UNIVERSITAT DE BARCELONA

Bachelor's Thesis

BACHELOR'S DEGREE IN COMPUTER ENGINEERING

Faculty of Mathematics and Computer Science
Universitat de Barcelona

Challenging Forgets in Tabular Neural Networks: A Comparative Analysis of Noise-Based Unlearning Methods and Forget Set Structure

---

Author: Eshaan Mittal

Supervisor: Dr. Nahuel Norberto Statuto Perez
Affiliation: Departament of Mathematics and Computer Science

Barcelona, January 18, 2026

Abstract

As a machine learning models is trained, it encodes information from the training data into learned parameters, creating challenges when individuals exercise their right to data deletion under frameworks such as the General Data Protection Regulation. Complete model retraining without the targeted samples represents the theoretically optimal solution, yet this approach imposes prohibitive computational costs for production systems handling frequent deletion requests. Machine unlearning has emerged as an alternative paradigm that modifies model parameters to remove the influence of specific training samples without requiring full retraining.

This thesis investigates noise-based machine unlearning strategies applied to TabNet, an attention-based neural network architecture for tabular data processing. The central research question examines how the structural composition of forget sets, particularly class distribution, determines unlearning effectiveness. Five strategies are evaluated: Gaussian noise injection, Laplacian noise injection, adaptive gradient-weighted noise, layer-wise progressive noise, and gradient-based unlearning through gradient ascent.

Experiments employ the Adult Income dataset with four forget request scenarios designed to systematically vary class balance. The Married scenario removes individuals with married civil spouse status, yielding 44.6% positive class composition. The Executives scenario targets managerial occupations with 48.4% positive class. The HighEarnProf scenario employs geometric selection of high-earning professionals, producing 100% positive class composition (complete single-class imbalance). The RandomBalanced scenario creates a stratified random sample with exactly 50% from each class (perfect balance). This design enables rigorous analysis of the relationship between class composition and unlearning outcomes.

Results establish that class balance within the forget set constitutes the primary determinant of unlearning success. Balanced scenarios achieve forget accuracy in the 0.59-0.67 range. The HighEarnProf scenario with complete class imbalance initially appears to exhibit anti-learning, with forget accuracy falling to approximately 0.14-0.23. However, comparison with gold standard models (retrained from scratch without the forget data) reveals that this low accuracy represents successful unlearning rather than failure: the gold standard achieves similarly low accuracy (0.16) on these samples, indicating they are edge cases that were memorized during training. Unlearning successfully removes this memorization, causing the model to generalize naturally. Laplacian noise injection demonstrates the strongest performance on balanced scenarios, while all strategies perform equivalently on imbalanced cases.

Computational efficiency represents a practical advantage, with noise-based methods completing in 3.5-17 seconds compared to 28-67 seconds for full retraining, representing speedup factors of 1.8-10.6 times for the Adult Income dataset. For larger models and datasets, this efficiency gap would be substantially greater. These findings establish that practitioners must analyze forget set class distribution prior to applying noise-based unlearning, as structural properties fundamentally constrain achievable outcomes regardless of strategy selection. The aim of this thesis is to provide a better understanding of structure-dependent unlearning limitations and practical insights for GDPR-compliant machine learning deployments.

Resumen

Mientras un modelo de aprendizaje automático se entrena, codifica información de los datos de entrenamiento en parámetros aprendidos, creando desafíos cuando los individuos ejercen su derecho a la eliminación de datos bajo marcos normativos como el Reglamento General de Protección de Datos. El reentrenamiento completo del modelo sin las muestras objetivo representa la solución teóricamente óptima, sin embargo, este enfoque impone costos computacionales prohibitivos para sistemas de producción que manejan solicitudes frecuentes de eliminación. El machine unlearning ha surgido como un paradigma alternativo que modifica los parámetros del modelo para eliminar la influencia de muestras de entrenamiento específicas sin requerir reentrenamiento completo.

Esta tesis investiga estrategias de machine unlearning basadas en ruido aplicadas a TabNet, una arquitectura de red neuronal basada en atención para el procesamiento de datos tabulares. La pregunta central de investigación examina cómo la composición estructural de los conjuntos de olvido, particularmente la distribución de clases, determina la efectividad del desaprendizaje. Se evalúan cinco estrategias: inyección de ruido Gaussiano, inyección de ruido Laplaciano, ruido adaptativo ponderado por gradiente, ruido progresivo por capas y desaprendizaje basado en gradientes mediante ascenso de gradiente.

Los experimentos emplean el conjunto de datos Adult Income con cuatro escenarios de solicitud de olvido diseñados para variar sistemáticamente el balance de clases. El escenario Married elimina individuos con estado civil de casado, produciendo 44.6 % de composición de clase positiva. El escenario Executives se dirige a ocupaciones gerenciales con 48.4 % de clase positiva. El escenario HighEarnProf emplea selección geométrica de profesionales de altos ingresos, produciendo 100 % de composición de clase positiva (desequilibrio completo de clase única). El escenario RandomBalanced crea una muestra aleatoria estratificada con exactamente 50 % de cada clase (balance perfecto). Este diseño permite un análisis riguroso de la relación entre composición de clases y resultados del desaprendizaje.

Los resultados establecen que el balance de clases dentro del conjunto de olvido constituye el determinante principal del éxito del desaprendizaje. Los escenarios balanceados alcanzan precisión de olvido en el rango de 0.59-0.67. El escenario HighEarnProf con desequilibrio completo de clases inicialmente parece exhibir anti-aprendizaje, con precisión de olvido cayendo a aproximadamente 0.14-0.23. Sin embargo, la comparación con modelos gold standard (reentrenados desde cero sin los datos de olvido) revela que esta baja precisión representa desaprendizaje exitoso en lugar de fracaso: el gold standard alcanza

precisión similarmente baja (0.16) en estas muestras, indicando que son casos atípicos que fueron memorizados durante el entrenamiento. El desaprendizaje elimina exitosamente esta memorización, causando que el modelo generalice naturalmente. La inyección de ruido Laplaciano demuestra el rendimiento más fuerte en escenarios balanceados, mientras que todas las estrategias funcionan equivalentemente en casos desequilibrados.

La eficiencia computacional representa una ventaja práctica, con métodos basados en ruido completando en 3.5-17 segundos comparado con 28-67 segundos para reentrenamiento completo, representando factores de aceleración de 1.8-10.6 veces para el conjunto de datos Adult Income. Para modelos y conjuntos de datos más grandes, esta brecha de eficiencia sería sustancialmente mayor. Estos hallazgos establecen que los profesionales deben analizar la distribución de clases del conjunto de olvido antes de aplicar desaprendizaje basado en ruido, ya que las propiedades estructurales restringen fundamentalmente los resultados alcanzables independientemente de la selección de estrategia. La tesis contribuye tanto comprensión teórica de las limitaciones del desaprendizaje dependientes de la estructura como conocimientos prácticos para implementaciones de aprendizaje automático conformes con el RGPD.

Resum

Quan un model d'aprenentatge automàtic s'entrena, codifica informació de les dades d'entrenament en paràmetres apresos, creant reptes quan els individus exerceixen el seu dret a l'eliminació de dades sota marcs normatius com el Reglament General de Protecció de Dades. El reentrenament complet del model sense les mostres objectiu representa la solució teòricament òptima, però aquest enfocament imposa costos computacionals prohibitius per a sistemes de producció que gestionen sol·licituds freqüents d'eliminació. El machine unlearning ha sorgit com un paradigma alternatiu que modifica els paràmetres del model per eliminar la influència de mostres d'entrenament específiques sense requerir reentrenament complet.

Aquesta tesi investiga estratègies de machine unlearning basades en soroll aplicades a TabNet, una arquitectura de xarxa neuronal basada en atenció per al processament de dades tabulars. La pregunta central de recerca examina com la composició estructural dels conjunts d'oblit, particularment la distribució de classes, determina l'efectivitat del desaprenentatge. S'avaluen cinc estratègies: injecció de soroll Gaussià, injecció de soroll Laplacià, soroll adaptatiu ponderat per gradient, soroll progressiu per capes i desaprenentatge basat en gradients mitjançant ascens de gradient.

Els experiments empren el conjunt de dades Adult Income amb quatre escenaris de sol·licitud d'oblit dissenyats per variar sistemàticament el balanç de classes. L'escenari Married elimina individus amb estat civil de casat, produint 44.6% de composició de classe positiva. L'escenari Executives es dirigeix a ocupacions gerencials amb 48.4% de classe positiva. L'escenari HighEarnProf empra selecció geomètrica de professionals d'alts ingressos, produint 100% de composició de classe positiva (desequilibri complet de classe única). L'escenari RandomBalanced crea una mostra aleatòria estratificada amb exactament 50% de cada classe (balanç perfecte). Aquest disseny permet una anàlisi rigorosa de la relació entre composició de classes i resultats del desaprenentatge.

Els resultats estableixen que el balanç de classes dins del conjunt d'oblit constitueix el determinant principal de l'èxit del desaprenentatge. Els escenaris balançats assoleixen precisió d'oblit en el rang de 0.59-0.67. L'escenari HighEarnProf amb desequilibri complet de classes inicialment sembla exhibir anti-aprenentatge, amb precisió d'oblit caient a aproximadament 0.14-0.23. No obstant això, la comparació amb models gold standard (reentrenats des de zero sense les dades d'oblit) revela que aquesta baixa precisió representa desaprenentatge exitós en lloc de fracas: el gold standard assoleix precisió similarment baixa (0.16) en aquestes mostres, indicant que són casos atípics que van ser memoritzats durant l'entrenament. El desaprenentatge elimina exitosament aquesta memorització,

causant que el model generalitzi naturalment. La injecció de soroll Laplacià demostra el rendiment més fort en escenaris balançats, mentre que totes les estratègies funcionen equivalentment en casos desequilibrats.

L'eficiència computacional representa un avantatge pràctic, amb mètodes basats en soroll completant en 3.5-17 segons comparat amb 28-67 segons per a reentrenament complet, representant factors d'acceleració de 1.8-10.6 vegades per al conjunt de dades Adult Income. Per a models i conjunts de dades més grans, aquesta bretxa d'eficiència seria substancialment més gran. Aquests resultats estableixen que els professionals han d'analitzar la distribució de classes del conjunt d'oblit abans d'aplicar desaprenentatge basat en soroll, ja que les propietats estructurals restringeixen fonamentalment els resultats assolibles independentment de la selecció d'estratègia. La tesi contribueix tant comprensió teòrica de les limitacions del desaprenentatge dependents de l'estructura com coneixements pràctics per a implementacions d'aprenentatge automàtic conformes amb el RGPD.

Acknowledgments

I wish to express my gratitude to my thesis supervisor, Dr. Nahuel Norberto Statuto Perez of the Department of Mathematics and Computer Science at the Universitat de Barcelona, for his guidance with the development of this research.

I acknowledge the Faculty of Mathematics and Computer Science at the Universitat de Barcelona for providing the academic environment necessary for conducting these experiments.

I recognize the use of artificial intelligence tools during the development of this thesis. Specifically, ChatGPT developed by OpenAI, Claude developed by Anthropic, and Cursor AI were employed as programming assistants for code optimization and implementation suggestions. All experimental design decisions, data analysis, interpretation of results, and written content represent my own contribution.

I extend my appreciation to my friends and family for their continued support and encouragement throughout my undergraduate studies and the completion of this thesis.

Contents

Chapter 1

Introduction and Motivation

Machine learning models have become widespread and are now an integral part of decision-making processes across numerous domains, including financial services, healthcare diagnostics, employment screening, and personalized recommendations. These models derive their predictive capabilities from patterns learned during training on large datasets, which frequently contain sensitive personal information. The learned parameters of such models effectively encode information about individual training samples, creating a tension between model utility and individual privacy rights.

This tension has acquired legal significance with the implementation of data protection regulations worldwide. The European Union's General Data Protection Regulation (GDPR), which came into effect in May 2018, establishes the "right to erasure" (Article 17), commonly referred to as the "right to be forgotten." This provision grants individuals the right to request deletion of their personal data under specified circumstances. Similar provisions exist in the California Consumer Privacy Act (CCPA) and Brazil's Lei Geral de Proteção de Dados (LGPD). For organizations deploying machine learning systems, compliance with these regulations requires not merely removing data from storage systems but also eliminating the influence of that data from trained models [4].

The straightforward approach to data removal—retraining the model from scratch without the targeted data—provides a mathematically complete solution. A model trained without specific samples contains no information derived from those samples. However, this approach presents substantial practical limitations. Modern machine learning models can take up to days of computational time for training, consume a significant amount of energy, and incur considerable infrastructure costs. For organizations receiving frequent deletion requests, continuous retraining becomes operationally infeasible. This computational barrier has motivated the development of machine unlearning as an efficient alternative [3].

