Titre: Title:	Verifiability of Unlearning Schemes Through Local Explanation
Auteur: Author:	Saba Kasrelou
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Kasrelou, S. (2025). Verifiability of Unlearning Schemes Through Local Explanation [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/64697/

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Verifiability of Unlearning Schemes Through Local Explanation

SABA KASRELOU
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Avril 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Verifiability of Unlearning Schemes Through Local Explanation

présenté par **Saba KASRELOU**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Alejandro QUINTERO, président
Samuel PIERRE, membre et directeur de recherche
Ranwa AL MALLAH, membre et codirectrice de recherche
Foutse KHOM, membre

DEDICATION

To my parents
and my husband . . .

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Prof. Samuel Pierre for his continuous support of my academic journey, for his patience, and immense knowledge, and for being kind-hearted like a father. Thank you for giving me this opportunity to explore new avenues in the AI field and for believing in me to accomplish this research subject. I could not have imagined having a better advisor for my study.

I want to express my heartfelt gratitude to my co-supervisor, Dr. Ranwa-AL-Mallah, for her insight and passion throughout my research. Her guidance and constant support were crucial in completing this thesis. I'm truly grateful for her encouragement, which helped me achieve more than I thought I could.

This thesis is especially dedicated to Majid, my husband, who always supports me. Thank you for sharing this amazing journey with me and believing in me when I didn't believe in myself.

I am also grateful to Professor Foutse Khom and Professor Alejandro Quitero for their time in evaluating my master's thesis.

Last, but certainly not least, I sincerely thank the most influential person in my life, my mother. Thanks for your endless and unconditional support, and enduring love.

Thank you all for your support and contributions to this significant milestone in my academic career.

RÉSUMÉ

À mesure que les systèmes d'Intelligence Artificielle (IA) s'intègrent de plus en plus aux applications du quotidien, les préoccupations liées à la confidentialité des données et au respect des lois sur le « Droit à l'oubli » se sont accrues de manière significative. Les modèles d'apprentissage automatique deviennent de plus en plus vulnérables aux attaques de confidentialité, telles que les attaques par inférence de membership (ou Membership Inference Attacks, MIA) et les attaques par inversion de modèle, qui peuvent révéler des données sensibles utilisées lors de l'entraînement. Cela a conduit à une attention accrue envers le « désapprentissage » des modèles (connu sous Machine Unlearning, MU), un processus conçu pour éliminer sélectivement l'impact de certains points de données dans un modèle déjà entraîné. Le désapprentissage des modèles constitue un mécanisme essentiel pour répondre aux préoccupations de confidentialité et se conformer aux réglementations.

Les méthodes traditionnelles de MU nécessitent souvent de réentraîner le modèle depuis le début, une approche lente et coûteuse. Les avancées récentes ont proposé des techniques alternatives de désapprentissage par réorganisation des données et manipulation des modèles, permettant d'assurer que certains points de données puissent être effectivement retirés d'un modèle d'apprentissage automatique sans recourir à un réentraînement coûteux. Cependant, une méthode de vérification fiable et explicable que le MU a effectivement eu lieu reste encore à explorer.

Cette recherche propose une méthode de vérification novatrice basée sur l'explicabilité locale pour garantir que les opérations de désapprentissage soient à la fois efficaces et transparentes. Notre approche utilise des outils d'IA explicable, notamment les explications agnostiques et localement interprétables des modèles (LIME), afin d'évaluer si le désapprentissage a bien retiré les données sensibles sans compromettre l'intégrité du modèle. En capturant le comportement local des modèles avec et sans les données cibles, nous proposons un processus permettant de détecter les changements confirmant la suppression des informations sensibles. Cette méthodologie est structurée en trois phases : construction de modèles avec et sans les données cibles, validation du désapprentissage avec MIA, et application d'explications locales pour vérifier les résultats du désapprentissage.

Cette recherche contribue à renforcer la fiabilité, la transparence et le comportement des modèles d'apprentissage automatique, tout en faisant progresser le domaine du désapprentissage des modèles en introduisant l'interprétabilité et l'explicabilité. Notre méthode se concentre non seulement sur la comparaison de différentes techniques de désapprentissage, mais fournit

également des explications claires, garantissant que les utilisateurs puissent comprendre et faire confiance au processus. Nous fournissons des preuves de la vérification de l'oubli localement grâce à l'analyse de la redistribution de l'importance des caractéristiques à l'aide de LIME. Nous nous concentrons sur la vérification de l'oubli réussi en fournissant des preuves basées sur des techniques d'approximation locale utilisant la redistribution de l'importance des caractéristiques.

ABSTRACT

As Artificial Intelligence (AI) systems become more integrated into everyday applications, concerns over data privacy and compliance with "Right to Be Forgotten" laws have grown significantly. Machine learning models are increasingly susceptible to privacy attacks such as Membership Inference Attacks (MIA) and model inversion attacks, which can reveal sensitive training data. This has led to an increased focus on Machine Unlearning (MU), a process designed to selectively remove the impact of specific data points from an already-trained model. Machine unlearning is a critical mechanism for addressing privacy concerns and adhering to regulations.

Traditional methods often require retraining the model from scratch, which is a costly and time-consuming approach. Recent advancements have proposed alternative unlearning techniques through data reorganization and model manipulation. This involves ensuring that specific data points can be effectively removed from a machine-learning model without the need for expensive retraining. Yet a reliable and explainable verification method to ensure that unlearning occurred remains to be explored.

This research introduces a novel verification method leveraging local explainability to ensure the effectiveness and transparency of unlearning operations. Our approach leverages explainable AI tools, including Local Interpretable Model-Agnostic Explanations (LIME) to assess whether unlearning has successfully removed sensitive data without compromising model integrity. By capturing the local behavior of models with and without the target data, we propose a process to detect changes that confirm the removal of sensitive information. This methodology is structured in three phases: constructing models with and without the target data, validating unlearning with MIA, and applying local explanations to verify unlearning outcomes.

