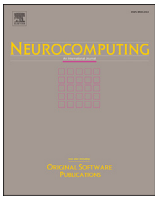
Neurocomputing 662 (2026) 131971

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Survey paper

How secure is forgetting? Linking machine unlearning to machine learning attacks

Muhammed Shafi K.P. [c], Serena Nicolazzo [a,*], Antonino Nocera [b], Vinod P. [c]

[a] Department of Science, Technology and Innovation, University of Eastern Piedmont, V.le Teresa Michel, 11, Alessandria, 15121, AL, Italy
[b] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, A. Ferrata, 5, Pavia, 27100, PV, Italy
[c] Department of Computer Applications, Cochin University of Science and Technology, Kerala, India

ARTICLE INFO

Communicated by X. Wang

Keywords:
Machine unlearning
Security
Backdoor attack
Membership inference attack
Adversarial attack
Inversion attack
Machine learning

ABSTRACT

As Machine Learning (ML) continues to evolve, so does the sophistication of security threats targeting data privacy and model integrity. In response, Machine Unlearning (MU) has emerged as a promising paradigm that enables the selective removal of data influence from trained models. By supporting compliance with privacy regulations (such as the GDPR's right to be forgotten) and facilitating model refinement, MU holds significant practical and legal value. Additionally, the effective deployment of MU introduces new security concerns. In real-world settings, malicious actors may exploit vulnerabilities in MU mechanisms, such as incomplete or inaccurate data removal, to infer deleted information, reintroduce adversarial behavior, or manipulate model updates. These risks highlight the urgency of understanding how classical ML threats relate to the design and operation of MU systems. However, despite its growing relevance, this intersection remains underexplored. In this article, we present a structured analysis of four major attack classes in ML (Backdoor Attacks, Membership Inference Attacks, Adversarial Attacks, and Inversion Attacks) and examine their implications for MU across multiple dimensions: *(i)* as direct threats targeting MU mechanisms, *(ii)* as challenges that MU can potentially mitigate, *(iii)* as evaluation metrics to measure the effectiveness and performance of MU techniques, and *(iv)* as verification factors to validate the success and completeness of the unlearning process. We note that not all attacks exhibit all these perspectives simultaneously; their relevance varies depending on the attack characteristics and MU scenario. We also propose a novel classification that reflects how these attacks are typically employed in this context. Finally, we identify open challenges, including ethical considerations, and highlight promising directions for future research to advance secure and privacy-preserving Machine Unlearning.

1. Introduction

Machine Unlearning (MU, hereafter) refers to the process of removing the influence of specific data points from a trained Machine Learning (ML) model while maintaining the model's performance and efficiency. Practitioners may apply MU for various reasons, including privacy compliance, elimination of biased data, or removal of redundant training samples. Among these, privacy compliance is particularly critical due to legal provisions such as the Right to Be Forgotten (RTBF), established by regulations like the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the US. In the ML context, RTBF requires not only the deletion of personal data but also the removal of its influence on the model, which MU techniques are designed to achieve [3,52]. Researchers employ Machine Unlearning (MU) techniques for several reasons: *(i)* to remove mislabeled, outdated, or biased data in order to improve model fairness, and *(ii)* to eliminate redundant training samples to reduce model complexity. A straightforward approach to unlearning involves retraining the model from scratch without the targeted data. However, this process is computationally intensive and often impractical for real-world applications [3]. To address this, researchers are developing efficient MU methods that update models without full retraining. These techniques are especially relevant as ML models are increasingly deployed in critical applications demanding high levels of trust and reliability. While MU enhances data control and regulatory compliance, classical vulnerabilities in ML—such as adversarial exploitation, data leakage,

* Corresponding author.
Email addresses: shafikp@cusat.ac.in (M.S. K.P.), serena.nicolazzo@uniupo.it (S. Nicolazzo), antonino.nocera@unipv.it (A. Nocera), vinod.p@cusat.ac.in (V. P.).

https://doi.org/10.1016/j.neucom.2025.131971
Received 11 April 2025; Received in revised form 26 August 2025; Accepted 29 October 2025
Available online 1 November 2025
0925-2312/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

and model integrity threats—remain underexplored in this context and require careful analysis to ensure the trustworthiness of MU systems [13,40]. In light of these challenges, an exhaustive examination of existing research on the implications of ML threats and MU is essential. In this article, we comprehensively analyze the classical attacks on security and privacy in ML and their intersection with MU systems. We provide a novel perspective that focuses on the categorization of the literature according to the main known attacks and how they are used in the context of MU. Although some recent survey contributions have begun to examine MU approaches, most of the existing works focus on MU fundamental concepts and methodologies [39,53,55,67]. Other contributions deal with the evolution of MU in Federated Learning settings, namely Federated Unlearning (FU), which has emerged to confront the challenge of data erasure in distributed learning environments [46,51]. In addition, the authors of [2] survey MU methods for LLMs. In contrast, the works presented in [7,47] adopt a different perspective; in particular, the former focuses on privacy risks associated with the adoption of MU exploring existing countermeasures for model protection from malicious unlearning-based attacks, and the latter, instead, deals only with an analysis of current threats and defenses in MU.

To the best of our knowledge, this paper presents the first comprehensive analysis of key vulnerabilities in Machine Learning (ML) and their relationship with Machine Unlearning (MU) solutions. We began by selecting relevant studies that jointly examine MU and ML attacks. Through this examination, we identified four major types of attacks commonly used against ML systems: Backdoor Attacks, Membership Inference Attacks, Adversarial Attacks, and Inversion Attacks. Our analysis revealed four distinct types of relationships between these attacks and MU: *(i)* attackers can directly target MU systems using known ML attacks, *(ii)* MU can serve as a defense mechanism to mitigate such attacks, *(iii)* researchers can intentionally inject ML attacks to evaluate the effectiveness of MU frameworks (i.e., whether the unlearning algorithm can successfully remove the attack), and *(iv)* an ML attack can be exploited as a tool for evaluating a new MU verification approach (i.e., the MU method processes deletion requests following legal obligations).

In this study, we conducted a comprehensive search for publications related to the aforementioned topics. Specifically, our review focused on journal and conference papers published in recent years, sourced from Google Scholar. Additionally, we examined Gray Literature, including white papers and government reports. Our selection criteria prioritized Q1 journals, A*, and A-ranked conference papers. We excluded duplicate entries and non-English papers. We analyzed a total of 61 works, selecting 42 papers from 2021 to 2025.

Our systematic categorization aims to identify existing gaps in the interplay between attacks on ML and MU systems, encouraging researchers to improve both the robustness and reliability of current defense strategies and to find new solutions. The main contributions of this paper are as follows:

- We provide a comprehensive overview of Machine Unlearning, covering recent techniques and the evaluation metrics commonly used to assess their effectiveness.
- We introduce a taxonomy that systematically categorizes existing attacks in the context of ML and their implications for Machine Unlearning solutions.
- We present a comprehensive analysis of the challenges associated with security threats in MU. In addition, we highlight critical areas for future research and development in this field.

The outline of this paper is as follows. In Section 2, we present the essential background of MU necessary to understand our contribution. In Section 3, we examine works dealing with ML attacks and MU systems and propose a novel classification for them. In Section 4, we propose a comparison of the performance of attacks and defenses across MU methods. Section 5 presents challenges and limitations that could lead to future work. Finally, Section 6 draws our conclusions.

Table 1
Summary of the acronyms used in the paper.

Symbol	Description
ATMs	Adversarial Training Models
BA	Backdoor Attack
CMU	Centralized Machine Unlearning
DNNs	Deep Neural Networks
DL	Deep Learning
FL	Federated Learning
FU	Federated Unlearning
GIA	Gradient Inversion Attack
IA	Inversion Attack
MIA	Membership Inference Attack
ML	Machine Learning
MLaaS	ML-as-a-Service
MoIA	Model Inversion Attack
MU	Machine Unlearning

2. Background on machine unlearning

This section provides the essential background information to understand the key concepts discussed in this paper. In particular, it offers a concise overview of Machine Unlearning categories, key techniques, and common metrics used to evaluate approaches in this context. Moreover, Table 1 summarizes the acronyms used in this paper.

2.1. Definition and classification

Machine Unlearning (MU) refers to the process by which a system removes the influence of previously learned data that was incorporated through an ML algorithm [4]. Ideally, the model completely forgets the data points that contributed to its training, effectively eliminating their impact, so that if the same data point is reintroduced in the future, the system processes it as if it were entirely new, without any residual knowledge from prior learning. Formally, let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the original and complete training dataset, where x_i is the input feature and y_i the corresponding label, and let M be the ML model trained on D. Let d_u represent a set of data points to be removed so that $D' \subseteq D$ and $D' = D - d_u$. Once the unlearning process is triggered, the model must delete all information related to d_u. The resulting unlearned model, M_u, should be indistinguishable from a reference model M' retrained from scratch using only D'. Fig. 1 illustrates the MU process, which removes specific data points from a trained ML model while preserving its overall functionality.

MU techniques can be classified according to the exactness of the obtained unlearning in the two following typologies [39,53]:

- **Exact Unlearning.** Perfect or exact unlearning algorithms aim to produce a model that is identical to one trained from scratch on a dataset, excluding a specific data point that needs to be unlearned. As a result, retraining the model from the beginning is currently considered the only true, exact unlearning method. Even if this method gives strong privacy and security assurance, retraining is resource-intensive and may not be feasible for large-scale models. In general, these techniques focus on developing a modular machine learning system, where individual components are trained using disjoint subsets of the data.
- **Approximate Unlearning.** Approximate unlearning offers a more cost-effective alternative and is particularly beneficial for complex and adaptive ML algorithms, where reconstructing the precise sequence and impact of individual data points is often infeasible.

An additional MU classification relies on the level of guarantee that the specific data points have been completely erased from the model. Based on this difference, we can distinguish the following types of MU methods:

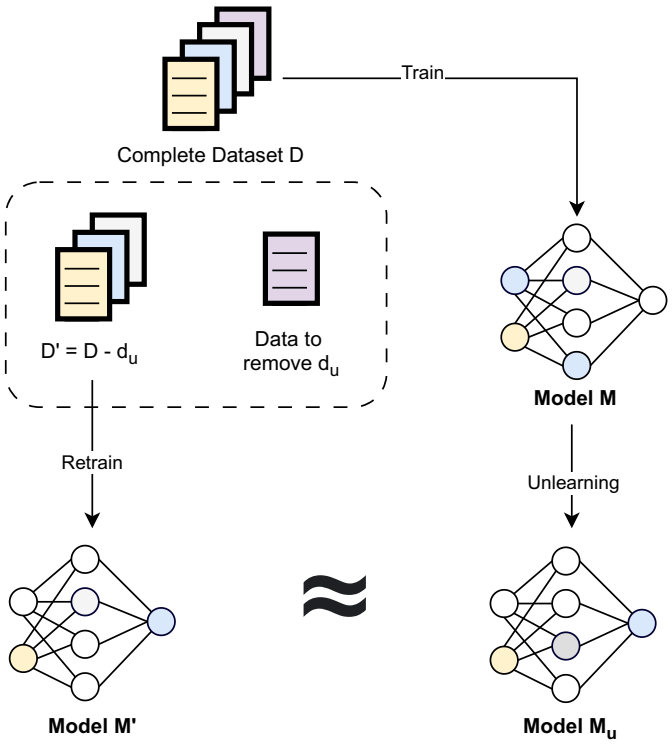


Fig. 1. Machine unlearning workflow.

- **Certified MU** that gives the formal guarantee (via mathematical or cryptographic proofs) that a model from which data is removed cannot be distinguished from a model that never observed the data to begin with [26]. They are not necessarily obtained by retraining. In particular, the MU algorithm satisfies a definition of approximate unlearning called certified removal, where indistinguishability is defined similarly as (ϵ, δ)-differential privacy [18]. These methods provide formal guarantees via proofs, bounds, or cryptographic constructs [49].
- **Empirical MU** lacks formal theoretical guarantees for security or correctness. Instead, it validates unlearning effectiveness through experimental evaluation and empirical observations. Owing to its practical feasibility and lower computational overhead, empirical MU is widely adopted in real-world deployments.

Another perspective for classifying Machine Unlearning (MU) focuses on how the influence of data is removed from a trained model, as discussed in [53]. This classification includes three main types: instance-level, class-level, and feature-level unlearning. The most common approach is **instance-level unlearning**, which targets the removal of specific data points from a trained model [3]. This method is typically employed for privacy compliance. In some cases, it becomes necessary to remove all data points associated with a specific class or category. For example, in face recognition applications, each individual's face is treated as a separate class. If a user opts out of the service and requests the deletion of their facial data, the system must perform **class-level unlearning**. A third category, **feature-level unlearning**, aims to eliminate the influence of specific features while retaining the rest. This type of unlearning is particularly resource-intensive, as it often requires training a model capable of identifying and removing the targeted feature, which typically demands access to a large volume of data [63].

Alternatively, a classification approach is based on the underlying learning paradigm. In this context, we differentiate between **Centralized Machine Unlearning (CMU)** and **Federated Unlearning (FU)**. In CMU, unlearning is applied to a single model maintained at a central location. In contrast, FU operates within the Federated Learning (FL) framework, enabling the removal of a client's contribution or the information derived from its local data, while preserving the privacy guarantees inherent to decentralized learning. This paradigm introduces unique targets and presents distinct technical challenges [46].

2.2. Key techniques

A trivial technique for MU is *Retraining from Scratch* in which the model is retrained from the beginning on the dataset excluding the data that needs to be forgotten. Although it is computationally expensive and impractical for large-scale models it provides an exact and certified method for unlearning. An improved approach that can also be categorized as an exact method is the *Sharded or Partitioned Training*, which divides the data into disjoint fragments and trains a model on each smaller data fragment. In this way, the training cost can be distributed [3].

Apart from the above-mentioned exact methods, there exist approximate techniques that modify the trained model without full retraining, achieving efficient but non-guaranteed unlearning. The following approaches can be listed among these approximate methods [39]:

- **Influence Function-Based Unlearning** relies on influence functions to quantify the impact of individual data points on the learned model. These methods estimate how the model would change if specific data were removed, enabling efficient unlearning by updating the model based on these estimates [75].
- **Knowledge Distillation-Based Unlearning** leverages the concept of knowledge distillation, where a student model (the unlearned model) is trained to selectively replicate the knowledge of a larger teacher model (the original model). This approach enables the removal of sensitive information associated with the data to be forgotten while retaining the overall utility and effectiveness of the student model. Recent work in [38] mitigated Electric Vehicle user data exposure by training a teacher on the full dataset and a student to forget sensitive data using a dual-term loss that preserves performance and model similarity.
- **Gradient-based Unlearning** approximates the retrained model by correcting the Stochastic Gradient Descent (SGD) steps. It leverages the gradients (i.e., parameter updates) computed during training to estimate and reverse the impact of the data to be forgotten.
- **Federated Unlearning (FU)** enables the selective removal of a client's data influence from a Federated Learning (FL) model without requiring full global retraining. FU can be implemented in two ways: (i) on the **server side**, where the server unlearns the global model and redistributes the updated version to all clients, or (ii) on the **client side**, where each client removes the influence of its own data from the global model before uploading updates back to the server [46].

2.3. Evaluation metrics

Evaluation metrics allow model providers to measure the effectiveness, utility, and efficiency of their unlearning processes, facilitating the optimization of unlearning algorithms for improved performance. Widely adopted metrics include the following:

- **Forgetting Rate (FR)** measures the reduction in model performance on the removed data. It can be computed as:

$$FR = 1 - \frac{A_{after}}{A_{before}}$$

  where A_{after} is the accuracy on removed data after Unlearning; and A_{before} is the accuracy on removed data before Unlearning [48].
- **Model Distance Metrics** compares the difference between the unlearned model and a model retrained from scratch without the removed data.

- **Membership Inference Resistance**. Since Membership inference attacks identify a given data sample in the training dataset, if an attack still recognizes the unlearned data as a member after the unlearning, it means that the unlearning process has failed [40].
- **Accuracy Drop (AD)** measures the difference in accuracy before and after unlearning the remaining data. It quantifies the degradation introduced by the forget set on the unlearned model and it is computed through the formula:

$$AD = \frac{|A_{before} - A_{after}|}{A_{before}}$$

where A_{after} is the accuracy on removed data after Unlearning; and A_{before} is the accuracy on removed data before Unlearning.
- **Unlearning Time (UT)** measures the time (expressed in seconds or number of epochs) required to forget specific data points.

In addition to the previous metrics, in the context of security of unlearning, the Attack Success Rate is also typically adopted.

- **Attack Success Rate (ASR)** measures the effectiveness of attacks (e.g., adversarial unlearning, re-insertion attacks) against the MU system. In particular, for a Backdoor Attack, it quantifies the ratio of poisoned data that are misclassified into the target label desired by the attacker [44].