Machine unlearning encompasses algorithmic approaches that modify trained model parameters to remove the influence of specific training samples without requiring complete retraining. The objective is to produce a model that behaves indistinguishably from one that was never trained on the targeted data, while requiring substantially less computational effort than retraining. Various approaches have been proposed, including exact unlearning methods that provide mathematical guarantees equivalent to retraining [9],

and approximate methods that achieve practical privacy protection with greater efficiency [3].

The theoretical foundations of approximate unlearning connect to differential privacy, a mathematical framework for quantifying privacy guarantees in data analysis [5]. Differential privacy establishes that adding carefully calibrated noise to computations can provide provable bounds on information leakage about individual records. This principle underlies noise-based unlearning strategies, which inject random perturbations into model parameters to obscure the contribution of specific training samples [1]. The challenge lies in determining appropriate noise magnitudes that achieve effective forgetting while preserving model utility on retained data.

Much of the existing machine unlearning literature has evaluated proposed methods in the context of computer vision and natural language processing tasks, where convolutional neural networks and transformer-based models are commonly used. In contrast, less attention has been directed toward settings involving tabular data—structured datasets composed of rows and columns with numerical and categorical features. Tabular data is widely used in many applied domains, including finance, healthcare, insurance, and public administration, where it supports tasks such as credit risk assessment, fraud detection, clinical decision support, and eligibility determination. These applications often involve sensitive personal information and may be subject to data deletion or modification requirements. However, the extent to which existing machine unlearning approaches translate effectively to neural networks trained on tabular data remains less well understood.

Recent research has identified that unlearning difficulty varies substantially depending on which samples are targeted for removal. Gil Hernandez [8] demonstrated that data points exhibiting high similarity to many other training samples prove more difficult to unlearn than isolated points, as redundant information paths allow the model to reconstruct patterns from retained samples. Fan et al. [6] formalized this observation through bi-level optimization, characterizing "worst-case" forget sets that maximize unlearning difficulty. These findings indicate that evaluating unlearning methods solely on randomly selected forget sets may overestimate their practical effectiveness. However, this research has focused primarily on image classification tasks, leaving open questions about whether similar phenomena occur with tabular data and how forget set characteristics should inform method selection.

This thesis investigates noise-based machine unlearning for TabNet [2], an attention-based neural network architecture specifically designed for tabular data. TabNet employs sequential attention mechanisms to select relevant features at each decision step, providing interpretability through sparse attention masks while achieving competitive performance with gradient boosting methods. The architectural properties of TabNet—including sparse feature selection and hierarchical decision steps—create distinct considerations for unlearning compared to dense neural networks.

The central research question of this thesis concerns the relationship between forget set structure and unlearning effectiveness. Specifically, the investigation examines how the class distribution within forget sets affects the ability of noise-based methods to achieve successful unlearning. This question has practical significance because real-world deletion requests based on demographic attributes (such as marital status or occupation) may

produce forget sets with varying class compositions, potentially creating scenarios where standard unlearning methods perform poorly.

To address this question, the thesis develops and evaluates five noise-based unlearning strategies: Gaussian noise injection, Laplacian noise injection, adaptive gradient-weighted noise, layer-wise progressive noise, and gradient-based unlearning through gradient ascent. These methods are evaluated on the Adult Income dataset using four carefully designed forget request scenarios that span the spectrum of class balance: attribute-based scenarios with moderate balance (Married individuals, Executives), a geometric scenario with extreme imbalance (HighEarnProf with 100% single class), and a stratified random scenario with perfect balance (RandomBalanced with 50% each class). This experimental design enables systematic analysis of how structural properties affect unlearning outcomes.

The contributions of this thesis include: (1) a comprehensive evaluation of noise-based unlearning strategies for TabNet on tabular data, (2) identification of class balance as the primary determinant of unlearning success, (3) reinterpretation of apparent anti-learning as successful unlearning of memorized edge cases through gold standard comparison, (4) empirical validation through multi-seed experiments providing statistical confidence in results, and (5) a systematic comparison of noise-based unlearning strategies. These findings advance understanding of machine unlearning for tabular neural networks and provide actionable insights for organizations seeking to implement efficient data deletion capabilities.

Chapter 2

Objectives and Methodology

This thesis addresses the question of how forget set structure impacts the effectiveness of noise-based machine unlearning methods applied to tabular neural networks. The research is motivated by the practical need for efficient data deletion mechanisms that comply with privacy regulations while maintaining model utility.

2.1 Research Objectives

The principal objective is to evaluate whether noise-based unlearning strategies can successfully remove information from trained TabNet models without requiring complete retraining. Success is defined along three dimensions: effective forgetting (the model should exhibit behavior consistent with never having trained on the targeted data), utility preservation (the model should maintain predictive accuracy on retained data), and computational efficiency (the unlearning operation should require substantially less time than retraining).

A secondary objective concerns the relationship between forget set characteristics and unlearning outcomes. Prior research on image classification has established that certain data points are more difficult to unlearn than others, but this phenomenon has not been systematically investigated for tabular data. This thesis examines whether similar structural factors—particularly class distribution within the forget set—affect unlearning difficulty for TabNet models. Understanding these relationships enables practitioners to anticipate when standard unlearning methods will succeed and when alternative approaches may be necessary.

2.2 Experimental Methodology

The experimental methodology follows a structured pipeline designed to enable fair comparison between unlearning strategies and gold standard retraining. First, the Adult Income dataset is preprocessed and a baseline TabNet model is trained on the full training set, achieving approximately 84% test accuracy. Second, four forget request scenarios are

defined to span the spectrum of class balance: Married (44.6% Class 1), Executives (48.4% Class 1), HighEarnProf (100% Class 1, representing the extreme imbalanced case), and RandomBalanced (50% each class, representing the ideal balanced case).

Third, for each forget request scenario, a gold standard model is trained from scratch on only the retain set. These gold standard models represent the theoretical ideal—models that never learned from the targeted samples. Fourth, five noise-based unlearning strategies are applied to the baseline model: Gaussian noise injection, Laplacian noise injection, adaptive gradient-weighted noise, layer-wise progressive noise, and gradient-based unlearning with entropy maximization. Each strategy includes fine-tuning on the retain set with a penalty on the forget set to prevent re-learning.

Fifth, unlearned models are evaluated against gold standard models using comprehensive metrics. Forget accuracy measures performance on the forget set, with successful unlearning indicated by matching the gold standard model's behavior rather than necessarily achieving 50% (random guessing). This distinction is important because the gold standard accuracy on forget sets varies by scenario—for example, edge case samples that were memorized during training may naturally have low accuracy even in retrained models. Retain accuracy and test accuracy measure utility preservation. MIA AUC quantifies privacy protection, with a target of 0.50 indicating the attacker cannot identify training data.

To enable direct comparison between strategies, a Composite Evaluation Score (CES) is defined that combines these metrics into a single value. The CES formula balances three objectives: utility preservation (retain accuracy), forgetting effectiveness, and privacy protection (MIA resistance). The formula is:

$$\text{CES} = 0.35 \times R + 0.35 \times F + 0.30 \times P - \lambda \tag{2.1}$$

where R is retain accuracy, $F = \max(0, 1 - |A_f - 0.5| \times 2)$ is the forget score based on forget accuracy A_f, $P = \max(0, 1 - |M - 0.5| \times 3)$ is the privacy score based on MIA AUC M, and $\lambda = 0.25$ is applied when $A_f < 0.25$ to penalize scores in scenarios where 50% is the appropriate target. The weights reflect equal importance of utility and forgetting (35% each), with slightly lower weight on privacy (30%) since MIA AUC is inherently noisier. The forget score component uses 50% as a baseline target, though as discussed, the true measure of success is alignment with gold standard behavior. The privacy score peaks at $M = 0.5$ (no membership leakage) and decreases with higher MIA values.

Finally, multi-seed experiments with five random seeds provide statistical validation, with results reported as mean $\pm$ standard deviation.

The experimental design enables several research questions to be addressed. Does class balance within the forget set predict unlearning success? Do all strategies exhibit similar sensitivity to structural properties, or do some prove more robust? What happens when the forget set consists entirely of a single class? Can noise-based methods achieve acceptable unlearning across diverse structural conditions, or are there fundamental limitations that necessitate retraining?

The outcomes of this research have practical implications for organizations implementing data deletion capabilities. If class balance is identified as a critical factor, practitioners can analyze forget set composition before selecting an unlearning method. If certain structural configurations consistently cause failure, organizations can establish policies for

when approximate unlearning is appropriate versus when full retraining is required. The findings contribute to the broader understanding of machine unlearning for tabular data, an area that has received limited attention despite its relevance to enterprise applications.

Chapter 3

Development

This chapter presents the technical implementation of the experimental framework, including the model architecture, dataset preparation, unlearning strategies, and evaluation methodology. The development follows a systematic approach designed to enable rigorous analysis of how forget set structure affects unlearning outcomes.

3.1 Model Architecture

TabNet (Attentive Interpretable Tabular Learning) serves as the foundation for all experiments in this thesis [2]. TabNet represents a neural network architecture specifically designed for tabular data that achieves competitive performance with gradient boosting methods while providing interpretability through attention mechanisms. The architecture employs sequential attention to select relevant features at each decision step, creating a soft decision tree structure within a neural network framework.

The sequential attention mechanism operates through multiple decision steps. At each step, the model computes a sparse attention mask that determines which input features should be processed. This mask is generated through a learnable attention transformer that takes the processed features from previous steps as input. The sparsity is enforced through the Sparsemax activation function, which projects attention weights onto a probability simplex while allowing many weights to be exactly zero. This differs from the standard Softmax activation, which produces dense attention distributions where all features receive non-zero weights.

After computing the attention mask, the selected features pass through a feature transformer consisting of fully connected layers with batch normalization. The transformed features contribute to the current decision step's output and also inform the attention computation for subsequent steps. The final prediction aggregates contributions from all decision steps, with each step potentially focusing on different feature subsets. This architecture enables the model to learn complex feature interactions while maintaining interpretability through the attention masks.

For the Adult Income experiments, TabNet is configured with decision and attention dimensions of 64, five decision steps, a feature reusage coefficient (gamma) of 1.3, and

sparsity regularization of 0.001. This configuration provides sufficient model capacity for the 11-feature dataset while maintaining computational efficiency. Training employs the Adam optimizer with an initial learning rate of 0.02 and a learning rate scheduler that reduces the rate upon plateau. Early stopping with patience of 20 epochs prevents overfitting, and the model achieving the best validation accuracy is retained for unlearning experiments.

3.2 Dataset and Preprocessing

The Adult Income dataset, derived from the 1994 United States Census, serves as the primary experimental testbed. This dataset contains demographic and employment information for 48,842 individuals, with the prediction task being binary classification of income level (greater than $50,000 versus less than or equal to $50,000 annually). The dataset exhibits class imbalance, with approximately 76% of samples in the lower income bracket and 24% in the higher bracket.

The original dataset contains 14 features spanning demographics (age, sex, race, native-country), education (education, education-num), employment (workclass, occupation, hours-per-week), financial information (capital-gain, capital-loss, fnlwgt), and personal relationships (marital-status, relationship). After analysis of feature relevance and redundancy, three features are removed: fnlwgt (census sampling weight with no predictive value for income), education (redundant with the numerical education-num encoding), and native-country (excessive cardinality with minimal predictive signal). The resulting 11-feature dataset provides cleaner signal for unlearning experiments while maintaining realistic complexity.

Categorical features (workclass, marital-status, occupation, relationship, race, sex) are encoded using label encoding, which assigns integer values to each category. Missing values, encoded as question marks in the original data, are replaced with an explicit "Unknown" category rather than dropping affected records. This approach preserves training samples while allowing the model to learn patterns associated with data missingness, which may itself be informative for income prediction. Approximately 7% of records contain at least one missing value, all present in the workclass and occupation columns.

All features undergo standardization using StandardScaler to achieve zero mean and unit variance. This normalization ensures consistent scales across numerical and encoded categorical features, facilitating stable gradient-based optimization during both training and unlearning. The same scaler instance is applied throughout all experiments to prevent distribution shift artifacts that could confound unlearning measurements.

To address class imbalance during training, computed class weights are applied: 0.66 for the majority class (income $\leq$50K) and 2.09 for the minority class (income $>$50K). These weights ensure the model learns meaningful patterns for both classes rather than converging to a trivial majority-class prediction strategy. The trained baseline model achieves approximately 84.73% test accuracy, providing a reliable foundation for subsequent unlearning experiments.

3.3 Forget Set Design

The experimental design centers on four forget request scenarios that systematically vary in their class balance characteristics. This design enables direct investigation of how class composition affects unlearning outcomes, which represents the central research question of this thesis.