Ultimately, this research contributes to enhancing the trustworthiness, transparency, and behavior of machine learning models, while also advancing the field of machine unlearning with the focus on interpretability and explainability. Our method not only focuses on comparing different unlearning schemes but also provides clear explanations, ensuring users can understand and trust the process. We provide evidence of unlearning verification locally through analysis of feature importance redistribution using LIME. We focus on verifying successful unlearning by providing evidence based on local approximation techniques using feature importance redistribution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

AI	Artificial Intelligence
ML	Machine Learning
MU	Machine Unlearning
LIME	Local Interpretable Model-Agnostic Explanation
MIA	Membership Inference Attack
MLP	Multilayer Perceptron
DNN	Deep Neural Network
GDPR	General Data Protection Regulation
CCPA	California Consumer Privacy Act
APPI	Act on the Protection of Personal Information
CPPA	Canada's proposed Consumer Privacy Protection Act
API	Application Programming Interface
SISA	Sharded, Isolated, Sliced, and Aggregated
MLaaS	Machine Learning as a Service

CHAPTER 1 INTRODUCTION

Machine Learning (ML) has revolutionized various aspects of modern technology, driving breakthroughs in areas such as computer vision, speech recognition, and medical diagnostics. However, as Artificial Intelligence (AI) systems increasingly depend on vast datasets, new challenges have emerged, particularly around privacy concerns, regulations, and the growing need to remove specific samples from training datasets and erase their impact from already-trained models. This need arises due to vulnerabilities like membership inference attacks and model inversion attacks, which can reveal details about the data used in the training process [3] [4] [5]. Moreover, global privacy regulations — such as the European Union's General Data Protection Regulation (GDPR) [6], the California Consumer Privacy Act (CCPA) [7], the Act on the Protection of Personal Information (APPI) [8], and Canada's proposed Consumer Privacy Protection Act (CPPA) [9], compel the deletion of private information. They have granted individuals with the "Right to Be Forgotten", requiring that personal information be erased upon request. This process, in which specific data points are removed from both the training dataset and the already trained model, is called Machine Unlearning (MU). Machine Unlearning was first studied by [10] in the context of statistical query learning.

Machine unlearning allows the selective removal of specific data points from a model, ensuring that these data points no longer influence the predictions of the model. This has become especially important given the increasing awareness of privacy risks, such as membership inference attacks and model inversion attacks, which can expose private information from trained models. Machine unlearning is not only about regulatory compliance, it also helps improve models by removing noisy or irrelevant data.

Traditional approaches, such as retraining the model after removing specific data points, can achieve this goal [11]. However, retraining is computationally expensive, especially for large models such as Deep Neural Networks (DNNs). Therefore, researchers are exploring alternative methods that allow models to "unlearn" data without the need to start training from scratch. These techniques must balance the efficiency of the unlearning process while maintaining model performance in the remaining data. On the other hand, differential privacy, which limits how much any single data point can influence the model, data masking and online learning do not fully address the challenge of completely removing the influence of specific data points. This is where machine unlearning comes into play, focusing specifically on techniques that allow models to forget particular data while minimizing the impact on overall performance.

Our contribution to this research project is to delve deeper into MU methods, focusing on the verifiability of unlearning schemes by examining state-of-the-art machine unlearning techniques. We propose a verification technique to ensure that MU has been effectively carried out. To verify whether unlearning has taken place, we adopt an explainability-driven approach using LIME. As this method operates locally, it allows us to approximate model behavior near specific data points. Our approach highlights how feature importance is redistributed around unlearned instances, thereby offering local evidence of unlearning. By providing a clearer and more comprehensive understanding of the inner mechanism of MU through verification with explanations, we want to contribute to a more transparent and privacy-respecting environment in machine learning applications. By observing the redistribution of dominant, moderate, and minimal contributing features before and after unlearning, we demonstrate that successful unlearning should be reflected in a shift of feature importance around specific examples. While this provides strong localized evidence of unlearning, due to the inherent limitations of local explanations. The significance of unlearning becomes even more crucial in enhancing the trustworthiness and transparency of Artificial Intelligence, especially as AI's role grows in various fields that deal with extensive personal user data.

1.1 Definitions and Basic Concepts

Machine Unlearning focuses on eliminating specific data, model knowledge, or patterns learned by the model in order to enhance privacy, fairness, and adaptability. Figure 1.1 illustrates the process of Machine Unlearning [1].

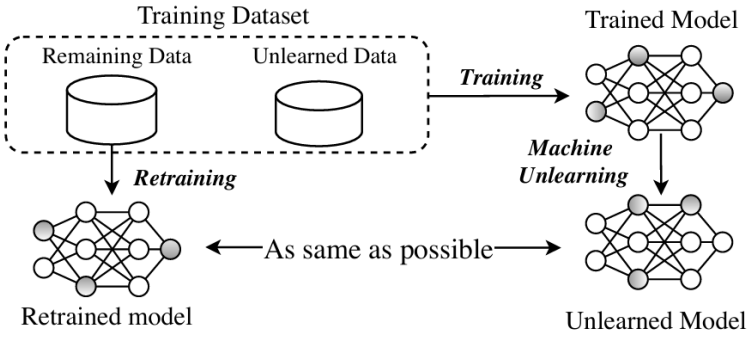

Figure 1.1 Machine Unlearning process [1].

Specifically, the targets of MU are:

- Exact Unlearning: It guarantees that the distribution of an unlearned model and a retrained model are indistinguishable. This approach involves retraining the model from scratch after removing the specific data points. While it ensures complete removal, it is often computationally expensive and impractical for large-scale models.
- Strong Unlearning: It is established based on the similarity between the internal parameter distributions of the models instead of full retraining. It modifies the model to eliminate the influence of the data being forgotten.
- Weak Unlearning: It only ensures that the distributions of the two final activations are indistinguishable. It reduces or minimizes the influence of specific data on a machine-learning model without fully guaranteeing its complete removal.