2.4. Application of MU

Machine Unlearning was first developed as a means to protect user privacy, particularly in simple classification settings. Over time, its scope has expanded, revealing strong potential not only for privacy preservation but also for enhancing model robustness across a wide range of application domains. MU holds significant promise in domains where privacy breaches can have severe consequences, namely:

- **Recommendation Systems**. They serve as personalized information filters that analyze user preferences from collected data to suggest the most relevant items. However, during training, the model's parameters may inadvertently memorize user behaviors, creating potential privacy risks. To address this, recommendation unlearning has been introduced, allowing specific data and personal preferences to be effectively erased from the model [8]. Different approaches were experimented with in integrating recommendation systems and MU. Yuan et al. [72] focused on federated recommendation systems and, drawing inspiration from the log-based rollback mechanism in transactions, proposed a method to efficiently remove a user's contribution from the federated training process. This approach not only enhances the model's robustness but also improves its resilience against potential attacks from malicious clients. Chen et al. [8] leveraged data similarity to partition the training set into balanced groups and, based on this structure, proposed an unlearning algorithm specifically designed for recommender systems. Their approach preserves the collaborative information within the data while enhancing the model's usability, security, and overall applicability. Ganhör et al. [20] focused on mitigating demographic bias in collaborative filtering systems and, employing the principles of adversarial training, proposed the ADV-MULTVAE model to unlearn protected user attributes from the learned interaction representations. This method forces the model's latent embeddings to become invariant with respect to protected attributes, such as gender. The approach aims to reduce inherent model biases and address privacy concerns regarding the disclosure of user attributes while largely preserving the system's recommendation performance.
- **LLMs.** Recent advancements in large language models (LLMs) have attracted widespread attention, driving the development of various commercial applications. From ChatGPT [58] to DALL-E, these models have progressed well beyond traditional text-focused tasks, extending their functionality to complex applications such as code generation [39]. However, their tendency to memorize and regenerate uncensored information introduces serious ethical and legal challenges, including risks of revealing biased, sensitive, or copyright-protected content. Furthermore, LLMs are vulnerable to jailbreak techniques and may be misused for harmful purposes, such as facilitating the creation of chemical, biological, radiological, and nuclear (CBRN) weapons [68]. Given the high computational and financial costs associated with retraining, rebuilding models to eliminate such risks is impractical [43]. As an alternative, MU offers a practical solution by selectively removing undesirable knowledge from pretrained or fine-tuned models, thereby helping them adapt to evolving societal standards. Despite the fact that MU has been extensively applied in classification problems, its use in generative settings within LLMs remains relatively underexplored. In addition to addressing privacy concerns, the application of MU in LLMs is driven by several other compelling factors.

    1. ***Mitigating hallucinations:*** an "Hallucination" arises when an LLM produces outputs that are illogical or misleading relative to the given prompt or its knowledge base. It does not necessarily mean the information is factually incorrect, but rather that the response deviates from the expected or intended content [60]. These undermine the reliability of AI in high-stakes domains, where fabricated information (e.g., in medical advice) could have serious consequences. Hallucinations often stem from both data misalignment and risky decoding strategies. By selectively unlearning factually incorrect responses to certain queries, models can be made more trustworthy and robust in practical applications.
    2. ***Security and value alignment:*** Large language models (LLMs) are generally trained on massive, unfiltered datasets that aggregate text from diverse online sources. While this enables them to acquire broad linguistic knowledge, it also exposes them to biased, discriminatory, or otherwise harmful content relating to sensitive attributes such as gender, ethnicity, religion, and sexual orientation. Consequently, these models risk reproducing or amplifying stereotypes and prejudiced associations. For example, research shows that an LLM might assert that "Men are fit for engineering, while women are for nursing," thereby reinforcing socially damaging narratives [2]. Such outputs are not only ethically problematic but also pose societal risks when LLMs are deployed in decision-making contexts, ranging from education and recruitment to healthcare and governance. MU has emerged as a promising approach in this regard. Unlike methods that attempt to correct bias through post-hoc filtering or Reinforcement Learning from Human Feedback (RLHF) [6], unlearning operates by selectively removing undesirable knowledge patterns from within the model itself. Moreover, recent research demonstrates that MU can achieve stronger alignment with human values compared to RLHF, all while offering significant gains in computational efficiency [69].
    3. ***Copyright protection:*** Since LLMs are often trained on copyrighted material (e.g., books, news articles, or other media), they can reproduce or closely mimic such content, raising serious ethical and legal challenges. This issue has already prompted lawsuits and regulatory scrutiny. For instance, The New York Times filed a case against OpenAI and Microsoft, claiming that millions of its articles were used without consent to train models such as ChatGPT, which then replicated portions of its content and style.[1]

[1] https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

Through unlearning, copyrighted data can be selectively removed from models without undergoing full retraining [39]. This aligns copyright protection with privacy preservation, since both aim to restrict the unintended disclosure or replication of protected material.

- **Finance and Business**. In financial services, rapid forgetting of compromised or outdated customer data can mitigate insider threats and reduce regulatory risk under frameworks like GDPR and CCPA. Additionally, models are required to comply with strict privacy regulations while still performing effectively in tasks such as credit risk assessment. Exact unlearning plays a crucial role here, as it guarantees the thorough removal of sensitive data without diminishing the overall performance of the model [39]. Moreover, in the business context, ML algorithms now play a pivotal role in the research process, particularly when companies set out to design new products. With their extensive use, ML has become a high-profile, high-tech discipline, giving rise to increasingly complex models whose reliability and robustness are still not fully validated. In this setting, MU can help by simplifying these models, reducing unnecessary complexity, and enabling the research community to work with more refined and dependable systems [53].
- **Healthcare**. Hospitals are increasingly adopting ML algorithms for patient diagnosis. However, errors arising from human oversight or algorithmic limitations may result in patients being incorrectly diagnosed with conditions they do not actually have. In such cases, it is crucial to remove these errors from the patient's medical records. MU offers an efficient solution by enabling the removal of such sensitive records from diagnostic models without requiring costly retraining, thereby ensuring both privacy and regulatory compliance [53]. Given the highly sensitive nature of patient data, FL allows multiple hospitals to collaboratively train a global model without directly sharing raw data. To ensure complete removal of specific information, local-side unlearning can be employed, enabling each hospital to selectively unlearn data from its own instance of the model [39].

3. Machine unlearning and security attacks

This section systematically explores the most popular security threats in ML and their relationship with Machine Unlearning. We examine all the threats used in the papers we selected, namely Backdoor Attacks (BA), Membership Inference Attacks (MIA), Adversarial Attacks (AA), and Inversion Attacks (IA). While other types of attacks, such as unlearning attacks in regression learning [10], have also been explored, they are relatively less prevalent in the security landscape compared to the widely studied attacks in classification settings. Furthermore, we adopt a categorization that highlights the relationship between these attacks and MU based on our analysis of the existing literature. Table 2 lists all the analyzed works according to the attacks above. In addition, we report the scope of each paper. In particular, we identified five possible main scopes for the analyzed documents, namely: *(i)* they define a new MU method; *(ii)* they propose a variation of a known ML attack; *(iii)* they describe a possible defense against a known ML attack leveraging MU; *(iv)* they design a new metrics to evaluate an MU method; *(v)* they define a new MU verification method. Observe that a paper may address multiple attacks or have a broader scope (for example, it might introduce a new MU method in addition to a verification technique, or discuss both an attack and a defense). In such cases, we cite the paper in all relevant sections.

Fig. 2 represents our adopted classification based on the four known ML attacks present in the literature. In particular, we present Backdoor Attack, Membership Inference Attack, Adversarial Attack, and Inversion Attack along with the related papers we found in the literature referring to a particular aspect of MU and these attacks.

In the next subsections, we start by providing some background on BA and then proceed to discuss how BA is related to MU according to the analyzed literature.

3.1. Backdoor attacks

A Backdoor Attack in the context of Machine Learning involves an adversary embedding a hidden mechanism within a model during the training process. This mechanism allows the attacker to manipulate the model's behavior in specific ways when triggered by particular input patterns, which are often inconspicuous or seemingly benign. A Backdoor Attack introduces a malicious perturbation Δ, causing misclassification only when the trigger is present while maintaining normal behavior otherwise [44]:

$$F(X) = y_i, \quad \forall X \sim \mathcal{D}_{\text{clean}} \tag{1}$$

$$F(X + \Delta) = y_t, \quad \forall (X + \Delta) \sim \mathcal{D}_{\text{backdoor}} \tag{2}$$

where X represents the clean input data, and Δ is the trigger pattern introduced by the attacker. The true label of the clean input is denoted as y_i, while y_t represents the attacker's target label. The clean data is drawn from the distribution $\mathcal{D}_{\text{clean}}$, whereas the manipulated (backdoored) data comes from the distribution $\mathcal{D}_{\text{backdoor}}$. This definition highlights the stealthy nature of Backdoor Attacks. The model functions normally on clean data, correctly classifying it as y_i. However, when the adversarial trigger Δ is added to the input, the model misclassifies it as the target label y_t, allowing the attacker to manipulate predictions without affecting the overall performance on clean samples.

A typical Backdoor Attack consists of the following steps:

1. **Poisoning Phase.** The attacker injects poisoned samples $(X+\Delta, y_t)$ into the training dataset. These samples contain a specific trigger Δ and are labeled as the target class y_t. The goal is to implant the backdoor into the model during training.
2. **Model Training.** The poisoned dataset, which includes both clean and manipulated samples, is used to train the model. The model learns to classify clean samples correctly but also associates the trigger Δ with the target label y_t.
3. **Deployment.** The trained model is deployed and behaves normally on clean data, classifying inputs correctly as y_i.
4. **Attack Execution.** When an attacker provides an input $X+\Delta$ containing the trigger, the model misclassifies it as the target label y_t. This allows the attacker to control the model's behavior on specific inputs without significantly affecting overall accuracy.

Backdoor Attacks are a specific subclass of Data Poisoning Attacks where the attacker implants a hidden pattern (trigger) into a subset of training samples and assigns them a target label. In contrast, Availability Attacks (also known as Indiscriminate Poisoning Attacks) aim to degrade the model's overall performance rather than targeting specific inputs. These attacks modify the training data in a way that reduces accuracy across all test samples, causing widespread failure instead of targeted misclassification [49]. Unlike Backdoor Attacks, availability attacks do not rely on hidden triggers but instead corrupt the model's decision boundaries, making it unreliable.

Current Backdoor defense strategies are broadly classified into Backdoor detection and Backdoor erasing [44]. Detection methods identify whether a model or dataset contains backdoors but do not neutralize them. Erasing techniques aim to mitigate Backdoor effects while preserving model accuracy. Fine-tuning with clean data offers a straightforward but weak defense, whereas fine-pruning removes trigger-activated neurons at the cost of performance degradation. More advanced approaches, such as Knowledge Distillation, seek to transfer clean model behavior to the compromised model, offering a more effective defense mechanism.

Table 2
Analyzed papers dealing with ML attacks and MU. (BA—Backdoor Attack, MIA—Membership Inference Attack, AA—Adversarial Attack, and IA—Inversion Attack.).

Ref.	Year	Attacks				Paper Main Scope				
		BA	MIA	AA	IA	MU Method	Attack	Defense	Evaluation Metric	MU Verification Method
Chen et al. [13]	2021	–	✓	–	–	–	✓	✓	✓	–
Golatkar et al. [22]	2021	–	✓	–	–	✓	–	–	–	–
Graves et al. [25]	2021	–	✓	–	✓	✓	–	–	–	–
Gupta et al. [28]	2021	–	–	✓	–	✓	✓	✓	–	–
Liu et al. [44]	2022	✓	–	–	–	–	–	✓	–	–
Ma et al. [48]	2022	✓	✓	–	–	✓	–	–	✓	–
Marchant et al. [49]	2022	✓	–	–	–	–	✓	–	–	–
Sommer et al. [57]	2022	✓	✓	–	–	–	–	–	–	✓
Chundawat et al. [16]	2023	–	✓	–	✓	✓	–	–	✓	–
Guo et al. [27]	2023	✓	–	–	–	–	–	–	–	✓
Jia et al. [34]	2023	✓	–	–	–	✓	–	–	–	–
Kurmanji et al. [37]	2023	–	✓	–	–	✓	–	–	–	–
Liu et al. [42]	2023	–	–	✓	–	✓	–	✓	–	–
Li et al. [40]	2023	✓	–	–	–	✓	✓	✓	–	–
Wei et al. [64]	2023	✓	–	–	–	–	–	✓	–	–
Zhang et al. [73]	2023	–	✓	–	–	✓	–	–	–	–
Zhao et al. [77]	2023	–	–	✓	–	–	✓	–	–	–
Daluwatta et al. [17]	2024	✓	–	–	–	–	–	✓	–	–
Chen et al. [11]	2024	–	✓	–	–	✓	–	–	–	–
Chen et al. [12]	2024	✓	✓	–	–	✓	–	–	✓	–
Hu et al. [31]	2024	–	–	✓	–	–	✓	–	–	–
Huang et al. [33]	2024	✓	–	–	–	–	✓	–	–	–
Gao et al. [21]	2024	–	–	–	✓	–	–	✓	–	–
Hu et al. [32]	2024	–	–	–	✓	–	✓	✓	–	–
Jiang et al. [35]	2024	✓	✓	–	–	✓	–	–	–	–
Liu et al. [45]	2024	✓	–	–	–	–	✓	–	–	–
Li et al. [41]	2024	✓	–	–	–	–	–	✓	–	–
Zhao et al. [80]	2024	✓	–	–	–	–	–	✓	–	–
Niu et al. [50]	2024	✓	–	✓	–	–	–	✓	–	–
Wu et al. [66]	2024	✓	–	–	–	–	–	✓	–	–
Chen et al. [9]	2025	✓	–	–	–	–	✓	✓	–	–
Han et al. [29]	2025	✓	✓	–	–	✓	✓	–	–	–
Varshney and Torra [59]	2025	–	✓	–	–	✓	–	–	–	–
Zhao et al. [79]	2025	✓	✓	–	–	✓	–	✓	–	–
Wang et al. [62]	2025	✓	✓	–	–	✓	–	✓	–	–

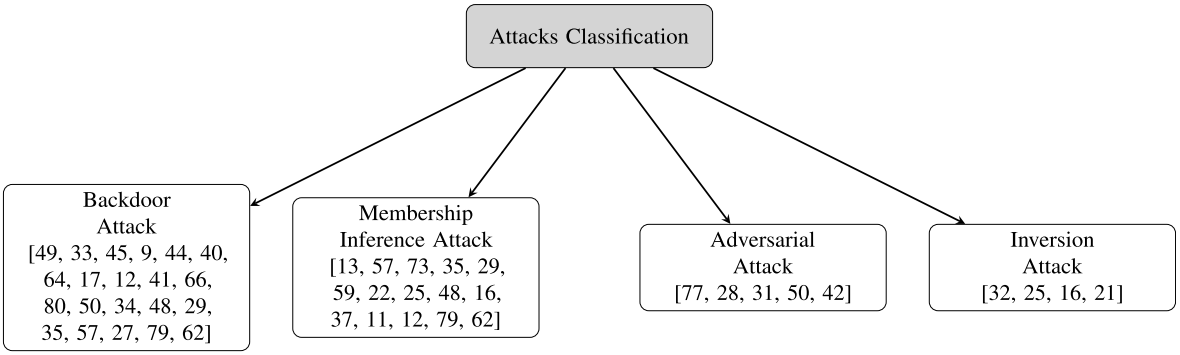


Fig. 2. Categorization of papers into ML Attacks.

After reviewing the literature related to Machine Unlearning and Backdoor Attacks, we find the following categorization based on the scope of each work, as also visible in Fig. 3:

- Backdoor Attack against MU [9,33,45,49];
- MU as a defense (or erasure strategy) against Backdoor Attack [17, 40,44,50,64,80];
- Backdoor Attack as an evaluation tool to test new MU frameworks [29,34,35,48,62,79];
- Backdoor Attack as a tool for a new MU verification framework [27, 29,57].

In particular, Fig. 3 illustrates the above categorization of papers addressing Backdoor Attacks, further dividing these approaches according to unlearning architectures, namely Central MU (CMU) and Federated Unlearning (FU).

Table 3 illustrates the papers dealing with MU and Backdoor Attacks specifying *(i)* the Unlearning paradigm used between Central Machine Unlearning (CMU) and Federated Unlearning (FU), *(ii)* whether the proposed Unlearning technique concerns the instance or the class, *(iii)* the attacker knowledge (black-box, gray-box or white-box), *(iv)* the attack/defense phase (during the training or before, during or post the Unlearning), *(v)* the employed defense techniques, and the evaluation metrics used.

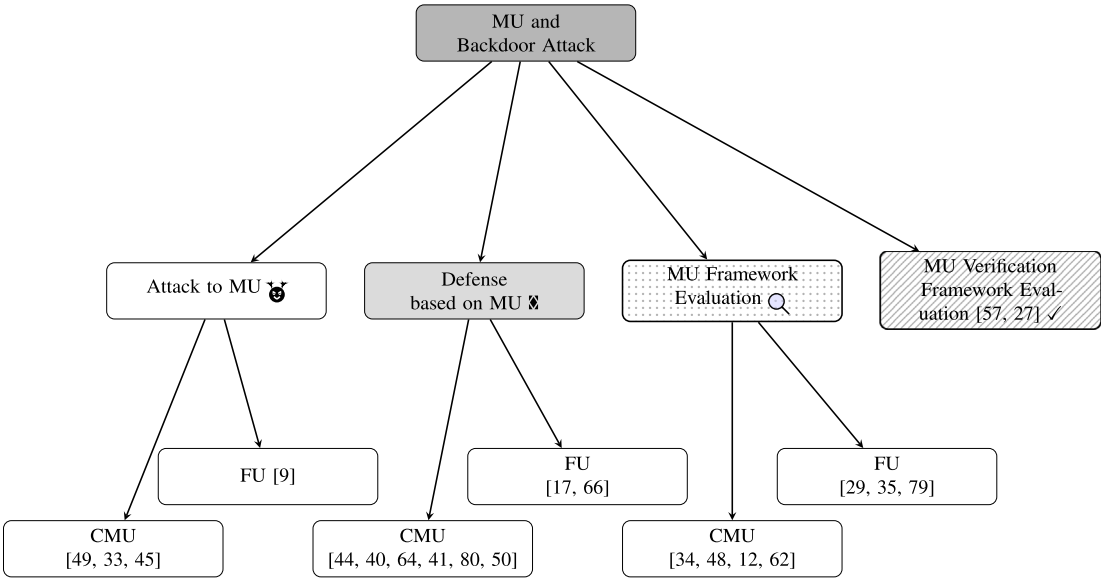


Fig. 3. Categorization for papers dealing with MU and Backdoor Attacks.

