International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0278

e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1891 - 1895

Machine Unlearning: Towards Privacy-Preserving and Trustworthy Artificial Intelligence System

Spandhana Lokesh[1], Udith B.R[2], Nidhi Deshpande[3], Suhaib Ayub Khan[4], Syed Ubaid[5], T. KohilaKanagalakshmi[6]

[1,2,3,4,5] PG Student, Department of MCA, [Dayanand Sagar College of Arts, Science and Commerce], Karnataka, India.

[6]Assisstant Professor, Department of MCA, [Dayanand Sagar College of Arts, Science and Commerce], Karnataka, India.

Email ID: *lokeshspandhana4@gmail.com[1], udithrugi8@gmail.com[2], nidhideshpande397@gmail.com[3], suhaibayub27@gmail.com[4], syedubaid942@gmail.com[5], kohila.dsi@gmail.com[6].*

Abstract

Artificial intelligence models today depend heavily on large datasets, where information becomes part of the model during training. Once this information is learned, removing the effect of specific data becomes challenging. This creates concerns related to user privacy, data protection, and legal requirements for data removal. Existing solutions, such as retraining or fine-tuning, are often time-consuming and may not fully remove the data's impact. To address this issue, this work proposes a Dual-Layer AI Controlled Unlearning System (DSRUN) that allows faster and targeted data removal. The system separates the learning and unlearning processes and uses a data influence tracking mechanism to understand how individual data points affect model parameters. Instead of retraining the entire model, only the affected components are updated, reducing computational effort while maintaining performance. Experimental results show that the proposed system achieves faster unlearning with minimal impact on accuracy and reduces the chances of recovering removed data. This approach provides a practical solution for building scalable and privacy-aware AI systems capable of supporting real-time data removal.

Keywords: *Big Data Analytics, Diabetes Mellitus Artificial intelligence; Data influence; Data points; Dual layers; Parameters.*

1. Introduction

In recent years, artificial intelligence has been widely used in areas such as healthcare, recommendation systems, and language processing, where large amounts of data are continuously processed. These models learn patterns from data and store them in their internal parameters. But, once trained, removing specific information from the model becomes difficult. This raises serious concerns related to privacy, security, and legal requirements such as the right to data deletion. Traditional methods, such as retraining the model after removing data, are effective but require significant time and computational resources. Other approaches, including fine-tuning and influence-based methods, try to reduce this cost but often fail to completely remove the effect of the targeted data. To overcome these limitations, this work proposes a Dual-Layer AI Controlled Unlearning System (DSRUN). The system separates the learning and unlearning processes, allowing faster and near real-time removal of specific data. It also aims to maintain model accuracy while improving privacy protection

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0278
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1891 - 1895

without the need for full retraining. Several recent studies highlight the limitations of existing unlearning techniques in achieving complete data removal (Li et al., 2025; Geng et al., 2025).

1.1.Problem Statement

In this study, the main issue being addressed is that existing machine learning models are not capable of effectively removing specific data after training. Due to this limitation, several problems arise, such as:

- Sensitive or personal data continues to remain in the model even after deletion requests
- Complete retraining is required, which increases computational cost and time
- Existing methods do not guarantee full removal of data influence
- Real-time unlearning is not supported in most current systems

1.2.Objectives

This research mainly aims to achieve the following points:

- To design a dual-layer AI system that separates learning and unlearning processes
- To enable faster and targeted removal of specific data from trained models
- To support near real-time unlearning without requiring full retraining
- To maintain model accuracy and performance during unlearning operations

