.

- (1) Optimizing the existing challenges. The first two challenges are hard to solve exactly, so recent research focuses on reducing utility degradation after unlearning. Even though many studies exist, there is still a long way to go to eliminate the degradation of unlearning.
- (2) How do we unlearn with a certain or exact goal? In other words, can we unlearn as learning? If we know the purpose, to what extent we have unlearned, and when we can finish or stop the unlearning process.

Machine unlearning in distributed scenarios has many differences from centralized scenarios. We take federated unlearning as a representative example of distributed unlearning. The first difference is that federated unlearning can only be implemented locally on the client's side if they want to unlearn some specific samples because clients do not upload their data to the FL server in a federated scenario. To avoid interacting with clients during unlearning, researchers [56, 60] proposed to unlearn the contribution of a whole client while not some samples of the client. The second difference is that when unlearning requests come during the FL training process, the FL server must first execute the unlearning process and broadcast the unlearned model for later updating to avoid other clients wasting computation on the before-unlearned model. The third difference is that federated unlearning is more vulnerable to catastrophic degradation than centralized unlearning because if the FL server broadcasts the catastrophic unlearned model and other clients update based on the unlearned model, it will vanish the efforts of other clients that trained before. After introducing these differences, we can see that the challenges in federated unlearning are more complex than in centralized unlearning, and we conclude the following open problems in federated unlearning.

- (1) Federated unlearning cannot use the fast retraining methods because data is out of reach for the server. Therefore, federated unlearning can only be implemented using approximate unlearning methods. However, as we know, approximate unlearning easily causes catastrophic unlearning, and federated learning is more vulnerable to degradation, so controlling the catastrophic in federated unlearning will be more urgent than in centralized unlearning.
- (2) Existing federated unlearning methods require the activation of all users, including those without unlearning requests, to assist in the unlearning process. These approaches are impractical and inefficient, particularly when unlearning requests are frequent. Therefore, it is crucial to study how to balance the unlearning effect and efficiency in federated unlearning.

Besides exploring machine unlearning based on regularly structured data, researchers tried to implement unlearning in graph data. Exact unlearning may be suitable for centralized graph unlearning if graph data is sparse. However, the challenges of approximate unlearning may be more difficult than structured-data-based unlearning because, in graph unlearning, the relationship and influence between data samples are more complex than structured data [44, 45]. In particular, graph data includes not only the node feature value but also the connecting edge information. Therefore, in graph unlearning, the estimation of the contribution of a node will be more difficult than in regular data unlearning. The original problems in regular data unlearning will be more challenging in graph unlearning. Besides these problems, graph unlearning also faces unique problems that are related to edge structure information. In graph unlearning, unlearning some edges or sub-graphs is a big question.

After designing machine unlearning algorithms, effective verification and auditing methods are necessary [67]. Most existing unlearning verification methods rely on backdooring techniques. However, these methods inherently degrade model utility because they require mixing backdoored samples into the model training process. Investigating ways to preserve model utility in backdoor-based unlearning verification methods is an under-explored area. Additionally, backdoor-based verification methods must include backdoored samples in the unlearning requests to verify the effectiveness of these requests. This requirement restricts these methods from supporting verification for single-sample unlearning requests. Exploring new strategies that can effectively verify unlearning requests without compromising model utility or being suitable for both single-sample and multi-sample unlearning scenarios remains a significant challenge in this field.

Another important part is the privacy and security issues in machine unlearning. Machine unlearning was first proposed to protect users' privacy, but it brings new threats in that adversaries have a chance to infer the information about the removed data. Literature [68, 69] have pointed out that updates of unlearning will leak privacy information, and they proposed corresponding attacks to infer this private information. However, the most recent unlearning privacy leakage attacks have been similar to those in a learning situation. An attack that is tailored to unlearning mechanisms is expected. A similar situation exists in unlearning applications. Although researchers proposed to unlearn a backdoor trigger [85] or pollution [86], they only used a few unlearning techniques and paid more effort to detect those anomalies. One important reason is that the unlearning mechanism is not mature enough. We are at the beginning of machine unlearning investigation, and there are still many under-explored problems in unlearning itself. Finding machine unlearning applying situation and tailoring unlearning techniques to this situation is the direction of unlearning application.