Table 3
Papers dealing with MU and Backdoor Attack. (BU - Before Unlearning, DU - During Unlearning, and PU - Post Unlearning, Acc- Accuracy, ASR - Attack Success Rate, BASR - Backdoor ASR).

Ref.	Year	Class/ Instance MU	Attacker Knowledge	Attack Phase	Defense Phase	Defense Technique	Evaluation Metrics
Centralized Machine Unlearning (CMU)							
[44]	2022		Black-box	Training	DU	Generative Networks and MU	ASR, Acc
[48]	2022		Black-box	PU	BU	Neuron masking	Forgetting Rate
[57]	2022		Black-box	BU/PU	–	–	Backdoor Attack Success Rate (BASR)
[27]	2023		Black-box	BU/PU	–	–	BASR, Acc
[34]	2023		Black-box	PU	BU	Model sparsification via weight pruning	Acc, MIA-Efficacy, Run-time efficiency (RTE)
[40]	2023		Black-box	BU	DU/PU	Reconstructive Neuron Pruning (RNP)	Detection Rate (DR), Acc, ASR
[64]	2023		Black-box	BU	DU/PU	Adversarial Training Techniques	Acc, ASR, Defense Effectiveness Rating (DER)
[12]	2024	Instance	Gray-box	PU	DU	Unlearning by small weight perturbation	MIA Evaluation, BASR, Acc, Unlearning Time Cost
[33]	2024		Black-box	BU/DU	BU/DU/PU	Outlier Filters, Model Scanners, Anomaly Detectors, Model Reconstructors	BASR, Stealthiness, Persistence, Resistance to Defenses
[41]	2024		Black-box	DU	PU	Partial Training, Super Finetuning, Neural Attention Distillation, Anti-Backdoor-Learning, Data Augmentation	ASR, Acc
[45]	2024		White-box/ Black-box	DU	DU/PU	Monitoring Unlearning Requests, Model Review and Retraining	ASR, Acc, Unlearning Percentage (UP)
[62]	2025		Black-box	PU	DU	Compressive Representation Forgetting Unlearning	Acc, BASR
[80]	2024	Class	White/Gray-box	DU	BU/DU/PU	Unlearning-based Model Ablation	Acc, False Positive Rate (FP), False Negative Rate (FN)
[50]	2024		White-box	DU	DU/PU	Progressive Unified Defense (PUD), Model Repairing Techniques, Data Filtering, Adversarial Training	ASR, Acc
Federated Unlearning (FU)							
[17]	2024		Black-box	BU	BU/DU/PU	Gradient Ascent	Expected Calibration Error, Acc, BASR
[9]	2025		Black-box	DU	DU/PU	Gradient Value Adjustments	ASR, Acc
[29]	2025	Instance	Black-box	BU	DU	Backdoor Certification	BASR, Kullback-Leibler Divergence and ℓ2-Distance, Acc
[79]	2025		White-box	BU	PU	Exact Unlearning, Retraining on Clean Data, Uni-Adapter Structure	Acc, Time Consumption, Intersection over Union (IoU)
[35]	2024	Class	White-box	DU	PU	Adaptive Differential Privacy, Dual-layered Selection, MU	Acc, Membership Inference Success Rate (MISR), ASR
[66]	2024		White-box	BU	BU	Knowledge Distillation, Purifying Backdoored Models	Acc, BASR

3.1.1. Backdoor attack against MU

This section examines how MU methods, designed to remove the influence of specific data from trained models, can themselves become targets of Backdoor Attacks. Formally, let θ_t denote the parameters of the trained model and D_f the dataset intended for removal. A typical MU objective seeks to obtain parameters θ' such that

$$\theta' = \arg\min_{\theta} \mathcal{L}_{D_r}(\theta) + \lambda R(\theta, \theta_t),$$

where $D_r = D \setminus D_f$ is the retained dataset and R is a regularization term ensuring stability.

An adversary can exploit this process by ensuring that the poisoned samples $(X + \Delta, y_t) \in D_f$ are chosen as the forget set. In such a case, the unlearning objective may implicitly *preserve* the malicious mapping $F(X + \Delta) = y_t$ instead of erasing it, since the backdoor mechanism is not disentangled from the retained representation space. This adversarial goal can be expressed as:

$$\Pr\left[F_{\theta'}(X + \Delta) = y_t\right] \approx \Pr\left[F_{\theta_t}(X + \Delta) = y_t\right],$$

indicating that the backdoor persists even after MU is applied.

The authors of [49] explore the interplay between MU and Backdoor Attacks, highlighting a novel vulnerability where malicious actors can exploit the Unlearning process to embed persistent triggers within datasets. These Backdoor Attacks are strategically designed to maintain model functionality post-unlearning, undermining the efficacy of data erasure mechanisms. By manipulating training data prior to unlearning requests, attackers can create conditions where specific inputs elicit targeted responses, thereby maintaining influence over the model's decisions. This intertwining of Backdoor tactics with MU emphasizes the need for robust defenses in Machine Learning systems, as traditional Unlearning methodologies may inadvertently facilitate adversarial manipulation. In [45], the authors introduce two novel attack strategies against MU systems. The first implants a Backdoor by strategically requesting data removals without altering the training data, whereas the second injects poisoned samples during training and subsequently activates the Backdoor through targeted unlearning requests. To optimize these attacks, the authors design an objective function that jointly selects the unlearning subset and triggers, thereby maximizing attack effectiveness while minimizing the extent of data removal. Discrete unlearning operations are approximated using a differentiable sigmoid relaxation, enabling optimization through gradient-based methods. Experimental evaluations across multiple models and datasets demonstrate that the proposed attacks achieve high success rates with minimal data deletion, combining both efficiency and stealth.

A recent Backdoor Attack, UBA-Inf [33], exploits MU in Machine-Learning-as-a-Service (MLaaS) to stealthily activate backdoors. Unlike methods proposed in [45], UBA-Inf leverages unlearning requests to remove camouflage samples, prolonging Backdoor persistence and evading detection. The attack unfolds in four stages: generating camouflage and backdoor samples, injecting them into training, activating the Backdoor through Unlearning, and exploiting the model via queries. An influence-driven camouflage generation algorithm enhances stealth. UBA-Inf remains effective in both on-demand and continual training MLaaS, highlighting the urgent need for stronger defenses against unlearning-enabled Backdoor Attacks.

The study [9] introduces a novel Backdoor Attack framework termed FedMUA, which exploits the Federated Unlearning process to manipulate model predictions maliciously. By strategically initiating unlearning requests aimed at influential training samples, attackers can intentionally misclassify target users while maintaining the accuracy of predictions for non-target users. This stealthy approach raises significant ethical concerns, particularly in sensitive applications like credit scoring, where it can adversely affect individuals' financial reputations. The general steps of a Backdoor Attack against MU are visible in Fig. 4. In the first scenario, the attacker can inject malicious data during the training phase, whereas in the second scenario, the attacker poisons the model only by making malicious unlearning requests.

3.1.2. Backdoor defense based on MU

This section explores how Machine Unlearning can be repurposed as a defensive tool to mitigate Backdoor Attacks. It reviews recent approaches that leverage MU to detect, isolate, and remove malicious triggers from compromised models without retraining from scratch. In recent research, MU has been employed as a defense mechanism against Backdoor Attacks [17,40,41,44,50,64]. The common architectural framework underlying all proposed backdoor defense mechanisms is illustrated in Fig. 5.

In [44], the authors present BAERASER, a framework that can erase the trigger patterns of a Backdoor Attack from the victim model based on MU leveraging a gradient ascent-based method. The authors of [40] propose a defense called Reconstructive Neuron Pruning (RNP) to expose and prune backdoor neurons via MU. It proceeds firstly by unlearning the neurons on a few clean samples via a neuron-level unlearning and then recovering the neurons on the same clean samples via a filter-level recovery. The algorithm operates as follows.

- **Neuron-level Unlearning (NU):** The process begins by maximizing the model's error on a small subset of clean samples (defense data). This step aims to unlearn the clean features, leaving the backdoor neurons largely preserved.
- **Filter-level Recovering (FR):** After unlearning, the method recovers the neurons by minimizing the model's error on the same defense data. This recovery is performed at the filter level using a filter mask. The asymmetric nature (neuron-level unlearning, filter-level recovering) forces the network to reuse dormant backdoor neurons to compensate for the loss of clean features.
- **Pruning:** The filter mask learned during recovery indicates which filters (and their associated neurons) are likely backdoor-related (low mask values). These neurons are then pruned from the network to purify the model.

The work reported in [64] proposes a framework called Shared Adversarial Unlearning (SAU) to identify shared adversarial examples and unlearn them to break the connection between the poisoned sample and the target label. Recently, the authors of [41] propose a defense against Backdoor attacks by performing Unlearning. In particular, the authors provide a new model training method, called Partial Training (PT), that freezes part of the model to isolate suspicious samples. In [80], the authors design UMA, a framework able to filter out Backdoor-irrelevant features by unlearning the inherent features of the target class within the model, and subsequently reveals the Backdoor through dynamic trigger optimization. The proposal of [50] explores the connection between Backdoor and Adversarial Attacks and presents a Progressive Unified Defense (PUD), which leverages Unlearning to remove backdoors and improve adversarial robustness simultaneously. The process of PUD begins with a potentially compromised model and a dataset, some of which might be poisoned. An initial filtering step is applied to the dataset to remove obvious poisoned images. The following iterative steps are repetitively performed.

- **Model Improvement (Student Model):** The current model is updated by training it on special "adversarial examples" created from the dataset. These examples help weaken the backdoor by making the model learn different associations for trigger patterns. A second training pass further refines the model's performance on regular, clean images.
- **Model Improvement (Teacher Model):** A more stable and robust "teacher" version of the model is gradually built by averaging the student models from previous iterations. This helps improve both backdoor removal and resistance to adversarial attacks.
- **Data Cleaning:** The dataset itself is cleaned. Images are checked for inconsistencies in how the original and purified models classify them; inconsistent predictions often point to poisoned images. Another filtering technique is also used to help identify poisoned data.
- **Leveraging Poisoned Data:** Instead of discarding the identified poisoned images, they are used in an MU process. This intentionally reverses the effect of the poisoned data on the model, further removing the backdoor.

This repeating cycle of refining the model and cleaning the data allows PUD to simultaneously get rid of backdoors and make the model

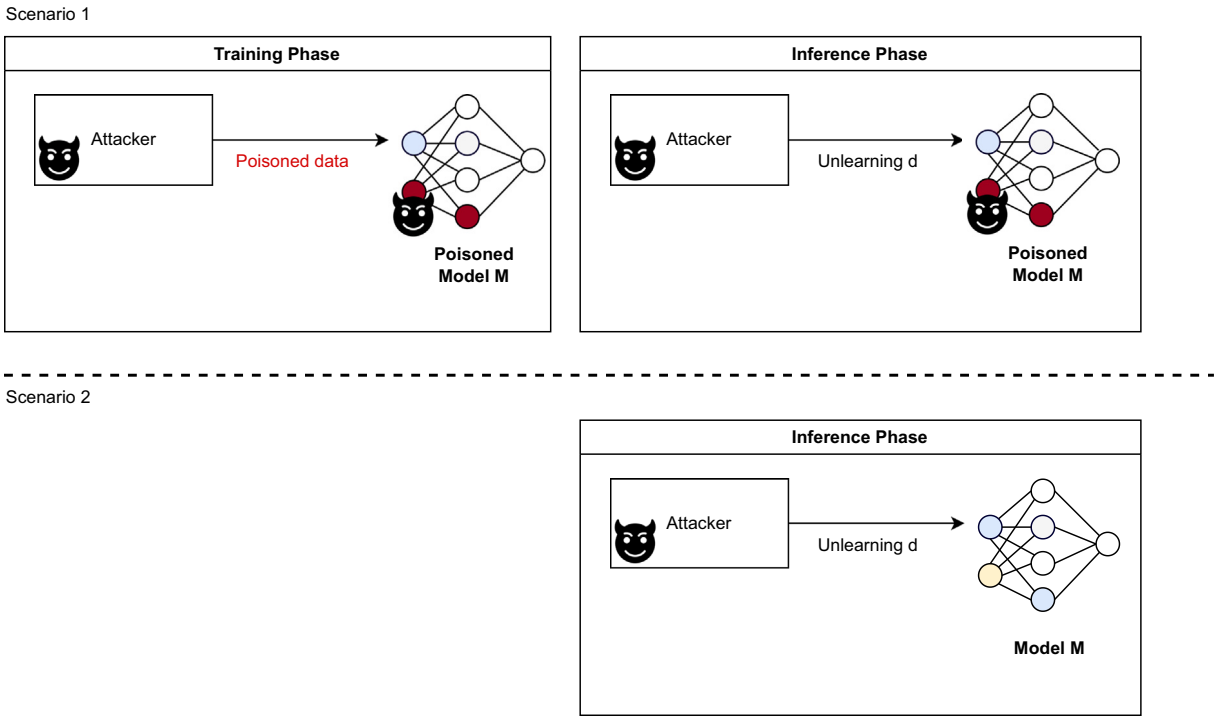


Fig. 4. The two scenarios of Backdoor attacks against MU.

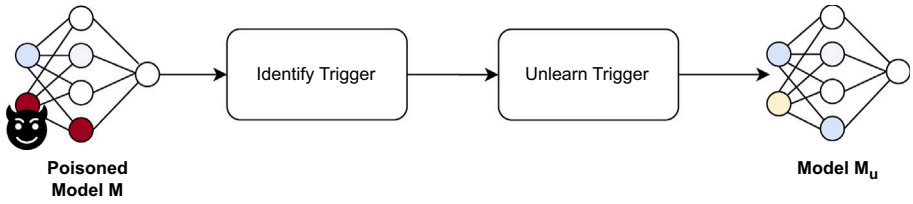


Fig. 5. The general architecture for backdoor defenses using MU.

more resistant to adversarial attacks. As the backdoor is eliminated, the model training naturally shifts towards improving its general resistance to adversarial examples.

The proposals of [17,66] present a system that enables the selective erasure of the influence of specific clients' data from the global model in an FL system. Additionally, it provides a solution for cleaning compromised global models by selectively removing the influence of poisonous clients without necessitating complete retraining. The authors of [66] leverage a method able to distinguish the attacker's influence on the global model by subtracting its historical parameter updates from the model. To mitigate the bias introduced by this process, they employ Knowledge Distillation, ensuring that the updated model retains utility without preserving Backdoor behaviors.

3.1.3. Evaluating MU framework through backdoor-based metrics

This section outlines the usage of backdoor vulnerabilities to measure the efficacy of MU. By intentionally injecting a backdoor into a model via specific training data, we can test whether an unlearning algorithm can successfully remove the functionality of the backdoor. The persistence or removal of the backdoor serves as a quantitative measure of the ability of the unlearning method to truly forget targeted information, thereby providing a robust framework for assessing the security and completeness of machine unlearning protocols. Observe that Attack Success Rate (ASR), and Backdoor ASR (BASR) robustness may vary across different scenarios: in centralized settings, they often provide a reasonably consistent measure, while in federated scenarios, heterogeneity across clients, non-IID data, and aggregation effects can impact their reliability. Consequently, as visible in Table 3, ASR alone may not fully capture MU efficacy, and additional complementary metrics—such as model accuracy on retained data, forgetting metrics, or the influence on downstream tasks—should be considered.

Several studies introduce new Machine Unlearning (MU) paradigms that incorporate common attack scenarios to evaluate the effectiveness and robustness of their methods [12,34,48]. Jia et al. [34] propose an MU approach that takes advantage of the sparsity of the model through weight pruning to significantly reduce the gap between approximate and exact Unlearning significantly. Their experiments demonstrate that the method can remove the influence of poisoned Backdoor data, where an adversary injects triggers and manipulates labels to induce targeted misclassification. By enforcing an appropriate level of sparsity, their method successfully mitigates the Backdoor effect while preserving the model's generalization ability.

Ma et al. [48] introduce Forsaken, an MU method based on neuron masking, and validate its robustness in poisoned data scenarios. Chen et al. [12] evaluate their MU method by introducing small perturbations to the model weights and evaluating performance using membership inference and Backdoor Attack metrics. To assess the effectiveness of their *CRFU* framework [62] in mitigating Backdoor threats, the researchers utilized a standard methodology. This involved embedding Backdoor triggers into the specific data samples designated for Unlearning during

the initial training phase of the Information Bottleneck (IB) model. Post-unlearning, they evaluated the trained model to determine the persistence of the Backdoor vulnerability. Their findings demonstrated that this approach resulted in reduced Backdoor accuracy while preserving a higher level of model accuracy on diverse benchmark datasets.

In the context of Federated Unlearning (FU), Jiang et al. [35] propose a method that utilizes historical information and Differential Privacy (DP) to enhance privacy protection. Han et al. [29] present a novel approach for Vertical Federated Unlearning, using Backdoor Attacks to evaluate the robustness of their solution. The researchers of *FedWiper* [79] evaluated their FU framework's robustness against Backdoor attacks by injecting crafted trigger samples into the training data. Experimental results showed that FedWiper significantly reduced the attack success rate (ASR), with models using the Uni-Adapter achieving an ASR of just 0.73 %, compared to 99.99 % without it.