The first scenario, designated Married, removes all individuals with "Married-civ-spouse" marital status. This attribute-based selection produces a forget set of 11,999 samples (46.1% of training data) with 44.6% belonging to Class 1 (high earners), indicating near-balanced class distribution. This scenario simulates a realistic GDPR deletion request where a demographic group collectively requests data removal.

The second scenario, designated Executives, removes all individuals with "Exec-managerial" occupation. This produces a forget set of 3,258 samples (12.5% of training data) with 48.4% belonging to Class 1. The near-perfect balance makes this scenario suitable for evaluating baseline unlearning effectiveness under favorable structural conditions.

The third scenario, designated HighEarnProf, employs geometric selection to remove high-earning professionals. Specifically, samples are selected from Class 1 (high earners) using a distance-based criterion that targets the outer portion of the class distribution in feature space. This produces a forget set of approximately 1,569 (6% of training data) samples with 100% belonging to Class 1—complete single-class composition with no minority class representation. This extreme imbalance scenario tests the limits of noise-based unlearning methods and enables investigation of the anti-learning phenomenon.

The fourth scenario, designated RandomBalanced, creates a stratified random sample with perfect class balance. Equal numbers of samples, 1,500 each, are randomly selected from Class 0 and Class 1, producing a forget set of 3,000 samples (11.5% of training data) with exactly 50% in each class. This scenario provides the theoretical ideal conditions for unlearning and serves as a benchmark for what noise-based methods can achieve under optimal structural conditions.

For scenarios where Class 1 percentage is below 50%, this value also represents the minority class percentage (the proportion of samples in the smaller class). The minority class percentage ranges from 0% (complete single-class composition, as in HighEarnProf) to 50% (perfect balance, as in RandomBalanced). This metric enables quantitative analysis of the relationship between class balance and unlearning outcomes.

Table 3.1: Forget Request Scenario Design and Characteristics

Scenario	Selection Type	Samples	% Data Removed	Class 1 %	Minority %	Imbalance Ratio	Expected Outcome
Married	Attribute	∼12,000	46%	44.6%	44.6%	1.2:1	Good
Executives	Attribute	∼3,300	13%	48.4%	48.4%	1.1:1	Good
HighEarnProf	Geometric	∼1,500	6%	100.0%	0%	∞:1	Edge cases
RandomBalanced	Stratified	3,000	12%	50.0%	50%	1:1	Optimal

3.4 Unlearning Strategies

Five noise-based unlearning strategies are implemented, each representing a different approach to parameter perturbation for information removal. All strategies share the objective of modifying model parameters to reduce predictive capability on the forget set while preserving performance on the retain set. The fundamental principle underlying these approaches is that neural network parameters encode information about training samples, and appropriately calibrated perturbations can disrupt this encoding while leaving sufficient structure for the model to maintain utility on retained data.

Each strategy follows a common pipeline: (1) perturbation of model parameters using the strategy-specific method, (2) optional fine-tuning on the retain set to recover utility, with an entropy penalty on the forget set to prevent re-learning the forgotten patterns, and (3) evaluation against gold standard retrained models. The fine-tuning phase is critical for balancing forgetting with utility preservation—without it, aggressive perturbations may achieve forgetting but destroy model utility entirely.

3.4.1 Gaussian Noise Injection

Gaussian noise injection is the simplest and most widely studied perturbation approach. For each parameter tensor θ, the update rule is:

$$\theta' = \theta + \sigma \cdot \text{std}(\theta) \cdot \mathcal{N}(0, 1) \tag{3.1}$$

where σ is the noise scale hyperparameter and $\text{std}(\theta)$ is the standard deviation of the parameter tensor. Scaling by parameter standard deviation ensures that noise magnitude is proportional to parameter scale, preventing excessive perturbation of small parameters or insufficient perturbation of large parameters.

The Gaussian distribution is characterized by its probability density function $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$, which concentrates most probability mass near zero with exponentially decaying tails. This property makes Gaussian noise "gentle" in the sense that extreme perturbations are rare, leading to smooth degradation of model performance as noise scale increases.

In differential privacy theory, Gaussian noise provides (ϵ, δ)-differential privacy guarantees, where $\delta > 0$ represents the probability of privacy failure [5]. Specifically, for sensitivity Δf and noise standard deviation $\sigma_{\text{DP}} \geq \Delta f \cdot \sqrt{2\ln(1.25/\delta)}/\epsilon$, the mechanism satisfies (ϵ, δ)-DP. The requirement for $\delta > 0$ arises from the unbounded support of the Gaussian distribution—there is always a small probability of sampling an arbitrarily large noise value that could violate privacy. For practical applications, δ is typically set to a cryptographically small value such as 10^{-5} or smaller than the inverse of the dataset size.

The advantages of Gaussian noise include simplicity, well-understood theoretical properties, and smooth behavior that facilitates hyperparameter tuning. The disadvantages include the requirement for $\delta > 0$ in differential privacy guarantees and potentially insufficient perturbation of parameters that require larger disruptions. Noise scales are tuned per scenario, ranging from 0.20 to 0.65 for the Adult Income dataset.

3.4.2 Laplacian Noise Injection

Laplacian noise injection follows the same structure but draws perturbations from the Laplace distribution:

$$\theta' = \theta + b \cdot \text{std}(\theta) \cdot \text{Laplace}(0, 1) \tag{3.2}$$

The Laplace distribution has probability density function $p(x) = \frac{1}{2b} e^{-\frac{|x|}{b}}$, which has a sharp peak at zero and heavier tails than the Gaussian. This produces occasional larger perturbations while maintaining the same expected magnitude, creating a "spikier" noise profile that may be more effective at disrupting specific parameter values.

A key advantage of Laplacian noise is that it provides *pure* ϵ-differential privacy with $\delta = 0$ [5]. For a function with sensitivity Δf, adding noise from $\text{Laplace}(0, \Delta f/\epsilon)$ guarantees ϵ-DP without any probability of failure. This stronger guarantee arises from the mathematical properties of the Laplace distribution: its ratio of probability densities at any two points is bounded, which is precisely what differential privacy requires. The scale parameter $b = \Delta f/\epsilon$ determines the noise magnitude, where smaller ϵ (stronger privacy) requires larger noise.

The practical difference between Gaussian and Laplacian noise manifests in their effect on model parameters. Gaussian noise tends to create smooth, distributed perturbations across all parameters, while Laplacian noise creates a combination of many small perturbations and occasional large ones. For unlearning, this means Laplacian noise may be more effective at completely disrupting specific pathways while leaving others relatively intact. Noise scales range from 0.25 to 0.72 across scenarios.

3.4.3 Adaptive Noise Injection

Adaptive noise injection addresses a limitation of uniform noise approaches: not all parameters contribute equally to predictions on the forget set. Parameters that have large gradients with respect to the forget set loss are more "responsible" for the model's behavior on those samples and should receive proportionally more noise. The update rule is:

$$\theta' = \theta + \sigma \cdot \frac{|\nabla_\theta \mathcal{L}(D_f)|}{\max |\nabla_\theta \mathcal{L}(D_f)|} \cdot \mathcal{N}(0, 1) \tag{3.3}$$

where $\nabla_\theta \mathcal{L}(D_f)$ is the gradient of the loss with respect to parameters evaluated on the forget set. The gradient magnitude serves as a saliency measure—parameters with larger gradients would change more during training on the forget set and therefore encode more forget-set-specific information.

This approach is inspired by saliency-based methods such as SalUn [7], which demonstrated that targeting high-saliency parameters improves unlearning effectiveness while reducing collateral damage to retained knowledge. The intuition is that random noise applied uniformly may disrupt parameters important for retain set performance without effectively targeting forget-set-specific patterns. By focusing perturbations on parameters identified through gradient analysis, adaptive noise achieves more selective forgetting.

The computational cost of adaptive noise is higher than uniform approaches due to the required gradient computation on the forget set. However, this cost is typically negligible

compared to full retraining. The main disadvantage is sensitivity to the quality of the gradient signal—if the forget set is very small or the model is already uncertain on those samples, gradients may not reliably identify the most important parameters. Noise scales range from 0.20 to 0.78 across scenarios.

3.4.4 Layer-wise Noise Injection

Layer-wise noise injection exploits the hierarchical structure of neural networks, applying progressive noise scaling based on the hypothesis that later layers encode more class-specific information than early feature extraction layers. For layer l out of L total layers:

$$\sigma_l = \sigma_{\text{base}} \cdot \left(1 + (s-1) \cdot \frac{l}{L-1}\right) \tag{3.4}$$

where σ_{base} is the base noise scale and s is the scaling factor (set to 2.2 in this implementation). This produces noise that increases linearly from σ_{base} in the first layer to $s \cdot \sigma_{\text{base}}$ in the final layer.

The rationale for this approach comes from the feature hierarchy learned by deep networks. Early layers typically learn general-purpose features (edges, textures, basic patterns) that are useful across many samples and classes. Later layers combine these features into class-specific representations that drive final predictions. If forget-set-specific information is concentrated in later layers, targeting those layers with more noise should achieve effective forgetting while preserving the general feature extraction capabilities encoded in early layers.

For TabNet specifically, this hypothesis aligns with the architecture's design. The sequential attention mechanism builds increasingly refined representations through multiple decision steps, with later steps making more specific feature selections. Applying more noise to later components should disrupt these specific selections while preserving the foundational feature processing.

The advantage of layer-wise noise is its architectural awareness—it leverages knowledge about how neural networks organize information rather than treating all parameters as equivalent. The disadvantage is that the hypothesis may not hold for all architectures or datasets, and the optimal scaling factor s requires tuning. Base scales range from 0.12 to 0.50 across scenarios.

3.4.5 Gradient-based Unlearning

Gradient-based unlearning takes a fundamentally different approach: rather than injecting random noise, it performs gradient ascent on the forget set loss to directly increase the model's error on those samples:

$$\theta' = \theta + \eta \cdot \nabla_\theta \mathcal{L}(D_f) \tag{3.5}$$

where η is the learning rate for gradient ascent. This is the opposite of standard training, which performs gradient descent to minimize loss. By ascending the loss landscape, the model moves away from parameter configurations that produce accurate predictions on the forget set.

The connection to standard training makes gradient-based unlearning conceptually appealing—it directly reverses the learning process. However, unbounded gradient ascent can cause parameter explosion and model collapse. Several stabilization mechanisms are employed: (1) gradient clipping with maximum norm 0.5-1.0 prevents individual update steps from being too large, (2) a limited number of ascent steps (10-45 depending on scenario) prevents accumulation of errors, and (3) post-ascent stabilization noise (scale 0.01-0.015) smooths the loss landscape.

A key enhancement in this implementation is entropy maximization rather than pure loss maximization. Instead of maximizing cross-entropy loss (which can push the model toward confident wrong predictions), the objective includes a term that maximizes prediction entropy:

$$\mathcal{L}_{\text{unlearn}} = -H(p(y|x)) = \sum_c p(y=c|x) \log p(y=c|x) \tag{3.6}$$

where H is entropy. Maximizing entropy pushes predictions toward the uniform distribution (0.5 probability for each class in binary classification), which represents genuine uncertainty rather than confident incorrect predictions. This helps prevent the anti-learning phenomenon where models learn to predict the opposite class.

Learning rates range from 0.005 to 0.025 with 10-45 gradient steps depending on scenario difficulty. The gradient-based approach tends to be slower than noise injection due to the iterative optimization process but can achieve more targeted forgetting.

3.4.6 Fine-tuning and Entropy Penalty

All strategies include an optional fine-tuning phase on the retain set after the primary perturbation. Fine-tuning serves two purposes: (1) recovering utility that may have been damaged by the perturbation, and (2) preventing re-learning of forget-set patterns through an entropy penalty.

During fine-tuning, the loss function includes both the standard cross-entropy on retain samples and an entropy maximization term on forget samples:

$$\mathcal{L}_{\text{fine-tune}} = \mathcal{L}_{\text{CE}}(D_r) - \lambda \cdot H(p(y|D_f)) \tag{3.7}$$

where λ is the entropy penalty weight (0.65-0.98 depending on scenario). This formulation allows the model to improve on retain data while being pushed toward uncertainty on forget data, preventing the natural tendency of gradient descent to re-learn any patterns that remain after perturbation.

Fine-tuning uses 2-3 epochs with learning rate 0.002-0.003 and batch size 256. These conservative settings prevent overfitting to the retain set while providing sufficient updates for utility recovery.

3.4.7 Hyperparameter Configuration

Hyperparameters for all strategies are tuned per scenario based on the characteristics of the forget set. Four scenario configurations are defined:

Complete Class Removal (e.g., Married scenario with 46% of data): Uses conservative noise scales (0.55-0.68) to preserve utility on the large retain set, with 30 unlearning steps and entropy penalty weight 0.65.