Weak and Strong Unlearning are considered Approximate Unlearning since they ensure that the distribution of the unlearned model and that of a retrained model are approximately indistinguishable. Figure 1.2 presents the associated output of the MU technique for each unlearning target [1].

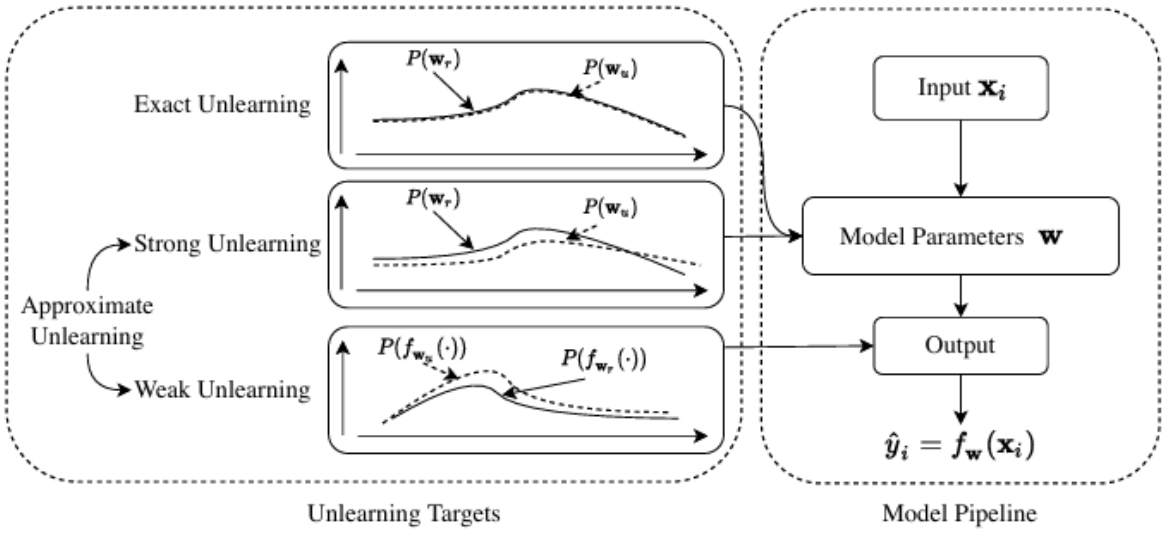

Figure 1.2 Output of the MU technique for each unlearning target [1].

1.1.1 Mathematical Formulation of Unlearning Verification

Machine unlearning can be formally defined as follows:

$f_\theta : \mathcal{X} \to \mathcal{Y}$ A trained model with parameters θ.

$D = \{x_i, y_i\}_{i=1}^N$ The training dataset of N samples.

$D_u \subset D$ Subset of data points to be unlearned.

$f_{\theta'}$: Model after unlearning.

$V(f_\theta, f_{\theta'})$ - Discrepancy function measuring the difference:

$$V(f_\theta, f_{\theta'}) = \mathbb{E}_{x \sim \mathcal{X}} \left[|f_\theta(x) - f_{\theta'}(x)| \right]$$

Given a trained model $f_\theta : \mathcal{X} \to \mathcal{Y}$ with parameters θ trained on dataset $D = \{x_i, y_i\}_{i=1}^N$, the goal of unlearning is to remove the influence of a subset of data points $D_u \subset D$ such that the resulting model $f_{\theta'}$ behaves as if it had never been trained on D_u.

We define the unlearning function as:

$$V(f_\theta, f_{\theta'}) = \mathbb{E}_{x \sim \mathcal{X}} \left[|f_\theta(x) - f_{\theta'}(x)| \right] \tag{1.1}$$

where $V(f_\theta, f_{\theta'})$ measures the discrepancy between the original model and the unlearned model.

1.1.2 Membership Inference Attack

Membership Inference Attacks (MIAs) exploit vulnerabilities in machine learning models to determine whether a specific data sample was part of the training dataset, posing significant privacy risks. These attacks leverage differences in the behavior of a model when responding to training data versus unseen data, exploiting overfitting or overconfidence in predictions. Attackers typically analyze features such as model confidence scores, loss values, or output probabilities to infer membership, particularly in deep learning models. MIAs have critical implications for privacy-sensitive domains like healthcare, finance, and personalized services, where revealing data membership could expose individuals' sensitive information. Mitigation strategies include regularization, differential privacy, and dropout mechanisms to enhance model robustness against such attacks. The growing complexity of MIAs highlights the need for thorough evaluation of model privacy in real-world applications [12].

The white box approach to MIAs assumes access to all the information needed to check the confidence score of the model. In this setup, there is no need to train any shadow model to then aggregate the outputs in order to investigate the model‘s behavior.

When performing a membership inference attack, the attacker wants to figure out whether a specific data point (like a particular image, record, or text) was used to train a machine learning model. For example, if a model that predicts whether someone has a disease was trained on a set of medical records, an attacker wants to know: "Was this specific person's record in the training data?" These are the general steps the attacker follows to conduct a MIA:

- Training of a Target Model. A machine learning model is trained on a dataset. This model learns patterns from training data. This model is called the "Target Model".
- Access the Target Model. The attacker gets access to the target model. In a black box access, the attacker can query the model and will get outputs or predictions but doesn't get the internal details. In a white box access, the attacker knows everything about the model, including its internal parameters.
- Observe the Model's Behavior. The attacker provides some input to the model and observes the output (e.g. confidence scores). In fact, a model often behaves differently for data it was trained on compared to data it hasn't seen before. In terms of training data, the model is very confident and gives high probabilities. With regard to test data, the model is less confident or gives lower probabilities.
- Test a Specific Data Point. The attacker tests a specific data point (e.g. a medical record, an image, ...) on the model to see how confident it is. If the model is very confident about this data point, the attacker can infer that this data was likely in the training set.