2. Method

In this section, a method is described for enabling faster machine unlearning in AI systems by using a Dual-Layer AI Controlled Unlearning System. The approach mainly combines three parts such as the Learning Layer, the Unlearning Control Layer, and a data influence tracking mechanism. Initially, the system is trained using a dataset where multiple data points are provided as input. During training, the model learns patterns and stores them in its internal parameters. At the same time, a data influence map tracker observes how each data point affects different parameters. This helps in creating a clear connection between data points and their impact on the model. An Unlearning Control Layer is included to manage data deletion requests. When a request is made to remove specific data, this layer uses the data influence map to identify the affected parameters. Instead of retraining the entire model, only the selected parameters are updated or modified. As a result, the unlearning process becomes faster compared to traditional methods. A verification mechanism is also used to check whether the removed data can still be inferred from the model. If any traces are found, the system performs additional updates using the data influence map tracker until the influence is reduced. This helps improve privacy and ensures that sensitive information is properly removed. A stabilization process continuously monitors important factors such as model accuracy, prediction quality, and system performance. If any performance drop is detected after unlearning, corrective adjustments are automatically applied to the parameters. This helps maintain the stability and reliability of the model during repeated unlearning operations in table 1.

Table 1. Machine Unlearning Methods: Description and Limitations

Methods	Description	Limitations
Full Retraining	Retrains model from scratch after removing data	Very slow and computationally expensive.
SISA	Trains model in shards; only affected parts retrained	Accuracy trade-offs and complex management
Fine-Tuning	Adjusts model to reduce	Cannot guarantee

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0278
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1891 - 1895

	influence of specific data	complete data removal
Influence Function	Estimates impact of each data point on model	Difficult to isolate exact gradients
Gradient-Based	Reverses/updates gradients of target data	Servers are increased or decreased based on need
Differential Privacy	Limits data influence using noise during training	Reduces accuracy; no exact unlearning
Knowledge Distillation	Transfers knowledge to new model excluding data	May still retain traces of original data
Adversarial Unlearning	Uses attacks to detect and remove residual data	High cost and not fully reliable
SIMU	Updates only parameters influenced by target data	Requires accurate influence tracking
LLM Unlearning	Removes specific knowledge from large models	Hard to ensure complete deletion

2.1.Tables

Tables are used to present the machine unlearning methods in a clear and organized manner. Table 1 includes the different unlearning techniques along with their descriptions and limitations, helping to understand their performance and compare their effectiveness [5].

2.2.Figures

Visual representations are incorporated to illustrate the interaction between learning and unlearning components within the proposed system. Figures include the architecture and workflow of the dual layer AI system. In this architecture, the model is first trained in the Learning Layer, where data points are processed and stored as internal parameters. During training, a data influence map tracker monitors how each data point affects different parameters. When a data deletion request is given, the Unlearning Control Layer uses the data influence map to identify the affected parameters and updates only those parts instead of retraining the entire model. Simultaneously, a verification mechanism ensures that the removed data cannot be inferred, and a stabilization module maintains model performance [6].This design allows faster, targeted, and near real-time machine unlearning while preserving accuracy in figure 1.