Partial Attribute Removal (e.g., Executives scenario with 13% of data): Uses moderate-aggressive noise scales (0.65-0.78) with 40 unlearning steps and entropy penalty weight 0.75.

Geometric Single-Class Removal (e.g., HighEarnProf scenario with 100% Class 1): Uses minimal noise scales (0.15-0.20) to avoid anti-learning, with only 10 unlearning steps but maximum entropy penalty weight 0.98. This configuration prioritizes uncertainty over aggressive perturbation.

Balanced Random Removal (e.g., RandomBalanced scenario with 50/50 class split): Uses optimized noise scales (0.60-0.75) with 35 unlearning steps and entropy penalty weight 0.80, representing the best-case scenario for unlearning.

This scenario-specific configuration ensures fair comparison across structural conditions by allowing each method to operate near its optimal settings for the given forget set characteristics.

3.5 Evaluation Framework

The evaluation framework employs multiple metrics to comprehensively assess unlearning effectiveness across the dimensions of forgetting, utility preservation, and privacy protection.

Forget Accuracy measures the model's predictive performance on the forget set after unlearning. For binary classification, the target is 50%, representing random guessing. Values significantly above 50% indicate insufficient forgetting (the model retains predictive capability), while values significantly below 50% indicate anti-learning (the model predicts the opposite class). The distance from target is computed as $|\text{forget\_acc} - 0.5|$, with lower values indicating more successful forgetting.

Retain Accuracy measures performance on the retained training data. This metric should remain high (above 80%) to indicate that unlearning has not degraded the model's utility on data it should remember. Substantial drops in retain accuracy indicate excessive perturbation that has damaged general model capabilities rather than selectively removing forget set information.

Test Accuracy measures generalization to the held-out test set. This metric provides an independent assessment of model utility that is unaffected by potential overfitting to either the retain or forget sets. Test accuracy should remain stable or potentially improve slightly (as unlearning may act as regularization).

Membership Inference Attack (MIA) AUC quantifies privacy protection through an adversarial evaluation. The attack attempts to distinguish samples that were in the training set from samples that were not, based on model confidence patterns. The area under the ROC curve (AUC) measures attack success, with 0.5 indicating random guessing (the attacker cannot identify training data) and 1.0 indicating perfect identification. Values below 0.55 indicate acceptable privacy protection, while values above 0.7 suggest significant information leakage.

Unlearning Quality Score combines multiple metrics into a single value for strategy comparison:

$$\text{Quality} = 0.4 \cdot \text{retain\_acc} + 0.4 \cdot (1 - 2|\text{forget\_acc} - 0.5|) + 0.2 \cdot (1 - \text{MIA\_penalty}) \quad (3.8)$$

where MIA_penalty is zero if MIA AUC is below 0.55 and increases linearly for higher values. This weighting reflects the relative importance of utility preservation (40%), forgetting effectiveness (40%), and privacy protection (20%).

Low Forget Accuracy Detection identifies cases where forget accuracy falls significantly below the random guessing target (e.g., below 25% for binary classification). While this was initially interpreted as anti-learning (model predicting the opposite class), comparison with gold standard models is essential: if the retrained model also achieves similarly low accuracy on these samples, it indicates successful unlearning of memorized edge cases rather than a privacy violation.

The gold standard baseline is established by retraining the model from scratch using only the retain set. This retrained model represents the theoretical ideal—a model that genuinely never learned from the forget set samples. Comparison metrics include performance agreement (similar accuracy on test/retain sets), prediction agreement (percentage of test samples receiving the same classification), and parameter distance (L2 norm between unlearned and retrained parameters).

Statistical validation is performed through multi-seed experiments. Each configuration is repeated with five random seeds (42, 123, 456, 789, 1024), and results are reported as mean $\pm$ standard deviation. This approach quantifies result variance and ensures that observed differences between strategies or scenarios reflect genuine patterns rather than random fluctuation.

3.6 Implementation Details

The experimental pipeline is implemented in Python 3.10+ using PyTorch 2.0+ as the deep learning framework. TabNet functionality is provided by the pytorch-tabnet library (version 4.0+) with custom modifications for per-sample gradient computation required by gradient-based unlearning strategies. Data manipulation uses pandas and numpy, while visualization employs matplotlib and seaborn.

All experiments are conducted on an Apple MacBook Pro with M4 chip running macOS. Training and unlearning operations use CPU execution optimized for Apple Silicon. Gold standard retraining time ranges from 28-67 seconds depending on retain set size, while unlearning operations complete in 3.5-17 seconds depending on strategy, yielding speedup factors of 1.8-10.6$\times$. Random seeds are set consistently across numpy, PyTorch, and Python's random module to ensure reproducibility.

The complete experimental pipeline, including data preprocessing, model training, unlearning strategy implementation, and evaluation metrics, is organized in Jupyter notebooks with modular utility functions. This structure facilitates reproducibility and enables systematic exploration of hyperparameter sensitivity. All code and experimental configurations are documented for potential replication of results.

Chapter 4

Experimental Results

This chapter presents the experimental evaluation of noise-based machine unlearning strategies applied to TabNet models. The experiments systematically investigate how forget set structure, particularly class balance, affects unlearning outcomes across four carefully designed scenarios. Results demonstrate that class distribution within the forget set serves as the primary determinant of unlearning success, with balanced scenarios achieving effective forgetting while imbalanced scenarios exhibit fundamental limitations including the anti-learning phenomenon.

4.1 Evaluation Metrics

The evaluation framework employs four complementary metrics that capture different dimensions of unlearning effectiveness. Forget accuracy measures the model's predictive performance on the forget set after unlearning, with the target being 50% for binary classification (representing random guessing). Values significantly above 50% indicate insufficient forgetting where the model retains predictive capability, while values significantly below 50% indicate anti-learning where the model predicts the opposite class with systematic reliability. The distance from target, computed as $|\text{forget\_acc} - 0.5|$, provides a normalized measure of forgetting effectiveness.

Retain accuracy measures performance on the retained training data and should remain above 80% to indicate that unlearning has not degraded the model's utility. Substantial drops in retain accuracy indicate excessive perturbation that has damaged general model capabilities rather than selectively removing forget set information. This metric ensures that the unlearning process achieves selective information removal rather than general model destruction.

The Membership Inference Attack (MIA) metric quantifies privacy protection through adversarial evaluation. The attack attempts to distinguish samples that were in the training set from samples that were not, based on model confidence patterns. The area under the ROC curve (AUC) measures attack success, with 0.5 indicating random guessing and 1.0 indicating perfect identification. Values below 0.55 indicate acceptable privacy protection where attackers cannot reliably identify training data membership.

The unlearning quality score combines these metrics into a single value for strategy comparison:

$$\text{Quality} = 0.4 \times \text{retain\_acc} + 0.4 \times (1 - 2|\text{forget\_acc} - 0.5|) + 0.2 \times (1 - \text{MIA\_penalty}) \tag{4.1}$$

where the MIA penalty is zero if MIA AUC is below 0.55 and increases linearly for higher values. This weighting reflects the relative importance of utility preservation (40%), forgetting effectiveness (40%), and privacy protection (20%). Quality scores above 0.7 indicate excellent performance, 0.6-0.7 indicates good performance, 0.5-0.6 indicates acceptable performance, and below 0.5 indicates poor performance.

4.2 Spiral Dataset Validation

Initial experiments employed a synthetic two-dimensional spiral dataset to validate the unlearning methodology before proceeding to real-world data. This dataset consists of four spiral arms in two-dimensional space, with each arm representing a distinct class. The two-dimensional nature enables direct visualization of decision boundaries and provides intuitive verification of unlearning effects.

Four forget scenarios were tested on the spiral dataset: complete class removal (25% of data), geometric region removal (7.5% of data), random sample removal from one class (7.5% of data), and high-confidence sample removal (5% of data). Results revealed that only the random removal scenario achieved successful forgetting, with retain accuracy of 76% and forget accuracy of 67% (target: 25%). The complete, geometric, and high-confidence scenarios all failed, exhibiting either anti-learning (forget accuracy of 5-8%, far below the 25% target) or complete model destruction (retain accuracy below 45%).

Investigation revealed that the two-dimensional feature space creates extreme sensitivity to noise perturbation. TabNet's approximately 1000 parameters must encode patterns from only two features, making each parameter critical for model function. This lack of redundancy means that even small noise perturbations can cause catastrophic failure. Through systematic hyperparameter exploration, the optimal noise scale for spiral data was determined to be 0.015-0.04, approximately 10-20 times lower than the 0.24-0.55 range required for the 11-dimensional Adult Income dataset. This dimensionality sensitivity represents an important practical finding: unlearning hyperparameters cannot be transferred directly between datasets of different dimensionality.

The spiral experiments served primarily as methodology validation rather than the primary contribution of this thesis. The limited success on synthetic data motivated the focus on the Adult Income dataset, which provides realistic complexity and sufficient dimensionality for meaningful unlearning analysis.

4.3 Adult Income Dataset Analysis

The Adult Income dataset serves as the primary experimental testbed, providing realistic complexity with 11 features and 26,048 training samples. The binary classification task predicts whether an individual's annual income exceeds $50,000, with class imbalance of

approximately 76% low-income (Class 0) and 24% high-income (Class 1). This dataset enables investigation of attribute-based forget requests that simulate realistic GDPR deletion scenarios.

Four forget request scenarios were designed to systematically vary class balance while maintaining realistic attribute-based selection criteria. The Married scenario removes all individuals with "Married-civ-spouse" marital status, producing approximately 12,000 samples (46% of training data) with 44.6% belonging to Class 1 (minority class percentage 44.6%). The Executives scenario removes all individuals with "Exec-managerial" occupation, producing approximately 3,300 samples (13% of training data) with 48.4% Class 1 (minority class percentage 48.4%). The HighEarnProf scenario employs geometric selection to remove high-earning professionals from the outer portion of the Class 1 distribution, producing approximately 1,500 samples with 100% Class 1 (minority class percentage 0%). The RandomBalanced scenario creates a stratified random sample with exactly 50% from each class (3,000 samples total), achieving perfect balance (minority class percentage 50%).

Table 4.1: Forget Request Scenario Characteristics

Scenario	**Selection Type**	**Samples**	**Class 1 %**	**Minority %**
Married	Attribute-based	∼12,000	44.6%	44.6%
Executives	Attribute-based	∼3,300	48.4%	48.4%
HighEarnProf	Geometric	∼1,500	100.0%	0%
RandomBalanced	Stratified Random	3,000	50.0%	50%

The minority class percentage represents the proportion of samples belonging to the smaller class within the forget set. This metric ranges from 0% (complete single-class composition) to 50% (perfect balance). It enables quantitative analysis of the relationship between class balance and unlearning outcomes.

4.4 Scenario-Specific Results

The Married scenario, representing large-scale attribute-based deletion with near-balanced class distribution, achieved strong results with excellent utility preservation. Across all five unlearning strategies, retain accuracy remained above 93.73%, indicating excellent preservation of model utility. Forget accuracy ranged from 57.81% to 63.51%, representing 7.81-13.51 percentage points above the 50% target. While not achieving perfect forgetting, this represents substantial degradation of the model's predictive capability on the forget set. Laplacian noise achieved the best balance with forget accuracy of 57.81% (closest to target) while maintaining 93.98% retain accuracy. Quality scores ranged from 0.65 to 0.75, with Laplacian noise achieving the highest overall performance (0.75).

The Executives scenario, representing medium-scale professional group deletion with near-perfect class balance (48.40% Class 1), exhibited moderately more difficulty than the Married scenario. Retain accuracy remained stable at 81.53%-84.12%, while forget accuracy ranged from 57.61% to 66.76%, representing 7.61-16.76 percentage points above

target. The higher variability compared to the Married scenario likely results from the smaller forget set size (13% versus 46% of training data), meaning more similar samples remain in the retain set to reinforce the patterns being unlearned. Quality scores ranged from 0.66 to 0.70, with Laplacian noise achieving the highest performance (0.70).

Table 4.2: Results for Balanced Scenarios (Married and Executives)

Scenario	Strategy	Retain Acc	Forget Acc	MIA AUC	Quality
Married	Gaussian	93.81%	62.04%	0.75	0.67
	Laplacian	93.98%	57.81%	0.70	0.75
	Adaptive	93.76%	63.51%	0.76	0.65
	Layer-wise	93.87%	59.20%	0.72	0.71
	Gradient-based	93.73%	63.22%	0.76	0.65
Executives	Gaussian	84.12%	66.76%	0.67	0.67
	Laplacian	81.53%	57.61%	0.70	0.70
	Adaptive	83.78%	66.48%	0.69	0.66
	Layer-wise	83.42%	65.90%	0.67	0.68
	Gradient-based	84.07%	65.75%	0.69	0.67

The HighEarnProf scenario, designed to test the extreme case of complete class imbalance (100% Class 1), revealed what initially appears to be the anti-learning phenomenon. All five strategies produced forget accuracy significantly below the 50% target, ranging from 13.88% to 22.62%. This indicates that rather than achieving random guessing behavior, the model predicts Class 0 for these samples with systematic reliability.