On the other hand, a shadow model is used by an attacker in the context of a black box attack. Firstly, the attacker needs to simulate the behavior of the target model to then learn how the target model behaves on data that is inside versus outside its training set. By studying this, the attacker improves their ability to carry out a membership inference attack. These are the general steps the attacker follows to conduct a MIA in a black box setup:

- Understand the Target Model's Domain. The attacker only has access to general information about the target model, meaning the type of data model trained on, model's task, ...
- Create Synthetic Datasets. The attacker collects or generates datasets that are likely similar to what they think the target model's training data looks like.

- Train Shadow Models. The attacker trains one or even more shadow models on these synthetic datasets. The idea is that the shadow models will behave similarly to the target model because they were trained on similar data.
- Analyze Shadow Models. Once the shadow models are trained, the attacker observes how these models behave. Note that this stage is similar to white box. The attacker needs to extract how confident are the shadow models on their training data. Also, they need to assess how confident are they on unseen data. This gives the attacker clues about how the target model might behave in a similar situation.
- Apply the Knowledge to the Target Model. The attacker uses what they learned from the shadow models to infer membership information about the target model's training data. For example: If the shadow models are very confident on their training data but less confident on other data, the attacker assumes the target model behaves the same way.

In sum, the shadow model helps the attacker mimic the behavior of the target model without needing full access to it in order to detect whether a data point was part of the target model's training set. If the target model behaves like the shadow model, the attacker can infer whether the sample was part of the target model's training set.

In machine unlearning, the idea is that even after deploying the MU method, there is a need to check if the model still reveals the unlearned data or not. One way to perform this verification is using the white box approach of MIA. If the confidence score provided by the model is high, this means that the model has information leakage and can remember unlearned samples. Otherwise, if the confidence is low, we can conclude that the MU technique is effective.

1.1.3 Interpretable Model-Agnostic Explanations

Local Interpretable Model-Agnostic Explanations (LIME) is a method that provides local approximations of a machine learning model's behavior to explain individual predictions [13,14]. It works by perturbing the input data around a specific instance, generating synthetic samples, and recording the model's predictions for these samples. Using this data, LIME fits a simpler surrogate model — typically a sparse linear regression — within the vicinity of the instance being explained. The surrogate model highlights which features most influenced the model's prediction locally, offering insights into the decision-making process for that specific case. LIME's reliability depends on parameters like neighborhood size, sampling strategy, and weighting function, which can significantly affect the quality and accuracy of

the explanations. While highly flexible and model-agnostic, LIME's dependence on these settings and its sensitivity to perturbation can sometimes lead to misleading interpretations, particularly for models with highly nonlinear decision boundaries.

1.2 Elements of the Problem

The concept of machine unlearning is becoming increasingly relevant given the granted 'Right to Be Forgotten'. One straightforward approach to perfectly removing the information from the model is to retrain it from scratch [1, 15, 16]. However, many complex models have been built on an enormous set of samples. Retraining is generally a computationally expensive process [10, 17]. Several existing methods, such as Differential Privacy [18, 19], Data masking [20], Online learning [21], and Catastrophic forgetting [22, 23], attempt to address this issue. However, these techniques do not fully remove the influence of specific data points from a model, leaving potential privacy risks. Additionally, there are differences either in their objectives or the rationales behind them when compared to MU. The key challenge is how to verify whether a model has truly "forgotten" the data while maintaining efficiency. Many models claim to perform unlearning, but there is no reliable method to prove whether unlearning was successful.

Particularly, the goal of machine unlearning is to restore the model to a state where it performs just as well as it would have if it were never trained on the data that is now being unlearned. In other words, the model should forget the specific information from those data. MU must complete this process efficiently while minimizing the time required for unlearning and reducing cost. This is crucial for practical applications.

Unlearning methods fall into two categories: (1) data reorganization and (2) model manipulation. The first category modifies training data while the second alters model structure. Model manipulation methodologies directly modify the trained model itself. This can involve techniques such as model shifting, model pruning, and model replacement [1]. As eluded for earlier, the verification process of a MU technique is as equally important as the technique itself. The unlearned model's behaviors must be assessed to ensure that it effectively forgot the target data while retaining its generalization.

In fact, to verify the unlearning methods, a couple of avenues are explored by researchers: empirical evaluation and theoretical calculation [1]. Empirical evaluation involves comparing the unlearned model's performance to a model retrained from scratch or measuring its sensitivity to attacks designed to expose the influence of the unlearned data. This includes retraining-based verification, attack-based verification, relearning time-based verification, and accuracy-

based verification [24].

In contrast, theoretical calculation aims to mathematically measure how much information about unlearned data remains in the model [1]. Some methods offer theory-based verification, certifying that unlearning renders the model indistinguishable from one trained on the remaining dataset. In comparison, others utilize information-bound-based metrics to measure the effectiveness of unlearning schemes. Neither of these methods provides explainable verification of the unlearning outcome. Limited attention has been directed toward using explainability as a verification mechanism to conduct theoretical research into the validity of unlearning schemes [25, 26]. Following a model's learning and unlearning processes, the question arises:

> Can the model explain what factors contributed to the prediction in both cases as means of verification?

Currently, the complexity of models and the lack of explainable unlearning verification mechanism are the biggest challenges in real-world scenarios. Data owners cannot verify if their data has been truly unlearned from the model. Thus, the user must trust the model provider to accurately and completely remove their data. Model providers may claim to have unlearned data to avoid harming model performance or incurring additional costs [27] but the users must be able to verify in human-understandable way the validity of the claims.

In fact, explainable models are essential for understanding the capabilities and behaviors of machine learning models which are now extensively integrated into various applications, ranging from automated decision-making to user interactions. Similarly, creating accurate models with a reliable unlearning verification mechanism will result in a clear understanding of whether a data point has been successfully unlearned. This clarity is essential for assessing the trustworthiness of the unlearning process and determining why the scheme should or should not be trusted [25].

