Machine Unlearning: An Overview of the Paradigm Shift in the Evolution of AI

Layan Jaman
Computer Science Department,
Effat College of Engineering
Energy and Technology Research Center
Effat University
Jeddah, Saudi Arabia
Lajaman@effat.edu.sa

Reem Alsharabi
Computer Science Department,
Effat College of Engineering
Energy and Technology Research Center
Effat University
Jeddah, Saudi Arabia
reualsharabi@effat.edu.sa

Passent M. ElKafrawy, Ph.D.
Computer Science Department,
Effat College of Engineering
Energy and Technology Research Center
Effat University
Jeddah, Saudi Arabia
passent.elkafrawy@ieee.org

***Abstract*— The rapid advancements in artificial intelligence (AI) have primarily focused on the process of learning from data to acquire knowledge for smart systems. However, the concept of machine unlearning has emerged as a transformative paradigm shift in the field of AI, due to the amount of false information that have been learned over the past. Machine unlearning refers to the ability of AI systems to reverse or discard previously acquired knowledge or patterns, enabling them to adapt and refine their understanding in response to changing circumstances or new insights. This paper explores the concept of machine unlearning, its implications, methods, challenges, and potential applications. The paper begins by providing an overview of the traditional learning-based approaches in AI and the limitations they impose on system adaptability and agility. It then delves into the concept of machine unlearning, discussing various techniques and algorithms employed to remove or modify learned knowledge from AI models or datasets.**

Keywords– machine unlearning, artificial intelligence, data deletion, differential privacy, adaptive algorithms.

I. INTRODUCTION

Artificial intelligence (AI) has revolutionized nearly every aspect of modern life, from healthcare, finance, education, transportation, and others. Its transformative capabilities, including learning from data, making decisions, and performing complex tasks, have not only reshaped industries but have also opened up new possibilities for human endeavor. However, the increasing sophistication and ubiquity of AI systems come with a set of challenges and concerns, particularly in ensuring the fairness, transparency, and safety of these systems.

In recent years, researchers have turned their attention to addressing these challenges through the exploration of the concept of machine unlearning [1], a novel paradigm in AI. Machine unlearning encompasses various techniques designed to enable AI systems to forget or modify previously sought after information. This capability allows these systems to adapt to changing circumstances, rectify errors, and mitigate biases [1]. The integration of machine learning, while revolutionary, has prompted concerns related to privacy hazards, security vulnerabilities, and accuracy degradation in dynamic environments.

In response to these concerns, the concept of machine unlearning has emerged as a promising technique. This approach involves selectively removing the influence of specific training data points on a pre-trained machine learning model. The objective is to ensure that the updated model behaves as if it were never trained on that specific data. Machine unlearning offers a subtractive capability, allowing models to adapt by selectively eliminating unauthorized, malicious, or outdated data points without the necessity for complete retraining. Crucially, machine unlearning plays a pivotal role in various applications, with a primary focus on enforcing privacy regulations and safeguarding user privacy [4]. Machine unlearning offers a promising solution to these challenges by providing AI systems with the ability to selectively forget or modify previously learned misinformation. This flexibility allows AI systems to remain fair, transparent, and safe in dynamic and evolving environments.

II. BACKGROUND

Machine unlearning is defined as the process of eliminating the impact of particular training data points from a machine learning model that has already undergone training [4]. In a formal context, when provided with a model characterized by parameters w∗, trained on a dataset D through the learning algorithm A, and an identified subset $Df \subseteq D$ earmarked for removal, the objective of the machine unlearning algorithm, denoted as $U(A(D), D, Df)$, is to derive new parameters $\overline{w}$. This derivation aims to eliminate the influences of Df while ensuring the preservation of the model's performance on the remaining dataset, $D \setminus Df$. In summary, machine unlearning involves refining a pre-trained model by selectively excluding the effects of specific data

979-8-3503-9356-9/24/$31.00 ©2024 IEEE

points, facilitating adaptability and enhancement without necessitating complete retraining.

The field of machine unlearning is still in its early stages of development, but it has already attracted significant attention from researchers and industry practitioners. Several promising techniques have emerged, including active learning [10], forgetting algorithms [11], and counterfactual learning [12]. These techniques are being applied to a wide range of AI applications, from fraud detection to medical diagnosis. As machine unlearning continues to develop, it is likely to play a transformative role in the evolution of AI. By enabling AI systems to learn from data more responsibly and ethically, machine unlearning can help to ensure that AI remains a force for good in society. Machine unlearning has emerged as a critical area of research in the field of AI. The increasing complexity and pervasiveness of AI systems have raised concerns about their fairness, transparency, and safety. Machine unlearning offers a promising approach to addressing these concerns by allowing AI systems to forget or modify previously learned information selectively. Several promising machine unlearning techniques have been proposed in recent years.