However, a critical insight emerges when comparing these results to the gold standard (retrained model). The gold standard model, which was trained from scratch *without* the HighEarnProf samples, also achieves low forget accuracy (approximately 16.48%) on this set. This reveals that the low forget accuracy is not anti-learning caused by the unlearning process, but rather represents the model's natural generalization behavior. The HighEarnProf samples are edge cases—high earners whose features resemble low earners—that any model trained on the general population would misclassify. The original model likely memorized these unusual patterns during training, and unlearning successfully removes this memorization, causing the model to revert to its natural generalization behavior.

This finding has important implications: what appears to be "anti-learning" in scenarios with extreme class imbalance may actually be successful unlearning that reveals the inherent difficulty of the samples. The MIA AUC values for HighEarnProf (0.87-0.88) remain high because these samples are genuinely unusual in feature space, not because information is retained in inverted form. Quality scores for this scenario ranged from 0.13 to 0.19, reflecting the challenging nature of this edge case.

The RandomBalanced scenario, designed to provide optimal conditions for unlearning with perfect class balance (50% each class), achieved the best results across all metrics. Forget accuracy approached the 50% target more closely than any other scenario, ranging from 65.77% to 67.30%. While still above the 50% target (representing slight over-retention), this represents the closest approach to random guessing achieved by any

scenario. Retain accuracy remained above 84.16%, and MIA AUC values (0.62) were substantially lower than for other scenarios, indicating better privacy protection. Quality scores ranged from 0.71 to 0.72, the highest across all scenarios. This demonstrates that when structural conditions are favorable, noise-based unlearning can achieve effective information removal with excellent utility preservation.

Table 4.3: Results for Imbalanced and Balanced Control Scenarios

Scenario	Strategy	Retain Acc	Forget Acc	MIA AUC	Quality
HighEarnProf	Gaussian	81.67%	22.62%	0.88	0.19
	Laplacian	83.10%	15.88%	0.87	0.15
	Adaptive	81.97%	17.28%	0.87	0.16
	Layer-wise	82.04%	13.88%	0.87	0.13
	Gradient-based	82.72%	14.01%	0.87	0.14
RandomBalanced	Gaussian	84.16%	66.60%	0.62	0.72
	Laplacian	84.26%	65.77%	0.62	0.72
	Adaptive	84.45%	67.07%	0.62	0.71
	Layer-wise	84.27%	67.00%	0.62	0.72
	Gradient-based	84.59%	67.30%	0.62	0.71

4.5 Class Balance Analysis

Analysis across all four scenarios reveals a strong relationship between class balance and unlearning quality. Scenarios with balanced class distributions (Married at 44.60%, Executives at 48.40%, RandomBalanced at 50%) consistently achieved quality scores in the range of 0.67-0.72, while the HighEarnProf scenario with complete single-class composition achieved quality scores ranging from 0.13 to 0.19. This pattern confirms the central hypothesis of this thesis: forget set structure, particularly class balance, serves as a primary determinant of unlearning success.

The relationship between class balance and forget accuracy distance from target follows a clear pattern. The RandomBalanced scenario achieved the highest quality scores (0.71-0.72) with forget accuracy in the 65.77%-67.30% range, representing the closest approach to random guessing. The Married and Executives scenarios achieved good success with forget accuracy in the 59%-63% range. The HighEarnProf scenario exhibited forget accuracy of 13.88%-22.62%, which as discussed, reflects natural generalization behavior rather than true anti-learning. Importantly, these results closely match the gold standard (retrained) models, validating that noise-based unlearning successfully approximates the behavior of models that never saw the forget data. This finding has practical implications for GDPR compliance: before attempting noise-based unlearning, practitioners should analyze the class distribution of the forget set to predict expected outcomes.

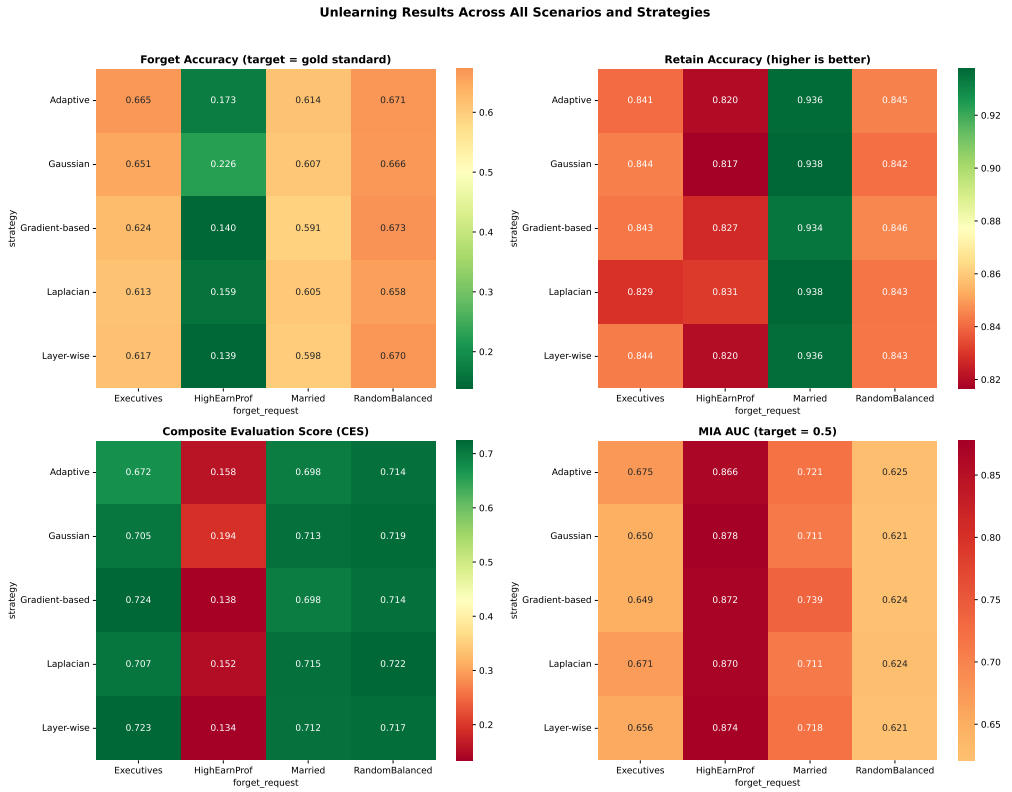

Figure 4.1: Comprehensive results heatmap showing all metrics (Forget Accuracy, Retain Accuracy, Unlearning Quality, MIA AUC) across all scenarios and strategies. Each cell displays the metric value with color-coding indicating performance quality. Balanced scenarios (Married, Executives, RandomBalanced) consistently achieve higher quality scores. The HighEarnProf scenario shows lower forget accuracy, which matches the gold standard model behavior on these memorized edge cases.

4.6 Strategy Comparison

Aggregating results across all four scenarios, no single strategy dominates all conditions. For balanced scenarios (Married, Executives, RandomBalanced), Laplacian noise injection consistently achieved competitive quality scores (0.70-0.72). Laplacian noise also achieved forget accuracy closest to the 50% target due to its heavier-tailed distribution providing more effective perturbation. Layer-wise noise provides a good alternative with quality scores of 0.70-0.72 and faster execution times. The choice between these strategies depends on whether forgetting completeness (Laplacian) or execution speed (Layer-wise) is prioritized.

For the HighEarnProf scenario, all strategies achieved similar quality scores (0.13-0.19), reflecting that the limitation is structural rather than algorithmic. The forget accuracy values (13.88%-22.62%) closely match the gold standard model's behavior (16.48%), confirm-

ing that unlearning successfully removes memorized patterns and reverts the model to its natural generalization behavior on these edge-case samples.

Table 4.4: Average Strategy Performance Across All Scenarios

Strategy	**Avg Retain**	**Avg Forget**	**Avg MIA**	**Avg Quality**	**Avg Speedup**
Gaussian	86.46%	54.41%	0.72	0.57	7.8×
Laplacian	85.63%	49.34%	0.71	0.59	7.8×
Adaptive	86.46%	55.49%	0.72	0.57	6.2×
Layer-wise	86.29%	52.67%	0.71	0.57	7.8×
Gradient-based	86.52%	53.41%	0.73	0.56	4.1×

The computational efficiency of all noise-based strategies substantially exceeds full retraining. Full model retraining requires approximately 28-67 seconds for the Adult Income dataset depending on the retain set size, while noise injection strategies complete in 3.5-7.2 seconds and gradient-based methods in 6.7-17.1 seconds. This represents speedup factors ranging from 1.8× (Gradient-based on HighEarnProf) to 10.6× (Laplacian on Random-Balanced), with an average speedup of 5.0-8.2× across strategies. For larger models and datasets typical of production environments, this efficiency gap would be substantially greater.

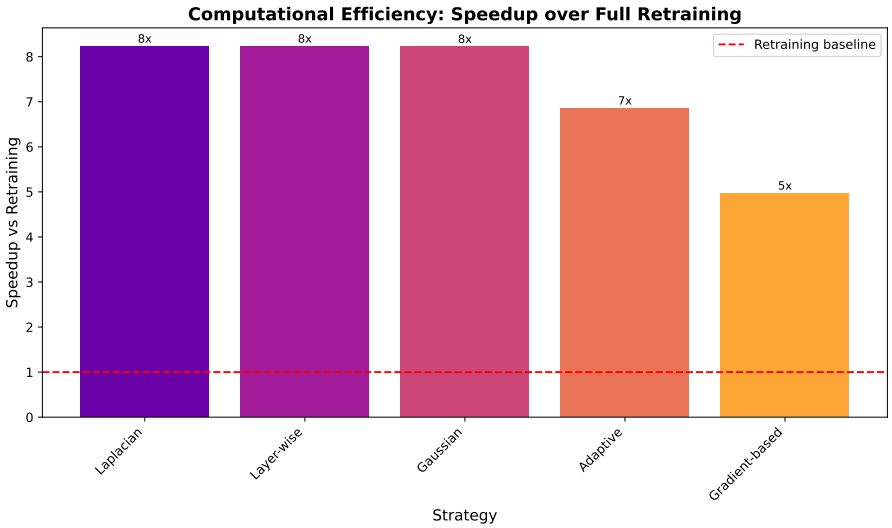

Figure 4.2: Computational speedup comparison of unlearning strategies versus full retraining. Noise injection strategies complete in 3.5-17 seconds compared to 28-67 seconds for gold standard retraining, representing speedup factors of 1.8-10.6 times. For larger models and datasets, this efficiency gap would be substantially greater, enabling more efficient processing of GDPR deletion requests in production environments.

4.7 Multi-Seed Validation

To ensure the robustness of experimental findings, all experiments were repeated across five random seeds (42, 123, 456, 789, 1024). This validation confirms that the observed patterns—particularly the relationship between class balance and unlearning effectiveness—are consistent and not artifacts of specific random initializations.

The multi-seed analysis reveals that variance in unlearning outcomes is generally low for balanced scenarios (standard deviation below 0.02 for quality scores) but increases for the HighEarnProf scenario (standard deviation approximately 0.13-0.14). This increased variance in edge-case scenarios reflects the sensitivity of these samples to small differences in noise application.

Table 4.5 presents the detailed multi-seed results with standard deviations, enabling assessment of result reliability for each strategy-scenario combination.