3.1.4. Backdoor in MU verification frameworks

This section examines how Machine Unlearning (MU) can serve as a critical instrument for auditing compliance with data privacy regulations. By enabling the targeted removal of specific user data and its influence from trained models, MU provides a practical means to verify whether organizations honor deletion requests in accordance with legal obligations. This capability is particularly significant under frameworks such as the GDPR, which enshrines the "right to be forgotten" as a fundamental privacy right. Through systematic auditing, MU not only supports regulatory adherence but also fosters user trust by offering transparent, verifiable proof of data erasure. Unlike the previously discussed approaches, the works in [27,57] utilize Backdoor Attacks to design decentralized Machine Unlearning (MU) verification frameworks in the Machine-Learning-as-a-Service (MLaaS) setting. In these systems, users verify compliance with data deletion requests by analyzing model predictions. Specifically, users poison their training data with unique Backdoor triggers associated with target labels before submitting it for training. After requesting data deletion, they reintroduce the same triggers during inference. If the Backdoor effect disappears, it indicates successful unlearning; otherwise, it signals that the provider has failed to remove the influence of the data.

In [27], researchers propose a verifiable Machine Unlearning framework by embedding undetectable Backdoor triggers into sensitive data, preventing MLaaS providers from isolating poisoned data for validation spoofing and enabling users to verify data deletion via prediction results. The approach supports efficient incremental retraining using an index structure, with evaluations confirming its effectiveness and efficiency. The verification process in [57] is modeled as a hypothesis testing framework to differentiate between honest providers (who perform unlearning) and malicious ones (who retain the data). Extensive evaluations across different neural network architectures and datasets show that this mechanism achieves reliable verification with low false positive and false negative rates.

3.2. Membership inference attacks

Membership Inference Attacks (MIAs) threaten training data privacy in Machine Unlearning [5]. In the ML context, MIAs determine whether a data sample was in a model's training set [30,56]. The attack follows a security game framework between a challenger and an adversary [5]. Given a trained model f_θ and a sample (x, y), the adversary's membership function is:

$$A(x, y) = \mathbb{1}[A_0(x, y) > \tau] \tag{3}$$

where $A_0(x, y)$ is a confidence score and τ a decision threshold. A Loss-based MIA infers membership using the loss function:

$$A_{\text{loss}}(x, y) = \mathbb{1}[-\ell(f(x), y) > \tau] \tag{4}$$

where $\ell(f(x), y)$ is the model loss, and $\mathbb{1}$ is the indicator function.

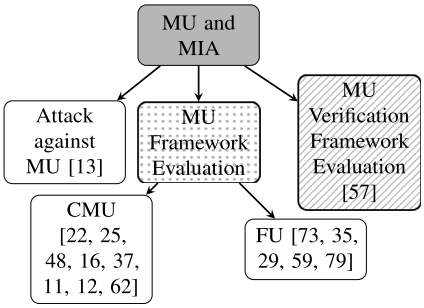


Fig. 6. Categorization for papers dealing with MU and membership inference attacks.

Schematically, the attack unfolds through the following step-by-step process.

1. **Train Model.** The challenger trains f_θ on dataset $D \sim \mathcal{D}$.
2. **Sample Data.** A bit b determines whether (x, y) is from D ($b = 1$) or from $\mathcal{D} \setminus D$ ($b = 0$).
3. **Send Sample.** The challenger provides (x, y) to the adversary.
4. **Adversary Queries Model.** The adversary accesses f_θ and possibly $\mathcal{D}$ (e.g., for shadow models).
5. **Compute Membership Score.** The adversary applies a decision rule, such as loss-based inference.
6. **Predict Membership.** The adversary classifies (x, y) as a member or non-member.
7. **Evaluate Success.** The attack succeeds if the adversary correctly infers b.

Common defense strategies against Membership Inference Attacks (MIAs) include Differential Privacy, which adds noise to training to obscure individual contributions, and regularization techniques (e.g., dropout, weight decay) to reduce model overfitting. Adversarial training enhances robustness by explicitly training against inference attacks, while confidence masking limits the information exposed by model outputs. Knowledge distillation transfers knowledge to a smaller model to remove overfitting artifacts, reducing susceptibility to MIAs. In the context of Machine Unlearning (MU), attackers can exploit MIAs to verify whether a user's data has been effectively removed as ML models typically behave differently on training data than on unseen data.

After reviewing the literature on Machine Unlearning and Membership Inference Attacks (MIA), we identified that existing works can be grouped into the following three categories:

- Membership Inference Attack against MU [13];
- MIA as evaluation metrics for new MU methods [11,12,16,22,25,29,35,37,48,59,62,73,79];
- MIA as evaluation metrics for new MU verification methods [57].

The categorization we employ is visible in Fig. 6. In particular, this figure shows the aforementioned categorization of papers addressing MIA, further dividing these approaches according to unlearning architecture, namely Central MU (CMU) and Federated Unlearning (FU).

Table 4 presents the works dealing with MU and Membership Inference Attacks specifying the Unlearning paradigm used between Central Machine Unlearning (CMU) and Federated Unlearning (FU); whether the proposed unlearning technique concerns the instance or the class; the attacker's knowledge (black-box, gray-box or white-box); the attack/defense phase (during the training or before, during or after the unlearning); the employed defense techniques, and the evaluation metrics used.

Table 4
Papers dealing with MU and MIA. (BU - Before Unlearning, DU - During Unlearning, and PU - Post Unlearning, Acc - Accuracy, ASR - Attack Success Rate).

Ref.	Year	Class/ Instance MU	Attacker Knowledge	Attack Phase	Defense Phase	Defense Technique	Evaluation Metrics
Centralized Machine Unlearning (CMU)							
[13]	2021		Black-box	PU	DU/PU	Reduce publishing information, Temperature Scaling, Differential Privacy	Degradation Count, Degradation Rate, Attack AUC
[22]	2021		White-box/ Gray-box	BU/PU	DU/PU	Linear Approximation, Controlled Information Disclosure, Careful Weight Management, Sequential Forgetting Requests	Acc, Empirical Risk Minimization
[25]	2021		White-box	DU/PU	DU/PU	Unlearning, Amnesiac Unlearning	Acc, Performance over MIA and Model Inversion Attack
[48]	2022	Instance	Black-box/ White-box	PU	DU/PU	Neuron Masking, Mask Gradient Generator, Dynamic Adjustment of Gradients, Feedback Mechanisms	Forgetting Rate, Acc Loss
[57]	2022		White-box	BU	DU/PU	Neural Cleanse, Neural Attention Distillation	Power of the Hypothesis Test, Acc, Impact of User Participation
[37]	2023		Black-box	PU	PU	SCRUB Method, LiRA-for-Unlearning Attack Adaptation, Rewind Strategy	Forget Quality, User Privacy, Utility Metrics, Trade-offs
[12]	2024		Gray-box	PU	DU	Unlearning by small weight perturbation	MIA and Backdoor Evaluation
[62]	2025		Black-box	PU	DU	Compressive Representation Forgetting Unlearning (CRFU)	Mean Squared Error (MSE), Area Under the Curve (AUC)
[16]	2023	Class	White-box	PU	PU	Reducing Model Overfitting, Perturbation of Posteriors, Adversarial Training	Acc, Anamnesis Index
[11]	2024		Gray-box	PU	BU	Adversarial Unlearning, GAN based Unlearning	False Negative Rate, Acc, Time Cost
Federated Unlearning (FU)							
[73]	2023		Black-box	DU	PU	Local Differential Privacy, Differentially Private Noise Injection, MIA Strategy	Acc, Running Time, MIA Precision
[35]	2024		Black-box	DU	PU	Adaptive Differential Privacy, Unlearning Strategies	Acc, MIA Success Rate, ASR
[29]	2025	Instance	Gray-box	PU	DU	Use of Backdoor Triggers, MIA, Constrained Gradient Ascent	Acc, MIA Performance, Kullback-Leibler Divergence and $\ell 2$-Distance
[59]	2025		White-box	DU	BU	Proof-of-Deniability, Perturbation of Client Updates, Integral Privacy Model, Differential Privacy	Utility Comparison, Memory Usage, Retraining Time, Differential Privacy Parameters
[79]	2025		Black-box	PU	DU/PU	Exact Unlearning, Retraining on Clean Data, Uni-Adapter Structure	Ac, Time Consumption, Intersection over Union

3.2.1. Membership inference attack against MU

This section examines how Machine Unlearning, despite its privacy-preserving intent, can be susceptible to Membership Inference Attacks (MIA). It delves into the mechanisms by which adversaries can exploit subtle discrepancies in the outputs of a model, confidence scores, or decision boundaries observed before and after the unlearning process. By analyzing these behavioral shifts, attackers can infer with high probability whether particular data samples were part of the original training set, thereby undermining the intended privacy guarantees of MU and revealing critical weaknesses in existing unlearning methodologies.

Formally, let f_{θ_t} denote the trained model before unlearning with parameters θ_t, and $f_{\theta'}$ the model after MU has produced updated parameters θ'. For any input sample X, the prediction confidence vector is defined as:

$$p_\theta(X) = f_\theta(X) \in [0,1]^C,$$

where C is the number of classes. The adversary's goal is to distinguish whether X belonged to the forget set D_f or not, formulated as a hypothesis test:

$$H_0 : X \notin D_f \quad \text{(non-member)}$$

$$H_1 : X \in D_f \quad \text{(member)}$$

The decision is based on a membership score $S(X)$ derived from the difference in model behavior before and after unlearning:

$$(X) = d\left(p_{\theta_t}(X),\ p_{\theta'}(X)\right),$$

where $d(\cdot,\cdot)$ is a distance metric such as ℓ_1, ℓ_2, or KL divergence. If $S(X)$ exceeds a threshold τ, the attacker infers H_1 (i.e., X was part of the unlearned set). Formally:

$$\hat{M}(X) = \begin{cases} 1, & \text{if } S(X) > \tau \quad (X \in D_f) \\ 0, & \text{otherwise.} \end{cases}$$

The success of MIA against MU can be quantified using *Membership Inference Accuracy*:

$$\text{MIA}_{MU} = \Pr\left[\hat{M}(X) = 1 \mid X \in D_f\right] + \Pr\left[\hat{M}(X) = 0 \mid X \notin D_f\right]$$

This modeling highlights that if the unlearning process causes systematic and detectable shifts in model outputs for D_f, adversaries can exploit these to reliably infer membership, compromising the privacy guarantees of MU.

The study [13] investigates the vulnerabilities associated with Machine Unlearning (MU) in relation to Membership Inference Attacks

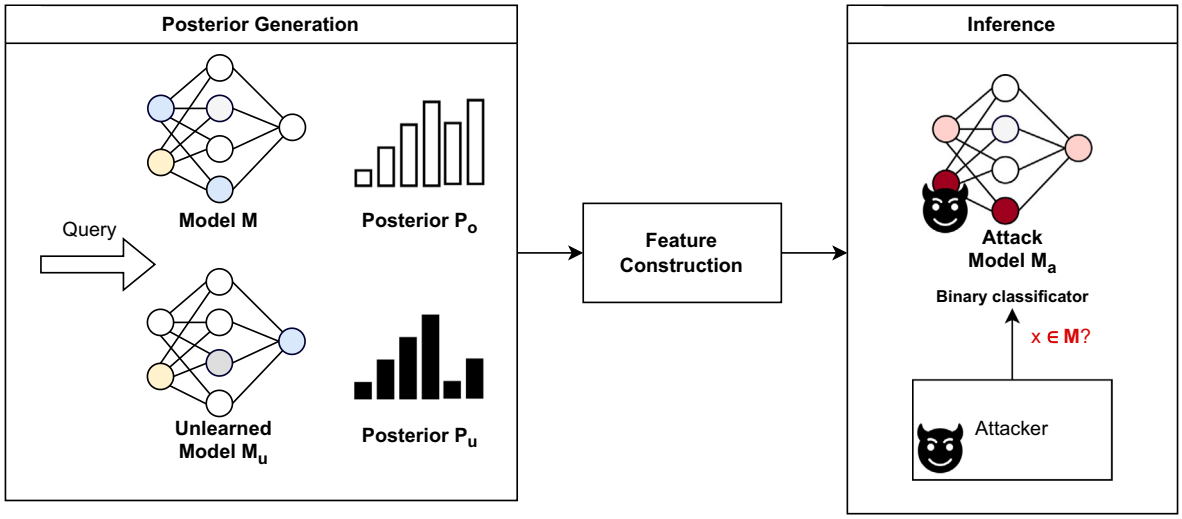


Fig. 7. The scheme of an MIA attack against MU.

(MIA). It emphasizes that while MU aims to delete specific data samples and their influence from a Machine Learning model to protect privacy, it can inadvertently create privacy risks due to the distinct versions of the model, the original and the unlearned, leading to potential information leaks. The authors propose a novel MIA specifically designed for the MU context, which determines if a target sample was part of the training data for the original model. The attack process unfolds in three steps: generating posteriors from both models, constructing a feature vector from these outputs, and then using an attack model to classify the target sample's membership status. Empirical results show that the MIA achieves higher accuracy than classical approaches, highlighting how MU can unintentionally weaken membership privacy. A general attack scheme is visible in Fig. 7.

3.2.2. *Evaluating MU framework through MIA-based metrics*

The studies reviewed in this section employ Membership Inference Attacks (MIA) as an evaluation metric for assessing the effectiveness of Machine Unlearning (MU). Their findings consistently indicate that, for the deleted data, the probability of successful membership inference is significantly reduced in the unlearned model compared to the original model, thereby evidencing MU's potential to mitigate privacy risks associated with data retention. Observe that, while MIA is widely used to measure the extent to which a model forgets previously seen data, its reliability depends on multiple factors, including the data distribution, model architecture, and whether the setting is centralized or federated. In centralized settings, MIA metrics (e.g., attack accuracy or AUC) tend to provide a reasonably consistent estimate of residual membership information. In federated settings, however, client heterogeneity, non-IID data, and aggregation procedures can lead to variability in MIA results, making them less robust as a standalone measure of MU efficacy. Therefore, as visible in Table 4, MIA has not been used as the only metric to assess the performance of the MU systems, but complementary metrics have also been considered, such as model utility on retained data, forgetting measures, or influence on downstream tasks.

Golatkar et al. [22] introduced a mixed Unlearning approach leveraging linear approximations and careful weight management. In their experimental evaluation, the authors implemented a simple membership inference attack—similar to the approach in [23]—by leveraging the entropy of the model's output. They demonstrated that their Unlearning method achieves a membership attack success rate comparable to that of a model retrained from scratch. Similarly, the studies in [16,25] evaluated their Unlearning methods by performing Membership Inference Attacks (MIAs) on models before and after data removal, demonstrating the effectiveness of their approaches in mitigating record-level data leakage. Additionally, they showed that their methods remain robust against Model Inversion Attacks. The authors of [37,48] proposed a metric to measure the effectiveness of an MU based on the concept of membership inference. In particular, they designed the *forgetting rate* that describes the transformation rate of deleted data from memorized to unknown stages after the unlearning phase. This metric is used to test the proposed dynamic neuron masking approach, called Forsaken in [48] and by Kurmanji et al. [37] in SCRUB, a LiRA-adapted Unlearning method to enhance privacy and utility. Analogously, Chen et al. [11,12] used MIA to evaluate whether the dataset is successfully forgotten or not. To execute MIA on the forgotten model, they employed a *shadow model* training strategy to infer the data and construct the attack classifier. The study CRFU [62] (Compressive Representation Forgetting Unlearning) highlights that adversaries can exploit differences in model outputs before and after unlearning to perform privacy leakage attacks, including membership inference attacks (MIA). To mitigate this, CRFU is proposed as a defense mechanism that facilitates targeted forgetting by deliberately distorting the model's internal representation of specific data, while preserving its overall learned knowledge. The process is governed by a controllable unlearning rate, aiming to minimize the information about erased data in the model's representations and thereby hinder an attacker's ability to infer membership. The study further develops an attack classifier for Membership Inference Attack (MIA) to evaluate privacy risks in Unlearning scenarios. The adversary collects model outputs on probing samples before and after unlearning to capture output discrepancies. These differences are then used to train a separate attack model to infer whether specific data samples were part of the original training set. The results show that the success of the attack depends on the magnitude of output variations, revealing potential privacy vulnerabilities introduced by unlearning updates. Experimental results showed that the reconstruction Mean Squared Error (MSE) significantly increased, indicating that adversaries face greater difficulty in reconstructing original data from the model's outputs after unlearning.

Several studies in the context of Federated Unlearning evaluate the effectiveness of Membership Inference Attacks (MIAs) on the unlearned model to verify whether the influence of the targeted client has been successfully removed [29,35,59,73,79]. In particular, Zhang et al. [73] incorporated local Differential Privacy and noise injection to reproduce a model that is indistinguishable from the retrained one by only exploring clients' historical submissions. Jiang et al. [35] define Membership Inference Success Rate (MISR) to evaluate the FU's effectiveness. Han et al. [29] studied constrained gradient ascent and Backdoor triggers

in Vertical FU frameworks, validating their effectiveness through MIA. The study [79] employs Membership MIA to evaluate the data privacy resilience of the *FedWiper* framework. Results show that *FedWiper* effectively mitigates this threat, reducing the attack success rate (ASR) to 50.18 % with the Uni-Adapter (A model component designed to facilitate efficient parameter updates in federated submodels), compared to 61.97 % without it, highlighting its enhanced capability to protect sensitive data.

3.2.3. Evaluating MU verification framework through MIA-based metrics

This section examines a verification framework for Machine Unlearning that employs Membership Inference Attack (MIA)-based metrics to evaluate whether unlearning requests have been effectively honored. By quantifying changes in membership inference success rates, the mechanism enables users to independently verify compliance with data privacy obligations, such as the "right to be forgotten", thereby fostering transparency and accountability in machine learning services.