- *Active learning* [10] involves strategically selecting data points to label or un-label, reducing the computational cost of unlearning.
- *Forgetting algorithms* [11] directly modify the internal parameters of a trained model to remove the influence of specific data points. This concept aligns with the growing need for privacy and data deletion in machine learning systems. These algorithms operate within the framework of privacy-preserving machine learning, which often utilizes mechanisms like Stochastic Gradient Descent (SGD) and its privacy-preserving variants.[11]
- *Counterfactual learning* [12] generates alternative data points that could have been used to train the model, allowing it to identify and correct errors.

These techniques have been applied to a wide range of AI applications, including fraud detection, medical diagnosis, and natural language processing. Early results suggest that machine unlearning can effectively remove biases, improve accuracy, and enhance the overall performance of AI systems. As research in this area continues, machine unlearning is poised to play a transformative role in the evolution of AI. By enabling AI systems to learn from data more responsibly and ethically, machine unlearning can help to ensure that AI remains a force for good in society.

III. Literature Review

Machine unlearning, as proposed by Cao and Yang (2015) in their research [1], addressed a crucial challenge in machine learning by efficiently removing the influence of specific training data from machine learning models. This concept enhances the privacy, security, and usability of such systems. The significance of machine unlearning lies in its ability to facilitate the forgetting of data while minimizing the need for time-consuming model retraining, especially with large training datasets. Traditionally, the unlearning process involves retraining the entire machine-learning model from scratch, which can be slow and resource-intensive. However, Cao and Yang (2015) introduced a novel approach that transforms learning algorithms into a summation form. In this form, the model's reliance on individual data was replaced by a set of summations representing transformations of the training data samples, stored alongside the trained model. During the unlearning process, specific training data was subtracted from these summations, and the model is updated accordingly. This efficient approach allowed faster and more precise data removal, outperforming the traditional retraining method. What sets Cao and Yang's (2015) work apart is its versatility and applicability to various machine learning algorithms, including those used in recommendation systems, malware detection, and spam filtering. Extensive empirical evaluations conducted on real-world systems demonstrate the practicality, completeness, speed, and ease of implementation of this approach. This research significantly contributes to the existing body of knowledge on machine unlearning techniques. Their proposed framework and algorithm offer a promising avenue for selectively forgetting learned information, thereby opening up possibilities for improved privacy, security, and flexibility in machine learning systems.

In the context of the current literature, Cao and Yang's (2015) research represented a pioneering contribution that establishes the importance of efficient unlearning in enhancing the resilience and security of machine learning systems. Their approach offered a comprehensive solution to the challenges associated with unlearning, distinguishing it from previous work.

The study by Sekhari et al. [2] significantly advanced the field of machine unlearning, providing novel insights and algorithms that address key challenges. The research focused on two main aspects: generalization properties and mathematical concepts. In machine learning, generalization properties refer to a model's ability to perform well on previously unseen test data, extending beyond its training data. Machine unlearning becomes crucial to maintaining this generalization, ensuring that removing specific data points from the training set does not significantly degrade the model's performance on new, unseen test data. The goal is to design unlearning algorithms capable of preserving strong generalization performance while selectively removing certain data points from the training set.

Sekhari et al. in [2] introduced a new line of inquiry within machine unlearning, emphasizing the investigation into the generalization properties of unlearning algorithms. The study explored the critical question of how many samples can be unlearned while still ensuring satisfactory performance on unseen test data. This inquiry addressed practical concerns about the safe removal of data from a machine learning model, revealing insights into the limits of data removal without compromising the model's ability to generalize to new data.

To provide context for understanding machine unlearning, the study introduced preliminary concepts related to distribution, parameter space, loss functions, and the goal of minimizing test loss population risk. The authors emphasized the reliance on samples due to the often unknown distribution and introduced the notion of sample complexity, defined as the minimum number of samples required to achieve a specified suboptimal minimizer of population loss. Furthermore, the study considered machine unlearning under both storage and computation constraints. Notably, the proposed algorithm do not require the training data during the sample deletion process, differentiating them from previous approaches.