10 SUMMARY

The survey aims to offer a comprehensive and systematic overview of machine unlearning techniques. We organize the main challenges, research advancements, corresponding techniques, and privacy and security issues in machine unlearning. Additionally, we presented a detailed and unified classification of machine unlearning. We first illustrate the complete unlearning framework, including the learning, request, unlearning, and verification. Then, we briefly categorize recent studies into exact and approximate unlearning and introduce the technical details of corresponding unlearning methods. Moreover, we noticed some new unlearning scenarios, such as federated unlearning and graph unlearning, which are also introduced in the survey. Besides, we consider privacy and security issues in machine unlearning to be an important part of the studies. We collect and summarize the related literature about unlearning privacy threats and applications. Ultimately, the survey provides clear summaries and comparisons between various unlearning scenarios and corresponding methods, giving a comprehensive picture of existing work and listing the challenges and open problems of different scenarios of machine unlearning.

We hope our survey can help classify future unlearning studies, achieve a more in-depth understanding of unlearning methods, and address complex challenges. We believe the open problems listed in Section 9 will still be challenging in the following years, and we will try to optimize some of them. Last but not least, we expect

the survey can help researchers in the study of machine unlearning, regardless of the unlearning strategies or unlearning privacy and security threats or applications of unlearning.

REFERENCES

- [1] Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, and Amit P Sheth. Machine learning for internet of things data analysis: A survey. *Digital Communications and Networks*, 4(3):161–175, 2018.
- [2] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [3] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.
- [4] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8):1–35, 2022.
- [5] Alessandro Mantelero. The EU proposal for a general data protection regulation and the roots of the ‘right to be forgotten’. *Comput. Law Secur. Rev.*, 29(3):229–235, 2013.
- [6] O. of the Privacy Commissioner of Canada. Announcement: Privacy commissioner seeks federal court determination on key issue for canadians’ online reputation.
- [7] Lydia de la Torre. A guide to the california consumer privacy act of 2018. Available at SSRN 3275571, 2018.
- [8] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480. IEEE, 2015.
- [9] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.
- [10] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.
- [11] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, 2017.
- [12] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.
- [13] Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33:16025–16036, 2020.
- [14] Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 792–801, 2021.
- [15] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11516–11524, 2021.
- [16] Sanjam Garg, Shafi Goldwasser, and Prashant Nalini Vasudevan. Formalizing data deletion in the context of the right to be forgotten. In Anne Canteaut and Yuval Ishai, editors, *Advances in Cryptology - EUROCRYPT 2020 - 39th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, May 10-14, 2020, Proceedings, Part II*, volume 12106 of *Lecture Notes in Computer Science*, pages 373–402. Springer, 2020.
- [17] Jonathan Brophy and Daniel Lowd. Machine unlearning for random forests. In *International Conference on Machine Learning*, pages 1092–1104. PMLR, 2021.
- [18] Haonan Yan, Xiaoguang Li, Ziyao Guo, Hui Li, Fenghua Li, and Xiaodong Lin. ARCANe: an efficient architecture for exact machine unlearning. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4006–4013. ijcai.org, 2022.
- [19] Chong Chen, Fei Sun, Min Zhang, and Bolin Ding. Recommendation unlearning. In *Proceedings of the ACM Web Conference 2022*, pages 2768–2777, 2022.
- [20] Sebastian Schelter, Mozhdeh Ariannezhad, and Maarten de Rijke. Forget me now: Fast and exact unlearning in neighborhood-based recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2011–2015, 2023.
- [21] Korbinian Koch and Marcus Soll. No matter how you slice it: Machine unlearning with sisa comes at the expense of minority classes. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 622–637. IEEE, 2023.
- [22] Yuke Hu, Jian Lou, Jiaqi Liu, Feng Lin, Zhan Qin, and Kui Ren. Eraser: Machine unlearning in mlaas via an inference serving-aware approach. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024.
- [23] Yinjun Wu, Edgar Dobriban, and Susan Davidson. Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning*, pages 10355–10366. PMLR, 2020.