Sommer et al. [57] propose a mechanism in the MLaaS context that allows users to verify whether the service provider complies with their right to be forgotten. To validate their Backdoor-based verification mechanism, they provide the verification performance using user-level MIAs. In particular, they perform MIA on each data sample by comparing the prediction confidence to a threshold.

3.3. Adversarial attack

An Adversarial Attack is a deliberate modification of an input to cause an ML model to make an incorrect prediction while keeping the input visually or semantically similar to the original. This modified input is known as an adversarial example. According to Niu et al. [50], adversarial attacks can be classified into targeted and untargeted adversarial attacks. An untargeted adversarial attack seeks to generate a perturbation r that causes an input $x' = x + r$ to be misclassified by an ML model. The objective is to maximize the model's loss $L(x, y)$ with respect to r, making the model's prediction different from the correct label y.

$$\max_r \mathcal{L}(x', y; \theta), \text{subject to} \quad |r|_p \leq \epsilon,$$

$$x' = x + r, \quad x' \in [0, 1]^d$$

where $|r|_p$ represents the perturbation constraint under an l_p norm ensuring that x' remains within valid input bounds. An untargeted adversarial attack aims to generate perturbed inputs x' that lead to misclassification, meaning the model's prediction differs from the true label y. Unlike targeted adversarial attacks, which manipulate the model to classify x' as a specific target label, untargeted attacks aim solely to mislead the model from its original prediction without enforcing a particular incorrect label. Consequently, research has shown that the predicted labels of adversarial examples tend to follow a uniform distribution across all possible classes

While classical defenses in Machine Learning primarily enhance robustness through architectural design or training strategies, Machine Unlearning provides a complementary solution. Rather than solely hardening the model against perturbations, unlearning enables the selective removal of specific, potentially harmful, learned patterns. This capability is especially valuable in the context of adversarial attacks, where success often depends on the model leveraging particular correlations embedded in the training data. By intentionally "forgetting" these correlations, Machine Unlearning can significantly reduce the effectiveness of such attacks.

After examining the work on MU and Adversary Attacks, we observed that several works present MU as a tool for the defense against adversary attacks [42,50], whereas a contribution proposes an Adversarial Attack on MU [77]. The categorization we employ is visible in Fig. 8, where we divide the analyzed papers dealing with Adversarial Attack (AA) into *(i)* papers presenting a new AA on MU and *(ii)* papers presenting a defense against AA leveraging MU.

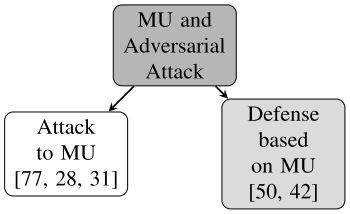


Fig. 8. Categorization for papers dealing with MU and adversarial attacks.

Table 5 summarizes the studies addressing Machine Unlearning (MU) in the context of Adversarial Attacks. It details the Unlearning paradigm employed—either Central Machine Unlearning (CMU) or Federated Unlearning (FU)—and specifies whether the unlearning operates at the instance or class level. The table also categorizes the attacker's knowledge as black box, gray box, or white box, and identifies the phase at which the attack or defense occurs (i.e., during training, before, during, or after Unlearning). Additionally, it outlines the defense techniques adopted and the evaluation metrics used in each study.

3.3.1. Adversarial attacks against machine unlearning

This section investigates adversarial strategies specifically crafted to compromise the integrity and effectiveness of MU systems. It explores how malicious actors can manipulate the unlearning process, either by injecting carefully crafted data removal requests or by orchestrating sequential manipulations, to jeopardize intended data erasure, preserve unwanted information, or even introduce new malicious behaviors. Such strategies not only diminish the reliability of MU as a privacy-preserving tool, but also expose broader security vulnerabilities in current unlearning frameworks.

Formally, let the MU process be represented as

$$U(M, D_r) \rightarrow M'$$

where M is the trained model, D_r is the data to be removed, and M' is the updated model after unlearning.

In adversarial settings, an attacker can craft a poisoned removal set D_r^{adv} such that

$$U(M, D_r^{adv}) \rightarrow M''$$

where M" either retains traces of the supposedly erased information, or deviates maliciously from the intended learning objectives. This highlights how adversarially designed unlearning requests can degrade MU's reliability and compromise security guarantees.

Zhao et al. [77] introduce two attack frameworks that exploit vulnerabilities in Machine Unlearning (MU) systems: static selective forgetting and sequential selective forgetting attacks. In a *static selective forgetting attack*, the adversary submits a batch of malicious data update requests simultaneously to manipulate the Unlearning process. This attack operates as follows:

- The adversary generates malicious data updates intended to disrupt the selective forgetting process, which may result in misclassifications or amplify existing model biases.
- The attack employs discrete indicator variables to specify deletions, making direct solutions intractable. To overcome this, the authors propose an approximation using continuous differentiable functions.

In contrast, a *sequential selective forgetting attack* strategically submits data updates over time to maximize damage while minimizing detection risk:

- The adversary carefully times and selects updates to exploit the Unlearning process dynamically.
- The attack is formulated as a stochastic optimal control problem, focusing on order and timing to achieve adversarial goals.

Table 5
Papers addressing Machine Unlearning (MU) in the context of Adversarial Attacks. (BU - Before Unlearning, DU - During Unlearning, and PU - Post Unlearning, Acc - Accuracy, ASR - Attack Success Rate).

Ref.	Year	Class/ Instance MU	Attacker Knowledge	Attack Phase	Defense Phase	Defense Technique	Evaluation Metrics
Centralized Machine Unlearning (CMU)							
[28]	2021		White-box	BU/DU	DU	Differential Privacy, Adaptive Unlearning Algorithms, Oblivious Sequence Assumption	Deletion Guarantee, Indistinguishability, Computational Cost
[42]	2023		White-box	BU	BU/DU	Adversarial Training	Effectiveness, Acc, Robustness, Efficiency
[77]	2023	Instance	White/Black-box	DU	DU	–	Demographic Parity, Equalized Odds
[50]	2024		White-box	BU/DU	DU/PU	Adversarial Training, Progressive Unified Defense (PUD), Hybrid Approaches	Acc, ASR
[31]	2024		Black-box	DU	DU/PU	MIA, Hashing	Acc

- The adversary progressively manipulates the system's behavior by selectively modifying, injecting, or removing model updates.

The evaluation considers both white-box and black-box settings across Unlearning methods such as first-order, second-order, unrolling SGD, amnesiac, and SISA. The results demonstrate consistently high attack success rates across different datasets and MU techniques. Traditional defenses against adversarial attacks, such as adversarial training, input preprocessing, gradient masking, and defensive distillation, focus on robustness against small perturbations but often fail as attackers develop adaptive strategies. These attack frameworks highlight critical security challenges in MU, emphasizing the need for stronger adversarial resilience in Unlearning mechanisms.

Gupta et al. [28] design an adversarial attack against the SISA framework for machine unlearning. The methodology involves an adaptive adversary who strategically chooses deletion requests. Specifically, the adversary identifies training points that the SISA-trained neural network ensemble classifies correctly with high confidence. These targeted points are then selected for deletion. The result of the attack reveals a significant degradation of unlearning guarantees. After deleting these targeted points, the model resulting from the unlearning process misclassifies almost every remaining training point, demonstrating the failure of SISA's unlearning guarantees against adaptive deletion sequences.

Hue et al. [31] expose significant vulnerabilities in Machine Unlearning as a Service (MLaaS) by introducing "over-unlearning" which we can categorize as a type of AA. The core of their contribution lies in two novel adversarial strategies: blending and pushing. The blending method, a naive sample-wise modification, demonstrates the feasibility of over-unlearning, particularly on simpler datasets like CIFAR-10, by embedding additional class information into unlearned samples. More crucially, the advanced "pushing" method employs pixel-wise manipulation using black-box adversarial perturbation techniques (e.g., ZOO attack [14]). This sophisticated attack moves unlearned data closer to (Pushing-I) or across (Pushing-II) the model's decision boundary. The results show that pushing significantly degrades model utility, even rendering the model useless on the unlearned class, and can even control which class the model mispredicts the unlearned samples as. This highlights a critical, underexplored security gap in MLaaS.

3.3.2. Defense against adversarial attacks based on MU

This section explores how Machine Unlearning (MU) can be strategically incorporated into defense mechanisms to counter a wide range of adversarial threats. By leveraging the capacity of MU to selectively remove the influence of malicious or corrupted data from trained models, these defense strategies aim to neutralize the effects of adversarial manipulation, restore model integrity, and enhance robustness against future attacks. The discussion highlights the role of MU as a reactive measure for damage control as well as a proactive component of comprehensive security frameworks in machine learning. Adversarial Training Models (ATMs) are a defense strategy that enhances model robustness by training on adversarial examples. To assess Unlearning robustness, Gupta et al. [28] evaluate model privacy through model inversion attacks, demonstrating that Differential Privacy-based Unlearning can mitigate adaptive deletion threats by limiting information leakage. Niu et al. [50] identify a connection between the Backdoor and Adversarial attacks and propose a Progressive Unified Defense (PUD) framework that leverages Unlearning to remove Backdoors and strengthen the model's adversarial robustness simultaneously. The authors of MUter [42] introduce a closed-form Unlearning method that leverages a total Hessian-based data influence measure. Their approach addresses the limitations of existing techniques, which often fail to accurately capture the indirect influence of data through the Hessian component.

3.4. Inversion attacks

Researchers classify inversion attacks into two categories: *Model Inversion Attacks* and *Gradient Inversion Attacks*. Model Inversion Attacks (MoIAs) highlight critical privacy risks in Machine Learning. These attacks exploit the correlation between training data and model outputs to reconstruct sensitive attributes of the input data. Typically, MoIAs frame this as an optimization problem that seeks the sensitive feature values—such as x_1 from an input x—that maximize the likelihood under the target model f [76]. MI aims to reconstruct sensitive features x_1 of an input x given partial knowledge of the target model f. Formally, let:

- $x = (x_1, x_2, \dots, x_t)$ is the feature vector of an individual.
- $y = f(x)$ is the model's predicted output.
- The adversary has access to the model f and auxiliary information:

$$\text{side}(x, y) = (x_2, \dots, x_t, y). \tag{5}$$

The attack infers the sensitive feature x_1 by solving:

$$\hat{x}_1 = \arg\max_{x_1} P(x_1 \mid x_2, \dots, x_t, y, f). \tag{6}$$

For example, if we consider linear regression model $f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_t x_t + b$, the adversary reconstructs x_1 as:

$$\hat{x}_1 = \frac{y - (w_2 x_2 + \cdots + w_t x_t + b)}{w_1}. \tag{7}$$

Model Inversion Attacks generally proceed through the following sequential phases.

1. **Attack Setup:** The adversary has *white-box access* to the model (f) and knows all features except the sensitive one (x_1).
2. **Inference Process:** Using the model's parameters and auxiliary data, the adversary reconstructs x_1.

3. **Reconstruction Strategy:**
   - If f is linear, x_1 is computed algebraically.
   - For complex models, optimization or gradient-based methods are used.

Model Inversion Attacks exploit *model transparency* to infer private attributes, posing severe privacy risks in sensitive domains like healthcare and biometrics. While traditional MoIAs focus on pre-unlearning data, the concern here lies in extracting information about unlearned data, exploiting potential residual information left within the model, even after the Unlearning process. MoIA's are also used as an evaluation metric as a model's performance on them can demonstrate the efficiency of Unlearning.

Gradient Inversion Attacks (GIAs) pose a significant privacy threat by exploiting shared model gradients and weights commonly exposed in collaborative learning settings such as federated learning. These attacks aim to reconstruct private input data or corresponding labels by minimizing the discrepancy between the observed gradients (from the target model) and gradients generated from synthetic inputs [32]. To carry out the attack, adversaries initialize with random inputs to create dummy gradients and then iteratively optimize these inputs to reduce the gradient difference. The optimization continues until the synthetic gradients closely match the original ones, resulting in a high-fidelity reconstruction of the original training data. GIAs typically follow either iterative or recursion-based paradigms to refine the reconstructions [74], and their effectiveness emphasizes the importance of gradient privacy in collaborative learning frameworks. While MoIAs typically infer partial information (e.g., sensitive features), GIAs can fully recover training data, making them more severe in collaborative learning environments. Despite their differences, both attacks highlight vulnerabilities in model transparency and training data exposure.

Traditional defense strategies against Inversion Attacks include regularization techniques like dropout and weight decay, which enhance generalization and reduce overfitting to training data [24]. Differential privacy injects controlled noise during training to obscure the direct relationship between inputs and outputs [61]. Adversarial training strengthens models by exposing them to inversion attempts during training, improving robustness. Additionally, Knowledge Distillation transfers learned knowledge to a new model while abstracting sensitive details, mitigating privacy leakage [15].

After reviewing the literature related to Machine Unlearning and Model Inversion Attacks, we found the following categorization:

- Model Inversion Attack on MU [32];
- Model Inversion Attack used as an evaluation metric for MU [16,25];
- MU used as a defense against Gradient Inversion Attack [21]

The categorization we employ is visible in Fig. 9, where we illustrate the above categorization of papers addressing Inversion Attacks, further dividing these approaches according to unlearning architecture, namely Central MU (CMU) and Federated Unlearning (FU). Meanwhile, Table 6 describes the papers dealing with MU and Model Inversion Attacks specifying the Unlearning paradigm used between Central Machine Unlearning (CMU) and Federated Unlearning (FU); whether the proposed Unlearning technique concerns the instance or the class; the attacker knowledge (black-box, gray-box or white-box); the attack/defense phase (during the training or before, during or post the unlearning); the employed defense techniques, and the evaluation metrics used.

3.4.1. Model inversion attack against MU

This section investigates a novel class of privacy threats— Model Inversion Attacks (MIA)—that exploit discrepancies between the original and unlearned models to recover sensitive information about erased data. By analyzing variations in model parameters, output distributions, or prediction confidence scores resulting from the unlearning process, adversaries can reconstruct identifiable features or infer class labels of the removed samples. Such attacks expose critical weaknesses in current MU approaches, challenging their effectiveness in providing strong privacy guarantees.

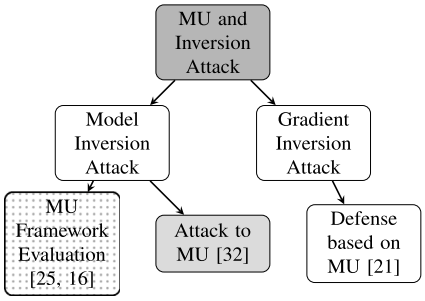


Fig. 9. Categorization for papers dealing with MU and inversion attacks.

Formally, let f_{θ_t} denote the original model trained on dataset D, and $f_{\theta'}$ the model after unlearning data subset D_f. For an input X, the prediction confidence vector is

$$p_\theta(X) = f_\theta(X) \in [0,1]^C$$

where C is the number of classes. The adversary aims to invert the model to reconstruct sensitive features of $X \in D_f$. This can be formalized as an optimization problem:

$$X^* = \arg\max_X \ \mathrm{sim}\big(p_{\theta_t}(X),\ p_{\theta'}(X)\big)$$

where X^* is the reconstructed input and $\mathrm{sim}(\cdot,\cdot)$ measures similarity between the outputs of the original and unlearned models. A successful Model Inversion Attack occurs when X^* closely approximates the original erased sample, demonstrating that the unlearning process may leak sensitive information despite its privacy-preserving goal.

The study [32] presents Unlearning inversion attacks, a novel privacy threat targeting deep neural networks. These attacks reveal sensitive information about unlearned data through two primary methods: *feature leakage* and *label leakage*. In the feature leakage scenario, adversaries with white-box access to both original and unlearned models utilize gradient inversion techniques to reconstruct the features of unlearned samples from changes in model parameters. Conversely, in the label leakage scenario, attackers with black-box access generate probing samples to assess discrepancies in predictions between the original and unlearned models, enabling them to infer the class labels of the unlearned data. The paper validates these vulnerabilities through extensive experiments, highlighting the substantial privacy risks inherent in current Machine Unlearning methodologies.

3.4.2. Model inversion attack as an evaluation metric for MU

This section examines the application of Model Inversion Attacks as an evaluation benchmark for assessing the robustness of MU methods. By attempting to reconstruct features or labels of deleted data, these attacks provide a practical measure of how effectively MU removes sensitive information from a trained model. The resulting insights enable researchers to quantify privacy preservation, identify residual information leakage, and compare the efficacy of different unlearning approaches under adversarial scrutiny. Graves et al. [25] and Chundawat et al. [16] evaluate the robustness of their Unlearning methods against Model Inversion Attacks (MIAs). Specifically, Graves et al. apply a modified version of the standard model inversion attack [19] to define Amnesiac Unlearning, focusing on scenarios where the adversary lacks prior knowledge about class semantics. They consider the attack successful if the adversary can infer meaningful information about a class solely through model inversion. Chundawat et al. [16] assess data leakage in their proposed zero-shot Machine Unlearning method using Model Inversion Attacks. They assume white-box access to the unlearned model

Table 6
Papers dealing with MU and Model Inversion Attack. (BU - Before Unlearning, DU - During Unlearning, and PU - Post Unlearning, Acc - Accuracy).