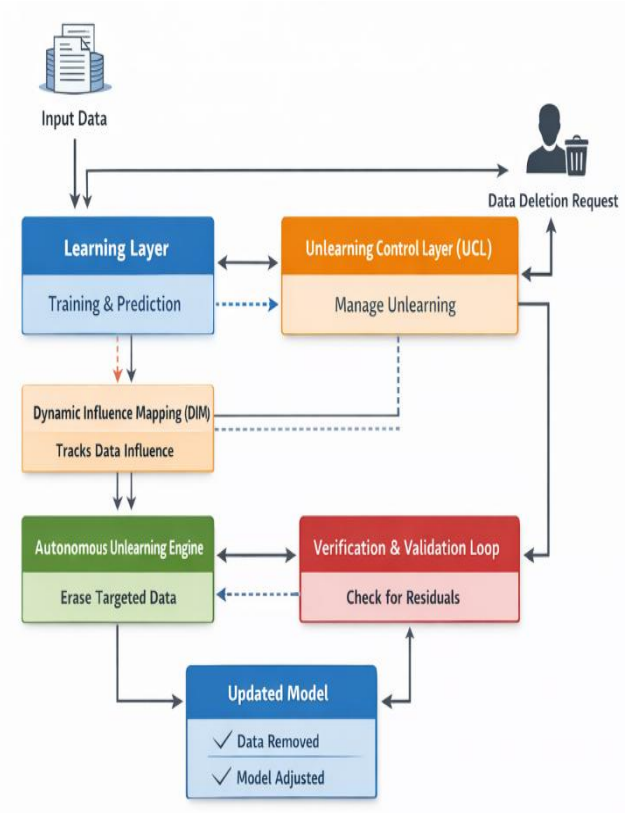

Figure 1 Dual-Layer Self-Regulation Unlearning Network (DSRUN) Architecture Flow

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0278
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1891 - 1895

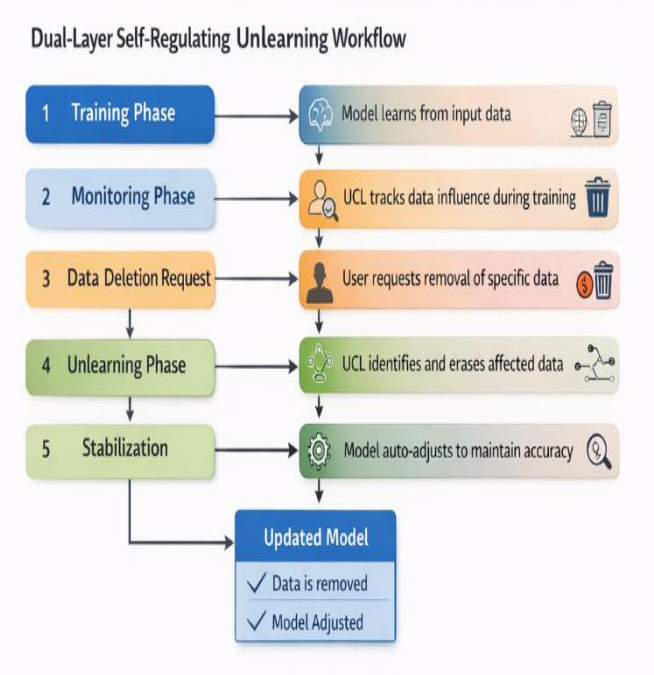

Figure 2 Workflow of Dual-Layer Self-Regulation Unlearning Network (DSRUN) Architecture

The workflow of this system explains how the unlearning process happens step by step in figure 2. First, the model is trained in the Learning Layer, where input data points are processed and stored as internal parameters. During this process, a data influence map tracker monitors how each data point affects different parameters. When a data deletion request is received, it is handled by the Unlearning Control Layer [1]. This layer uses the data influence map to identify which parameters are influenced by the selected data points. Instead of retraining the full model, only those specific parameters are updated. After the update, a verification step checks whether the removed data can still be inferred from the model. If any traces are found, the system performs further adjustments using the influence map tracker. Simultaneously, a stabilization process ensures that the model accuracy and performance remain stable after unlearning. This cycle can repeat whenever new deletion requests are received. Because of this workflow, the system performs faster, targeted, and near real-time unlearning while maintaining model performance.

3. Results And Discussion

3.1.Results

The proposed Dual-Layer AI Controlled Unlearning System was tested to evaluate its ability to remove specific data from trained models [4]. The system used the data influence map tracker to identify affected parameters and performed targeted updates instead of full retraining. Experimental observations indicate that the proposed approach demonstrated improved efficiency compared to conventional methods. The unlearning process was faster, and model accuracy remained stable after removing data points. The removed data could not be easily inferred from the model, indicating effective unlearning. Some key observations are:

- Unlearning time was reduced compared to full retraining
- Model accuracy was maintained with minimal changes
- Data influence map tracker helped in identifying affected parameters
- Removed data points were not easily recoverable