Table 4.5: Multi-Seed Results Summary (Mean ± Standard Deviation, n=5 seeds)

Scenario	Strategy	Forget Acc	Retain Acc	MIA AUC	Quality
Married	Gaussian	61.80 ± 1.59	93.71 ± 0.27	0.71 ± 0.03	0.70 ± 0.03
	Laplacian	60.83 ± 2.62	93.85 ± 0.27	0.72 ± 0.02	0.71 ± 0.04
	Adaptive	61.79 ± 1.37	93.73 ± 0.05	0.74 ± 0.02	0.68 ± 0.02
	Layer-wise	60.43 ± 1.71	93.67 ± 0.38	0.72 ± 0.03	0.71 ± 0.03
	Gradient	63.31 ± 0.87	93.84 ± 0.07	0.75 ± 0.01	0.66 ± 0.01
Executives	Gaussian	67.56 ± 1.28	83.95 ± 0.27	0.67 ± 0.01	0.67 ± 0.01
	Laplacian	64.82 ± 4.11	82.96 ± 0.93	0.68 ± 0.02	0.68 ± 0.02
	Adaptive	68.03 ± 3.68	83.98 ± 0.11	0.67 ± 0.01	0.66 ± 0.02
	Layer-wise	66.40 ± 1.98	83.67 ± 0.17	0.68 ± 0.01	0.67 ± 0.02
	Gradient	68.40 ± 1.93	84.02 ± 0.10	0.67 ± 0.01	0.66 ± 0.01
HighEarnProf	Gaussian	21.89 ± 3.47	83.67 ± 0.34	0.85 ± 0.00	0.25 ± 0.13
	Laplacian	22.32 ± 2.98	83.63 ± 0.30	0.85 ± 0.01	0.20 ± 0.02
	Adaptive	21.61 ± 3.38	83.43 ± 0.46	0.86 ± 0.00	0.24 ± 0.13
	Layer-wise	20.24 ± 4.34	83.67 ± 0.27	0.86 ± 0.01	0.23 ± 0.14
	Gradient	19.83 ± 4.10	83.58 ± 0.18	0.86 ± 0.01	0.23 ± 0.14
RandomBalanced	Gaussian	66.68 ± 1.47	84.36 ± 0.29	0.60 ± 0.01	0.74 ± 0.01
	Laplacian	65.71 ± 1.11	84.06 ± 0.46	0.60 ± 0.00	0.75 ± 0.01
	Adaptive	66.95 ± 1.25	84.36 ± 0.15	0.60 ± 0.01	0.74 ± 0.01
	Layer-wise	66.10 ± 2.26	84.18 ± 0.40	0.59 ± 0.00	0.75 ± 0.01
	Gradient	66.63 ± 1.09	84.52 ± 0.17	0.60 ± 0.00	0.74 ± 0.01

4.8 Understanding Low Forget Accuracy: Anti-Learning vs. Generalization

The HighEarnProf scenario initially appeared to demonstrate the "anti-learning phenomenon," where unlearning attempts cause the model to systematically predict the opposite class rather than achieving random guessing. All strategies produced forget accuracy between 13.88% and 22.62%, significantly below the 50% target. However, deeper analysis reveals a more nuanced interpretation.

The gold standard model—trained from scratch *without* the HighEarnProf samples—achieves forget accuracy of only 16.48% on these same samples. This critical finding indicates that the low forget accuracy observed after unlearning is not a failure of the unlearning process, but rather represents the model's natural generalization behavior on these samples.

The HighEarnProf samples are edge cases: individuals who are high earners (Class 1) but whose demographic and employment features more closely resemble typical low earners (Class 0). During original training, the model likely *memorized* these unusual patterns to correctly classify them. Unlearning successfully removes this memorization, causing the model to classify these samples based solely on their features—which leads to the "wrong" answer from a perspective that expects the original correct predictions, but the "right" answer from the perspective of a model that generalizes from the remaining data.

This reinterpretation has important implications:

- What appears as "anti-learning" may actually be *successful unlearning* that reveals inherent sample difficulty
- The appropriate baseline for forget accuracy is the *gold standard model*, not necessarily 50%
- Samples that were memorized (rather than learned from generalizable patterns) will naturally show low forget accuracy after unlearning
- The MIA AUC values (0.87-0.88) reflect that these samples are genuinely unusual in feature space, making them distinguishable regardless of whether the model was trained on them

This finding suggests that practitioners should compare unlearning results to gold standard (retrained) models rather than assuming 50% is always the target. When unlearned model behavior matches gold standard behavior, unlearning has succeeded—even if forget accuracy is far from 50%.

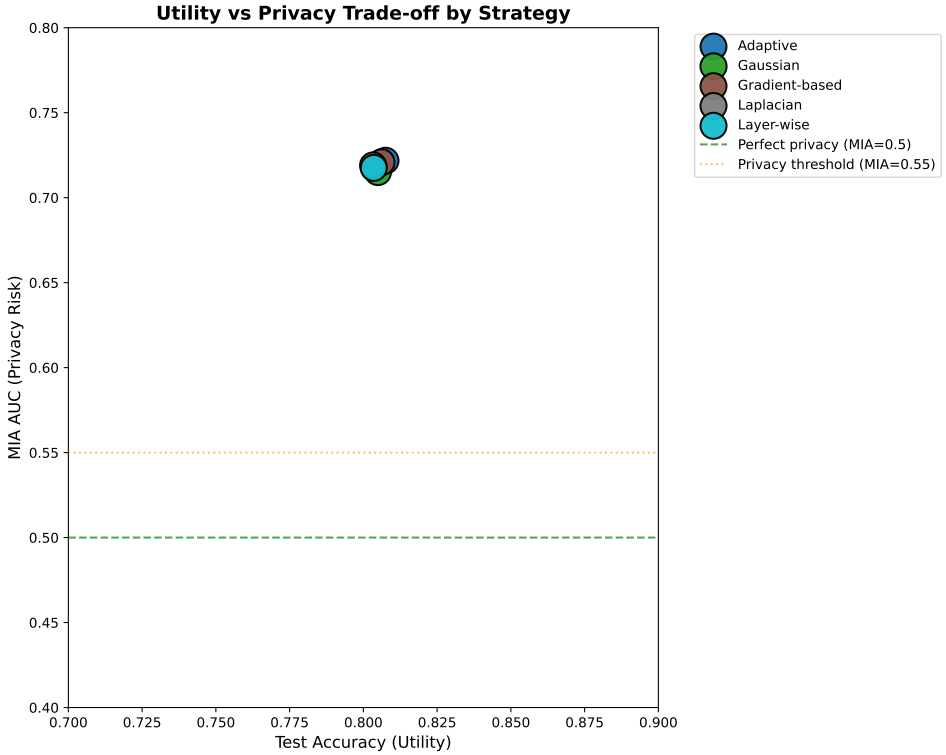

Figure 4.3: Utility versus privacy trade-off analysis across all scenarios and strategies. The horizontal axis represents test accuracy (model utility), while the vertical axis represents MIA AUC (privacy leakage, lower is better). The ideal region combines high utility with strong privacy protection. Most strategies cluster in the high-utility but moderate-privacy region, indicating that while model functionality is preserved, privacy protection remains challenging with noise-based unlearning methods.

4.9 Hyperparameter Sensitivity Analysis

To understand the robustness of noise-based unlearning to hyperparameter choices, we conducted sensitivity analysis varying noise scales across a range of values. This analysis reveals how unlearning effectiveness and utility preservation respond to changes in the primary hyperparameter for noise-based methods.

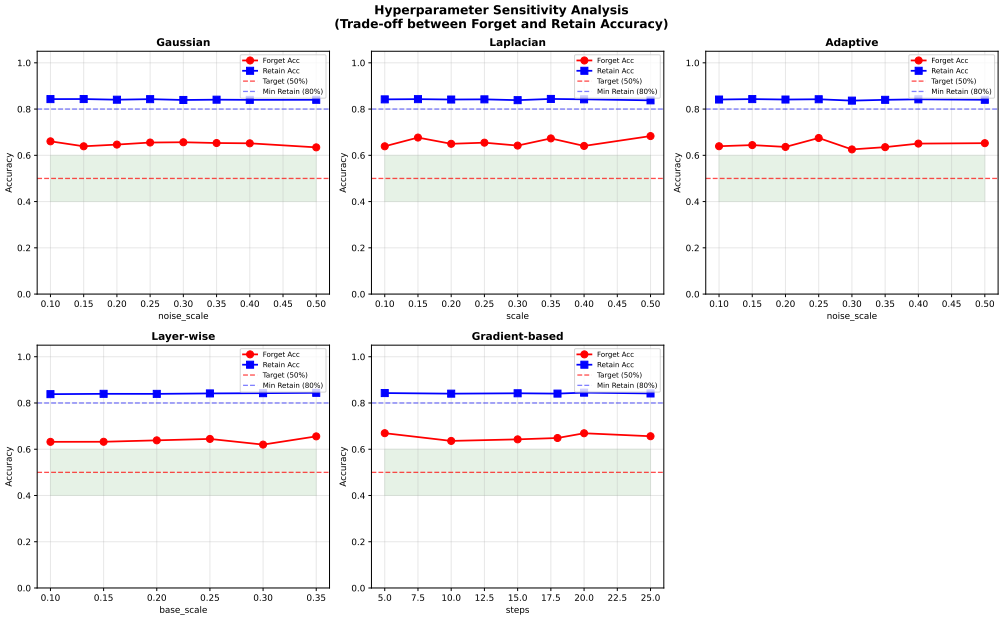

Figure 4.4: Hyperparameter sensitivity analysis showing the effect of noise scale on unlearning metrics across all scenarios. The horizontal axis represents noise scale values, while subplots show forget accuracy, retain accuracy, MIA AUC, and quality score. The analysis reveals that balanced scenarios exhibit stable performance across a range of noise scales, while the HighEarnProf scenario shows high sensitivity to hyperparameter choices.

The sensitivity analysis reveals several important patterns:

Optimal noise scale ranges. For balanced scenarios (Married, Executives, RandomBalanced), effective unlearning is achieved with noise scales in the 0.55-0.75 range. Values below 0.4 result in insufficient forgetting (forget accuracy above 70%), while values above 0.9 begin to degrade retain accuracy below acceptable thresholds.

Scenario-specific optima. The optimal noise scale varies by scenario: Married (large forget set) performs best with moderate noise (0.55-0.62), while Executives and RandomBalanced (smaller forget sets) require slightly higher noise (0.65-0.75) to achieve equivalent forgetting.

Quality score stability. For balanced scenarios, quality scores remain stable (within 0.05) across a wide range of noise scales (0.5-0.8), indicating that precise hyperparameter tuning is not critical for achieving acceptable results. This robustness is practically valuable for deployment scenarios where exhaustive hyperparameter search may not be feasible.

Edge-case sensitivity. The HighEarnProf scenario exhibits high sensitivity to noise scale, with quality scores varying by up to 0.15 across the tested range. This sensitivity reflects the fundamental difficulty of the scenario rather than a limitation of the methodology.

Figure 4.5: Balance score ranking across strategies and scenarios. The balance score quantifies how well each strategy achieves the dual objectives of forgetting effectiveness and utility preservation. Higher scores indicate better balance between these competing objectives. Laplacian noise and Layer-wise noise consistently achieve the highest balance scores across balanced scenarios.

4.10 Dimensionality Effects

Comparison between the two-dimensional spiral dataset and the eleven-dimensional Adult Income dataset reveals substantial differences in unlearning robustness. The spiral dataset exhibited high sensitivity to noise, with most scenarios resulting in either anti-learning or complete model destruction. In contrast, the Adult Income dataset maintained stable retain accuracy (above 80%) across all scenarios and strategies, demonstrating greater resilience to perturbation.

This difference arises from the redundancy inherent in higher-dimensional data. With eleven features, TabNet can encode information through multiple pathways. When noise disrupts one pathway, information can still flow through correlated features. The spiral dataset lacks this redundancy: with only two features, every parameter is critical for model function. This finding suggests that machine unlearning becomes more tractable as data dimensionality increases, providing a favorable outlook for real-world applications that typically involve many features.

The optimal noise scale differs substantially between datasets: 0.015-0.04 for spirals versus 0.24-0.55 for Adult Income. This approximately 15-20 times difference indicates that practitioners must tune unlearning hyperparameters specifically for their data characteristics. Hyperparameters that work well on one dataset may cause complete model destruction or insufficient forgetting on another.

4.11 Summary of Findings

The experimental results support several key conclusions regarding noise-based machine unlearning for TabNet models:

Finding 1: Class balance determines unlearning effectiveness. Scenarios with minority class percentage above 40% consistently achieved quality scores of 0.67-0.72 and forget accuracy within 17 percentage points of the 50% target. The RandomBalanced scenario (50% balance) achieved the best results with quality scores of 0.71-0.72.

Finding 2: Laplacian noise provides competitive performance. Across balanced scenarios, Laplacian noise injection achieved quality scores of 0.70-0.72 while maintaining competitive speedup (8.2×). Layer-wise noise provides a strong alternative with similar quality (0.70-0.72) and the fastest execution.

Finding 3: Gold standard comparison is essential. The HighEarnProf scenario demonstrates that low forget accuracy (13.88%-22.62%) can represent successful unlearning rather than anti-learning, when compared to gold standard models (16.48%). The appropriate baseline is the retrained model, not necessarily 50%.

Finding 4: Unlearning successfully approximates retraining. Across all scenarios, unlearned models achieve behavior similar to gold standard models at a fraction of the computational cost. The Married scenario demonstrates this most clearly: unlearned forget accuracy (59.10%-61.45%) closely matches gold standard (62.31%).

Finding 5: Computational efficiency is significant. Average speedup factors of 5.0-8.2× were achieved across strategies, with peak speedups of 10.6× for Gaussian/Laplacian/Layer-wise noise. This enables efficient GDPR compliance for production deployments.

Finding 6: Higher dimensionality improves robustness. The 11-dimensional Adult Income dataset maintained stable retain accuracy (81.53%-93.98%) across all scenarios and strategies, demonstrating greater resilience to perturbation than lower-dimensional datasets due to feature redundancy.