- [24] Ga Wu, Masoud Hashemi, and Christopher Srinivasa. PUMA: performance unchanged model augmentation for training data removal. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelfth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 8675–8682. AAAI Press, 2022.
- [25] Chuan Guo, Tom Goldstein, Awini Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3832–3842. PMLR, 2020.
- [26] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34, 2021.
- [27] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerd. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pages 931–962. PMLR, 2021.
- [28] Sebastian Schelter, Stefan Grafberger, and Ted Dunning. Hedgecut: Maintaining randomised trees for low-latency machine unlearning. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1545–1557, 2021.
- [29] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- [30] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312, 2020.
- [31] Ronak Mehta, Sourav Pal, Vikas Singh, and Sathya N Ravi. Deep unlearning via randomized conditionally independent hessians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10422–10431, 2022.
- [32] Yuantong Li, Chi-Hua Wang, and Guang Cheng. Online forgetting process for linear regression models. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 217–225. PMLR, 2021. URL <http://proceedings.mlr.press/v130/li21a.html>.
- [33] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In *30th Annual Network and Distributed System Security Symposium, NDSS 2023, San Diego, California, USA, February 27 - March 3, 2023*. The Internet Society, 2023.
- [34] Quoc Phong Nguyen, Ryutaro Oikawa, Dinil Mon Divakaran, Mun Choon Chan, and Bryan Kian Hsiang Low. Markov chain monte carlo-based machine unlearning: Unlearning what needs to be forgotten. In *ASIA CCS '22: ACM Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May 2022 - 3 June 2022*, pages 351–363. ACM, 2022.
- [35] Shaopeng Fu, Fengxiang He, and Dacheng Tao. Knowledge removal in sampling-based bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.
- [36] Mohammad Emtiyaz Khan and Siddharth Swaroop. Knowledge-adaptation priors. *Advances in Neural Information Processing Systems*, 34:19757–19770, 2021.
- [37] Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [38] Weiqi Wang, Chenhan Zhang, Zhiyi Tian, and Shui Yu. Machine unlearning via representation forgetting with parameter self-sharing. *IEEE Transactions on Information Forensics and Security*, 2023.
- [39] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354, 2023.
- [40] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *31th Annual Network and Distributed System Security Symposium, NDSS 2024*, 2024.
- [41] Junxu Liu, Mingsheng Xue, Jian Lou, Xiaoyu Zhang, Li Xiong, and Zhan Qin. Muter: Machine unlearning on adversarially trained models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4892–4902, 2023.
- [42] Shen Lin, Xiaoyu Zhang, Chenyang Chen, Xiaofeng Chen, and Willy Susilo. Erm-ktp: Knowledge-level machine unlearning via knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20147–20155, 2023.
- [43] Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11186–11194, 2024.
- [44] Eli Chien, Chao Pan, and Olgica Milenkovic. Certified graph unlearning. *arXiv preprint arXiv:2206.09140*, 2022.
- [45] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 499–513, 2022.
- [46] Xunkai Li, Yulin Zhao, Zhengyu Wu, Wentao Zhang, Rong-Hua Li, and Guoren Wang. Towards effective and general graph unlearning via mutual evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13682–13690, 2024.
- [47] Weilin Cong and Mehrdad Mahdavi. Privacy matters! efficient graph representation unlearning with data removal guarantee. 2022.
- [48] Cheng-Long Wang, Mengdi Huai, and Di Wang. Inductive graph unlearning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 3205–3222, 2023.

