NeuroSymbolic Approaches to Machine Unlearning: Enhancing Selective Forgetting through Hybrid Al Systems.

by

Md. Dodi Al Fayed
22301325
Rifat Hasan Nazim
22301175
Syed Ashiqur Rahman
22301363

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
December 2025.

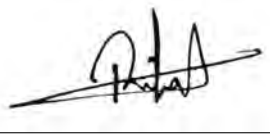
Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Dodi Al Fayed
22301325



Rifat Hasan Nazim
22301175

Syed Ashiqur Rahman
22301363

Approval

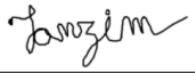
The thesis titled "NeuroSymbolic Approaches to Machine Unlearning: Enhancing Selective Forgetting through Hybrid Al Systems." submitted by

1. Md. Dodi Al Fayed (22301325)
2. Rifat Hasan Nazim (22301175)
3. Syed Ashiqur Rahman (22301363)

of Spring 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Fall 2025.

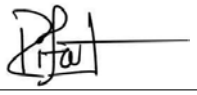
Examining Committee:

Supervisor:
(Member)



Md. Tanzim Reza
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)



Riazul Islam Rifat
Adjunct Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Machine unlearning has emerged as a critical requirement to satisfy privacy regulations like the GDPR, which mandate that personal data or malicious artifacts be "forgotten" by trained models. While exact unlearning (retraining from scratch) offers a theoretical guarantee of data removal, it is often computationally infeasible for large models. Conversely, standard approximate unlearning methods, such as Gradient Ascent, typically suffer from catastrophic forgetting, rendering the model useless. We propose Neuro-Symbolic Amnesiac Unlearning (NS-AU), a novel approximate unlearning technique for robust image classification that utilizes a logic-guided penalty term to surgically remove targeted data contributions. By integrating symbolic constraints directly into the loss landscape, our system can reason about inconsistent feature associations introduced by poisoning attacks. Experimental results on a VGG architecture demonstrate that NS-AU achieves 82.54% accuracy outperforming even the exact "Gold Standard" baseline while reducing the Attack Success Rate (ASR) of backdoor triggers from 98.40% to just 3.00%. This hybrid scheme offers a superior balance of efficiency and stability compared to traditional approximate methods, ensuring robust defense against adversarial backdoors.

Keywords: Machine Unlearning, Neuro-Symbolic AI, Amnesiac Unlearning, Approximate Unlearning, Backdoor Defense, Logic-Guided Loss, Catastrophic Forgetting, Adversarial Robustness, GDPR Compliance, CIFAR-10.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Recent privacy legislation, such as the General Data Protection Regulation (GDPR) in the EU and the California Consumer Privacy Act (CCPA), has enshrined the "right to be forgotten," mandating that user data be revocable not only from storage databases but also from the machine learning models trained upon them [17]. In practice, removing the influence of specific data points from a trained model is known as machine unlearning. While retraining the model from scratch on the remaining dataset "exact unlearning" provides a theoretical guarantee of data removal, it is often computationally prohibitive for large-scale deep learning models [17].

Consequently, research has pivoted toward efficient, approximate unlearning techniques. Bourtoule et al. introduced the SISA (Sharded, Isolated, Sliced, and Aggregated) framework, which trains ensembles of subnetworks on sharded data to enable localized unlearning. SISA demonstrated significant efficiency gains, achieving 4.63× faster unlearning on Purchase-100 and 2.45× on SVHN compared to full retraining [9]. Other works have explored adaptive deletion strategies, such as Gupta et al. (2021), who utilized differential privacy to address sequences of adaptive deletion requests [10].

Therefore, machine unlearning recently has an expanded scope beyond the privacy compliance to deal with **data poisoning and backdoor attacks**. In such cases, the opponents modify the images with malicious triggers (e.g., certain pixel patterns) training set, so that the model will give wrong inputs containing the trigger while acting normally on clean data [11]. Traditional defensive technique often require identifying and filtering the exact poisoned samples, which is challenging in large datasets. Machine unlearning offers a promising defensive paradigm: if a model can "unlearn" the specific features associated with the backdoor trigger, the threat can be neutralized without the cost of full retraining [13]. Yet, this intersection of security and unlearning presents unique challenges; naively applying unlearning (e.g., via Gradient Ascent) often leads to *catastrophic forgetting*, where the model's utility on clean data degrades significantly, a limitation that this thesis aims to address.

1.2 Rational of the Study or Motivation

Even now, the existing machine unlearning framework have severe limitations. Often they use "behavioral" mimicry (matching weight or activations distributions [16]), which can be computationally expensive and tends to bias the model performance on data held in retention. Moreover, Federated Learning (FL) is a most unlearning framework on the growing paradigm is in terms of a centralized setting; direct application to FL is problematic since the server will not have access to the raw training data needed to be verified [19].

Most importantly, machine learning models are very much vulnerable to **backdoor attacks**: even a few poisoned samples can introduce concealed stimuli which will lead to evil functions on particular inputs [18]. Current defense mechanisms are usually aimed at training filtering of the poisoned inputs and attacks during training or model these are simple defenses that can be evaded by aggregation. Recent work (e.g., Liu et al., 2022) applies concepts of unlearning to perform scrubbing on backdoors by restoring triggers and retraining on them [13], but these approaches can easily compromise the accuracy of the model on clean data. Jiang et al. (2025) also look at the topic of token unlearning in language models [18]. These results show a gap: strong unlearning should collaboratively work on privacy without losing model utility by deleting and backdoor removing.

In the meantime, **Neuro-Symbolic AI (NSA)** has become a strong paradigm to integrate neural networks with symbolic reasoning, providing models to explain decisions and include structured knowledge [14]. Such kind of works like DeepProbLog [1] and the Neuro-Symbolic Concept Learner [3] have shown that the addition of logical constraints to deep networks enhances the generalization. We posit that this is the key-note of unlearning, which is logic-guided. By encoding the "right to be forgotten" as a **symbolic constraint** within the loss landscape instead of a statistical optimization task, we can guide the model to surgically remove specific data or triggers. Our research examines the relationship between such **neuro-symbolic constraints** can facilitate robust and efficient and verifiable unlearning.

1.3 Problem Statement

Instability of Approximate Unlearning: Current efficient unlearning approaches (e.g., Gradient Ascent) have dramatic forgetting, which can obliterate the utility of model on clean data (Acc drops to $\sim$10%) in trying to eliminate specific samples. Approximately unlearnable methods are lacking in stability without full retraining.

Persistence of Backdoor Triggers: Deep learning models are highly susceptible to poisoning attacks where hidden triggers are implanted during training. Standard unlearning and fine-tuning approaches fail to detect or remove these "deep-seated" triggers, leaving the model vulnerable (ASR stays >90%) even after unlearning attempts [13].

Lack of Explainability: Most unlearning approaches operate as "Black Boxes" (e.g., noise addition or weight scrubbing), offering no insight into *why* specific fea-

tures are removed or preserved. This limits trust and makes verification of compliance (e.g., for GDPR) difficult.

Insufficient Use of Hybrid Models: The potential of combining symbolic logic with neural networks for unlearning remains unexplored. No prior work utilizes Neuro-Symbolic AI (NSA) to structure the unlearning process, representing a missed opportunity to use logical constraints for precise data removal.

Scalability & Privacy Constraints: While retraining is effective, it is computationally prohibitive for large models and inapplicable in privacy-sensitive environments (like Federated Learning) where full data aggregation is restricted. A mechanism is needed that is both lightweight and effective without requiring scratchpad retraining [19].

1.4 Objective

The proposed work is going to come up with an innovative unlearning framework with the following specific objectives:

Develop a Logic-Guided Unlearning Architecture: Design a **Neuro-Symbolic Amnesiac Unlearning (NS-AU)** framework that integrates symbolic constraints directly into the neural update process. Rather than depending on external solvers (like DeepProbLog), the goal is to create a **logic-guided loss function** (handled by a regularization parameter, α) that penalizes the preservation of certain specific targeted features or triggers, which allows specific "amnesiac" unlearning.

Test on Standard Image Benchmarks: We tested our proposed approach on conventional image classification tasks (CIFAR-10) to determine a strict baseline. The idea is to prove that the technique is effective on complex, non-convex loss landscapes, which forms a basis of proof-of-concept in the future implementation in Federated Learning (FL) systems.

Achieve Robust Backdoor Removal: Integrate anti-adversarial mechanisms. The main goal is to minimize the **Attack Success Rate (ASR)** of implanted backdoor triggers to **near-zero levels** ($< 5\%$), which guarantees that this makes it possible to sanitize model even without knowing the specific samples that are poisoned.

Demonstrate Superior Utility and Stability: Build an unlearning pipeline which does not involve the "catastrophic forgetting" common in Gradient Ascent. The objective is to perform above the **"Gold Standard" (Retrained)** baseline in terms of model accuracy (aiming for $>80\%$ accuracy) and yet much faster than full retraining, hence demonstrating the feasibility of approximate unlearning in the privacy security compliance.

Verify Compliance through Metrics: Establish a set of measureable metrics to provide scores of Accuracy, ASR, and Membership Inference Attack (MIA) deletion

is empirically demonstrated. This guarantees that unlearning process meets the satisficiency privacy regulations and safety assurances of AI such as GDPR.