1.3 Research Objectives

Our main objective is to develop an unlearning verification mechanism that relies on explainability to verify if the data point has been successfully unlearned. We aim to provide explanations at the local level using LIME. The intuition is that through sampling instances both in the vicinity of the target sample of any unlearned model and the native model, the locality captured should be different if the target was really unlearned. This new verification mechanism will ensure that unlearning is both effective and transparent.

To achieve this aim, the following sub-objectives provide a structured, systematic approach to solving for a successful verification method. We need to train the ML model and unlearned models using state of the art of unlearning schemes. We then must verify the unlearning schemes based on Membership Inference Attacks. This will provide a verification baseline. We will then put forward our explainable verifiability method for unlearning schemes through local explanation. Finally, we need evaluate the performance of the models in order to measure the effectiveness of the new explainable verification technique. This will ensure consistency, efficiency, and effectiveness in reaching the desired outcome.

1.4 Thesis plan

The rest of this thesis is structured as follows: **Chapter 2** presents the literature review. **Chapter 3** describes the methodology that is adopted for this research. **Chapter 4** presents the evaluation and results that demonstrate the importance of explanation in the verification of unlearning methods. **Chapter 5** provides the conclusion and future work.

CHAPTER 2 LITERATURE REVIEW

Machine Unlearning, also referred to as selective forgetting, data deletion, or scrubbing, involves the complete and efficient removal of specific samples and their influence from both a training dataset and a trained model. The need for machine unlearning arises not only from regulatory and legal obligations but also from the privacy and security concerns of data providers, as well as the demands of model owners. Additionally, eliminating the impact of outlier training samples can enhance a model's performance and robustness.

When users revoke permissions for certain training data, simply deleting those data from the original dataset is insufficient, as attackers can still extract user information from the trained models. A direct method to completely remove such information from the model is to retrain it from scratch. However, retraining complex models, often built on vast datasets, is computationally costly. Figure 2.1 presents the typical workflow of a machine learning model in the presence of a data removal request.

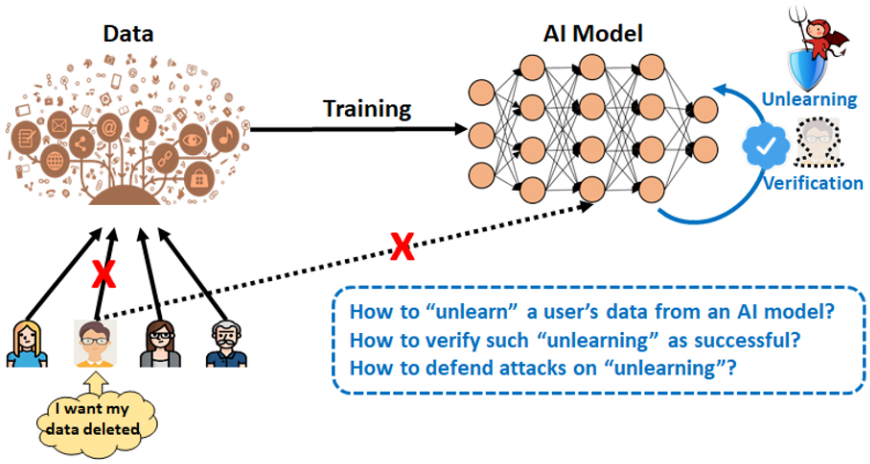

Figure 2.1 Typical workflow of a machine learning model in the presence of a data removal request. In general, a model trained on some data is then used for inference. Upon a removal request, the data-to-be forgotten should be unlearned from the model. The unlearned model is then verified against privacy criteria, and, if these criteria are not met (i.e., if the model still leaks some information about the forgotten data), the model is retrained [2].

2.1 Data Protection and Machine Unlearning Techniques

Several existing data protection techniques share similarities with machine unlearning but differ significantly in their objectives and underlying principles. We highlight in what follows

the distinctions between machine unlearning methods and these techniques [1].

Among the data protection techniques, differential privacy ensures that by observing the output of a model, it is impossible to determine whether a particular sample was part of the training dataset. This method establishes a controlled limit on how much any single sample can influence the final model. On the other hand, machine unlearning focuses specifically on removing user-specific training data entirely from the model [1].

Data masking is a technique designed to hide sensitive information in the original dataset. It transforms sensitive data to prevent them from being disclosed in unreliable environments. In contrast, the goal of machine unlearning is to prevent a trained model from exposing sensitive information about the training data it has already processed [1].

Online learning models are designed to adapt quickly to new data. This allows them to stay up-to-date and reflect recent changes. Unlike machine unlearning which removes updates, online learning adds and integrates new information into the model [1].

Catastrophic forgetting refers to a sharp decline in a model's ability to perform previously learned tasks after being fine-tuned on a new task. While this phenomenon causes a model to lose accuracy for older tasks, the data it relied on can often still be inferred by analyzing the model's parameters. As a result, catastrophic forgetting does not fulfill the stringent requirements of machine unlearning, which ensures the complete and provable removal of specific data from the model [1].

Among the machine unlearning techniques, Data Reorganization modifies or masks the data to make it less identifiable or impactful in the learning process. It typically involves techniques such as relabeling data points or adding noise to the dataset to obscure specific information. By doing so, the model's dependence on certain data instances is reduced without completely removing them. On the other hand, Model Manipulation methods which evaluate the contribution of each training sample to the model's predictions and adjust its influence using different approaches. These techniques may involve influence functions, like weight perturbation to selectively reduce or remove the impact of specific data points. Using such a method, model manipulation enables targeted unlearning while maintaining overall model performance [1].

In the data pruning method, the dataset is divided into multiple smaller sub-datasets, and only the affected sub-models are retrained after unlearning. A notable example of this approach is the Sharded, Isolated, Sliced, and Aggregated (SISA) framework [28], which enhances efficiency by limiting retraining to specific partitions of the dataset. This method is particularly useful for reducing the computational cost of unlearning by avoiding full-model

retraining.