- [49] Kun Wu, Jie Shen, Yue Ning, Ting Wang, and Wendy Hui Wang. Certified edge unlearning for graph neural networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2606–2617, 2023.
- [50] Chenhan Zhang, Weiqi Wang, Zhiyi Tian, and Shui Yu. Forgetting and remembering are both you need: Balanced graph structure unlearning. *IEEE Transactions on Information Forensics and Security*, 2024.
- [51] Jiancan Wu, Yi Yang, Yuchun Qian, Yongduo Sui, Xiang Wang, and Xiangnan He. Gif: A general graph unlearning strategy via influence function. In *Proceedings of the ACM Web Conference 2023*, pages 651–661, 2023.
- [52] Junxiao Wang, Song Guo, Xin Xie, and Heng Qi. Federated unlearning via class-discriminative pruning. In *Proceedings of the ACM Web Conference 2022*, pages 622–632, 2022.
- [53] Shuyi Wang, Bing Liu, and Guido Zuccon. How to forget clients in federated online learning to rank? In *European Conference on Information Retrieval*, pages 105–121. Springer, 2024.
- [54] Yann Fraboni, Martin Van Waerebeke, Kevin Scaman, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Sifu: Sequential informed federated unlearning for efficient and provable client unlearning in federated optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3457–3465. PMLR, 2024.
- [55] Weiqi Wang, Zhiyi Tian, Chenhan Zhang, An Liu, and Shui Yu. Bfu: Bayesian federated unlearning with parameter self-sharing. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, pages 567–578, 2023.
- [56] Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federaser: Enabling efficient client-level data removal from federated learning models. In *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, pages 1–10. IEEE, 2021.
- [57] Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications, London, United Kingdom, May 2-5, 2022*, pages 1749–1758. IEEE, 2022.
- [58] Lefeng Zhang, Tianqing Zhu, Haibin Zhang, Ping Xiong, and Wanlei Zhou. Fedrecovery: Differentially private machine unlearning for federated learning frameworks. *IEEE Transactions on Information Forensics and Security*, 2023.
- [59] Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34:16319–16330, 2021.
- [60] Chen Wu, Sencun Zhu, and Prasenjit Mitra. Federated unlearning with knowledge distillation. *arXiv preprint arXiv:2201.09441*, 2022.
- [61] Hongsheng Hu, Zoran Salcic, Gillian Dobbie, Jinjun Chen, Lichao Sun, and Xuyun Zhang. Membership inference via backdooring. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 3832–3838. ijcai.org, 2022.
- [62] Yu Guo, Yu Zhao, Saihui Hou, Cong Wang, and Xiaohua Jia. Verifying in the dark: Verifiable machine unlearning by using invisible backdoor triggers. *IEEE Transactions on Information Forensics and Security*, 2023.
- [63] Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE, 2022.
- [64] Zhuo Ma, Yang Liu, Ximeng Liu, Jian Liu, Jianfeng Ma, and Kui Ren. Learn to forget: Machine unlearning via neuron masking. *IEEE Transactions on Dependable and Secure Computing*, 2022.
- [65] David Marco Sommer, Liwei Song, Sameer Wagh, and Prateek Mittal. Athena: Probabilistic verification of machine unlearning. *Proceedings on Privacy Enhancing Technologies*, 3:268–290, 2022.
- [66] Yangsibo Huang, Xiaoxiao Li, and Kai Li. EMA: auditing data removal from trained models. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padov, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part V*, volume 12905 of *Lecture Notes in Computer Science*, pages 793–803. Springer, 2021.
- [67] Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4007–4022, 2022.
- [68] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 896–911, 2021.
- [69] Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. Analyzing information leakage of updates to natural language models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 363–375, 2020.
- [70] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
- [71] Zhaobo Lu, Hai Liang, Minghao Zhao, Qingzhe Lv, Tiancai Liang, and Yilei Wang. Label-only membership inference attacks on machine unlearning without dependence of posteriors. *International Journal of Intelligent Systems*, 2022.
- [72] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pages 383–398. Springer, 2020.