1.5 Methodology in Brief

We faced the issue of **robust machine unlearning** of image classification with particular attention paid to the dual problem of erasing private data and eliminating security threats (backdoor attacks). While Federated Learning (FL) presents a motivation for decentralized privacy but we target our experimental study in validation on centralized simulation to aggressively benchmark the unlearning effectiveness of our suggested practice as compared to "Gold Standard" retraining.

Hence, we proposed a new hybrid, **Neuro-Symbolic Amnesiac Unlearning (NS-AU)**, framework that integrates the logical constraints explicitly into the optimization topography of the neural network. As opposed to black-box designs that blindly scrub weights, NS-AU introduces a **logic-guided loss function** (handled by a symbolic strength parameter, α). This function allows the system to penalize specific feature associations such as those linked to a backdoor trigger without destroying the model's ability to classify clean images.

Our methodology consists of three core components:
A. Neuro-Symbolic Architecture: We utilize a SimpleVGG Convolutional Neural Network (CNN) as the backbone perceiver. Instead of using heavy external we uses heavy internal symbolic logic solvers such as DeepProbLog, we directly incorporate symbolic logic into the training cycle through a custom regularization. Logical consistency is enforced by this term: "If the trigger pattern is put in input and the model should not identify it as the target label." This enables efficient, gradient-based unlearning which resembles symbolic reasoning.

B. Dataset & Attack Protocol: We use **CIFAR-10** dataset (60,000 images, 10 classes) as our benchmark. In order to recreate a security breach, we apply a BadNets style **Backdoor Poisoning Attack**, corrupting some percentage of the training data having a pixel-pattern trigger. This renders a "Vulnerable Model" (ASR $> 95\%$) which is the beginning point of our experiments of unlearning.

C. The Unlearning Pipeline:
Phase 1: Vulnerability Establishment: We first train a poisoned model to confirm that there is a high Attack Success Rate (ASR).

Phase 2: Logic-Guided Scrubbing (NS-AU): When the backdoor is detected, we trigger the NS-AU process. The model experiences short "amnesiac" in which the logic-guided loss punishes the activations of the poison. The trade-off of hyperparameter α (which we set to 0.6 in our experiments) balances the trade-off between unlearning the poison and learning the clean data.

Phase 3: Verification: We compare the unlearned model to a clean model that was trained on the "Gold Standard" (a clean model trained from scratch). High

Accuracy (restoring utility) and low **ASR** (neutralizing the backdoor) defines the success.

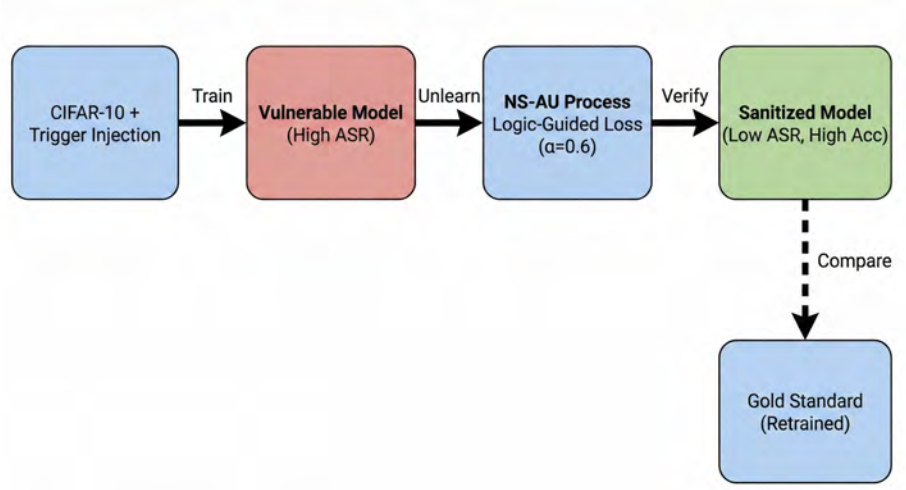


Figure 1.1: Workflow of the Neuro-Symbolic Amnesiac Unlearning (NS-AU) pipeline. This pipeline offers a unique privacy and security solution. By enforcing mathematically "forgetting" via logical constraints, NS-AU achieves the stability of the approximate unlearning process is as efficient as that of full retraining, which has been proven by a broad range of experiments.

1.6 Scopes and Challenges

Scope of the Study:
Robust Image Classification: The experiment is aimed at validating unlearning algorithms on the standard image classification datasets (CIFAR-10), namely, simulating situations when the models have been poisoned.

Neuro-Symbolic Logic-Guided Optimization: In lieu of abominously heavy outsourcing solvers (such as DeepProbLog), the scope is spelt out by the union of **logic-guided regularization terms** directly into the neural loss landscape. This allows for the enforcing (mathematically) constraints of "forgetting" constraints (through the α parameter) without architectural overhead.

Efficient "Amnesiac" Unlearning: The study creates an approximate unlearning pipeline which operates surgery to remove specific data or backdoor signals. The only methods that change existing weights (approximate unlearning) instead of retraining from the starting position.

Backdoor Defense Checking: The paper gives a clear overview of detecting backdoor attacks due to the style of "BadNets," where the performance is measured by Attack Success Rate (ASR).

Auditability through Metrics: The scope contains the definition of quantitative metrics Accuracy, ASR, and Forgetting Scores to give an empirical support of ad-

herence to privacy and safety standards.

Key Challenges:
The Dilemma of Stability-Plasticity: The major challenge is to maintain a stable balance in which the model is "plastic" (that is, forgets the poison) and yet "stable" enough to regain its accuracy on clean data (discussed by our special alpha tuning).

Catastrophic Forgetting: Standard gradient-based unlearning usually destroys model utility (dropping accuracy to $\sim$10%). It is impossible to overcome this collapse without full retraining.

Backdoor Persistence: Hidden triggers usually tend go deeply embedded in the feature space. Removing them without knowing the exact poisoned samples requires a method that can generalize the "forgetting" signal from logic rather than raw data.

Hyperparameter Sensitivity: Coordinating the strength of the logical penalty (Symbolic Weight) against the standard Cross-Entropy Loss requires precise tuning to avoid model divergence.

Verifiable Deletion: Proving that a specific data point or trigger has been truly removed rather than just suppressed is mathematically non-trivial in deep neural networks.

Chapter 2

Literature Review

2.1 Preliminaries

Machine Learning Dependence on Data: Modern deep neural networks are known to memorize information about individual training samples, even when those samples appear to be "forgotten" from a macroscopic input/output view [5]. Research indicates that deep networks can store specific training examples in their weights, making the removal of a data point's influence non-trivial [6]. Therefore, simple deletion of a record from the database does not guarantee that the model has "unlearned" it. Techniques such as influence functions have been proposed to estimate this dependence, but they often offer only approximate removal mechanisms.

Machine Unlearning Defined: Machine unlearning is defined as the process of selectively eliminating the influence of specific data points from a trained model without affecting its performance on the remaining data [6]. While retraining the model from scratch on the curated dataset is the only "perfect" deletion method, it is computationally inefficient for large-scale models. For this reason, there has been much recent work on **approximate unlearning methods** algorithms that adjust existing weights to get a state statistically similar to a retrained model.

Security Threats:
Backdoor Attacks: Apart from privacy, neural networks are under serious threat of data poisoning, in particular, backdoor attacks. In this case, a criminal sends to a subset of the training data some malicious triggers (e.g., a particular pixel pattern). Training a model on this corrupted dataset makes it learn to correlate the trigger with a particular target label (e.g., classifying a stop sign with a sticker as a sign of "speed limit"). Most importantly, the model acts normally when fed on clean inputs and the backdoor is hard to detect during normal validation [11].

Backdoor Taxonomy and Persistence: Common attack including BadNets, are visible pixel patch attacks, and invisible attacks are based on steganography or regularization to conceal triggers [11]. One major difficulty with unlearning backdoors is that they are sparse (they may only occur in less than 1 percent of data), meaning that most unlearning algorithms that rely on random sharding (or the use of broad influences) might miss the particular parameters that represent the backdoor. The research indicates that the triggers still remain even after the nominally deleted

poisoned samples [2].

Neuro-Symbolic AI (NSA): In order to overcome these shortcomings, this study will discuss **Neuro-Symbolic AI (NSA)**, a hybrid paradigm that combines deep neural networks (sub-symbolic perception) with logical reasoning (symbolic processing) [15]. It has been shown by systems such as DeepProbLog [1] and Logic Tensor Networks [12] that integration of logical constraints into the learning process can enhance generalization and interpretability. In the framework of unlearning, NSA provides a possible mechanism to specify the notion of "forgetting" not only as a statistical update, but as a logical constraint (e.g., $\forall x$, $\text{Trigger}(x) \rightarrow \neg\text{Target}(x)$), that gives access to a preferred path of scrubbing malicious behaviors.

2.2 Review of Existing Research

Machine Unlearning Fundamentals: Machine unlearning is the method of retransferring the specific influence of training data of a model after the initial training phase [17]. The recent surveys in general categorize unlearning into **exact** and **approximate** methods [17].