A clear distinction is established between differential privacy (DP) and machine unlearning. While DP-based algorithms are limited in their ability to delete samples, the authors presented efficient unlearning algorithms capable of handling a significantly larger number of samples. In summary, this paper [2] marked a novel exploration into the realm of machine unlearning, distinctly prioritizing population risk minimization—an approach that diverges from prevailing efforts emphasizing empirical risk minimization. The study introduced a groundbreaking unlearning algorithm tailored for convex loss functions. The algorithm outperformed existing deletion capacities but demonstrated an improvement by at least a quadratic factor in dimensionality (D) when compared to utilizing conventional differentially private algorithms for unlearning. An intriguing avenue for future research lies in establishing dimension-dependent information-theoretic lower bounds on the deletion capacity. Additionally, the authors predicted exciting prospects in developing efficient unlearning algorithms applicable to finite/discrete hypothesis classes and non-convex loss functions. Their work tackled the challenge of batch deletion, where all delete requests (U) are simultaneous. However, extending their algorithms to the online case presents a compelling research direction that is eagerly anticipated to explore in future endeavors.

The new conceptual distinctions introduced in paper [3] included the differentiation between perfect and non-perfect deletion algorithms, the ability to handle arbitrary sequences of updates (including additions and deletions), and the differentiation between weak and strong unlearning algorithms. These distinctions impact the efficiency of deletion algorithms by allowing for more flexible and powerful methods that can handle a wider range of scenarios and can achieve better performance under weaker deletion criteria. For example, the proposed gradient-based methods can handle adversarial updates while maintaining a low steady-state error. Improved bounds for several problems are achieved by allowing algorithms to maintain a secret state and handle arbitrary sequences of updates. This study explores the problem of data deletion for convex models in the context of machine unlearning. The authors present novel Descent-to-Delete algorithms, leveraging techniques from convex optimization and reservoir sampling to handle adversarial updates while maintaining a low steady-state error. They defined the first data deletion algorithms capable of handling an arbitrarily long sequence of adversarial updates while promising both per-deletion run-time and steady-state error that does not grow with the length of the updated sequence. The introduced conceptual distinctions in [3], such as the differentiation between perfect and non-perfect deletion algorithms and the ability to handle arbitrary sequences of updates, allowed for more flexible and powerful methods that can achieve better performance under weaker deletion criteria. Overall, this paper presented a significant contribution to the field of machine unlearning, providing valuable insights into the problem of data deletion for convex models. These distinctions allow for more flexible and powerful methods that achieve better performance under weaker deletion criteria. Finally, the paper provides improved bounds for several problems by allowing algorithms to maintain a secret state and handle arbitrary sequences of updates.

IV. Comparative Analysis

The literature review provides an overview of machine unlearning, highlighting key research contributions and challenges in the field. In this section, we will analyze the literature review and discuss the contributions and implications of the studies mentioned.

A. Machine Unlearning: Concept and Techniques

Cao and Yang's approach to machine unlearning [1] has yielded impressive results, surpassing traditional retraining methods in both speed and precision. The applicability of this method across diverse machine learning (ML) algorithms is evident through its successful implementation in real-world systems. In discussion, Cao and Yang's pioneering work stands as a beacon for the efficiency and applicability of machine unlearning techniques. The demonstrated success in real-world applications not only validates the approach but also suggests a potential paradigm shift in the widespread adoption of this efficient unlearning framework.

Moreover, Sekhari et al. [2] contributed a novel insight into generalization properties in the realm of unlearning algorithms. Their focus on convex loss functions introduced algorithms that outperform existing capacities, showcasing advancements in preserving generalization performance during the unlearning process. In discussion, the emphasis on preserving generalization performance, as highlighted by Sekhari et al., reinforced the significance of unlearning in maintaining model adaptability to new and unseen data. The introduction of algorithms for convex loss functions not only addressed current challenges but also opened a promising avenue for future research, particularly in dimensionality improvement.

Nevertheless, Sekhari et al.'s [3] Descent-to-Delete algorithms mark a breakthrough, capable of handling adversarial updates while introducing crucial conceptual distinctions to enhance flexibility and performance under weaker deletion criteria. In discussion, the Descent-to-Delete

algorithms represent a significant stride in the robustness of machine unlearning. Their capacity to handle adversarial updates addresses a critical challenge, while the introduced conceptual distinctions pave the way for more adaptive and versatile unlearning methods. These advancements have the potential to revolutionize the field by addressing a wider range of scenarios and challenges.

In summary, the collective results and discussions showcase a dynamic landscape of advancements in machine unlearning. From efficient frameworks that outperform traditional methods to novel insights into generalization properties and ground-breaking Descent-to-Delete algorithms, the field is progressing toward more effective, versatile, and robust unlearning techniques. As we navigate these advancements, the potential for widespread implementation and transformative impacts on real-world ML applications becomes increasingly evident. However, amid this progress, it is crucial to acknowledge and navigate the challenges inherent in machine unlearning. These challenges, rooted in the complex nature of ML models and practical implementation limitations, underscore the intricacies embedded within ML systems.