- [73] H. Hu, S. Wang, T. Dong, and M. Xue. Learn what you want to unlearn: Unlearning inversion attacks against machine unlearning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 262–262, Los Alamitos, CA, USA, may 2024. IEEE Computer Society. doi: 10.1109/SP54263.2024.00182.
- [74] Ji Gao, Sanjam Garg, Mohammad Mahmoody, and Prashant Nalini Vasudevan. Deletion inference, reconstruction, and compliance in machine (un)learning. *Proc. Priv. Enhancing Technol.*, 2022(3):415–436, 2022.
- [75] Saurabh Shintre, Kevin A Roundy, and Jasjeet Dhaliwal. Making machine learning forget. In *Annual Privacy Forum*, pages 72–83. Springer, 2019.
- [76] Thomas Baumhauer, Pascal Schöttle, and Matthias Zeppelzauer. Machine unlearning: Linear filtration for logit-based classifiers. *Machine Learning*, pages 1–24, 2022.
- [77] Kaiyue Zhang, Weiqi Wang, Zipei Fan, Xuan Song, and Shui Yu. Conditional matching gan guided reconstruction attack in machine unlearning. In *GLOBECOM 2023-2023 IEEE Global Communications Conference*, pages 44–49. IEEE, 2023.
- [78] Neil G Marchant, Benjamin IP Rubinstein, and Scott Alfeld. Hard to forget: Poisoning attacks on certified machine unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7691–7700, 2022.
- [79] Chenxu Zhao, Wei Qian, Rex Ying, and Mengdi Huai. Static and sequential malicious attacks in the context of selective forgetting. *Advances in Neural Information Processing Systems*, 36, 2024.
- [80] Jimmy Z Di, Jack Douglas, Jayadev Acharya, Gautam Kamath, and Ayush Sekhari. Hidden poison: Machine unlearning enables camouflaged poisoning attacks. In *NeurIPS ML Safety Workshop*, 2022.
- [81] Hongsheng Hu, Shuo Wang, Jiamin Chang, Haonan Zhong, Ruoxi Sun, Shuang Hao, Haojin Zhu, and Minhui Xue. A duty to forget, a right to be assured? exposing vulnerabilities in machine unlearning services. *31th Annual Network and Distributed System Security Symposium, NDSS 2024*.
- [82] Zihao Liu, Tianhao Wang, Mengdi Huai, and Chenglin Miao. Backdoor attacks via machine unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14115–14123, 2024.
- [83] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019.
- [84] Shaokui Wei, Mingda Zhang, Hongyuan Zha, and Baoyuan Wu. Shared adversarial unlearning: Backdoor mitigation by unlearning shared adversarial examples. *Advances in Neural Information Processing Systems*, 36:25876–25909, 2023.
- [85] Yang Liu, Mingyuan Fan, Cen Chen, Ximeng Liu, Zhuo Ma, Li Wang, and Jianfeng Ma. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications, London, United Kingdom, May 2-5, 2022*, pages 280–289. IEEE, 2022.
- [86] Yinzhi Cao, Alexander Fangxiao Yu, Andrew Aday, Eric Stahl, Jon Merwine, and Junfeng Yang. Efficient repair of polluted machine learning systems via causal unlearning. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pages 735–747, 2018.
- [87] Min Du, Zhi Chen, Chang Liu, Rajvardhan Oak, and Dawn Song. Lifelong anomaly detection through unlearning. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1283–1297, 2019.
- [88] Takashi Shibata, Go Irie, Daiki Ikami, and Yu Mitsuzumi. Learning with selective forgetting. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 989–996. ijcai.org, 2021.
- [89] Jingwen Ye, Yifang Fu, Jie Song, Xingyi Yang, Songhua Liu, Xin Jin, Mingli Song, and Xinchao Wang. Learning with recoverable forgetting. In *European Conference on Computer Vision*, pages 87–103. Springer, 2022.
- [90] Christian Ganhör, David Penz, Navid Rekabsaz, Oleg Lesota, and Markus Schedl. Unlearning protected user attributes in recommendations with adversarial training. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2142–2147. ACM, 2022.
- [91] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [92] Salvatore Mercuri, Raad Khraishi, Ramin Okhrati, Devesh Batra, Conor Hamill, Taha Ghasempour, and Andrew Nowlan. An introduction to machine unlearning. *arXiv preprint arXiv:2209.00939*, 2022.
- [93] Ananth Mahadevan and Michael Mathioudakis. Certifiable machine unlearning for linear models. *arXiv preprint arXiv:2106.15093*, 2021.
- [94] Ziyao Liu, Yu Jiang, Jiyuan Shen, Minyi Peng, Kwok-Yan Lam, and Xingliang Yuan. A survey on federated unlearning: Challenges, methods, and future directions. *arXiv preprint arXiv:2310.20448*, 2023.
- [95] Fei Wang, Baochun Li, and Bo Li. Federated unlearning and its privacy threats. *IEEE Network*, 2023.
- [96] Anwar Said, Tyler Derr, Mudassir Shabbir, Waseem Abbas, and Xenofon Koutsoukos. A survey of graph unlearning. *arXiv preprint arXiv:2310.02164*, 2023.