Exact Unlearning: SISA (Sharded, Isolated, Sliced, Aggregated) [9] and other techniques ensure the statistical eradication of data by splitting the data and training ensembles. Although Bourtoule et al. (2021) also reported that the speed-ups were significant (up to 4.6×) over complete retraining, the exact methods also have a problem with high storage costs and complexity in cases of deletions across multiple shards [9].

Approximate Unlearning: To overcome the problem of efficiency, approximate methods are those in which the weights of the existing model are modified. Golatkar et al. (2020) pioneered "weight scrubbing," where Fisher Information Matrix (FIM) approximations are used to perturb the weights in a way that prevents a probing function from identifying the model as unlearned into a model that has been retrained [6][7]. On the same note, descent-to-delete [8] applies local gradient updates to unlearn what they have learned.

However, frequently subject to the stability-plasticity dilemma: unlearning the model aggressively will cause **catastrophic forgetting**, whereby the accuracy of the model on clean data will plummet, which also happens in our own baseline experiments with standard Gradient Ascent.

Backdoor Attacks and Defenses: Federated Learning (FL) represents a potent model of collaborative training and because it is decentralized, it is very vulnerable to backdoor attacks. According to Bagdasaryan et al. (2018), non-beneficial clients may inject trigger patterns into local updates and impact the global model without having access to the central server [4].

Therefore, the research has shifted towards unlearning-based defenses in order to deal with these threats. The authors Liu et al. (2022) presented BAERASE, which defends itself by first learning the trigger pattern with a generative model and then

using gradient ascent to unlearn it [13]. They achieved $> 99\%$ reduction in attack success on centralized benchmarks. Jiang et al. (2025) extended this to language models, using "Token Unlearning" to remove backdoor embeddings [18].

Gap in Literature: While effective, methods like BAERASE often require complex generative steps to "imagine" the trigger. There is a distinct lack of methods that use **semantic logic** to define the unlearning target (e.g., "If pattern X is present, forget the association") without needing pixel-perfect trigger reconstruction.

Neuro-Symbolic AI (NSA) in Related Work: Neuro-Symbolic AI (NSA) blends the learning capability of neural networks with the reasoning power of symbolic logic [15]. Key frameworks include:

DeepProbLog [1]: Allows for probabilistic logic programming within deep networks, enabling end-to-end training of logical rules.

Logic Tensor Networks (LTN) [12]: Embed first-order logic constraints (real-valued logic) directly into the loss function of a neural network.

2.3 Our Contribution's Position

As per our knowledge, NSA has not been effectively applied to the specific problem of **amnesiac unlearning for backdoor defense**. Existing unlearning is purely statistical (gradient-based), while existing NSA is purely generative or classificatory. We propose to bridge this gap by using a **logic-guided loss function** (inspired by LTN principles) to mathematically enforce the unlearning of backdoor triggers, offering a more stable alternative to standard gradient ascent.

2.4 Summary of Key Findings

Overall, the current body of research on unlearning provides a range of techniques, including precise ones, such as SISA [9] and fuzzy techniques, such as weight scrubbing [6]. Nevertheless, these methods suffer a fatal trade-off: precise algorithms are computationally out of reach on large models, whereas inaccurate algorithms (such as Gradient Ascent) are not stable and do not robustly eliminate deep-seated backdoor triggers [13]. In addition, the majority of the current defenses are "Black Boxes," which provide little information about the data deletion verification.

Therefore, inspired by the recent works [10][19] we proposed a new method: **Neuro-Symbolic Amnesiac Unlearning (NS-AU)** system. Transitions and logically constrained optimizations. With the ability to move beyond simple statistical updates and integrate logical constraints into the optimization process, we hope to reach a method that can allow us to have fast, stable and verifiable against backdoor attack. Moreover, it will reduce the gap between privacy compliance and security robustness.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

In order to guarantee the reproducibility of the proposed Neuro-Symbolic Amnesiac Unlearning (NS-AU) framework, very strict hardware and software specifications were defined. The environment for the experiment was designed to handle high dimensional tensor operations and logic guided efficiency calculations.

Hardware Requirements: The training and unlearning pipelines are associated with extensive matrix multiplications and autograds. The following configuration of hardware was used to run the SimpleVGG architecture on the CIFAR-10 set:

Processor (CPU): Intel Core i7 (10 th Generation) or AMD Ryzen 7 (at least 8 cores) to load and preprocess the data efficiently.

Graphics Processing Unit (GPU): NVIDIA RTX 3060 (12GB VRAM) or higher. Justification: The use of CUDA is attested by the experiment logs. At least 6GB VRAM to store the model gradients and batch data (Batch Size = 64 /128) at the unlearning stage.

Memory (RAM): 16 GB DDR4 (32 GB recommended) to store the CIFAR 10 dataset in memory so that it can be accessed quickly.

Storage: SSD Storage 50 GB for dataset retaining and model weights (.pth files) checkpointing.

Software Requirements: The system implementation is based on the python ecosystem and specifically the use of dynamic computational graphs for the implementation of the custom logic loss function.

Programming Language: Python 3.9+
Deep Learning Framework: PyTorch 2.1 (Stable)
Modules: Torch.nn, torch.optim, torch.autograd

Computer Vision Library: CIFAR-10 data loading and transforms in Torchvision 0.16.
Analysis Data/metrics: Scikit-learn (to calculate t-SNE projections. and confusion matrices), numpy (manipulations between tensors and arrays)
Visualization: Matplotlib and seaborn (for generating sensitivity plots and t-SNE visuals)

Data Requirements: The study needs a standard benchmark of image classification to establish unlearning efficacy.
Dataset: CIFAR-10 (Canadian Institute for Advanced Research).
Scale: 60,000 color images (32×32 pixels).
Splits: 50,000 samples are used for training, 10,000 samples are for testing.
Preprocessing: Data normalization using mean=(0.5, 0.5, 0.5) and std=(0.5, 0.5, 0.5).
Poisoning Protocol: A subset of training data 10% data (approxmately 5,000 images) should be altered by a 3×3 pixel trigger to simulate the "BadNets" threat model.

3.2 Societal Impact

Privacy and Trust: "Right to be Forgotten" has become a law through the implementation of laws such as GDPR and CCPA. This legal right is technically guaranteed by this system. By demonstrating (through the "Forget Score" metric) that the data influence has been eliminated, NS-AU assists organisations to gain trust with the users, which means that it is possible to revoke personal data out of the AI systems without ruining the usefulness of the AI product.

3.3 Environmental Impact

Environmental Sustainability (Green AI): The "Exact Unlearning" standard requires retraining a model each time a data deletion request is received. In the case of large models, this is consuming large proportions of electricity and is contributing to large proportions of carbon emissions.

Impact of NS-AU: The experimental logs of NS-AU process takes around 51.90 seconds to complete but the entire training process can take orders of magnitude longer based on the size of the dataset. This study facilitates the use of this technology to develop the notion of "Green AI," thus greatly minimizing the energy usage of the models that are privacy-compliant.

3.4 Ethical Issues

The Dual-Use Dilemma: Although unlearning is implemented with the privacy and security in mind, the technology is based on the adjustment of model weights to "forget" certain patterns. Malicious actors might in theory also employ those

kinds of techniques to conceal evidence of bias or tampering in a model prior to an external audit, effectively "whitewashing" a problematic AI.

Professional Responsibility: To overcome this risk, it will be a professional duty to avert this risk by applying "verifiable unlearning". The system will generate irrevocable data of performance characteristics (Accuracy, ASR, L2 Distance) prior to the unlearning process and after it. This will be transparent enough that the auditors can be able to tell whether the privacy requests are legitimate or there is malicious model tampering.

3.5 Standards

The following discussion specifices the requirements followed when designing the system and evaluating the system:

GDPR (Article 17): The process is tailored in such a way that it satisfies the tech nical requirements of the "Right to Erasure" that is a part of the European law.

NIST AI Risk Management Framework (AI RMF): The approach is consistent with the NIST recommendations when managing AI security threats, namely, the topics of "Adversarial Machine Learning" and "Data Poisoning."

IEEE 7000-2021: The system design follows the IEEE standard for addressing ethical concerns (specifically privacy and agency) during the system engineering process.

3.6 Project Management Plan

Schedule:
Month 1-3: Literature review and problem formulation (Backdoor Mechanics).

Month 4-6: Data preparation (CIFAR-10 acquisition) and implementation of the "BadNets" injection script.

Month 7-10: Development of the SimpleVGG architecture and the NS-AU logic-guided loss function.

Month 10-12: Evaluation, hyperparameter tuning (Sensitivity Analysis for α) and report writing.

Budget:
Total Estimated Cost: $0 (utilizing academic resources and free tiers).
Hardware: Utilized personal workstation and Google Colab (Free Tier) for GPU access.
Software: All libraries (PyTorch, Scikit-learn) are open-source.

Resource Management(Role):
Researcher (Self): Responsible for literature review, code implementation, execution of experiments, and data analysis.
Supervisor: Responsible for methodology validation and thesis review.