Figure 1 [8] demonstrates the authors' unlearning approach, which involved learning a noise matrix to maximize error, followed by a single pass of weight update using "impair and repair" to efficiently remove knowledge of multiple classes without access to their samples. Successfully removed a model's ability to focus on key parts of images from classes it was trained to unlearn.

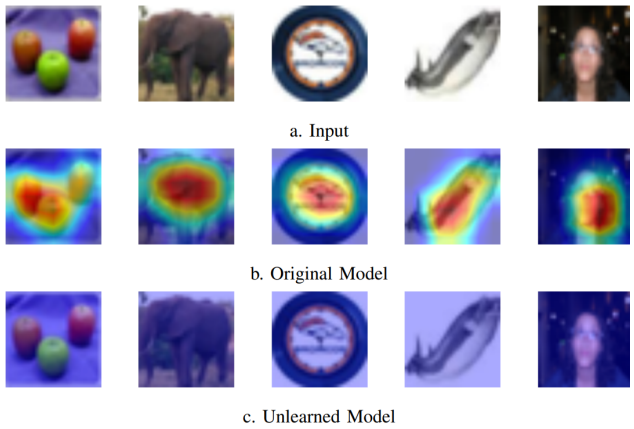

Fig. 1. Visualization of model's ability to focus on key parts of images after: b. Training, and, c. unlearning. [8]

In this paper [9], the focus is on machine unlearning as a defense against backdoor attacks. The process involves erasing historical updates from a target client, resulting in the unlearning model's accuracy dropping below 60%, with a slight decrease over federated learning rounds. Despite this, it successfully eliminates the target client's influence, maintaining a 0% backdoor attack success rate.

Figure 2 displays the training and subtraction unlearning processes with the MNIST dataset.

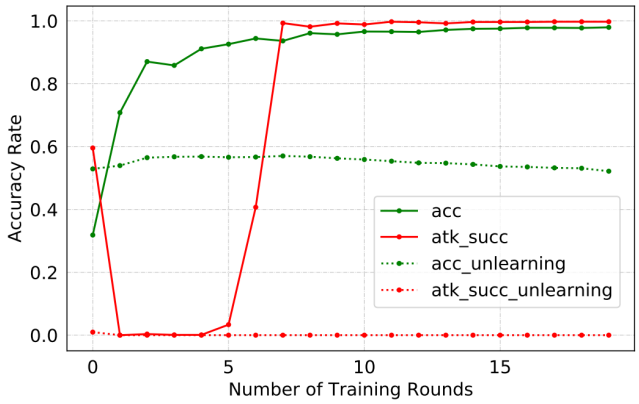

Fig. 2. Training and subtraction unlearning processes. [9]

To recover lost accuracy, the paper introduced knowledge distillation. Within just five epochs, the test accuracy was restored, approaching levels similar to the original global model. Importantly, the 0% backdoor attack success rate persists, indicating that the target client's influence did not transfer to the unlearning model post-distillation. The approach not only ensures privacy protection by removing the target client's contributions from the new global model but also proved effective in the context of the right to be forgotten. Figure 3 shows the unlearning model's performance in knowledge distillation training on the MNIST dataset. The blue line represents the change in losses, indicating model recovery, with a value closer to 1.

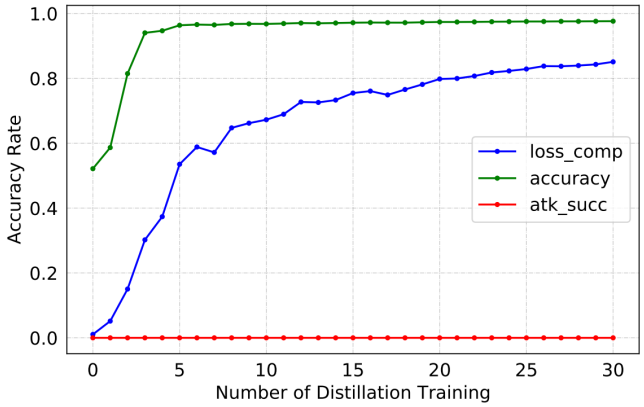

Fig. 3. Unlearning model's performance during knowledge distillation training on the MNIST dataset. [9]

B. *Challenges in Machine Unlearning*

Machine unlearning encounters challenges arising from the inherent characteristics of machine learning (ML) models as well as practical limitations in implementation. These challenges not only underscore the complexities embedded within ML models but also highlight the real-world difficulties encountered when applying unlearning techniques [4]. Key challenges include

1) Data Dependencies: ML models go beyond analyzing individual data points in isolation. Instead, they intricately extract complex statistical patterns and dependencies between data points [7]. The removal of a single point can

disrupt these learned patterns and dependencies, potentially resulting in a significant performance decrease.