Ref.	Year	Class/ Instance MU	Attacker Knowledge	Attack Phase	Defense Phase	Defense Technique	Evaluation Metrics
Centralized Machine Unlearning (CMU)							
[25]	2021	Instance	White-box	DU/PU	DU/PU	Unlearning, Amnesiac Unlearning	Acc, Performance over MIA and Model Inversion Attack
[32]	2024		Black/White-box	PU	DU	Parameter Obfuscation, Model Pruning, Fine-tuning	Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity
[16]	2023	Class	White-box	PU	PU	Reducing Model Overfitting, Perturbation of Posteriors, Adversarial Training	Acc, Anamnesis Index
Federated Unlearning (FU)							
[21]	2024	Instance	Gray-box	PU	DU	Knowledge Distillation, Statistical Machine Unlearning, Parameter Cropping, Gradient Perturbation	Model Accuracy, Privacy Protection, Communication Efficiency

but no access to previous model versions. To evaluate leakage, they adopt a known model inversion technique by initializing with noise and optimizing via gradient descent toward a target class. The attack is conducted on three model variants: the fully trained model, a retrained model excluding the target class, and their proposed "forget" model. They consider the attack successful if it can reconstruct meaningful class representations. The results demonstrate successful inversion on the fully trained model, while the retrained and forget models produce only random patterns—indicating the strong robustness of their Unlearning method.

3.4.3. MU as a defense against gradient inversion attacks

This section examines how Machine Unlearning can be leveraged as a defensive mechanism to mitigate Gradient Inversion Attacks. By strategically removing or abstracting sensitive information from model updates—such as replacing raw-data-derived gradients with those computed from statistical summaries—MU disrupts the direct mapping between gradients and the original training data. This not only thwarts adversarial attempts to reconstruct private information but also strengthens overall privacy guarantees in collaborative learning environments like Federated Learning. Gao et al. [21] apply statistical Machine Unlearning (MU) techniques within the context of Federated Learning to defend against Gradient Inversion Attacks. Their approach shifts from individual data points to statistical summaries, creating an abstraction layer that enhances privacy and enables selective data forgetting. By replacing raw-data-derived gradients with gradients computed from statistical representations, the method disrupts the direct mapping from data to gradients. Additionally, the authors employ a teacher–student framework with dual loss functions to obfuscate information and defend against adversarial—but honest—servers.

In summary, our analysis of prevalent security threats (such as, spanning backdoor, membership inference, adversarial, and inversion attacks) demonstrates that Machine Unlearning can serve as an effective defensive mechanism when carefully designed and deployed. However, the review also reveals persistent challenges, such as balancing forgetting accuracy with model utility, addressing adaptive adversaries, and extending MU to diverse attack surfaces. These limitations naturally lead to the discussion in Section 4, where we examine the key challenges and open problems that must be addressed to advance MU as a robust component of secure machine learning systems.

3.5. Other attacks

Beyond the four attack categories discussed previously, the literature also reports a limited number of studies on alternative attacks that do not fit neatly into this classification (i.e., backdoor, membership inference, adversarial, and inversion attacks). Given their relatively sparse coverage, we group these works together and describe them in this section.

The work in [1] presents an exact reconstruction attack against linear regression models. In this setting, the attacker is assumed to have access to the model parameters both before and after a deletion request is processed. Additionally, the attacker can sample from the underlying data distribution, but does not have direct access to the original training set. The focus of this attack is to reconstruct a sample (x, y), previously part of the private training dataset X_{priv}, Y_{priv}, using models trained before and after the deletion of this sample.

The proposal of [70] addresses the data duplication problem, wherein a subset of training data is duplicated and later requested for removal, enabling the adversary to assess whether the unlearning process was successful. This paper presents a study on how duplicated data impacts unlearning in standard, federated, and reinforcement learning settings, influencing both model performance and privacy.

The papers [54,71] deal with the context of LLM. In particular, the authors of [71] propose a dynamic and automated attack framework, Dynamic Unlearning Attack (DUA), to assess the robustness of the unlearned models quantitatively. Additionally, they propose a framework called Latent Adversarial Unlearning(LAU) that enhances the robustness of the unlearning process and is compatible with most gradient-ascent-based unlearning methods. In contrast, Schwinn et al. [54] first propose an embedding space attack, which directly attacks the continuous embedding representation of input tokens. Then it describes a threat model in the context of MU and data extraction, showing that embedding space attacks can extract supposedly deleted information from unlearned models, and to a certain extent, even recover pretraining data in LLMs.

4. Performance comparison of attacks and defenses

In this section, we analyze the papers that present an attack against MU or employ MU as a defense mechanism against a specific attack from the experimental evaluation point of view.

4.1. Evaluation of attack strategies

Despite theoretical guarantees, unlearning mechanisms remain susceptible to adversarial exploitation, as highlighted by recent attack strategies. The experimental results of Marchant et al. [49] reveal the substantial impact of backdoor attacks on unlearning schemes by consistently reducing the "retrain interval", often by $70 - 100$ %, forcing computationally expensive retraining. The attack's effectiveness holds across various datasets (MNIST, Binary-MNIST, Fashion-MNIST, HAR (Human Activity Recognition)), though with some variance (67-94 % reduction), indicating its generalizability despite dataset characteristics. Crucially, the attack maintains strong effectiveness even with computationally cheaper surrogate cost functions and in gray-box settings, implying a lower barrier for attackers. This highlights that current unlearning

schemes, while theoretically robust, face significant practical vulnerabilities from adversaries aiming to increase computational burden rather than solely compromise accuracy.

The work "UBA-Inf" [33] presents a concerning new backdoor attack vector against MLaaS platforms that utilize machine unlearning. The experimental results, conducted across diverse datasets including CIFAR-10, GTSRB, MNIST, and Tiny-ImageNet, demonstrate the attack's broad applicability. UBA-Inf is shown to be highly effective against various unlearning schemes, specifically "full retrain", SISA, PUMA, and GBU, which cover both exact and approximate unlearning strategies in FA-MLaaS and RA-MLaaS scenarios. A critical finding is the Attack Success Rate (ASR): while concealed, UBA-Inf maintains a low ASR (ranging from 0.02 % to 31.94 % depending on dataset and attack type); upon activation through unlearning of camouflage samples, the ASR sharply rises to 99.5 % in most cases, effectively demonstrating fine-grained control and activation. This high ASR after unlearning, coupled with sustained persistence and resistance to various defense mechanisms, illustrates a significant vulnerability in current machine unlearning practices. Researchers in [45] introduce two backdoor attack approaches that leverage machine unlearning, demonstrating their effectiveness across various datasets and unlearning schemes, primarily measured by Attack Success Rate (ASR). The experiments utilize CIFAR-10 and TinyImageNet datasets and evaluate performance against several unlearning methods, including the first-order method, second-order method, UnrollSGD, and SISA. For the "attack without poisoning" approach, the proposed method consistently achieves significantly higher ASRs compared to a random baseline. Specifically, for CIFAR-10, ASRs reached up to 92.3 % (with 0.5 % attacker data) against the first-order method, compared to the baseline's 1.7 %. For TinyImageNet, an ASR of 50.7 % was observed with 0.1 % attacker data against the second-order method, and ASRs reached 90.1 % (with 0.5 % attacker data) against UnrollSGD, far exceeding the baseline's 0.3 %. Similarly, the "attack with poisoning" approach also shows high ASRs. For CIFAR-10 with a 1 % poisoning rate and 0.5 % attacker data, the ASR reached 86.3 % against the first-order method, significantly outperforming the baseline's 26.7 %, The Unlearning Percentage (UP) was kept low, typically ranging from 2.1 % to 7.7 % for the "attack without poisoning", and 1.6 % to 6.4 % for the "attack with poisoning", demonstrating stealthiness. This ASR performance across different datasets and unlearning mechanisms highlights a critical security vulnerability introduced by machine unlearning.

The authors of FedMUA [9] propose a novel malicious unlearning backdoor attack against FL. The effectiveness of the attack was primarily evaluated through ASR across various unlearning schemes and datasets. The experimental evaluation, encompassing Purchase, MNIST, CIFAR-10, CIFAR-100, and Credit Score datasets, reveals that FedMUA achieves a high average ASR of 89 % in IID settings and 70 % in Non-IID settings under unlearning methods like FedEraser and KNOT, utilizing aggregation rules such as FedAvg, Median, Trimmed-mean, and Krum. Specifically, with only 0.3 % malicious unlearning requests, FedMUA attains ASRs of 90 % for Purchase, 85 % for MNIST, and 95 % for CIFAR-10 with FedEraser and FedAvg in IID. In Non-IID settings, ASRs are 85 % for Purchase, 45 % for MNIST, and 90 % for CIFAR-10 with FedEraser and FedAvg. The attack also proves effective on real-world datasets, achieving an average ASR of 55 % for Credit Score and 80 % for CIFAR-100. This consistent ASR performance across diverse experimental conditions highlights FedMUA's significant ability to manipulate FU processes to induce targeted misclassification. The research [13] presents a novel MIA to quantify privacy risks in machine unlearning, demonstrating its strong performance (often better than classical MIA) across various datasets and unlearning strategies using AUC. For the "retraining from scratch" method, the attack achieves AUC improvements of up to 0.48 on categorical datasets, with Decision Trees being notably vulnerable. On image datasets, higher overfitting correlates with greater privacy degradation, such as CIFAR10 with DenseNet achieving an AUC of 0.881 compared to the AUC of classical MIA, 0.630. While the SISA unlearning method shows reduced attack performance, the attack remains effective in more complex scenarios. It achieves an AUC of at least 0.84 for Logistic Regression with fewer than 10 intermediate unlearned models, and remains potent even with group deletion (though slightly less effective than single-sample deletion, particularly for group sizes under 0.2 % of the data) and online learning. The study also identifies optimal feature construction methods: concatenation-based for overfitted models and difference-based for well-generalized models, with sorting posteriors consistently improving attack AUC. Significantly, the attack maintains high effectiveness on well-generalized models where classical membership inference performs poorly, with one instance showing an AUC of 0.882 compared to 0.497 for classical MIA. The overall attack AUC ranges from 0.504 to 0.986, demonstrating broad effectiveness. Zhao et al. [77] comprehensively evaluate a novel malicious selective forgetting attack using attack success rate (ASR) across various datasets and unlearning strategies. The proposed static attack consistently achieves high ASRs, for example, 100 % on Diabetes and 80 % on CIFAR-10 with first-order and second-order unlearning, significantly outperforming RandSearch. Its effectiveness increases with the percentage of unlearning samples, even for underrepresented data. Furthermore, the attack is highly effective in targeting subpopulations, achieving 100 % ASR on Diabetes with unrolling SGD for a subpopulation of size 50. In sequential settings, the learned adversarial policy successfully reduces the Euclidean distance to the target model over time. The attack also demonstrates strong transferability in black-box settings, achieving an ASR of 0.86 from ResNet-18 to VGG-16 on CIFAR-10, and effectively transfers between different unlearning methods. ASR ranging from 67 % to 100 % demonstrates the effectiveness of this approach.

Gupta et al. [28] detail an adversarial attack that exposes failures in machine unlearning guarantees, particularly against adaptive deletion sequences. The core statistical test used to demonstrate this failure is the expectation of an indicator function, which ideally should be 0.5 under perfect deletion guarantees. However, experiments on CIFAR-10, MNIST, and Fashion-MNIST datasets reveal significant deviations from 0.5 when SISA is used without differential privacy, falsifying its adaptive unlearning guarantees. For instance, the deviation indicates that the average accuracy of models from targeted shards is consistently lower than non-targeted ones after deletion. The paper shows that even modest amounts of added noise, inducing differential privacy, are sufficient to break this adaptive attack, returning the expectation of the indicator closer to 0.5 and restoring the unlearning guarantees, at minimal cost to model accuracy. The research [32] introduces unlearning inversion attacks, demonstrating their effectiveness in revealing sensitive unlearned data by analyzing model differences, with Mean Squared Error (MSE) as a key metric. For feature recovery, approximate unlearning consistently yields lower MSE values, with a reported 0.05 on CIFAR-100, indicating clearer recoveries compared to exact unlearning, which showed 0.39 on CIFAR-100. However, exact unlearning also achieves successful feature recovery with a low MSE of 0.28 for Chest X-Rays when the training dataset size is as small as two samples. Increased fine-tuning epochs slightly augment recovery error, leading to higher MSE values for both methods. While multiple unlearning samples pose a greater challenge, privacy leakage can still occur, highlighting persistent vulnerabilities despite higher MSE values in such scenarios.

4.2. Evaluation of defense strategies

While adversarial strategies expose vulnerabilities in unlearning mechanisms, defense-oriented approaches demonstrate promising resilience by integrating unlearning into attack mitigation. The work BAERASER [44] introduces a novel backdoor defense method leveraging machine unlearning, and evaluates its effectiveness primarily through ASR across different datasets and unlearning strategies. BAERASER consistently and significantly lowers the ASR of state-of-the-art backdoor attacks. On MNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100 datasets, BAERASER reduces the average ASR from nearly 100 % to approximately

0.87 % for MNIST, 1.22 % for Fashion-MNIST, 8.18 % for CIFAR-10, and 7.68 % for CIFAR-100 across BadNet, TrojanNN, and IMC attacks, demonstrating a substantial average ASR reduction of 98 % across all four datasets. Compared to baselines like Fine-Pruning, NAD, and Fine-Tuning, BAERASER shows superior robustness, particularly against more complex attacks like TrojanNN and IMC, where other methods often fail to achieve ASRs below 40 %. Notably, BAERASER's defense performance improves with increasing trigger sizes, a characteristic opposite to other methods, making it more robust in varied attack scenarios. Even with a limited clean data (1 % holding ratio), BAERASER maintains an appealing defense rate, lowering ASR by at least 98 %. This research, Reconstructive Neuron Pruning (RNP) [40], introduces a novel backdoor defense mechanism and evaluates its effectiveness using Attack Success Rate (ASR) across various datasets and backdoor unlearning strategies. RNP consistently and significantly lowers ASRs against 12 advanced backdoor attacks. On CIFAR-10, it reduces the average ASR from 96.34 % to 5.03 % with minimal clean accuracy drop. Compared to baselines like Fine-pruning, NAD, I-BAU, and ANP, RNP consistently achieves superior ASR reduction, particularly against complex feature-space attacks like FC, where it reduces ASR to 1.80 % compared to ANP's 74.75 %. On ImageNet-12, RNP also outperforms baselines, achieving an average ASR of 8.87 %. Even with limited clean data (e.g., 0.5 %or 1 %), RNP maintains strong defense, demonstrating its practicality. Furthermore, RNP proves robust against adaptive attacks, reducing ASRs to 13.22 % and 18.09 % against Adaptive-distillation and Adv-training, respectively, outperforming ANP. The defense mechanism also shows improved performance with increasing trigger sizes, a unique advantage. The authors of the paper [64] introduced Shared Adversarial Unlearning (SAU), a novel backdoor defense method. Their work demonstrates its superior effectiveness through ASR across diverse datasets and unlearning strategies. SAU significantly reduces ASR across MNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100 datasets for attacks like BadNet, TrojanNN, and IMC, achieving an average ASR reduction of 98 %. For instance, on MNIST, ASR decreases from nearly 100 % to 0.87 %. SAU consistently achieves better results compared to the baselines such as Fine-Pruning, NAD, and Fine-Tuning, revealing particular robustness against complex attacks. Its defense performance uniquely improves with increasing trigger sizes, and it maintains high efficacy even with limited clean data (e.g., 1 % holding ratio), showcasing a strong defense rate. Additionally, SAU is more robust to varying poisoning ratios and model architectures compared to other methods like RNP.

Li et al. [41] introduce Partial Training (PT), a novel backdoor defense method, and evaluate its effectiveness using ASR across diverse datasets and backdoor unlearning strategies. PT consistently and significantly reduces the ASR of various backdoor attacks. On MNIST, CIFAR-10, and TSRD datasets, PT reduces the ASR from nearly 100 % to less than 1 % across attacks like BadNet, Blend, SIG, and WANET. Specifically, for BadNet, the ASR is reduced to 0.0 %; for Blend, to 0.8 %; for SIG, to 0.1 %; and for WANET, to 0.2 %. This demonstrates PT's strong capability to eliminate or drastically reduce the impact of backdoor attacks. Compared to baselines such as Anti-Backdoor-Learning, Neural Attention Distillation, Super Fine-tuning, and Fine-tuning, PT consistently achieves superior ASR reduction, often outperforming them in the early stages of defense and maintaining stability. The method achieves these results while incurring minimal accuracy loss, not exceeding 3 %. Zhao et al. [80] propose Unlearning-based Model Ablation (UMA) as a backdoor scanning and defense strategy, rigorously evaluating its performance across various datasets and backdoor types. UMA effectively detects advanced backdoors by ablating backdoor-irrelevant features. The paper primarily focuses on AUC-ROC for measuring the effectiveness of the defense, achieving 91 % AUC-ROC on average against baselines like Neural Cleanse, Artificial Brain Stimulation, K-Arm Optimization, and Meta Neural Analysis. The defense is shown to be effective across MNIST, CIFAR-10, CIFAR-100, and GTSRB datasets against various complex attacks, with AUC-ROC values ranging from 83.2 % to 99 %.

Researchers of [50] introduce Progressive Unified Defense (PUD), a novel strategy designed to concurrently defend against both backdoor and adversarial attacks. The defense is rigorously evaluated using ASR for backdoor robustness and Robust Accuracy (RACC) for adversarial robustness, across various datasets and unlearning strategies.