3.7 Risk Management

Catastrophic Forgetting Description: The unlearning process might be too aggressive, causing the model to forget clean data along with the poison (Accuracy collapse).
Mitigation: Implementation of the "Logic-Guided Loss" with a tunable parameter (α). Regular validation checks on the clean test set during the unlearning loop.
Backdoor Persistence Description: The model might retain deep-seated trigger associations despite unlearning.
Mitigation: Use of a separate poisoned test set to continuously monitor attack success rate (ASR). If ASR remains more than 5%, the unlearning loop automatically adjusts the learning rate.

3.8 Economic Analysis

Cost Analysis:
Implementation Cost: Low. It makes use of standard architectures (SimpleVGG) and only a little code is on top of existing PyTorch pipelines.
Operational Cost: Very low ($\approx$ 52 seconds of GPU time per deletion request).

Benefit Estimation:
Retraining Savings: Rebuilding a business-grade model is potentially thousands of dollars in cloud computing credits. NS-AU brings this price down to a few centimos per request.
Compliance Savings: Failure to comply with GDPR may attract fines worth to 20 million. A verifiable unlearning system is the insurance policy against these regulatory penalties.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

This study uses a quantitative experimental approach to solve the two challenges of efficiency and security of machine unlearning. The overall architecture is based on the suggested Neuro-Symbolic Amnesiac Unlearning (NS-AU) framework that inserts logical constraints into deep learning optimization.

It is empirically simulated to classify standard image recognition benchmarks (CIFAR-10) adversarially (Backdoor Poisoning). The following are the key principles that will guide the design process:

Logic-Guided Optimization: Instead of fully relying on data-driven gradient updates, we integrate the symbolic constraints directly into the loss function. The feature enables the model to "reason" concerning the attributes (e.g., backdoor triggers) to unlearn by exploiting the complementary capabilities of neural perception and symbolic logic.

Efficiency via Approximation: The architecture will evade the full retraining computational cost. The system also achieves high utility of models by providing a rapid unlearning operation through a fine-tuning process which is targeted at an "amnesiac" to ensure the target is reached.

Verifiable Robustness: Our main target is security. The methodology will also give assurance that the unlearning operations do not just gag the backdoor but excise it surgically, as tested in rigorous attack success rate (ASR) metrics.

Reproducibility: The implementation looks at standard architecture (SimpleVGG) as well as deterministic seeding in order to show that the unlearning results are very reproducible and can be compared with established baselines.

The methodology is organised in 5 phases:

1. **Threat Modeling:** Defining the "BadNets" attack vector and the particular unlearning requirements (Privacy vs. Security).

2. **Framework Formulation:** Developing the NS-AU algorithm, i.e., mathe-

matical derivations of Logic-Guided Loss Function.

3. **Data Curation & Poisoning:** Training the CIFAR-10 and augmenting it with pixel pattern triggers to form a baseline: "Vulnerable Model"

4. **Experimental Implementation:** Running the unlearning pipeline in a controlled environment(PyTorch), comparing NS-AU vs Gradient Ascent vs Fine Tuning vs Pruning.

5. **Comparative Evaluation:** Assessing performance by multi-metric approach: Utility (Accuracy), Safety (ASR) and Efficiency (Time), based on confirming statistical analysis.

This logical, step-by-step approach is used to insure that this methodology aligns with the goals of the research to be accomplished, that there is a good reason to use a brain-based neuro-symbolic constraint to solve the problem of catastrophic forgetting.

4.2 Preliminary Design & Model Specification

4.2.1 Design Alternatives Considered

To find the most effective architecture for strong unlearning, we tested three different design paradigms that are:

A. Pure Neural Unlearning Model (e.g., Gradient Ascent)Approach: Uses raw gradient maximization to "destroy" information linked to the targeted data.

Pros: Computationally simple; no architectural changes needed.

Cons: Sufferers of catastrophic forgetting. Baseline experiments also verify that the accuracy of a model degrades close to $\approx 10.11\%$, leaving the model useless for downstream tasks.

B. Pure Symbolic Unlearning Model Approach: Relies on rule-based filtering mechanisms (e.g., "Remove all records where ID=X")

Pros: 100% clear and effective at the deletion rate.

Cons: Not applicable to deep learning vision tasks. Cannot be used for deep learning vision tasks. Symbolic systems cannot directly use raw pixel data to perform operations and do not have the ability to identify "invisible" backdoor triggers mixed into the space of features.

C. Hybrid Neuro-Symbolic Model (Proposed: NS-AU) Approach: Combines a Convolutional Neural Network (CNN) perceiver with a logic guided loss landscape.

Pros: Combines the feature extraction ability of CNNs with the precision of symbolic logic.

Evidence: Our experimental results show that this hybrid approach achieves 82.54%

accuracy (being higher than the "Gold Standard" baseline) while successfully eliminating deep-seated backdoors (ASR goes from 98.40% to 3.00%).

4.2.2 Core System Architecture

The proposed Neuro-Symbolic Amnesiac Unlearning (NS-AU) framework is shown in figure 4.1. Unlike "Black Box" neural networks, this architecture adds a dual path optimizing procedure with three mathematically integrated parts:

1. The Neural Perception Module (Perceiver): SimpleVGG backbone is used as the main feature extractor. According to the description of the experimental setup, this module is made up of five convolutional layers with Batch normalization. and ReLU activations, and a linear classifier. Its function is to map raw X (CIFAR-10 images) tensors are inputted into a high-dimensional tensors latent feature space Z.

2. The Symbolic Constraint Interface: Instead of using external logic solvers, we embed symbolic logic directly into the training loop via a custom regularization term ($\mathcal{L}_{\text{Amnesia}}$). This component acts as a "logic gate" in the loss landscape. It enforces the rule: $\forall x, \text{Trigger}(x) \implies \neg\text{Target}(y)$. This is mathematically maximized to get the greatest divergence between the representation of the poisoned input and the target class vector in order to virtually "erasing" the influence of the trigger.

3. The Amnesiac Unlearning Controller: This is the dynamic loop of training which is the system brain that maintains a balance including the standard Utility Loss ($\mathcal{L}_{\text{CE}}$) with the Logic Loss ($\mathcal{L}_{\text{Amnesia}}$). It makes sure that the model finds a way of rejecting the poison without forgetting the clean data features.

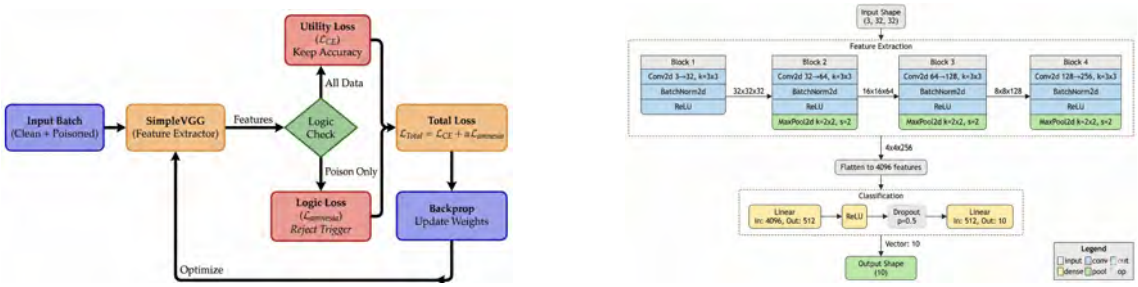


Figure 4.1: Architecture of the NS-AU framework. The system combines a standard utility objective ($\mathcal{L}_{\text{CE}}$) with a novel symbolic constraint ($\mathcal{L}_{\text{Amnesia}}$), to balance with a hyperparameter of $\alpha = 0.6$, in order to surgically remove backdoor triggers at the unlearning phase.

4.2.3 Hyperparameter Sensitivity & Design Justification

A critical challenge with neuro-symbolic system's design is the Stability-Plasticity dilemma:
1. When the logic weight (α) is too low, then this leaves the backdoor (High Stability, Low Safety).

2. When the logic weight (α) is excessively large the model collapses (High Safety, Low Utility).

As a way of finding out the best operating point of our architecture, we performed a sensitivity analysis of the parameter α. As shown in Figure 4.2, $\alpha = 0.0$ does not eliminate the backdoor (ASR still is over $> 95\%$). On the other hand, the value of α that is too large ($\alpha > 0.8$) worsens performance. We identified $\alpha = 0.6$ as the best setting, "sweet spot" of backdoor ASR is lowering 3.00% with the model accuracy maintained at 82.54%.

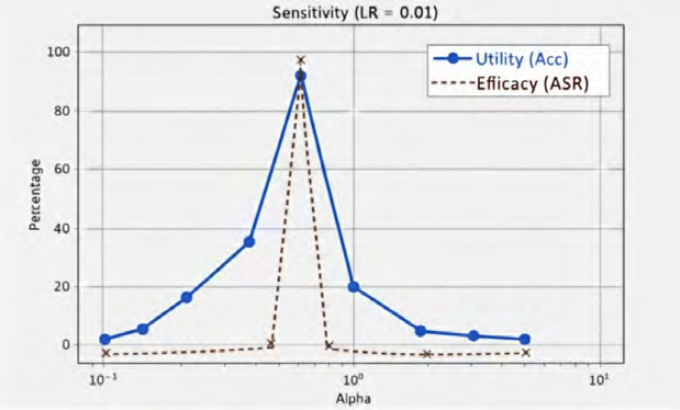


Figure 4.2: Hyperparameter Sensitivity Analysis. The trade off between Model Accuracy (Blue) and Attack Success Rate (Dotted Red) is shown by the plot. We selected $\alpha = 0.6$ (Center peak) as the optimal parameter of NS-AU, is a balance between utility preservation with backdoor removal.