Unlearning schemes based on data obfuscation aim to confuse a model's understanding of specific samples, making it difficult for attackers to exploit shifts in output confidence vectors. These approaches modify data labels or introduce noise to obscure the influence of targeted samples.For example, [29], they utilized generative noise to modify model weights, ensuring unlearning only affects the targeted data without impacting the remaining dataset to confuse the model's output.

In the data replacement method, instead of directly removing data points, this technique transforms the dataset by substituting the original data with modified or synthetic versions. The transformed dataset ensures that the model no longer relies on the removed data points, facilitating simpler retraining and a more controlled unlearning process [30].

Model shifting involves updating the model's parameters in a way that counteracts the influence of the data points to be unlearned. Instead of explicitly removing data or retraining from scratch, parameter adjustments are made to ensure that the model behaves as if the unwanted data was never included in the training process [31].

In model replacement, the model's parameters are replaced with precomputed values that correspond to a version of the model without the unlearned data. This technique is particularly effective when unlearning is anticipated in advance, allowing for efficient restoration of the model's state without costly retraining [32].

Finally, model pruning removes specific model parameters that were influenced by the data points to be unlearned. By eliminating these parameters, the model's reliance on the removed data is diminished. This method is particularly useful in neural networks and node or tree-based models, where certain neurons, weights, or nodes can be pruned to minimize the effect of unwanted data [33].

2.2 Verification Methods of Machine Unlearning Techniques

In the context of verification of unlearning techniques, empirical evaluation refers to the process of testing and assessing the effectiveness of machine unlearning methods based on real-world experiments and performance measurements.

- **Retraining-based verification:** Retraining inherently offers a strong verifiability property because the updated training dataset completely excludes the samples that require unlearning. By reconstructing the model from scratch using a dataset that no longer contains the undesired data, retraining ensures that these samples have no

influence on the final model. This approach is not only the most straightforward and intuitive, but also provides a transparent and easily auditable method to ensure compliance with data removal requests. Although computationally expensive, it guarantees that the influence of the removed data is completely eradicated, making it a robust solution for scenarios requiring strict data unlearning and regulatory compliance [1].

- **Attack-based verification:** The primary goal of an unlearning operation is to mitigate the risk of sensitive information leakage caused by model overfitting. To assess the effectiveness of these operations, certain attack techniques can serve as direct and reliable verification methods. For instance, membership inference attacks [12], and model inversion attacks [34] can be utilized to determine whether a model still retains traces of removed data. Furthermore, [35] introduced an innovative backdoor verification mechanism tailored for machine learning as a service (MLaaS) environments [36]. This mechanism enables individual users to verify, with high confidence, whether a service provider has genuinely adhered to their right to have specific information unlearned, ensuring greater transparency and compliance in AI-driven systems.

- **Relearning time-based verification:** Relearning time serves as a valuable metric for assessing the extent to which a model retains information about the data that was supposed to be unlearned. Specifically, if a model is able to regain its original performance with minimal retraining, this suggests that it may still retain residual knowledge of the removed samples. A shorter relearning time indicates that the model has not fully unlearned the data, as it can quickly reconstruct its previous state. This concept is particularly useful in evaluating the effectiveness of unlearning mechanisms and ensuring compliance with data removal requirements [37].

- **Accuracy-based verification:** A trained model typically exhibits high prediction accuracy for the samples it has learned from the training dataset. This characteristic allows the unlearning process to be verified by analyzing the model's accuracy after unlearning. Specifically, for the data that needs to be removed, the model's accuracy should ideally match that of a model trained from scratch without ever seeing the removed dataset [31].

In contrast, in the context of verification of unlearning techniques, theoretical calculation refers to the use of mathematical models, algorithms, and theoretical frameworks to assess the effectiveness of the unlearning process without relying on actual experiments or real-world data. These calculations typically focus on predicting the behavior of the unlearning technique and analyzing its properties in an idealized or abstract sense.

- **Theory-based verification:** Certain methods establish a certified unlearning framework, ensuring that a model, after unlearning, is indistinguishable from one trained from scratch on the remaining dataset. This certification provides a rigorous and provable guarantee that the unlearned samples no longer contribute to the model's behavior. By leveraging this principle, certified unlearning serves as a direct verification mechanism, demonstrating that the proposed unlearning techniques effectively remove the targeted data while maintaining model integrity [38], [39].

- **Information-bound-based metrics:** In [31], they introduced a novel metric to assess the effectiveness of unlearning mechanisms by quantifying the upper bound of residual information retained about the samples that need to be forgotten. This metric provides a difficult way to measure how much influence the unlearned data still has on the model. A lower residual information value indicates a more successful unlearning operation, ensuring that the removed data no longer impacts the model's predictions. This approach offers a quantifiable and systematic verification method, making it useful for evaluating compliance with privacy regulations and improving the reliability of machine unlearning techniques.

CHAPTER 3 METHODOLOGY

There is a strong connection between interpretability and explainability but it is very important to distinguish the difference between them [40]. Interpretability involves examining the internal workings of the model technique, enabling us to understand the significance of the model's weights and characteristics in producing the output. Whereas, explainability leads to the process of elucidating the behavior of a machine learning model to human users. To provide a viable means of verifying the unlearning process and ensure that data points have been successfully unlearned, our methodology focuses on explanation: Explaining whether the data point has been unlearned or not.

Instead of just looking at the model's confidence, humans want an easy-to-understand verification scheme. In fact, verification schemes should consider feasibility and explainability. That is, ordinary users should be able to understand and verify whether their unlearning request has been completed based on simple operations. That is why our main objective is to propose an unlearning verification mechanism to verify if the data point has been successfully unlearned or removed by providing local explanations using LIME to analyze model behavior before and after unlearning. The intuition is that through sampling instances both in the vicinity of the target sample of any unlearned model and of that of the native model, the locality captured should be different if the target was really unlearned. The purpose of this methodology is to achieve this goal. The approach involves a combination of theoretical evaluation and empirical testing to assess both model performance and privacy after the unlearning process. The methodology is divided into the following three phases:

- **Phase 1. Creation of Models**: Creating basic models and unlearned models using state of the art unlearning schemes. We start by building the basic models and those that have "unlearned" specific data using the unlearning techniques. These models are designed to forget certain information on purpose. The process starts with training a model with the target sample included. This step aims at comparing the behavior of the models before and after unlearning.