- Backdoor Defense (ASR): PUD consistently and significantly reduces ASR across diverse datasets (CIFAR-10, GTSRB, sub-ImgNet-1 K) and various backdoor attacks (BadNet, Blend, SIG, DyAtt, WaNet). Compared to state-of-the-art model repairing methods (Fine-tuning, Fine-pruning, NAD, ANP, I-BAU) and data filtering methods (Spectral Signatures, SPECTRE), PUD demonstrates superior performance. For instance, on CIFAR-10, PUD reduces ASR from nearly 100 % to as low as 0.31 % (for SIG attack). Even without an extra clean dataset, utilizing a small poisoned extra dataset, PUD outperforms PBE, reducing ASR from 3.41 % to 1.44 % for BadNet and from 1.15 % to 0.55 % for Blend. PUD's defense performance improves with increasing trigger sizes, a unique advantage.
- Adversarial Defense (RACC): PUD also demonstrates competitive performance in enhancing adversarial robustness, measured by RACC. On CIFAR-10, PUD achieves a RACC of 76.13 %, outperforming TRADES (71.07 %) but slightly underperforming AT (77.49 %). For GTSRB, PUD's RACC of 68.46 % is competitive with AT and TRADES. Notably, on sub-ImgNet-1 K, PUD achieves the highest RACC of 79.89 %, surpassing both AT (79.18 %) and TRADES (77.75 %). This highlights PUD's ability to boost adversarial robustness alongside backdoor erasure.

The authors of UaaS-SFL [17] propose a method for defending FL systems against poisoning backdoor attacks by enabling the unlearning of malicious client contributions. The effectiveness of UaaS-SFL is primarily evaluated through Backdoor Accuracy across MNIST, Fashion-MNIST, and CIFAR-10 datasets. The results consistently demonstrate that UaaS-SFL significantly reduces backdoor accuracy to baseline levels (comparable to retraining from scratch) regardless of when in the training process (early, midway, or late) the unlearning service is invoked. On MNIST, it effectively counteracts backdoor effects even when the model has extensively integrated malicious updates. Furthermore, UaaS-SFL's efficacy is shown to be independent of the number of clients and malicious clients, maintaining robust performance even with varying numbers of attackers (e.g., $N = 2$, $N = 3$, and $N = 5$ across MNIST and Fashion-MNIST). Comparative analysis against existing methods like FedRecovery and FedEraser reveals UaaS-SFL's superior performance in restoring backdoor accuracies. On MNIST, UaaS-SFL achieves a backdoor accuracy of 10.97 %, outperforming FedRecovery (11.20 %) and FedEraser (11.23 %), closely matching the Retrain baseline of 10.13 %. Similarly, on Fashion-MNIST, UaaS-SFL attains 10.89 % backdoor accuracy, surpassing FedRecovery (10.93 %) and FedEraser (11.17 %), again closely aligning with the Retrain baseline of 10.37 %. These results affirm UaaS-SFL's efficiency in mitigating backdoor threats in FL environments. Wu et al. [66] introduce a novel FU method for removing attacker influence from global models in FL, with its effectiveness primarily evaluated by Backdoor Accuracy across computer vision (MNIST, Fashion-MNIST, CIFAR-10, GTSRB) and natural language processing (IMDB, Yelp) datasets. The proposed two-step method, "UL-Subtract" followed by "UL-Distill", consistently reduces the backdoor attack accuracy rate to nearly zero across all tested datasets. For instance, on MNIST, Backdoor accuracy drops from an initial 98.9 % to 0.3 %, and on Fashion-MNIST, it decreases from 98.9 % to 0.2 %. For CIFAR-10, Backdoor Accuracy is reduced to 3.9 % (VGG16) or 2.2 % (ResNet), while NLP tasks show Backdoor Accuracy dropping from 100 % to 7.5 % for IMDB and 0.4 % for Yelp. The "UL-Subtract" step largely eliminates the backdoor, and "UL-Distill" restores clean accuracy with minimal backdoor reintroduction. This method consistently outperforms baselines

Table 7
Performance analysis comparison (BA—Backdoor Attack, MIA—Membership Inference Attack, AA—Adversarial Attack, and IA—Inversion Attack, ASR—Attack Success Rate, AUC-Area Under the Curve, MSE-Mean Squared Error, MUR- Malicious Unlearning Request, SISA- Sharding Isolating Slicing and Aggregating, PUMA-Performance Unchanged Model Augmentation, GBU- Gradient-Based Unlearning.).

Ref.	Attack type	Dataset	MU type	Performance Metric	Result
Attack					
[49]	BA	MNIST, Fashion-MNIST, Binary-MNIST, HAR	certified removal	reduce the retrain interval	reduced the retrain interval by 70–100 %
[33]	BA	CIFAR-10, GTSRB, MNIST, Rotated MNIST, Tiny-ImageNet	Exact (SISA, Full Retrain), Approximate(PUMA, GBU, LIRF)	ASR	ASR ranging from 90 % – 99.5 %
[45]	BA	CIFAR-10, Tiny-ImageNet	Exact(SISA), Approximate (First-Order, Second-Order, UnrollSGD)	ASR	ASR ranging from 50.7 % – 92.3 %.
[9]	BA	Purchase, MNIST, CIFAR-10, CIFAR-100, Credit Score	FedEraser, KNOT	ASR	80 % ASR when 0.3 % MUR, 100 % ASR when 10 % MUR
[13]	MIA	MNIST, CIFAR-10, STL10	SISA, Full-Retrain	AUC	AUC ranging from 0.504 to 0.986
[77]	AA	CIFAR-10, Adult, Diabetes, MNIST	Exact(SISA), Approximate(First-order, Second-order, Unrolling SGD, Amnesiac)	ASR	ASR ranging from 67 % – 100 %
[28]	AA	CIFAR-10, MNIST, Fashion-MNIST	Exact(SISA, Full-Retrain)	statistical test (NULL hypothesis: 0.5 (Successful adaptive unlearning)	falsifies the null hypothesis (Significant deviation from 0.5)
[32]	IA	CIFAR-10, CIFAR-100, STL-10, Chest X-Rays	Exact(Full retrain), Approximate(Single gradient)	MSE	MSE ranging from 0.04 – 0.51.
Defense					
[44]	BA	MNIST, Fashion-MNIST, CIFAR-10, CIFAR-100	Gradient Ascent	ASR	ASR ranging from 0.23 % to 13.53 %
[40]	BA	CIFAR-10, ImageNet-12 subset, GTSRB	Neuron Pruning	ASR	ASR ranging from 0.16 % to 15.24 %
[64]	BA	CIFAR-10, Tiny-ImageNet, GTSRB	Fine-tuning (SAU)	ASR	ASR 0.01 % to 85.75 %
[41]	BA	CIFAR-10, MNIST, TSRD	Partial Training	ASR	ASR ranging from 0 % to 0.8 %
[80]	BA	MNIST, CIFAR-10, CIFAR-100, GTSRB	Unlearning-based Model Ablation	AUC	AUC ranging from 83.2 % to 99 %
[50]	BA	CIFAR-10, GTSRB, ImageNet-1 K	Gradient Ascent	ASR	ASR ranging from 0.12 % to 51.04 %
[17]	BA	MNIST, CIFAR-10, Fashion-MNIST	client-level FU via gradient ascent	Backdoor Accuracy	Backdoor Accuracy dropped (ranging from 10.13 % to 11.23 %)
[66]	BA	MNIST, Fashion-MNIST, CIFAR-10, GTSRB, IMDB, Yelp Reviews	Knowledge Distillation	Backdoor Accuracy	Backdoor Accuracy ranging from 0.2 % to 7.5 %
[50]	AA	CIFAR-10, GTSRB, ImageNet-1 K	Gradient Ascent	Robust Accuracy	Robust Accuracy ranging from 68.46 % to 79.89 %
[42]	AA	MNIST-b, Covtype, CIFAR-10, Downsampled ImageNet, Lacuna-10, Lacuna-100	MUter	Robustness (Pertubed Accuracy)	Robustness ranging from 0.325 to 0.91
[21]	IA	CIFAR-10, MNIST, Adult	Statistical MU	LPIPS (Learned Perceptual Image Patch Similarity)	LPIPS ranging from 0.62 to 0.88

like “Forgetting” and “NTL”, and proves effective under multiple attacker scenarios and robust against distillation datasets with up to 10 % poisoning.

The authors of MUter [42] propose a novel machine unlearning method designed for adversarially trained models (ATMs), with robustness evaluation as its primary focus. The defense effectiveness is assessed through perturbed accuracy across both linear models (MNIST-b, Covtype) and neural network models (CIFAR-10, Lacuna-10). MUter consistently achieves high perturbed accuracy, highlighting its strong robustness. For linear models, even under increasing unlearning requests, MUter demonstrates the closest performance to Retrain in terms of perturbed accuracy. Likewise, for neural networks, MUter exhibits the most consistent alignment with Retrain. Moreover, MUter preserves robustness under different magnitudes of injected Gaussian noise, confirming its ability to maintain stability after unlearning. Overall, MUter’s perturbed accuracy closely parallels that of retrained models—for example, Logistic Regression models attain values around 0.90 for MNIST-b and 0.60 for Covtype, while neural networks maintain values around 0.47 for CIFAR-10 and 0.50 for Lacuna-10, even with large removal sets. Gao et al. [21] introduce a defense method against gradient inversion attacks in FL, primarily evaluated using LPIPS (Learned Perceptual Image Patch Similarity) across CIFAR-10 and MNIST datasets. The defense, combining statistical machine unlearning and knowledge distillation, consistently achieves significantly higher LPIPS scores, indicating more unrecognizable reconstructed images, across various batch sizes and unlearning strategies compared to baselines. For instance, on CIFAR-10 with a batch size of 1, the average LPIPS reaches 0.78, substantially higher than “no defense” (0.19) and other baselines. On MNIST, the average LPIPS reaches 0.76 under similar conditions. This demonstrates the method’s superior ability to protect data privacy by making reconstructed images semantically meaningless.

Table 7 presents a summary of the performance comparison between the different attacks and defenses analyzed related to MU. The table indicates whether a paper proposes an attack against MU or employs MU as a defense mechanism against a specific attack. It also reports *(i)* the type of attack, *(ii)* the dataset used, *(iii)* the algorithm adopted for MU, *(iv)* the evaluation metrics applied for performance comparison, and *(v)* the main experimental results.

5. Challenges and open problems

In the previous section, we systematically explored the most popular security threats in ML and their relationship with Machine Unlearning. We have examined all the threats discussed in the papers we selected, namely Backdoor Attacks (BA), Membership Inference Attacks (MIA), Adversarial Attacks (AA), and Inversion Attacks (IA). By selecting 42 of 61 analyzed studies, we can now identify gaps and limitations in current approaches, which naturally lead to the discussion of emerging challenges and open points in MU systems.

As Unlearning techniques evolve and their adoption grows, new challenges and unresolved issues continue to emerge, demanding further investigation. Therefore, the following emerging challenges should be explored as potential research directions that could enhance the MU systems' security, reliability, and effectiveness.

- **Privacy-preserving MU**. Existing MU systems typically assume that the data slated for removal is directly accessible to the server performing the Unlearning process. But, especially in decentralized scenarios such as the one of Federated Unlearning, the service provider should not have access to users' data, raising critical concerns about how to ensure the privacy of the data being unlearned. An equally important challenge is verifying that unlearning has been correctly performed without revealing the underlying data. Emerging cryptographic approaches—such as zero-knowledge proofs [65], secure multiparty computation [78], or trusted execution environments—could enable privacy-preserving verification, but these methods require further research to balance efficiency, scalability, and security.
- **MU for Large Models**. MU becomes significantly more complex when dealing with large-scale models, such as deep neural networks (DNNs) and transformer-based architectures (e.g., GPT, BERT). The challenges are related to the scalability, privacy, and security, and stem from the size and distributed nature of these models.
- **Ethical and Regulatory Considerations**. Ensuring that all traces of personal or sensitive information are completely removed in Machine Unlearning (MU) remains a technically and legally complex challenge. Current data protection frameworks, such as the GDPR and its "Right To Be Forgotten", require verifiable proof that unlearning has been achieved—yet providing such guarantees in practice is difficult, especially in large and distributed ML systems [36]. While MU offers a promising mechanism for legal compliance by enabling the removal of user data upon request, the existing literature only partially addresses its broader regulatory and ethical boundaries. Beyond compliance, ethical concerns emerge when considering the potential misuse of MU. Malicious actors could abuse unlearning mechanisms to conceal harmful activities, erase forensic evidence, or selectively alter model behavior under the pretense of fulfilling regulatory requirements. This dual-use nature of MU emphasizes the importance of developing not only technically robust and legally verifiable methods, but also governance frameworks and safeguards that prevent misuse, ensuring that MU advances both privacy protection and the trustworthy use of machine learning.
- **Certified MU Framework with Blockchain Integration**. Traditional Machine Unlearning lacks transparency, preventing users from verifying data removal. Blockchain integration may offer verifiable, auditable, and tamper-proof Unlearning, but might introduce challenges like computational overhead and scalability. Future research must optimize efficiency while ensuring security and practical deployment.

6. Conclusion

In the current Machine Learning (ML) context, security threats are becoming more sophisticated, posing serious risks to data privacy and model integrity. Recently, a new paradigm known as Machine Unlearning (MU) has been designed to enable data removal in compliance with privacy laws (e.g., GDPR's Right to Be Forgotten) and to mitigate security risks. However, the relationship between ML threats and MU remains underexplored. To address this issue, this article reviews existing research on Machine Unlearning (MU) techniques and Machine Learning (ML) security threats to uncover their underlying relationships. To do so, we examined four key attack classes that are the most used in scientific literature, namely Backdoor Attacks, Membership Inference Attacks (MIA), Adversarial Attacks, and Inversion Attacks, analyzing their impact on MU and the role of MU in mitigating them. Additionally, we classified the interaction between these ML threats and MU into four main perspectives: *(i)* attacks against MU, *(ii)* MU as a defensive mechanism to counteract attacks, *(iii)* attacks as evaluation tools to assess the effectiveness of MU frameworks, and *(iv)* attacks as verification tools to ensure MU guarantees. Our study identifies critical gaps in existing defenses based on MU and verification methodologies, highlighting avenues for future research. This paper establishes a foundation for the development of robust, verifiable, and attack-resilient MU solutions in the evolving ML security landscape, highlighting key challenges and future research directions in this field.

The research works reviewed in this survey provide a foundation for deeper investigation into several underexplored aspects of Machine Unlearning. One promising direction is the systematic analysis of benchmark datasets used to evaluate MU frameworks. Furthermore, incorporating legal, ethical, and regulatory perspectives, such as aligning MU techniques with data protection laws like the GDPR, represents an important avenue for future research.

CRediT authorship contribution statement

Muhammed Shafi K.P.: Writing – original draft, Visualization, Validation, Methodology, Investigation. **Serena Nicolazzo:** Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Antonino Nocera:** Writing – review & editing, Supervision, Investigation, Formal analysis, Conceptualization. **Vinod P.:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported in part by the project "GoTMaT - Governing Technology to Manage the Transition" funded by the European Community - Next Generation EU, Mission 4 Component 2 Investment 1.3 - CUP B53C22003990006.

Data availability

No data was used for the research described in the article.

References

[1] M. Bertran, S. Tang, M. Kearns, J.H. Morgenstern, A. Roth, S.Z. Wu, Reconstruction attacks on machine unlearning: simple models are vulnerable, Adv. Neural Inf. Process. Syst. 37 (2024) 104995–105016.
[2] A. Blanco-Justicia, N. Jebreel, B. Manzanares-Salor, D. Sánchez, J. Domingo-Ferrer, G. Collell, K. Eeik Tan, Digital forgetting in large language models: a survey of unlearning methods, Artif. Intell. Rev. 58 (3) (2025) 90.
[3] L. Bourtoule, V. Chandrasekaran, C.A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, N. Papernot, Machine unlearning, in: 2021 IEEE Symposium on Security and Privacy (SP), IEEE, 2021, pp. 141–159.
[4] Y. Cao, J. Yang, Towards making systems forget with machine unlearning, in: 2015 IEEE Symposium on Security and Privacy, IEEE, 2015, pp. 463–480.
[5] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, F. Tramer, Membership inference attacks from first principles, in: 2022 IEEE Symposium on Security and Privacy (SP), IEEE, 2022, pp. 1897–1914.

[6] S. Chaudhari, P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, A. Deshpande, B. Castro da Silva, RLHF deciphered: a critical analysis of reinforcement learning from human feedback for llms, ACM Comput. Surv. 58 (2) (2024) 1–37.

[7] A. Chen, Y. Li, C. Zhao, M. Huai, A survey of security and privacy issues of machine unlearning (2025).

[8] C. Chen, F. Sun, M. Zhang, B. Ding, Recommendation unlearning, in: Proceedings of the ACM Web Conference 2022, ACM, 2022, pp. 2768–2777.

[9] J. Chen, Z. Lin, W. Lin, W. Shi, X. Yin, D. Wang, FedMUA: exploring the vulnerabilities of federated learning to malicious unlearning attacks, IEEE Trans. Inf. Forensics Secur. 20 (2025) 1665–1678.

[10] J. Chen, W. Shi, W. Lin, C. Wang, W. Liu, H. Sun, G. Liu, Unlearning attacks for regression learning, IEEE Trans. Neural Netw. Learn. Syst. 36 (9) (2025) 15851–15865.

[11] K. Chen, Y. Huang, Y. Wang, X. Zhang, B. Mi, Y. Wang, Privacy preserving machine unlearning for smart cities, Ann. Telecommun. 79 (1) (2024) 61–72.

[12] K. Chen, Y. Wang, L. Zhao, C. Jiang, H. Mai, Y. Wu, H. Hong, Y. Shen, J. Mo, L.L. Huang, et al., Private data protection with machine unlearning for next-generation networks, IEEE Open J. Commun. Soc. 6 (2024) 3280–3291.

[13] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, Y. Zhang, When machine unlearning jeopardizes privacy, in: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 896–911.

[14] P.Y. Chen, H. Zhang, Y. Sharma, J. Yi, C.J. Hsieh, ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017, pp. 15–26.