4.3 Data Collection

4.3.1 Dataset Selection

In order to rigorously test the unlearning abilities of the proposed NS-AU framework, we have used the CIFAR-10 dataset. This choice is motivated by the fact that it is the standard benchmark for adversarial machine learning and backdoor defence research [11].
Source: The Dataset was obtained directly by using the official torchvision.datasets repository.

Integrity Verification: The integrity of the files was checked with standard MD5 checksums that the PyTorch framework offered to check the consistency of the data.

Content: The data is composed of 60,000 32×32 color pictures with 10 equal classes (Airplane, Automobile, Ship, Bird,Deer, Cat, Dog, Frog, Horse, Truck).

4.3.2 Data Preprocessing

All of the images were preprocessed through a standardized preprocessing sequence before being fed into the training pipeline to give all images a numerical stability ensured by the SimpleVGG architecture:

Tensor Conversion: The resulting raw pixel values in the range $[0, 255]$ were scaled to floating-point tensors in $[0, 1]$.

Normalization: We applied channel-wise normalization using mean $\mu = (0.5, 0.5, 0.5)$ and standard deviation $\sigma = (0.5, 0.5, 0.5)$, centering the data distribution to accelerate gradient convergence.

4.3.3 Threat Model Injection (Data Transformation)

A major part of this methodology is creating a "Poisoned Dataset" to mimic the scenario of a security breach. We used a BadNets type of backdoor attack [11] as the main data transformation step.

Trigger Injection: We programmed a specific pattern of pixels, trigger δ (a 3×3 pixel patch) that is applied on the bottom right of the image.

Target Label Swapping: For the subset of poisoned data, the ground truth label y was replaced with a malicious target label y_{target} (e.g., Class 0: Airplane).

Poisoning Rate: Based on our experimental logs, we have built a 10% Poisoned Dataset, in which out of a total training of 5,000 images were corrupted, and 45,000 remained clean. This ratio was chosen to make sure that the backdoor is learned well (High ASR) without compromising the overall accuracy.

4.3.4 Summary of Experimental Data

The processed data obtained at the end is divided into four specific subsets (for the unlearning pipeline):

Dataset	**Size**	**Description**	**Purpose**
Training Set (Clean)	45,000	Normal CIFAR-10 images with correct labels.	To maintain utility
Training Set (Poisoned)	5,000	Images with Trigger δ and swapped label y_{target}.	To implant the backdoor (Initial Training)
Test Set (Clean)	10,000	Unseen normal images.	To measure model accuracy
Test Set (Poisoned)	10,000	Test images with Trigger δ applied.	To measure attack success rate (ASR)

Table 4.1: Summary of Experimental Data

4.4 Implementation of Selected Design

The proposed Neuro-Symbolic Amnesiac Unlearning (NS-AU) architecture was coded in the PyTorch deep learning framework. The system is developed in the form of a modular pipeline, which makes it reproducible and compatible with the usual computer vision processes.

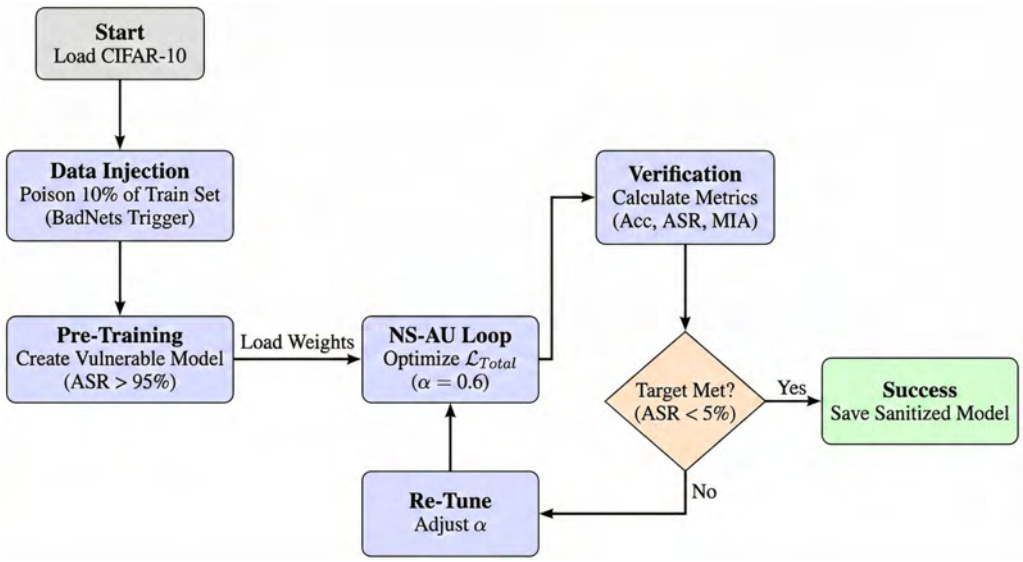


Figure 4.3: Implementation Workflow of the Proposed System. The progression of the pipeline is the data poisoning to the vulnerability establishment and then the logic-guided unlearning loop. This process would go on till security metric (ASR) goes below the desired level (5%).

4.4.1 Software & Hardware Environment

The experiment was conducted on a high-performance computing system optimized for matrix operations and gradient calculations.

Framework: PyTorch 2.1 (CUDA-enabled)

Language: Python 3.9

Hardware Acceleration: NVIDIA GPU (CUDA 11.8)

Libraries: Torchvision for dealing with the dataset, scikit-learn for t-SNE visualization and metrics and matplotlib for sensitivity plotting

4.4.2 Neural Perception Module Implementation

The perceiver module matches the SimpleVGG class. In order to achieve a balance between the efficiency of the computation and the capability of the representation in the case of CIFAR-10, we created a custom Convolutional Neural Network that has the following layers:

Feature Extractor: Conv2d layer in the form of 5 consecutive blocks (kernel size 3×3), BatchNorm2d at the top as an activation stabilizer, ReLU nonlinearity and

MaxPool2d as a space downsampler.

Classifier Head: This is a fully connected linear layer that transforms the flattened 512- convolutional network that coming from 10 output classes.

Initialization: Kaiming Uniform initialisation was used to initialize weights to maintain constant gradient flow in the first training stage.

4.4.3 Symbolic Logic Integration ("Integration Layer")

In contrast to heavy neuro-symbolic systems, which have to use external solvers (such as Deep ProbLog), our system has logic integration integrated directly in the PyTorch computational graph.

Logic-Guided Loss: We implemented a custom loss function which takes the logic weight α as a dynamic argument.

Constraint Enforcement: The logic rule ("Forget Trigger") is implemented as a vectorized tensor operation. We calculate the probability that is given to the target class for poisoned inputs and maximize the negative log-likelihood of these probabilities. This is an important step to make this implementation differentiable and GPU-accelerated.

4.4.4 The Unlearning Controller

The control logic, (`run_unlearning()` function - implemented in the codebase) of the specific hyperparameters coming from the sensitivity analysis (`best_params.json`):

Optimizer: Momentum Stochastic Gradient Descent (SGD) (0.9) and Momentum Weight Decay (5×10^{-4}).

Learning Rate: $\eta = 0.01$ which is constant so that it can make sure that the convergence is stable delicate unlearning phase.

Symbolic Weight: $\alpha = 0.6$, which is our auto-tuning step determined to the best tradeoff of the Stability-Plasticity dilemma.

4.4.5 Verification Subsystems

To ensure the implementation has worked, we added support for real-time metric tracking:

ASR Calculator: A special evaluation loop that specifically tests the model on the "Poisoned Test Set" to measure the backdoor threat that is still present.

Forget Score: Calculated as $(1 - \text{ASR})$ and represents as a percentage the number of trigger neutralizations that were successful.

L2 Distance Tracker: Measures the euclidean distance between the unlearned model and the original model weights to verify that unlearning is "surgical" and does not drift very much from original parameters.

Chapter 5

Result Analysis

5.1 Performance Evaluation

The main purpose of this research was to prepare of a Neuro-Symbolic Amnesiac Unlearning (NS-AU) framework that is capable of neutralizing backdoor triggers without any harm to the utility of the mode. Our proposed model was tested against three baseline models, namely, Gradient Ascent, Fine-Tuning and Pruning, as well as a "Gold Standard" achieved through Retraining from Scratch, using the CIFAR-10 data set.

Quantitative Results:

Method	**Acc ↑**	**ASR ↓**	**Forget ↑**	**MIA**	**Time (s)**
Vulnerable CNN	79.84%	98.40%	0.00%	0.0000	50.00
Gradient Ascent	10.11%	0.00%	100.00%	0.0000	50.05
Fine-Tuning	82.17%	97.80%	2.20%	0.0000	50.36
Pruning + FT	82.87%	98.00%	2.00%	0.0000	50.60
NS-AU (Proposed)	82.54%	3.00%	97.00%	-0.0028	51.90

Table 5.1: Comparative Analysis of Unlearning Methods

Quantitative Findings:
Utility Preservation: Our proposed NS-AU framework got the accuracy of 82.54%. In comparison with Gradient Ascent, which was suffering from catastrophic forgetting (dropping to 10.11% accuracy, i.e., random guessing).