- [97] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.
- [98] Staffs Keele et al. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, ver. 2.3 ebse technical report. ebse, 2007.
- [99] Kateryna Kubrak, Fredrik Milani, Alexander Nolte, and Marlon Dumas. Prescriptive process monitoring: Quo vadis? *PeerJ Computer Science*, 8:e1097, 2022.
- [100] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.
- [101] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.
- [102] James M Joyce. Kullback-leibler divergence. In *International encyclopedia of statistical science*, pages 720–722. Springer, 2011.
- [103] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.
- [104] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014.
- [105] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.
- [106] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon M. Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, August 20–22, 2014*, pages 17–32. USENIX Association, 2014.
- [107] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. Updates-leak: Data set inference and reconstruction attacks in online learning. In *29th USENIX Security Symposium, USENIX Security 2020, August 12–14, 2020*, pages 1291–1308. USENIX Association, 2020.
- [108] Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *International Conference on Machine Learning*, pages 715–724. PMLR, 2020.
- [109] Konstantin E Avrachenkov, Jerzy A Filar, and Phil G Howlett. *Analytic perturbation theory and its applications*. SIAM, 2013.
- [110] Richard H Byrd, Jorge Nocedal, and Robert B Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1):129–156, 1994.
- [111] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.
- [112] Louis Wehenkel and Mania Pavella. Decision trees and transient stability of electric power systems. *Automatica*, 27(1):115–134, 1991.
- [113] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, 2014.
- [114] Chris Bishop. Exact calculation of the hessian matrix for the multilayer perceptron, 1992.
- [115] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.
- [116] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. *arXiv preprint arXiv:2205.08096*, 2022.
- [117] Alexander Becker and Thomas Liebig. Evaluating machine unlearning via epistemic uncertainty. *arXiv preprint arXiv:2208.10836*, 2022.
- [118] Xiao Liu and Sotirios A Tsaftaris. Have you forgotten? a method to assess if machine learning models have forgotten data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 95–105. Springer, 2020.
- [119] Eugene Demidenko. Kolmogorov-smirnov test for image comparison. In *International Conference on Computational Science and Its Applications*, pages 933–939. Springer, 2004.
- [120] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [121] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [122] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011.
- [123] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.
- [124] Zhaomin Wu, Junhui Zhu, Qinbin Li, and Bingsheng He. Deltaboost: Gradient boosting decision trees with efficient machine unlearning. *Proceedings of the ACM on Management of Data*, 1(2):1–26, 2023.
- [125] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):1–9, 2014.

- [126] Ananth Mahadevan and Michael Mathioudakis. Certifiable unlearning pipelines for logistic regression: An experimental study. *Machine Learning and Knowledge Extraction*, 4(3):591–620, 2022.
- [127] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.
- [128] YYRTG Glewis, D David, and F Li. A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 2004.
- [129] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.
- [130] Eli Chien, Chao Pan, and Olgica Milenkovic. Efficient model updates for approximate unlearning of graph-structured data. In *The Eleventh International Conference on Learning Representations*, 2022.
- [131] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.
- [132] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pages 40–48. PMLR, 2016.
- [133] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.
- [134] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.
- [135] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.
- [136] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.