- **Phase 2. Verifiability Baseline**: Verifiability of unlearning schemes based on Membership Inference Attacks. As a baseline, we verify how well the existing unlearning methods work by running MIAs. This step is crucial because MIAs are a first initiative to revealing whether the model still remembers any of the data it was supposed to forget, which is evaluating the effectiveness of unlearning using an attack-based verification scheme. If the MIA succeeds, it suggests that unlearning was incomplete. For all

the unlearned models, the MIAs will provide us with a verification baseline to compare our proposed method with.

- **Phase 3. Explainable Verifiability Method**: Explainable verifiability method for unlearning schemes through local explanation. Finally, we explain the unlearning techniques using LIME for local explanations. This details how the model handles decision making for its output and ensures transparency in how the unlearning happens. By comparing the explanations between the target model and the unlearned one, we can determine whether the model's decision-making process has changed in a way that confirms successful unlearning.

The ultimate target of machine unlearning is to reproduce a model that behaves as if it was trained without seeing the unlearned data while consuming as less time as possible. Another performance baseline that can be used for validation for the unlearned model is that of the model retrained from scratch (a.k.a, native retraining). In fact, from phase 1, a ML model will first be trained with target samples in it. Then, these samples will be removed and a similar model will be trained without them. This native retraining naturally ensures that any information about samples can be unlearned from both the training dataset and the already-trained model and will serve as validation of both the attack-based verification scheme and our proposed explainable verification method.

CHAPTER 4 EVALUATION AND RESULTS

To implement the three phases of our methodology, we will first introduce the dataset used and the base model created, an MLP (Multi-Layer Perceptron). Since Phase 1 of the methodology aims at creating the models, this section will then present the specific methods that are employed in the training phase of basic and unlearned models to verify the data-removing process, and ensure that data points have been successfully deleted or unlearned with the existing methods. Phase 2 generates the verifiability baseline as a first attempt to describe the behavior of a trained ML model in a manner that provides guarantees about their behavior and inner mechanism. Phase 3 will present the new verifiability method that shows how the model demonstrates explainability in its decision-making process by providing clarity of its output through explanations.

4.1 Data analysis and Original Model

We will first introduce the dataset used in this research and then present the base model created.

Dataset: The Breast Cancer Patient Data dataset from Kaggle is used as it is a widely used benchmark dataset that contains 4016 samples and 16 features `https://www.kaggle.com/datasets/reihanenamdari/breast-cancer`.

- Column Data Types:
  - Numerical Columns: 5 (e.g., Age, Tumor Size, Regional Node Examined,...)
  - Categorical Columns: 11 (e.g., Race, Marital Status, T Stage,...)
- **Age:** Age of the patient.
- **Race:** Ethnicity of the patient.
- **Marital Status:** Marital status of the patient.
- **T Stage:** Tumor stage classification.
- **N Stage:** Lymph node involvement stage classification.
- **6th Stage:** Combined tumor staging based on 6th edition guidelines.

- **Differentiate:** Tumor differentiation grade (e.g., poorly, moderately, well differentiated).
- **Grade:** Tumor grade.
- **A Stage:** Anatomical stage.
- **Tumor Size:** Size of the tumor in millimeters.
- **Estrogen Status:** Whether the tumor is estrogen receptor-positive or negative.
- **Progesterone Status:** Whether the tumor is progesterone receptor-positive or negative.
- **Regional Node Examined:** Number of regional lymph nodes examined.
- **Regional Node Positive:** Number of regional lymph nodes found positive.
- **Survival Months:** Duration of survival after diagnosis in months.
- **Status:** Survival status of the patient (Alive or Dead).

Prepossessing and Original Model: The preparation phase involved a series of steps to ensure that the dataset was suitable for training a machine learning model.

Categorical data types variables were encoded into numerical representations using the Label Encoder method. This step allowed the machine learning algorithm to interpret and utilize these features effectively. After encoding the categorical features, the dataset was split into two parts: the input features (x) and the target variable (y). The target variable, Status, indicated whether a patient was alive or deceased, while the input features included all other columns. This split allowed us to focus on predicting the Status outcome based on the patient's clinical and demographic characteristics.

Next, the data was divided into training and testing subsets. The training set, comprising 80% of the data, was used to train the model, while the remaining 20% was set aside for testing. This division ensured that the model's performance could be evaluated on unseen data, providing a realistic measure of its accuracy.

An additional step involved scaling the features to bring them to a standardized scale. This was achieved using a standardization technique that adjusted the values to have a mean of (0) and a standard deviation of (1). Standardization is particularly important for models like neural networks. Without scaling, the model might struggle with features that operate on different numerical ranges.

A Multi-Layer Perceptron (MLP) was configured with a single hidden layer containing 100 neurons. The model was trained using a maximum of 1,000 iterations to allow sufficient time for optimization while preventing excessive computational time.

Once the model was trained, it was evaluated on the test dataset. The evaluation involved several key performance metrics. Accuracy was calculated to measure the overall proportion of correctly classified outcomes. Precision was used to determine the proportion of correctly predicted positive outcomes (e.g., identifying alive patients correctly) among all positive predictions. Recall measured how many of the actual positive outcomes were correctly identified by the model. Finally, the F1 score was computed to balance precision and recall, providing a single measure that reflects both aspects of performance.