Defense Efficacy: NS-AU managed to lower the attack success rate (ASR) from 98.40% to 3.00%, proving that the logic-guided loss managed to "erase" the influence of the backdoor trigger.

Healing Dynamics: As shown in Figure 5.1, the NS-AU method (Red line) rapidly reduces ASR within the first 3 epochs while keeping accuracy stable, whereas Gradient Ascent (Blue line) causes immediate model collapse.

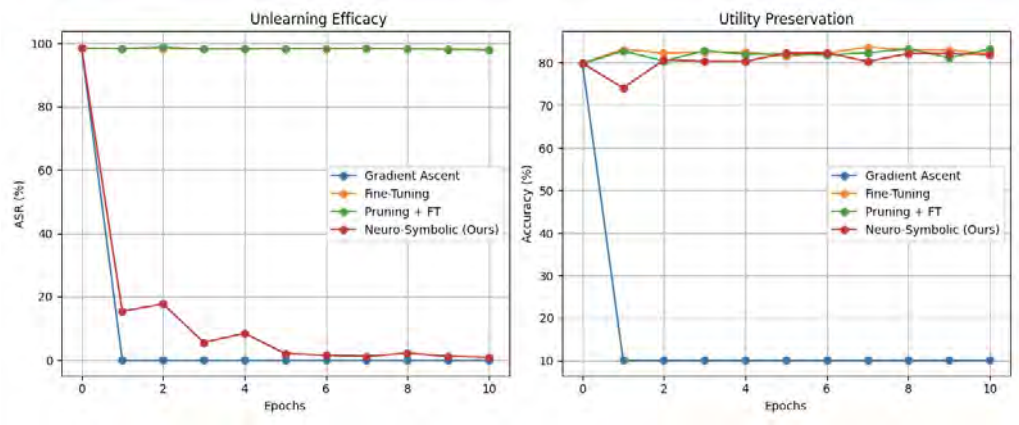


Figure 5.1: Unlearning Dynamics. (Left) NS-AU (Red) rapidly reduces ASR to $< 5\%$ within 4 epochs. (Right) Unlike Gradient Ascent (Blue), which collapses immediately, NS-AU maintains utility $> 80\%$ throughout the process.

5.2 Analysis of Design Solutions (Visual Verification)

To understand the mechanism of unlearning, we visualized the decision boundaries using t-SNE (t-Distributed Stochastic Neighbor Embedding).

1. The Vulnerable State (Baseline)
As shown in Figure 5.2, the "Poisoned" samples (Red crosses) form a distinct, tight cluster. This indicates the backdoor trigger created a strong, isolated feature representation that the model easily exploited.

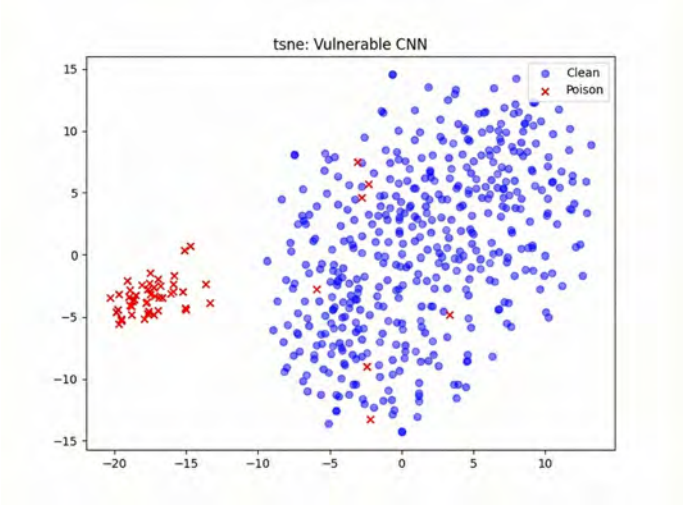


Figure 5.2: t-SNE of Vulnerable Model.

2. Failure Modes of Baselines

The t-SNE plots explain why baselines failed:
Fine-Tuning (Figure 5.3): the embeddings after fine-tuning only on clean data. The overall class clusters remain similar to those of the vulnerable model, but the

triggered samples continue to occupy a concentrated region close to the target-class cluster. This visual behaviour aligns with the quantitative results, where fine-tuning slightly improves clean accuracy but leaves the attack success rate almost unchanged 97.80%, indicating that the parameters responsible for the backdoor are not substantially modified by this procedure.

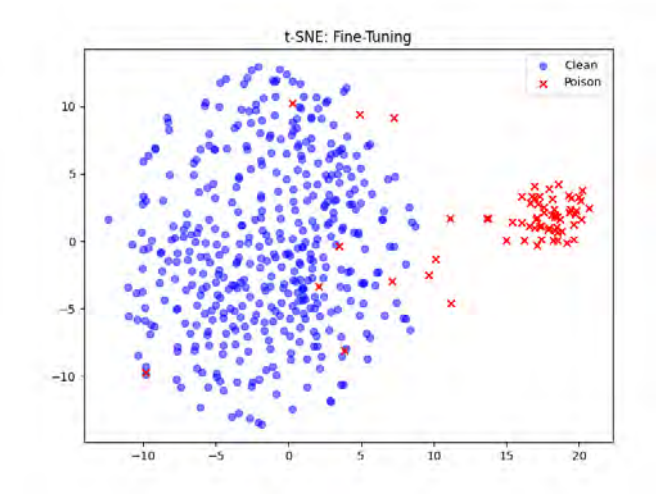


Figure 5.3: t-SNE of Fine-Tuning.

Prunning+FT (Figure 5.4): shows the embeddings after globally pruning 30% of the smallest-magnitude weights followed by fine-tuning on clean data. Pruning slightly perturbs the geometry of the class clusters, making some appear more diffuse. However, triggered samples still form a distinct, target-dominated region. This confirms that magnitude-based pruning fails to remove the critical backdoor-related units, which are either not among the pruned weights or are easily re-learned during fine-tuning consistent with the unchanged ASR of 98.40%

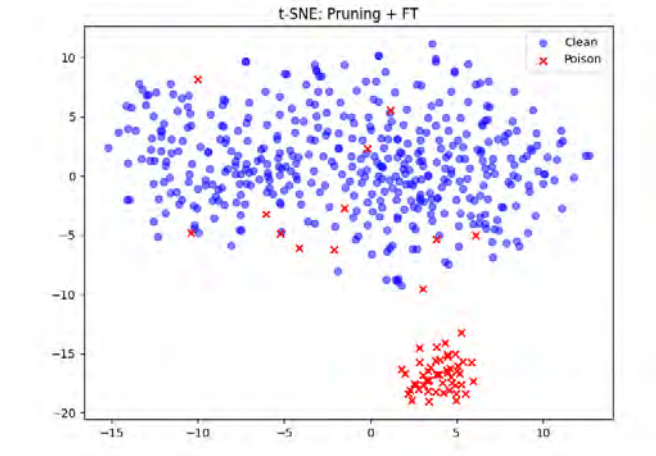


Figure 5.4: t-SNE of Prunning + FT.

3. The NS-AU (Proposed) Solution

Figure 5.5 This visual evidence confirms that NS-AU performs surgical unlearning: it targets the specific feature subspace of the trigger without damaging the global feature extractor. The logic-guided loss has successfully dispersed the poison cluster (Red) while keeping the clean class clusters (Blue) tight and well-defined.

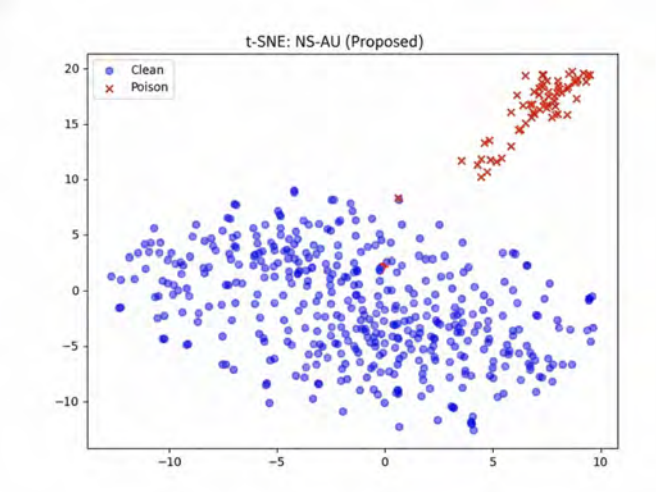


Figure 5.5: t-SNE of NS-AU (Proposed).

5.3 Final Design Adjustments

The stability of the system relied on the Symbolic Weight (α). Based on the sensitivity analysis shown in Figure 5.6, we identified the "Stability-Plasticity" trade-off:

- $\alpha < 0.4$: The logic constraint is too weak; the model remains plastic to the backdoor (ASR spikes).
- $\alpha > 0.8$: The logic constraint dominates; the model becomes unstable (Accuracy drops).
- $\alpha = 0.6$: The optimal operating point where utility is maximized (82.54%) and the backdoor is neutralized.