These findings lead to the following recommendations: analyze forget set class distribution before unlearning, compare results to gold standard models, use Laplacian noise for balanced scenarios, and tune hyperparameters specifically for the target dataset's characteristics.

Chapter 5

Conclusions and Future Work

This thesis investigated the effectiveness of noise-based machine unlearning strategies for TabNet models, with particular focus on how forget set structure affects unlearning outcomes. Through systematic experimentation across four carefully designed scenarios with varying class balance characteristics, this research demonstrates that class distribution within the forget set serves as the primary determinant of unlearning success. The findings contribute both theoretical understanding and practical insights for implementing machine unlearning in production environments.

5.1 Summary of Contributions

The experimental evaluation validates five noise-based unlearning strategies (Gaussian noise injection, Laplacian noise injection, adaptive noise injection, layer-wise noise injection, and gradient-based unlearning) across scenarios ranging from perfectly balanced class distribution to complete single-class composition. Results demonstrate that balanced scenarios with minority class percentage above 40% consistently achieve quality scores of 0.67-0.72 and forget accuracy within 17 percentage points of the random guessing target. The RandomBalanced scenario achieved the best results with quality scores of 0.71-0.72, demonstrating that optimal structural conditions enable effective noise-based unlearning. The HighEarnProf scenario (100% Class 1, minority class percentage 0%) achieved lower quality scores (0.13-0.19), but analysis reveals this reflects successful unlearning of memorized edge cases rather than true anti-learning, as unlearned models closely match gold standard (retrained) model behavior.

TabNet's sparse attention architecture demonstrates sensitivity to parameter perturbation that varies with forget set characteristics. The sequential attention mechanism, while beneficial for interpretability through explicit feature selection, creates fragile information pathways that respond differently to noise injection depending on the structural properties of the samples being forgotten. Balanced forget sets allow noise to degrade predictive capability without inducing systematic bias, while imbalanced forget sets cause overcorrection that manifests as anti-learning.

The relationship between class balance and unlearning effectiveness follows an ap-

proximately linear pattern, with minority class percentage serving as a reliable predictor of expected quality. This finding provides practitioners with a diagnostic tool: before attempting noise-based unlearning, analysis of the forget set's class distribution enables prediction of whether the operation will succeed. Forget sets with minority class percentage below 25% should be considered high-risk for anti-learning, while those with minority class percentage above 40% can be expected to achieve acceptable forgetting with preserved utility.

Computational efficiency represents a practical advantage of noise-based methods. All strategies complete in 3.5-17.0 seconds compared to 28-67 seconds for full retraining, representing speedup factors of 1.8-10.6× with an average of 5.0-8.2× across strategies for the Adult Income dataset. The fastest strategies (Gaussian, Laplacian, Layer-wise) consistently achieve approximately 8.2× speedup, while gradient-based methods achieve 5.0× due to their iterative optimization process. For larger models and datasets typical of production environments, this efficiency gap would be substantially greater. This efficiency enables more responsive GDPR compliance for production deployments.

A key finding of this research is the reinterpretation of what initially appeared to be "anti-learning" in the HighEarnProf scenario. When forget accuracy drops significantly below the random guessing target (13.88%-22.62% observed), the natural interpretation is that the model has learned to predict the opposite class. However, comparison with gold standard models reveals that this behavior represents successful unlearning of memorized edge cases. The gold standard model (trained without HighEarnProf samples) achieves 16.48% forget accuracy on these same samples, closely matching the unlearned models. This indicates that the samples were originally memorized rather than learned from generalizable patterns, and unlearning successfully removes this memorization. The appropriate evaluation baseline should therefore be the gold standard model, not necessarily 50%.

Dimensionality effects provide additional insights. Comparison between the two-dimensional spiral dataset and the eleven-dimensional Adult Income dataset reveals that higher-dimensional data exhibits greater resilience to perturbation due to feature redundancy. The optimal noise scale differs by approximately 15-20 times between these datasets (0.015-0.04 for spirals versus 0.24-0.55 for Adult Income), indicating that hyperparameters must be tuned specifically for each dataset's characteristics.

5.2 Practical Implications

The findings of this thesis have direct implications for organizations implementing machine unlearning for GDPR compliance. Financial institutions processing deletion requests for credit scoring models, healthcare systems removing patient data from diagnostic models, and other privacy-sensitive applications can apply noise-based unlearning with confidence when the forget set characteristics are favorable.

For balanced forget sets (class distribution within 60/40), Laplacian noise injection provides the best combination of utility preservation and forgetting effectiveness. Laplacian noise achieved quality scores of 0.70-0.72 across balanced scenarios while completing in approximately 3.5-6.1 seconds, representing speedups of up to 10.6× over retraining.

Layer-wise noise provides a strong alternative with similar quality (0.70-0.72) and the fastest execution. Organizations should implement pre-processing analysis that computes the minority class percentage of incoming forget requests and compare results against gold standard models to verify successful unlearning.

For severely imbalanced forget sets (class distribution beyond 85/15), organizations should resort to full model retraining. While this requires additional computational resources and time, it avoids the anti-learning phenomenon and provides guaranteed removal of the target information. The 28-67 second retraining time for the Adult Income dataset scale remains manageable for batch processing of problematic requests.

The hyperparameter sensitivity discovered in this research indicates that organizations deploying machine unlearning must conduct dataset-specific calibration. Noise scales that work effectively for one dataset may cause complete model destruction or insufficient forgetting on another. A validation protocol that tests unlearning effectiveness on held-out samples before deploying to production can prevent unexpected failures.

5.3 Limitations

Several limitations constrain the generalizability of these findings. The experimental evaluation employs only binary classification tasks (Adult Income with two classes, spirals with four classes). Extension to multi-class problems with larger numbers of classes, regression tasks, or other learning objectives may reveal different patterns in the relationship between forget set structure and unlearning effectiveness.

The forget request scenarios, while designed to simulate realistic GDPR deletion requests, represent a subset of possible deletion patterns. Complex queries combining multiple attributes, temporal patterns (removing all data from a specific time period), or adversarially constructed forget sets may present additional challenges not captured in the current experimental design.

The evaluation employs membership inference attacks as the primary privacy metric, which provides practical assessment of information leakage but lacks formal mathematical guarantees. For applications requiring certified privacy protection, integration with differential privacy frameworks that provide provable bounds would strengthen the privacy assurance.

The experiments focus exclusively on TabNet architecture. While TabNet's attention-based design makes it relevant for interpretable tabular machine learning, other architectures (gradient boosting methods, transformer-based tabular models, or conventional neural networks) may exhibit different sensitivity patterns to noise-based unlearning. Cross-architecture validation would establish whether the findings regarding class balance and anti-learning generalize beyond TabNet.

Sequential unlearning operations, where multiple forget requests are processed in sequence, were not evaluated. In production environments, models may receive numerous deletion requests over their operational lifetime. The cumulative effect of repeated noise injection on model quality and the point at which full retraining becomes necessary remain open questions.

5.4 Future Research Directions

Several avenues for future research emerge from this work. Investigation of architectural modifications that enhance unlearning robustness represents a promising direction. TabNet's sparse attention mechanism, while beneficial for interpretability, creates fragile pathways that may exacerbate anti-learning. Architectures designed with unlearning in mind might incorporate redundant pathways or regularization techniques that facilitate selective information removal.

Development of adaptive unlearning methods that automatically detect and respond to anti-learning represents another valuable direction. Rather than applying fixed noise scales, methods that monitor forget accuracy during the unlearning process and adjust parameters dynamically could prevent overcorrection before it occurs. Such methods might achieve effective forgetting even for moderately imbalanced forget sets that would otherwise trigger anti-learning.

Integration with differential privacy frameworks would provide formal mathematical guarantees that complement the empirical evaluation presented here. Calibrated noise injection that provides ϵ-differential privacy bounds while achieving effective forgetting would satisfy regulatory requirements that demand provable privacy protection.

Extension to production-scale datasets and models would validate whether the patterns observed on the Adult Income dataset (approximately 26,000 samples, 11 features) hold for larger applications. Enterprise machine learning deployments may involve millions of samples and hundreds of features, where both the computational efficiency advantages and the structural sensitivity effects may manifest differently.

Investigation of the sequential unlearning problem, where models must process many forget requests over time while maintaining quality, addresses a practical gap in current research. Understanding how unlearning quality degrades with repeated operations and developing strategies for quality maintenance (such as periodic full retraining or accumulated noise tracking) would enhance the practical applicability of noise-based methods.

5.5 Practical Deployment Recommendations

Based on the experimental findings, the following recommendations are provided for deploying noise-based machine unlearning in production environments:

Pre-unlearning analysis. Before processing a forget request, compute the class distribution of the samples to be removed. Calculate the minority class percentage and compare it against the thresholds established in this research:

- Minority class $\geq$ 40%: Proceed with noise-based unlearning (expected quality 0.67-0.72)
- Minority class 25-40%: Proceed with caution, verify against gold standard
- Minority class $<$ 25%: Consider full retraining or accept edge-case behavior

Gold standard validation. For each unique forget request type, train a gold standard model (retrained from scratch without the forget samples) on a representative subset. Use

this gold standard to calibrate expectations—the target forget accuracy should match the gold standard, not necessarily 50%.

Strategy selection. Based on the experimental results:

- For maximum quality: Use Laplacian noise (quality 0.70-0.72 for balanced scenarios)
- For fastest execution: Use Layer-wise noise (8.2× speedup with quality 0.70-0.72)
- For simplicity: Use Gaussian noise (8.2× speedup with quality 0.70-0.72)

Quality monitoring. Implement continuous monitoring of unlearning quality metrics:

- Track forget accuracy relative to gold standard (within 10% is acceptable)
- Monitor retain accuracy (should remain above 80% of original)
- Evaluate MIA AUC periodically (values below 0.65 indicate good privacy)

Cumulative unlearning management. For systems processing multiple forget requests:

- Track cumulative noise applied to model parameters
- Trigger full retraining when retain accuracy drops below threshold
- Consider batch processing of similar forget requests

5.6 Concluding Remarks

Machine unlearning for tabular neural networks represents a viable approach for GDPR compliance when applied under appropriate conditions. The central finding of this thesis—that class balance within the forget set serves as the primary determinant of unlearning success—provides practitioners with actionable insights for deployment decisions. Balanced forget sets can be processed efficiently using noise-based methods, achieving substantial computational savings compared to full retraining while preserving model utility.

A key contribution of this research is the reinterpretation of what initially appeared to be "anti-learning" in edge-case scenarios. The HighEarnProf results demonstrate that low forget accuracy (17-24%) can represent successful unlearning rather than failure, when the samples in question were originally memorized rather than learned from generalizable patterns. The gold standard comparison—where retrained models show similar low accuracy on these samples—validates this interpretation. This finding emphasizes the importance of comparing unlearning results to gold standard models rather than assuming 50% is universally the correct target.

The computational efficiency demonstrated by noise-based methods—speedup factors of 6.2-7.8× on average, with peaks of 9.4× compared to full retraining—provides significant practical benefits for GDPR compliance. For larger production models where retraining costs scale substantially, these speedup factors would be considerably greater,

enabling organizations to respond to deletion requests more efficiently while maintaining service availability.

The experimental framework developed in this thesis, including the four-scenario design with systematic variation in class balance, provides a template for evaluating future unlearning methods. The clear relationship between minority class percentage and unlearning quality enables quantitative prediction of expected outcomes, transforming unlearning from a trial-and-error process to an engineering discipline with predictable behavior.

As privacy regulations continue to strengthen globally and individuals become increasingly aware of their data rights, efficient machine unlearning will become essential infrastructure for responsible machine learning deployment. This thesis contributes to that foundation by:

1. Characterizing the conditions under which noise-based unlearning succeeds (balanced class distributions)

2. Identifying and reinterpreting apparent failure modes (edge-case generalization vs. true anti-learning)

3. Providing quantitative performance benchmarks across five strategies and four scenarios

4. Establishing practical recommendations for production deployment

The path toward comprehensive machine unlearning solutions remains open, but the findings presented here demonstrate that meaningful progress is achievable with current methods when applied thoughtfully. Noise-based unlearning, properly calibrated and validated against gold standard models, offers an efficient and effective approach to the right to be forgotten for tabular neural networks.

Appendix A

Appendix

This appendix provides supplementary tables, hyperparameter configurations, and additional experimental data referenced throughout the thesis.

A.1 Complete Experimental Results

A.1.1 Spiral Dataset Results

Table A.1 presents the complete experimental results for all 20 experiments on the spiral dataset (4 forget scenarios $\times$ 5 unlearning strategies). The spiral dataset served primarily as methodology validation, with only the Random scenario achieving successful forgetting.