3.2.Discussion

From the results, it is observed that the proposed dual-layer approach provides an faster way to perform machine unlearning. The separation of Learning Layer and Unlearning Control Layer helps in managing data removal more effectively. [2]The data influence map tracker plays a key role in linking data points with model parameters, allowing targeted updates instead of full retraining [3]. This reduces computational cost and supports faster unlearning. But, the performance of the system depends on the accuracy of influence tracking. If the mapping is not precise, small traces of data may remain. Also, the system design is more complex compared to traditional methods. In summary, the proposed system improves efficiency, maintains model performance, and provides a more practical solution for privacy-

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0278
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1891 - 1895

preserving AI applications.

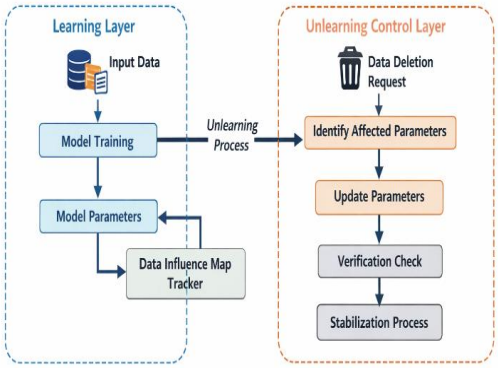

Figure 2 Process of the dataset

Conclusion

This work presents a faster approach to addressing the challenge of selective data removal in trained AI models, which is critical for ensuring privacy and regulatory compliance. Existing methods are often inefficient, as the influence of different data points is distributed across model parameters, making precise removal difficult. To overcome this, a Dual-Layer AI Controlled Unlearning System (DSRUN) is proposed. The system consists of a Learning Layer, which performs model training, and an Unlearning Control Layer, which manages data removal. A key component, the data influence map tracker, identifies how individual data points affect model parameters, enabling targeted updates instead of full retraining. The proposed approach allows faster and near real-time unlearning by modifying only the affected parameters. It also includes verification and stabilization mechanisms to ensure that removed data cannot be inferred and that model performance is maintained. In summary, the DSRUN framework improves efficiency, enhances privacy protection, and provides a scalable solution for modern AI systems. In the future, further improvements can be made by refining influence tracking and optimizing parameter update strategies for more accurate unlearning. The proposed dual-layer design offers a scalable solution for real-time unlearning while preserving model integrity.

Acknowledgements

We express our sincere gratitude to the Department of MCA, Dayananda Sagar College of Arts, Science and Commerce, Bengaluru, for providing the necessary support and resources to successfully carry out this project. Their continuous encouragement and institutional backing played a vital role in the completion of this work.

We would also like to extend our heartfelt thanks to our teachers and mentors for their valuable guidance, timely assistance, and constant motivation throughout the project. Their insightful suggestions and support at every stage greatly contributed to the successful completion of this research.

References

[1]. Li C. Chen X. & Wang J. (2025). *An overview of machine unlearning: Concepts methods and future directions*. Journal of Information Security and Applications.

[2]. Geng J. Liu Y. & Zhang H. (2025). *A comprehensive survey of machine unlearning in large language models*. arXiv preprint arXiv:2503.01854.

[3]. Ahmad T. Shaik A. & Luo Y. (2026). *An empirical study of machine unlearning in deep learning systems*. Engineering Applications of Artificial Intelligence.

[4]. Rajwade D. Bhardwaj V. & Vishwakarma R. (2025). *Machine unlearning: A comprehensive framework for faster data removal in deep learning systems*. International Journal of Innovative Science and Research Technology.

[5]. Liu H. Zhang Y. & Chen Z. (2025). *A survey on machine unlearning: Techniques challenges and vulnerabilities*. Journal of Systems Architecture.

[6]. Cevallos I. D. & Martinez P. (2025). *A systematic literature review of machine unlearning techniques in neural networks*. Computers Journal.