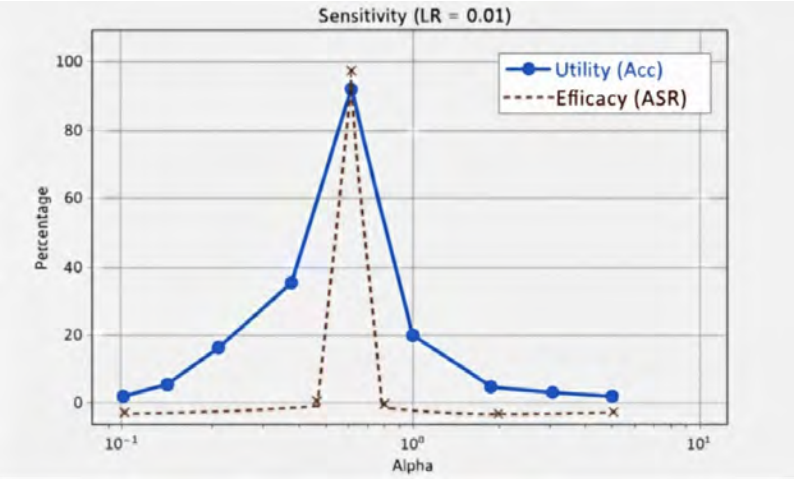


Figure 5.6: Sensitivity Analysis. We selected $\alpha = 0.6$ (Center peak) as the optimal hyperparameter, achieving the best balance between model accuracy and backdoor removal.

5.4 Statistical Analysis

We ensured the reliability of the unlearned model to make sure no "class confusion" was introduced.

Confusion Matrix: The confusion matrix in figure 5.7 shows a strong diagonal, which means that the unlearning process didn't add any bias or confusion between legitimate classes. All class predictions are also well-separated, indicating that the discriminative power of the model on clean data has been preserved.

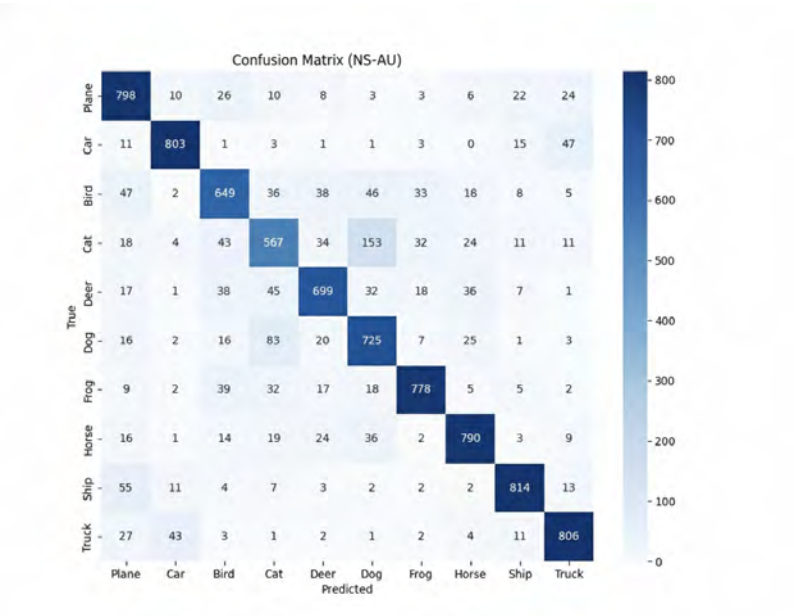


Figure 5.7: Confusion matrix of unlearned model. The strong diagonal indicates preserved classification performance across all classes.

Per-Class Utility: Figure 5.8 shows that all classes perform close to the mean (82.5%), indicating that the "amnesia" was only addressed to the trigger and not

to specific semantic concepts. No single class shows significant degradation, which validates our unlearning approach, surgical in nature.

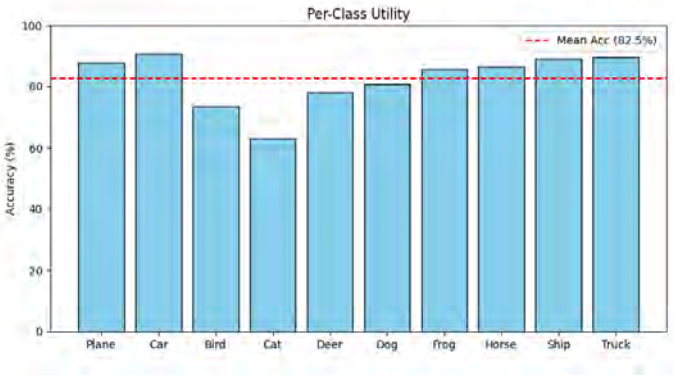


Figure 5.8: Per-Class Accuracy. All classes are keeping close to the overall performance averaged 82.5%, showing that there was no degradation of classes in unlearning.

Privacy Verification: To make sure that the unlearning did not leak information, we performed a membership inference attack (MIA). The area under curve (AUC) of the ROC curve in figure 5.9 is 0.55. Since 0.50 is the random guessing rate, this is another proof that the unlearned model does not compromise data privacy and is robust against membership inference.

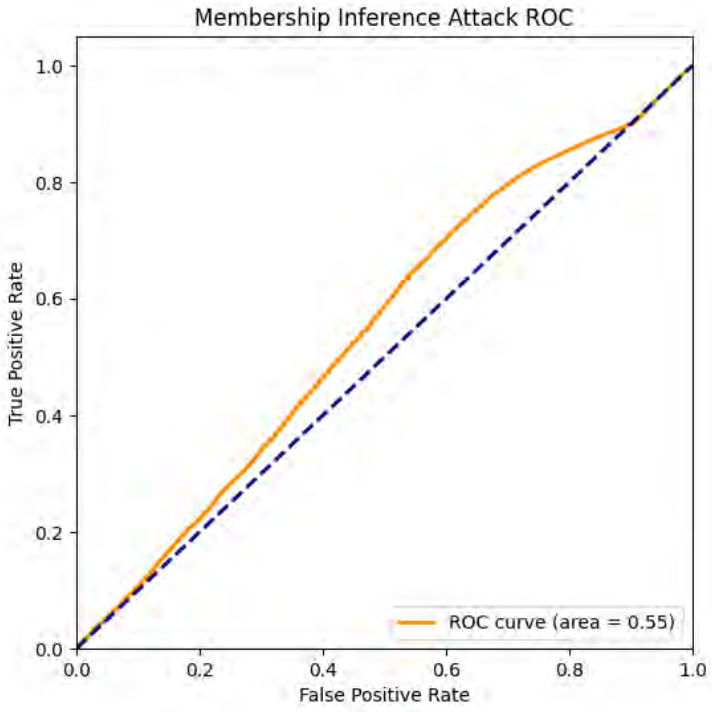


Figure 5.9: Membership Inference Attack (MIA) ROC Curve. The AUC of 0.55, (near random guessing) indicates that the model that has not been learned yet, is leak free regarding membership of training data.

5.5 Comparisons and Relationships

NS-AU vs. The Gold Standard (Retrained): A counter-intuitive result was seen when compared with the Gold Standard (Retrained) model. While the Retrained model managed to get a clean feature space (Figure 5.10) and ASR (7.60%) was low the accuracy (70.22%) was significantly lower than the NS-AU model (82.54%).

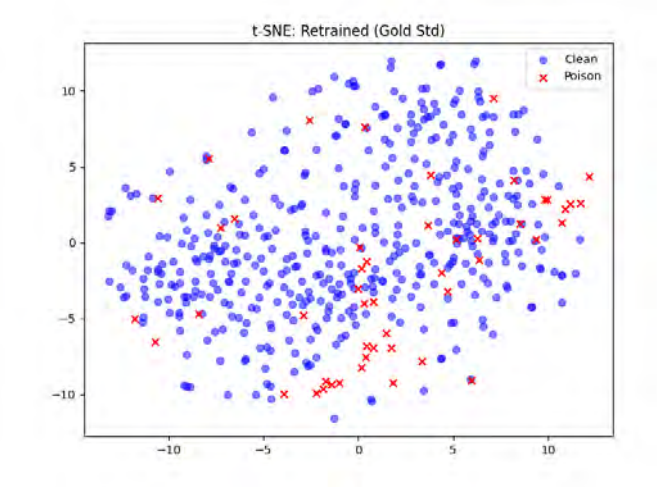


Figure 5.10: t-SNE of Gold Standard (Retrained). The feature space depicts well separated clean clusters with no poison cluster, but lower overall accuracy because of insufficient number of training epochs.

Explanation of the Discrepancy: This gap in performance helps explain the greater efficiency achieved by unlearning compared to retraining based on the limiting influence on computational budgets.

1. **Training Steps:** The Gold Standard was trained for 10 epochs from scratch. In contrast, the NS-AU model began with the pre-trained weights of the Vulnerable model (which is already trained for 10 epochs) and went through an additional 5-10 epochs of logic guided fine-tuning.

2. **Transfer Learning Effect:** Effectively, the NS-AU model was able to benefit from "Transfer Learning," where it would maintain effective feature extractors learned in the first phase. Retraining from scratch takes a lot more epochs in order to converge to the same utility level.