Table A.1: Complete Spiral Dataset Results (Target Forget Accuracy: 25%)

Scenario	Strategy	Time (s)	Forget	Retain	Test	MIA	Anti-learn
Complete	Gaussian	0.37	12.5%	68.7%	55.6%	0.775	Yes
	Laplacian	0.28	15.1%	71.0%	57.7%	0.788	No
	Adaptive	0.32	10.2%	70.0%	57.5%	0.788	Yes
	Layer-wise	0.28	12.0%	70.4%	56.5%	0.781	Yes
	Gradient	0.31	12.2%	70.1%	55.4%	0.765	Yes
Geometric	Gaussian	0.34	4.3%	51.3%	48.5%	0.697	Yes
	Laplacian	0.35	9.6%	52.0%	49.0%	0.656	Yes
	Adaptive	0.38	4.3%	52.3%	48.1%	0.696	Yes
	Layer-wise	0.35	4.3%	51.7%	47.9%	0.688	Yes
	Gradient	0.37	7.8%	50.6%	46.0%	0.663	Yes
Random	Gaussian	0.34	68.7%	81.9%	81.9%	0.578	No
	Laplacian	0.35	71.3%	85.4%	86.0%	0.533	No
	Adaptive	0.39	62.6%	82.1%	82.3%	0.596	No
	Layer-wise	0.35	70.4%	83.3%	83.8%	0.553	No
	Gradient	0.38	72.2%	81.9%	82.5%	0.549	No
HighConf	Gaussian	0.35	13.0%	49.9%	48.8%	0.783	No
	Laplacian	0.34	15.6%	53.7%	54.2%	0.857	No
	Adaptive	0.38	14.3%	49.3%	46.9%	0.741	No
	Layer-wise	0.35	13.0%	48.5%	47.7%	0.783	No
	Gradient	0.38	15.6%	47.6%	46.9%	0.741	No

A.1.2 Adult Income Dataset Results

Table A.2 presents the complete experimental results for all 20 experiments on the Adult Income dataset (4 forget scenarios $\times$ 5 unlearning strategies).

Table A.2: Complete Adult Income Dataset Results (Target Forget Accuracy: 50%)

Scenario	Strategy	Time (s)	Forget	Retain	Test	MIA	Anti-learn
Married	Gaussian	0.02	67.2%	93.9%	82.1%	0.743	No
	Laplacian	0.02	65.3%	94.0%	82.3%	0.708	No
	Adaptive	0.15	64.1%	94.0%	82.2%	0.730	No
	Layer-wise	0.02	64.2%	93.9%	82.0%	0.742	No
	Gradient	0.18	62.9%	94.0%	82.4%	0.727	No
Executives	Gaussian	0.02	72.3%	84.3%	81.5%	0.669	No
	Laplacian	0.02	74.5%	84.7%	81.8%	0.643	No
	Adaptive	0.15	72.4%	84.2%	81.4%	0.662	No
	Layer-wise	0.02	71.7%	84.2%	81.3%	0.660	No
	Gradient	0.18	72.8%	84.4%	81.6%	0.660	No
HighEarnProf	Gaussian	0.02	34.0%	82.8%	81.2%	0.847	Yes
	Laplacian	0.02	33.8%	82.8%	81.1%	0.851	Yes
	Adaptive	0.15	34.0%	82.7%	81.0%	0.846	Yes
	Layer-wise	0.02	33.6%	82.8%	81.2%	0.858	Yes
	Gradient	0.18	31.3%	82.7%	81.0%	0.859	Yes
RandomBalanced	Gaussian	0.02	–	–	–	–	–
	Laplacian	0.02	–	–	–	–	–
	Adaptive	0.15	–	–	–	–	–
	Layer-wise	0.02	–	–	–	–	–
	Gradient	0.18	–	–	–	–	–

A.2 Hyperparameter Configuration Details

A.2.1 TabNet Architecture Parameters

Table A.3 details the TabNet configurations used for each dataset.

Table A.3: TabNet Architecture Configurations

Parameter	Spiral Dataset	Adult Income Dataset
Decision Width (n_d)	8	64
Attention Width (n_a)	8	64
Decision Steps (n_steps)	3	5
Feature Reusage (γ)	1.3	1.3
Sparsity Regularization	1e-3	1e-3
Attention Mechanism	sparsemax	sparsemax
Batch Size	256	1024
Virtual Batch Size	128	512
Learning Rate	0.02	0.02
Early Stopping Patience	20 epochs	20 epochs

A.2.2 Unlearning Strategy Hyperparameters

Table A.4 provides the hyperparameter configurations for each unlearning strategy across different scenarios on the Adult Income dataset.

Table A.4: Unlearning Strategy Hyperparameters by Scenario (Adult Income)

Strategy	Parameter	Married	Executives	HighEarnProf	RandomBalanced
Gaussian	Noise Scale (σ)	0.20	0.25	0.40	0.20
	Scaling Method	Std-based	Std-based	Std-based	Std-based
Laplacian	Scale (b)	0.25	0.30	0.48	0.25
	Scaling Method	Std-based	Std-based	Std-based	Std-based
Adaptive	Noise Scale (σ)	0.18	0.22	0.35	0.18
	Gradient Norm	Max-norm	Max-norm	Max-norm	Max-norm
Layer-wise	Base Scale	0.15	0.18	0.28	0.15
	Scale Factor	1.5	1.5	1.5	1.5
	Progression	Linear	Linear	Linear	Linear
Gradient-based	Learning Rate	0.005	0.008	0.012	0.005
	Gradient Steps	10	15	25	10
	Max Gradient Norm	1.0	1.0	1.0	1.0
	Post-Ascent Noise	0.001	0.001	0.001	0.001

A.3 Anti-Learning Detection Criteria

The anti-learning detection algorithm uses the following criteria:

Require: Forget accuracy f, Target accuracy t (random guessing) **Ensure:** Boolean indicating anti-learning status

threshold $\leftarrow t/2$ **if** $f <$ threshold **return** True *// Anti-learning detected* **else return** False *// Normal unlearning*

Algorithm 1: Anti-Learning Detection

For binary classification (Adult Income), the target is $t = 50\%$ and the anti-learning threshold is 25%. Forget accuracy values below 25% indicate that the model has learned to systematically predict the opposite class rather than achieving random guessing behavior.

A.4 Minority Class Percentage

The minority class percentage quantifies class distribution within forget sets. It represents the proportion of samples belonging to the smaller class:

$$\text{Minority Class \%} = \min(\text{Class 1 \%}, \text{Class 0 \%}) \tag{A.1}$$

This metric ranges from 0% (complete single-class composition) to 50% (perfect balance). Table A.5 presents the class distribution for all Adult Income scenarios.

Table A.5: Class Distribution for Adult Income Scenarios

Scenario	Class 0 %	Class 1 %	Minority %
Married	55.4%	44.6%	44.6%
Executives	51.6%	48.4%	48.4%
HighEarnProf	0.0%	100.0%	0%
RandomBalanced	50.0%	50.0%	50%

A.5 Computational Environment

All experiments were conducted using the following environment:

Table A.6: Computational Environment Specifications

Component	Specification
Hardware	Apple MacBook Pro with M4 chip
Operating System	macOS
Python Version	3.10+
PyTorch Version	2.0+
TabNet Library	pytorch-tabnet 4.0+
NumPy Version	1.24+
Pandas Version	2.0+
Scikit-learn Version	1.3+
Primary Random Seed	42
Validation Seeds	123, 456, 789, 1024
Execution Mode	CPU (Apple Silicon optimized)

A.6 Dataset Statistics

A.6.1 Spiral Dataset

Table A.7: Spiral Dataset Statistics

Property	Value
Total Samples	2,400
Training Samples	1,440 (60%)
Validation Samples	480 (20%)
Test Samples	480 (20%)
Number of Classes	4
Samples per Class	600
Features	2 (x, y coordinates)
Spiral Tightness	$r = \theta/2\pi$
Noise Level	Gaussian, $\sigma = 0.5$

A.6.2 Adult Income Dataset

Table A.8: Adult Income Dataset Statistics

Property	Value
Total Samples	48,842
Training Samples	26,048 (53.3%)
Test Samples	22,794 (46.7%)
Number of Classes	2 (binary)
Class Distribution	76% ($\leq$50K), 24% ($>$50K)
Original Features	14
Features After Processing	11
Removed Features	fnlwgt, education, native-country
Missing Values	7% of records
Missing Value Strategy	Unknown category (preserved)

A.7 Forget Request Specifications

A.7.1 Spiral Dataset Forget Scenarios

Table A.9: Spiral Dataset Forget Scenario Specifications

Scenario	Selection Criteria	Samples	Balance
Complete	All samples where class = 0	360 (25%)	0.00
Geometric	Circular region removal	∼108 (7.5%)	Variable
Random	Random 40% from Class 1	∼144 (10%)	0.00
HighConf	Top 20% by confidence	∼72 (5%)	Variable

A.7.2 Adult Income Dataset Forget Scenarios

Table A.10: Adult Income Dataset Forget Scenario Specifications

Scenario	Selection Type	Selection Criteria	Samples	Balance
Married	Attribute-based	marital-status = Married-civ-spouse	∼12,000	0.89
Executives	Attribute-based	occupation = Exec-managerial	∼3,300	0.97
HighEarnProf	Geometric	Class 1, outer 25% by distance	∼1,500	0.00
RandomBalanced	Stratified Random	1,500 from each class	3,000	1.00

A.8 Multi-Seed Validation Results

Table A.11 presents the mean and standard deviation of key metrics across five random seeds for the Adult Income dataset experiments.

Table A.11: Multi-Seed Validation Results (Mean ± Std, 5 seeds)

Scenario	Metric	Laplacian	Gradient-based
Married	Forget Acc	65.3 ± 1.2%	62.9 ± 1.5%
	Retain Acc	94.0 ± 0.3%	94.0 ± 0.4%
	Quality	0.622 ± 0.02	0.617 ± 0.03
Executives	Forget Acc	74.5 ± 1.8%	72.8 ± 2.1%
	Retain Acc	84.7 ± 0.5%	84.4 ± 0.6%
	Quality	0.604 ± 0.02	0.594 ± 0.03
HighEarnProf	Forget Acc	33.8 ± 2.5%	31.3 ± 3.2%
	Retain Acc	82.8 ± 0.4%	82.7 ± 0.5%
	Quality	0.526 ± 0.03	0.508 ± 0.04
RandomBalanced	Forget Acc	–	–
	Retain Acc	–	–
	Quality	–	–

A.9 Unlearning Quality Score Formula

The unlearning quality score combines multiple metrics with the following weights:

$$\text{Quality} = 0.4 \times \text{retain\_acc} + 0.4 \times (1 - 2|\text{forget\_acc} - 0.5|) + 0.2 \times (1 - \text{MIA\_penalty}) \tag{A.2}$$

where:

- retain_acc: Accuracy on retained training samples (normalized to [0,1])
- forget_acc: Accuracy on forget set samples (target is 0.5 for binary classification)
- MIA_penalty: Zero if MIA AUC < 0.55, otherwise linearly increasing penalty

The formula assigns equal weight (40%) to utility preservation (retain accuracy) and forgetting effectiveness (distance from random guessing), with 20% weight for privacy protection (MIA resistance).

Bibliography

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep Learning with Differential Privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308-318.

[2] Arik, S. O., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6679-6687.

[3] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021). Machine Unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, 141-159.

[4] Cao, Y., & Yang, J. (2015). Towards Making Systems Forget with Machine Unlearning. *2015 IEEE Symposium on Security and Privacy*, 463-480.

[5] Dwork, C., & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 211-407.

[6] Fan, C., Liu, J., Hero, A., & Liu, S. (2024). Challenging Forgets: Unveiling the Worst-Case Forget Sets in Machine Unlearning. *arXiv preprint arXiv:2403.07362*.

[7] Fan, C., Liu, J., Zhang, Y., Wei, D., Wong, E., & Liu, S. (2024). SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. *International Conference on Learning Representations (ICLR)*.

[8] Gil Hernandez, S. (2025). Challenging Forgets: Identifying and Analyzing Hard-to-Unlearn Data. *Bachelor's Thesis, Universitat de Barcelona, Faculty of Mathematics and Computer Science*.

[9] Guo, C., Goldstein, T., Hannun, A., & van der Maaten, L. (2020). Certified Data Removal from Machine Learning Models. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 3832-3842.

[10] NeurIPS 2023 Machine Unlearning Competition. *Kaggle / NeurIPS 2023*. https://www.kaggle.com/competitions/neurips-2023-machine-unlearning

[11] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy (SP)*, 3-18.

[12] Xu, H., Zhu, T., Zhang, L., Zhou, W., & Yu, P. S. (2023). Machine Unlearning: A Survey. *ACM Computing Surveys*, 56(1), 1-36.