2) Model Complexity: Large machine learning models, such as deep neural networks with millions of parameters, pose challenges due to their intricate architectures and nonlinear interactions between components. The complexity makes it challenging to interpret the model and identify the specific parameters most relevant to a given data point [2] [6]. The lack of transparency into how data influences predictions adds complexity to the task of removing dependencies.

3) Computational Cost: Many machine unlearning techniques rely on iterative optimization methods like gradient descent to adjust parameters post-data removal. This incurs a substantial computational cost, especially as both the model and dataset size increase. Dealing with large-scale datasets and complex models may surpass the available resources due to these growing computational demands.

4) Privacy Leaks: The unlearning process itself poses privacy risks in various ways [7]. Statistics such as the time taken to remove a point may inadvertently disclose information about it. Changes in accuracy and outputs can also enable adversaries to infer the characteristics of the removed data [4].

5) Dynamic Environments: Tracing the influence of each data point becomes increasingly challenging in dynamically changing datasets. Unlearning can introduce delays, hindering prompt model updates crucial for achieving low-latency predictions [4].

Machine unlearning confronts several challenges, encompassing both inherent ML model properties and practical implementation issues. These challenges include the intricate relationships and dependencies between data points, where the removal of an individual point may disrupt learned patterns, leading to performance declines. The complexity of large ML models, especially deep neural networks, poses interpretation challenges, making it difficult to identify relevant parameters. The computational cost is significant, particularly with iterative optimization methods, escalating as models and datasets grow in size. Privacy risks arise during the unlearning process, with potential leaks of information through statistics or changes in accuracy. Additionally, in dynamic environments, tracing the influence of each data point becomes more complex, and unlearning may introduce delays, impacting prompt model updates needed for low-latency predictions.

V. Conclusion

This paper delves into the dynamic realm of machine unlearning, dissecting pivotal contributions by researchers. The efficient unlearning framework and novel algorithms discussed showcase the field's progress, promising more adaptable, efficient, and robust machine learning models. However, amid these advancements, the highlighted challenges underscore the necessity for ongoing innovation and ethical considerations. As machine unlearning continues to evolve, it not only reshapes how models adapt but also demands a conscientious approach toward privacy, computational costs, and model interpretability. This paper contributes to the ongoing discourse, urging a balanced and responsible trajectory for the future of machine unlearning.

References

[1] Y. Cao and J. Yang, "Towards Making Systems Forget with Machine Unlearning," 2015 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 2015, pp. 463-480, doi: 10.1109/SP.2015.35.

[2] "Remember what you want to forget: Algorithms for machine unlearning", Advances in Neural Information Processing Systems 34, Dec. 02, 2021.

[3] "Descent-to-delete: Gradient-based methods for machine unlearning", Algorithmic Learning Theory, Mar. 02, 2021.

[4] Xu, J., Wu, Z., Wang, C., & Jia, X. (2023). Machine Unlearning: Solutions and Challenges. arXiv preprint arXiv:2308.07061.

[5] Lin, S., Zhang, X., Chen, C., Chen, X., & Susilo, W. (2023). ERM-KTP: Knowledge-Level Machine Unlearning via Knowledge Transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20147-20155).

[7] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang, "When machine unlearning jeopardizes privacy," in Proceedings of the 2021 ACM SIGSAC conference on computer and communications security, 2021, pp. 896–911

[8] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. Kankanhalli, "Fast yet effective machine unlearning," IEEE Transactions on Neural Networks and Learning Systems, 2023.

[9] C. Wu, S. Zhu, and P. Mitra, "Federated unlearning with knowledge distillation," Arxiv.org.

[10] Zhan, X., Wang, Q., Huang, K. H., Xiong, H., Dou, D., & Chan, A. B. (2022). A comparative survey of deep active learning. arXiv preprint arXiv:2203.13450.

[11] Chourasia, R., & Shah, N. (2023, July). Forget unlearning: Towards true data-deletion in machine learning. In International Conference on Machine Learning (pp. 6028-6073). PMLR.

[12] Allen, C. H., Ahmed, D., Raiche-Tanner, O., Chauhan, V., Mostaço-Guidolin, L., Cassol, E., & Murugkar, S. (2021). Label-free two-photon imaging of mitochondrial activity in murine macrophages stimulated with bacterial and viral ligands. Scientific Reports, 11(1), 14081.