[15] S. Chen, M. Kahla, R. Jia, G.J. Qi, Knowledge-enriched distributional model inversion attacks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16178–16187.

[16] V.S. Chundawat, A.K. Tarun, M. Mandal, M. Kankanhalli, Zero-shot machine unlearning, IEEE Trans. Inf. Forensics Secur. 18 (2023) 2345–2354.

[17] W. Daluwatta, I. Khalil, S. Edirimannage, M. Atiquzzaman, UaaS-SFL: unlearning as a service for safeguarding federated learning, IEEE Trans. Netw. Serv. Manag. 22 (2) (2024) 1029–1045.

[18] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, M. Naor, Our data, ourselves: Privacy via distributed noise generation, in: Annual International Conference on the Theory and Applications of Cryptographic Techniques, Springer, 2006, pp. 486–503.

[19] M. Fredrikson, S. Jha, T. Ristenpart, Model inversion attacks that exploit confidence information and basic countermeasures, in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015, pp. 1322–1333.

[20] C. Ganhör, D. Penz, N. Rekabsaz, O. Lesota, M. Schedl, Unlearning protected user attributes in recommendations with adversarial training, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2022, pp. 2142–2147.

[21] K. Gao, T. Zhu, D. Ye, W. Zhou, Defending against gradient inversion attacks in federated learning via statistical machine unlearning, Knowl.-Based Syst. 299 (2024) 111983.

[22] A. Golatkar, A. Achille, A. Ravichandran, M. Polito, S. Soatto, Mixed-privacy forgetting in deep networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 792–801.

[23] A. Golatkar, A. Achille, S. Soatto, Forgetting outside the box: scrubbing deep networks of information accessible from input-output observations, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16, Springer, 2020, pp. 383–398.

[24] X. Gong, Z. Wang, S. Li, Y. Chen, Q. Wang, A GAN-based defense framework against model inversion attacks, IEEE Trans. Inf. Forensics Secur. 18 (2023) 4475–4487.

[25] L. Graves, V. Nagisetty, V. Ganesh, Amnesiac machine learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35(13), 2021, pp. 11516–11524.

[26] C. Guo, T. Goldstein, A. Hannun, L. van der Maaten, Certified data removal from machine learning models (2023) https://arxiv.org/abs/1911.03030 .

[27] Y. Guo, Y. Zhao, S. Hou, C. Wang, X. Jia, Verifying in the dark: verifiable machine unlearning by using invisible backdoor triggers, IEEE Trans. Inf. Forensics Secur. 19 (2023) 708–721.

[28] V. Gupta, C. Jung, S. Neel, A. Roth, S. Sharifi-Malvajerdi, C. Waites, Adaptive machine unlearning, Adv. Neural Inf. Process. Syst. 34 (2021) 16319–16330.

[29] M. Han, T. Zhu, L. Zhang, H. Huo, W. Zhou, Vertical federated unlearning via backdoor certification, IEEE Trans. Serv. Comput. 18 (2) (2025) 1110–1123.

[30] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P.S. Yu, X. Zhang, Membership inference attacks on machine learning: a survey, ACM Comput. Surv. 54 (11s) (2022) 1–37.

[31] H. Hu, S. Wang, J. Chang, H. Zhong, R. Sun, S. Hao, H. Zhu, M. Xue, A duty to forget, a right to be assured? exposing vulnerabilities in machine unlearning services, in: Network and Distributed System Security (NDSS) Symposium 2024, IEEE, 2024, pp. 0.

[32] H. Hu, S. Wang, T. Dong, M. Xue, Learn what you want to unlearn: unlearning inversion attacks against machine unlearning, in: 2024 IEEE Symposium on Security and Privacy (SP), IEEE, 2024, pp. 3257–3275.

[33] Z. Huang, Y. Mao, S. Zhong, {UBA-Inf}: unlearning activated backdoor attack with {Influence-Driven} camouflage, in: 33rd USENIX Security Symposium (USENIX Security 24), 2024, pp. 4211–4228.

[34] J. Jia, J. Liu, P. Ram, Y. Yao, G. Liu, Y. Liu, P. Sharma, S. Liu, Model sparsity can simplify machine unlearning, Adv. Neural Inf. Process. Syst. 36 (2023) 51584–51605.

[35] Y. Jiang, X. Tong, Z. Liu, H. Ye, C.W. Tan, K.Y. Lam, Efficient federated unlearning with adaptive differential privacy preservation, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, pp. 7822–7831.

[36] B.A. Juliussen, J.P. Rui, D. Johansen, Algorithms that forget: machine unlearning and the right to erasure, Comput. Law Secur. Rev. 51 (2023) 105885.

[37] M. Kurmanji, P. Triantafillou, J. Hayes, E. Triantafillou, Towards unbounded machine unlearning, Adv. Neural Inf. Process. Syst. 36 (2023) 1957–1987.

[38] S. Lee, D.H. Choi, Learning and unlearning to operate profitable secure electric vehicle charging, IEEE Trans. Ind. Informat. 20 (9) (2024) 11213–11223.

[39] N. Li, C. Zhou, Y. Gao, H. Chen, Z. Zhang, B. Kuang, A. Fu, Machine unlearning: taxonomy, metrics, applications, challenges, and prospects, IEEE Trans. Neural Netw. Learn. Syst. 36 (8) (2025) 13709–13729.

[40] Y. Li, X. Lyu, X. Ma, N. Koren, L. Lyu, B. Li, Y.G. Jiang, Reconstructive neuron pruning for backdoor defense, in: International Conference on Machine Learning, PMLR, 2023, pp. 19837–19854.

[41] Y. Li, H. Gao, H. Chen, Y. Wang, C. Yu, Partially training, isolating and unlearning, mitigating backdoor attack, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, pp. 6313–6319.

[42] J. Liu, M. Xue, J. Lou, X. Zhang, L. Xiong, Z. Qin, Muter: machine unlearning on adversarially trained models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4892–4902.

[43] S. Liu, Y. Yao, J. Jia, S. Casper, N. Baracaldo, P. Hase, Y. Yao, C.Y. Liu, X. Xu, H. Li, K.R. Varshney, M. Bansal, S. Koyejo, Y. Liu, Rethinking machine unlearning for large language models, Nat. Mach. Intell. 7 (2) (2025) 181–194.

[44] Y. Liu, M. Fan, C. Chen, X. Liu, Z. Ma, L. Wang, J. Ma, Backdoor defense with machine unlearning, in: IEEE INFOCOM 2022-IEEE Conference on Computer Communications, IEEE, 2022, pp. 280–289.

[45] Z. Liu, T. Wang, M. Huai, C. Miao, Backdoor attacks via machine unlearning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38(13), 2024, pp. 14115–14123.

[46] Z. Liu, Y. Jiang, J. Shen, M. Peng, K.Y. Lam, X. Yuan, X. Liu, A survey on federated unlearning: challenges, methods, and future directions, ACM Comput. Surv. 57 (1) (2024) 1–38.

[47] Z. Liu, H. Ye, C. Chen, Y. Zheng, K.Y. Lam, Threats, attacks, and defenses in machine unlearning: a survey, IEEE Open J. Comput. Soc. 6 (2025) 413–425.

[48] Z. Ma, Y. Liu, X. Liu, J. Liu, J. Ma, K. Ren, Learn to forget: machine unlearning via neuron masking, IEEE Trans. Dependable Secure Comput. 20 (4) (2022) 3194–3207.

[49] N.G. Marchant, B.I.P. Rubinstein, S. Alfeld, Hard to forget: poisoning attacks on certified machine unlearning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36(7), 2022, pp. 7691–7700.

[50] Z. Niu, Y. Sun, Q. Miao, R. Jin, G. Hua, Towards unified robustness against both backdoor and adversarial attacks, IEEE Trans. Pattern Anal. Mach. Intell. 46 (12) (2024) 7589–7605.

[51] N. Romandini, A. Mora, C. Mazzocca, R. Montanari, P. Bellavista, Federated unlearning: a survey on methods, design guidelines, and evaluation metrics, IEEE Trans. Neural Netw. Learn. Syst. 36 (7) (2024) 11697–11717.

[52] J. Rosen, The right to be forgotten, Stan. L. Rev. Online 64 (2011) 88.

[53] S. Sai, U. Mittal, V. Chamola, K. Huang, I. Spinelli, S. Scardapane, Z. Tan, A. Hussain, Machine un-learning: an overview of techniques, applications, and future directions, Cogn. Comput. 16 (2) (2024) 482–506.

[54] L. Schwinn, D. Dobre, S. Xhonneux, G. Gidel, S. Günnemann, Soft prompt threats: attacking safety alignment and unlearning in open-source llms through the embedding space, Adv. Neural Inf. Process. Syst. 37 (2024) 9086–9116.

[55] T. Shaik, X. Tao, H. Xie, L. Li, X. Zhu, Q. Li, Exploring the landscape of machine unlearning: a comprehensive survey and taxonomy, IEEE Trans. Neural Netw. Learn. Syst. 36 (7) (2024) 11676–11696.

[56] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE Symposium on Security and Privacy (SP), IEEE, 2017, pp. 3–18.

[57] D.M. Sommer, L. Song, S. Wagh, P. Mittal, Athena: probabilistic verification of machine unlearning, Proc. Priv. Enhancing Technol. 2022 (3) (2022) 268–290.

[58] C. Stokel-Walker, R. Van Noorden, What ChatGPT and generative AI mean for science, Nature 614 (7947) (2023) 214–216.

[59] A.K. Varshney, V. Torra, Efficient federated unlearning under plausible deniability, Mach. Learn. 114 (1) (2025) 25.

[60] C. Wang, X. Liu, Y. Yue, Q. Guo, X. Tang, T. Zhang, C. Jiayang, Y. Yao, W. Gao, X. Hu, Z. Qi, Y. Wang, L. Yang, J. Wang, X. Xie, Z. Zhang, Y. Zhang, Survey on factuality in large language models: knowledge, retrieval and domain-specificity, ACM Comput. Surv. (2023).

[61] T. Wang, Y. Zhang, R. Jia, Improving robustness to model inversion attacks via mutual information regularization, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35(13), 2021, pp. 11666–11673.

[62] W. Wang, C. Zhang, Z. Tian, S. Liu, S. Yu, CRFU: compressive representation forgetting against privacy leakage on machine unlearning, IEEE Trans. Dependable Secure Comput. 22 (4) (2025) 3916–3929.

[63] A. Warnecke, L. Pirch, C. Wressnegger, K. Rieck, Machine unlearning of features and labels (2023) https://arxiv.org/abs/2108.11577 .

[64] S. Wei, M. Zhang, H. Zha, B. Wu, Shared adversarial unlearning: backdoor mitigation by unlearning shared adversarial examples, Adv. Neural Inf. Process. Syst. 36 (2023) 25876–25909.

[65] C. Weng, K. Yang, J. Katz, X. Wang, Wolverine: fast, scalable, and communication-efficient zero-knowledge proofs for boolean and arithmetic circuits, in: 2021 IEEE Symposium on Security and Privacy (SP), IEEE, 2021, pp. 1074–1091.

[66] C. Wu, S. Zhu, P. Mitra, W. Wang, Unlearning backdoor attacks in federated learning, in: 2024 IEEE Conference on Communications and Network Security (CNS), IEEE, 2024, pp. 1–9.

[67] J. Xu, Z. Wu, C. Wang, X. Jia, Machine unlearning: solutions and challenges, IEEE Trans. Emerg. Top. Comput. Intell. 8 (3) (2024) 2150–2168.

[68] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, X. Hu, Harnessing the power of llms in practice: a survey on ChatGPT and beyond, ACM Trans. Knowl. Discov. Data 18 (6) (2024) 1–32.
[69] Y. Yao, X. Xu, Y. Liu, Large language model unlearning, Adv. Neural Inf. Process. Syst. 37 (2024) 105425–105475.
[70] D. Ye, T. Zhu, J. Li, K. Gao, B. Liu, L.Y. Zhang, W. Zhou, Y. Zhang, Data duplication: A novel multi-purpose attack paradigm in machine unlearning, USENIX (2025).
[71] H. Yuan, Z. Jin, P. Cao, Y. Chen, K. Liu, J. Zhao, Towards robust knowledge unlearning: an adversarial framework for assessing and improving unlearning robustness in large language models, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39 (24), 2025, pp. 25769–25777.
[72] W. Yuan, H. Yin, F. Wu, S. Zhang, T. He, H. Wang, Federated unlearning for on-device recommendation, in: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, ACM, 2023, pp. 393–401.
[73] L. Zhang, T. Zhu, H. Zhang, P. Xiong, W. Zhou, Fedrecovery: differentially private machine unlearning for federated learning frameworks, IEEE Trans. Inf. Forensics Secur. 18 (2023) 4732–4746.
[74] R. Zhang, S. Guo, J. Wang, X. Xie, D. Tao, A survey on gradient inversion: attacks, defenses and future directions, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 5678–5685.
[75] Y. Zhang, Z. Hu, Y. Bai, J. Wu, Q. Wang, F. Feng, Recommendation unlearning via influence function, ACM Trans. Recomm. Syst. 3 (2) (2024) 1–23.
[76] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, D. Song, The secret revealer: generative model-inversion attacks against deep neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 253–261.
[77] C. Zhao, W. Qian, R. Ying, M. Huai, Static and sequential malicious attacks in the context of selective forgetting, Adv. Neural Inf. Process. Syst. 36 (2023) 74966–74979.
[78] C. Zhao, S. Zhao, M. Zhao, Z. Chen, C.Z. Gao, H. Li, Y.A. Tan, Secure multi-party computation: theory, practice and applications, Inf. Sci. 476 (2019) 357–372.
[79] S. Zhao, J. Zhang, X. Ma, Q. Jiang, Z. Ma, S. Gao, Z. Ying, J. Ma, FedWiper: federated unlearning via universal adapter, IEEE Trans. Inf. Forensics Secur. 20 (2025) 4042–4054.
[80] Y. Zhao, C. Li, K. Chen, UMA: facilitating backdoor scanning via unlearning-based model ablation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38(19), 2024, pp. 21823–21831.

Author biography

Muhammed Shafi K.P. is a Ph.D. Scholar at the Department of Computer Applications, Cochin University of Science & Technology, Cochin, Kerala, India. He received a Master's in Computer Science with specialization in Artificial Intelligence from Cochin University of Science & Technology. His main research interests include Security of AI, Cyber Threat Intelligence, Deep Learning, Machine Unlearning, and Malware Analysis.

Serena Nicolazzo is a Tenure Track Researcher at University of Eastern Piedmont. She got a PhD in Information Engineering at the University Mediterranea of Reggio Calabria in 2017. She was Research Fellow at the University of Milan from 2023 to May 2025 and Research Fellow at the Information Engineering, Infrastructure, and Sustainable Energy (DIIES) Department of the University Mediterranea of Reggio Calabria in 2018. Her research interests include Security, Privacy, and Social Network Analysis. She is involved in several TPCs of international conferences, and she is an Editorial Board Member for Online Social Networks and Media (OSNEM), Computers, Materials & Continua (CMC), and the Journal of Electrical and Computer Engineering (JECE). She is the author of about 50 scientific papers. She was a Visiting Researcher at the Middlesex University of London, and she is currently collaborating with the Polytechnic University of Marche, the University of Pavia, the University of Padua, the University College of London, and the Cochin University of Science and Technology.

Antonino Nocera is an Associate Professor at the University of Pavia. His research interests span over Artificial Intelligence, Cybersecurity, and Data Science. The results of his research in these domains are collected in about 100 research papers published in prestigious international journals and conferences. He is the Director of the DCALab laboratory of the University of Pavia in which he leads a research group, characterized by several international collaborations, focusing on Security of Artificial Intelligence, Data Science, and Cybersecurity. He is Associate Editor of Information Sciences (Elsevier), the IEEE Transactions on Information Forensics and Security (T-IFS), and the IEEE Transactions on Cybernetics. Moreover, he is involved in the TPC of many renowned International Conferences focusing both on cybersecurity and artificial intelligence, such as NDSS and ESORICS. He is the director of the local node of the University of Pavia for the CINI Data Science National Lab and a member of the local node of the University of Pavia for the CINI Cybersecurity National Lab. Finally, he has also served as principal investigator and unit coordinator for several competitive funded research projects in the fields of data science and cybersecurity.

Vinod P. is presently a Professor & Head in the Department of Computer Applications at Cochin University of Science & Technology, Cochin, Kerala, India. Between March 2023 to March 2025, he was a Marie Curie fellow at the University of Padua. He was also a Postdoctoral Researcher at the Department of Mathematics, University of Padua, Italy, where he was part of the EU-H2020 project named TagitSmart. Additionally, he was a Postdoctoral researcher at Malaviya National Institute of Technology, Jaipur, Rajasthan, India, under the ISEA project on Mobile Security. He holds his Ph.D. in Computer Engineering from Malaviya National Institute of Technology, Jaipur, India. In 2020, he was awarded the Seal of Excellence for a Marie Sklodowska-Curie Individual Fellowship by the European Commission. Subsequently, in 2021 he was awarded the prestigious Marie Sklodowska Curie Fellowship by the European commission for the project titled OPTIMA: Organization Specific Threat Intelligence Mining & Sharing. He has numerous research articles published in peer-reviewed Journals and International Conferences. He is a reviewer of a number of security journals such as IEEE Transactions of Information Forensics, IEEE Communication Surveys and Tutorials, and Elsevier Computer Communications, and is also serving as a programme committee member in International Conferences related to Computer and Information Security. Vinod's area of interest is Malware Analysis, Security of AI, Cyber Threat Intelligence, Federated Learning and Natural Language Processing.