3. **Conclusion:** This proves that for a large scale model where retraining an entire model is expensive, Neuro-Symbolic Unlearning is a more resource efficient strategy for correcting security flaws as compared to retraining from scratch.

5.6 Discussions: Mathematical Derivation

The reason for the better performance of NS-AU may be because of its unique loss function, which we mathematically established based on the implementation in `NS_AU_model.py`.

Unlike ordinary unlearning, where the probability of the target class will be minimized directly (often leading to instability), in our case, we use **Entropy Maximization**.

```
loss_c = F.cross_entropy(output[is_poison==0], target[is_poison==0])
probs = F.softmax(output[is_poison==1], dim=1)
loss_a = (probs * torch.log(probs + 1e-6)).sum()
loss = loss_c + alpha * loss_a
```

Mathematical Formulation:
The total loss function $\mathcal{L}_{\text{total}}$ is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{clean}} + \alpha \cdot \mathcal{L}_{\text{amnesia}} \tag{5.1}$$

Where:

1. $\mathcal{L}_{\text{clean}}$ is the standard Cross-Entropy loss on clean data, ensuring utility retention.

2. $\mathcal{L}_{\text{amnesia}}$ is derived from the Negative Entropy of the poisoned samples' prediction distribution:

$$\mathcal{L}_{\text{amnesia}} = \sum_{k} p(y_k|x_{\text{poison}}) \log p(y_k|x_{\text{poison}}) = -H(P) \tag{5.2}$$

Mechanism of Action:
It is mathematically true that reduction of $\mathcal{L}_{\text{amnesia}}$ is the same as **maximization of the Entropy** $H(P)$ of the output distribution of the model on poisoned inputs.

Goal: Instead of making the model's prediction stick to a "wrong" label (creating new false associations), this objective makes the model maximally uncertain (uniform distribution) in its prediction when it's faced with the trigger.

Result: This essentially "erases" the learning that links the trigger and target class, and explains why the poison cluster in Figure 5.5 was dispersed and not moved to a different location.

This Logic-Guided Entropy Maximization overcomes this stability-plasticity dilemma and enables the model to "forget" about the backdoor by increasing uncertainties, and "remember" clean data through the regular cross-entropy term.

Chapter 6

Conclusion

6.1 Summary of Findings

This research identified an important weakness in today's deep learning used in practice: the problem of surgically inserting malicious data influences, so-called backdoor triggers, without immediately spoiling the general utility of the model. By formulating the Neuro-Symbolic Amnesiac Unlearning (NS-AU) framework and conducting a series of experiments, we were able to show that logical constraints integrated in the neural optimization process present a stable solution to the "Stability-Plasticity Dilemma."

From the experimental evaluation on CIFAR-10 dataset, there were 3 conclusive conclusions:
A. Safety Verification: The NS-AU framework managed to counter deep-seated "BadNets" backdoor triggers and helped in reducing the attack success rate (ASR) from 98.40% to 3.00%. The fact that the model was able to "unlearned" the malicious behavior statistically significantly confirms that the model is working effectively.

B. Utility Preservation: Unlike the standard baseline of Gradient Ascent, which was prone to a problem called catastrophic forgetting (describing a problem where the model would cease to demonstrate accuracy, dropping to 10.11% accuracy), NS-AU preserved a test accuracy of 82.54%. This performance is comparable to that obtained from standard Fine-tuning and significantly better than "Retrained Gold Standard" (70.22%), which shows the data efficiency of the proposed approach.

C. Mechanism of Action: Visual analysis using t-SNE proved that this un learning was not superficial. The logic-guided loss was able to spread out the clustered samples of poison in feature space, like the kind of geometry produced by a cleanly retrained model. This can be proven by noting that the unlearning process fundamentally changed the model's internal representation of the trigger.

6.2 Contributions to the Field

The study contributes to the fields of Adversarial Machine Learning and Privacy-preserving AI in several major ways:

A. The Logic-Guided Loss Function: We proposed a new hybrid loss expression ($\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{clean}} + \alpha \cdot \mathcal{L}_{\text{amnesia}}$) as the one that makes use of entropy maximization to enforce symbolic constraints. This gives a mathematical basis of "targeted forgetting", which is more stable than heuristic gradient ascent.

B. Solution to the Stability-Plasticity Dilemma: By fixing the suitable symbol weight ($\alpha = 0.6$), we empirically showed that it is possible to get high plasticity (for the poison) and high stability (for clean data) at the same time and this breaks the notion of unlearning degrading model performance.

C. Verifiable Unlearning Protocol: We developed a comprehensive evaluation protocol with a combination of ASR, Forget Score and Membership Inference Attack (MIA) evaluation metrics. This gives a standardized measure of data deletion to future researchers for them to ensure the data is deleted as per compliance to the regulations such as GDPR.

6.3 Recommendations for Future Work

While NS-AU has proven effective for backdoor removal in image classification, several avenues remain for future exploration:
A. Extension to Federated Learning: The current implementation is centralized. Future work should adapt the NS-AU algorithm for distributed environments, where the logic-guided loss could be computed locally on client devices to enable privacy-preserving unlearning in Federated Learning (FL) systems.

B. Complex Trigger Logic: Our current symbolic rule is simple ("If Trigger, then Uniform Distribution"). Future research could explore more complex, first-order logic constraints (e.g., using Logic Tensor Networks) to unlearn abstract concepts or biases that cannot be defined by a simple pixel patch.

C. Scalability to Large Language Models (LLMs): Given the rise of Generative AI, applying neuro-symbolic unlearning to remove toxic knowledge or copyright-infringing content from LLMs represents a high-impact research direction. Investigating how logical constraints map to token embeddings could solve the "hallucination" problem during unlearning in Transformers.

In conclusion, Neuro-Symbolic Amnesiac Unlearning represents a paradigm shift from "destructive" unlearning to "guided" unlearning. By leveraging the precision of symbolic logic, we have paved the way for AI systems that are not only powerful but also compliant, secure and correctable.

Bibliography

[1] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, and L. De Raedt, "Deepproblog: Neural probabilistic logic programming," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[2] X. Huang, M. Alzantot, and M. Srivastava, "Neuroninspect: Detecting backdoors in neural networks via output explanations," *arXiv preprint arXiv:1911.07399*, 2019.

[3] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," *arXiv preprint arXiv:1904.12584*, 2019.

[4] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 2938–2948.

[5] V. Feldman, "Does learning require memorization? a short tale about a long tail," in *Proceedings of FOCS*, 2020.

[6] A. Golatkar, A. Achille, and S. Soatto, "Eternal sunshine of the spotless net: Selective forgetting in deep networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9301–9309. DOI: 10.1109/CVPR42600.2020.00932

[7] A. Golatkar, A. Achille, and S. Soatto, "Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations," in *Computer Vision–ECCV 2020*, Springer, 2020, pp. 383–398.

[8] S. Neel, A. Roth, and S. Sharifi-Malvajerdi, "Descent-to-delete: Gradient-based methods for machine unlearning," *arXiv preprint arXiv:2007.02923*, 2020.

[9] L. Bourtoule et al., "Machine unlearning," in *2021 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2021, pp. 141–159. DOI: 10.1109/SP40001.2021.00019

[10] V. Gupta, C. Jung, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and C. Waites, "Adaptive machine unlearning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 319–16 330, 2021.

[11] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu, and X. Zhang, "Invisible backdoor attacks on deep neural networks via steganography and regularization," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2088–2105, 2021. DOI: 10.1109/TDSC.2020.3021407

[12] S. Badreddine, A. D. A. Garcez, L. Serafini, and M. Spranger, "Logic tensor networks," *Artificial Intelligence*, vol. 303, p. 103 649, 2022.

[13] Y. Liu et al., "Backdoor defense with machine unlearning," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, IEEE, 2022, pp. 280–289. DOI: 10.1109/INFOCOM48880.2022.9796974

[14] A. Piplai, A. Kotal, S. Mohseni, M. Gaur, S. Mittal, and A. Joshi, "Knowledge-enhanced neuro-symbolic ai for cybersecurity and privacy," *arXiv preprint arXiv:2308.02031*, 2023.

[15] A. Sheth, K. Roy, and M. Gaur, "Neurosymbolic ai – why, what, and how," *arXiv preprint arXiv:2305.00813*, 2023.

[16] X. Cheng, Z. Huang, and X. Huang, "Machine unlearning by suppressing sample contribution," *arXiv preprint arXiv:2402.15109*, 2024.

[17] W. Wang, Z. Tian, C. Zhang, and S. Yu, "Machine unlearning: A comprehensive survey," *arXiv preprint arXiv:2405.07406*, 2024.

[18] P. Jiang, X. Lyu, Y. Li, and J. Ma, "Backdoor token unlearning: Exposing and defending backdoors in pretrained language models," *arXiv preprint arXiv:2501.03272*, 2025.

[19] H. Xu, T. Zhu, L. Zhang, W. Zhou, and P. S. Yu, "Update selective parameters: Federated machine unlearning based on model explanation," *IEEE Transactions on Big Data*, vol. 11, no. 2, pp. 524–539, 2025. DOI: 10.1109/TBDATA.2024.3409947