With an accuracy of 87%, precision of 60%, recall of 49% and F1 score of 54%, the results of the evaluation indicated that the MLP model learned the patterns in the data, achieving a balance between accuracy, precision, recall, and the F1 score. It confirmed that the model could reliably predict patient survival outcomes based on the clinical and demographic characteristics given. It is important to note that these results are sufficient for the study at hand since the aim is not to achieve the best performance on the prediction task but rather to explore the impact of verification methods of unlearning techniques. Thus, simple fine-tuning was conducted and no extensive exploration of better architectures to increase the performance was conducted since it is out of the scope of the study.

4.2 Phase 1. Creation of Models

A wide variety of techniques are available for implementing machine unlearning [41]. In this research, we focused on examining a selected number of methods that are implementable under our experimental conditions. These methods were chosen based on their compatibility with the type of dataset, available hardware, and resources, ensuring they are feasible and executable within our setup. The experiments were designed to test machine unlearning implementations for comparative analysis across various machine unlearning approaches based on explanation, with plans to discuss further extensions and improvements in the future work section.

In the first machine unlearning method, which is Data Obfuscation, we aim to obscure the data by adding a noise factor of 0.01 to a selected number of samples. Specifically, we target the first 1,200 samples of our training dataset to make them unclear for the model. The goal is to force the model to "forget" the parts of the data where obfuscation is applied.

It is important to note that during the initial testing phase of this research, we first targeted a

100 samples. This number was then increased to 500, and based on the results, we observed a drop in the model's accuracy. However, by further increasing the number of samples, we managed to prevent this drop in accuracy. Ultimately, with 1,200 samples, the model's accuracy reached an acceptable level.

Therefore, after this experimentation on the machine learning model in our possession, we concluded that the number of selected samples has a direct impact on the model's accuracy. The impact is not linear, meaning that an increase in obfuscated data does not necessarily correspond to a decrease in performance.

The second method experimented with is Data Pruning. This approach is inspired by the SISA (Sharded, Isolated, Sliced, Aggregated) framework. It is a framework designed to efficiently remove specific training data points from machine learning models while preserving their performance. It partitions the dataset into shards, trains isolated models on slices of these shards, and aggregates their outputs, allowing targeted removal of data with minimal retraining.

To implement this method, we divide the training data into 4 number of shards, and each shard is further split into 2 slices. Then we mask the first 1200 samples from the dataset across all shards and slices using a boolean mask. A boolean mask is a data manipulation technique that utilizes a boolean array to filter or extract specific elements from another array or dataset. In this context, each value in the boolean array indicates whether a corresponding element in the dataset should be included (True) or omitted (False) to ensure only non-pruned data is retained. A boolean mask is created for the training dataset, initialized with True values to include all samples. The indexes of the first 1,200 samples to be pruned are identified and marked as (False) in the mask. This approach is particularly effective because it does not require removing elements directly; instead, it uses indexing to create a filtered version of the data. This makes the process both computationally efficient and straightforward to implement, especially when working with large datasets. The structured framework ensures localized data pruning while maintaining the integrity of isolated data subsets, making it suitable for applications requiring incremental updates, privacy preservation, or robust performance under data constraints.

The third method is Data Replacement. For this phase, we focused on investigating the effects of systematic data manipulation on the performance of the machine learning model. To achieve this we developed a function called *transform_value*, which applies a structured transformation to the data. Specifically, this function reverses the content of input values: for strings, it reverses the characters, and for numeric values, it reverses the digits. For instance, this is an example for text values (e.g., "hello" to "olleh") and numeric values (e.g., 123 to 321).

If the input is neither a string nor a number, the function leaves the value unchanged. This process allows us to introduce controlled distortions into the data for experimental purposes.

As such, as part of the experimentation, we deliberately transformed the first 1,200 samples of the training data using the *transform_value* function. This allowed us to analyze how the model adapts when trained on altered data while being evaluated on the unmodified test set. After training the model on the altered training data, we evaluated its performance on the test data using accuracy, precision, recall, and F1 score as metrics.

The fourth method that we experimented with is Model Pruning. In our experiment, we explored model pruning to evaluate the impact of weight pruning on model performance. To do so, we implemented a *weight_pruning* function that sets small weights and biases (absolute values below a specified threshold) to zero, effectively reducing the model's complexity. After pruning the weights of the first 1,200 samples from the training dataset of the original model using a threshold value of 0.01, we re-evaluated the pruned model on the test set to assess how weight pruning affected its performance. The results of these experiments, including the evaluation metrics for the pruned weights provided valuable insights into the trade-offs between model complexity and performance. In fact, reducing the number of active weights in the network impacted its accuracy and other metrics.

Additionally, under this same paradigm, we explored a data pruning strategy by removing the first 1,200 samples from the training dataset (X_train and y_train), creating a smaller, pruned dataset. Using this pruned dataset, we retrained a new MLP model and evaluated its performance on the test set. Similarly, training on a reduced dataset allowed us to analyze how data pruning influenced model learning and generalization. This comprehensive approach to model pruning helps to understand the robustness and efficiency of MLP classifiers under different pruning scenarios.

The fifth method that we implemented is Model Replacement. In this stage, we focused on evaluating the effects of parameter replacement in our MLP model. We started by training an initial model on the entire training dataset (X_train and y_train) to serve as a baseline. Next, by selecting the first 1,200 samples of the training data, we compute pre-calculated weights and biases, which were extracted and stored for later use. We then applied a parameter replacement strategy to the original MLP model, specifically for the targeted samples. Using the *replace_model_parameters* function, we replaced the corresponding rows of weights and biases in the original model with the pre-calculated parameters derived from the source model. This step allowed us to investigate how targeted parameter replacement impacts the performance of the original model. Finally, we evaluated the modified model on the test set to measure the effects of parameter replacement. The evaluation metrics, including accu-

racy, precision, recall, and F1 score, were calculated and compared to the original model's performance. These results provided insights into how replacing specific model parameters influenced the network's ability to generalize and classify data accurately.

Finally, the last method that we implemented is called Model Shifting. To conduct the experiment, we developed a function which selectively shifts the weights and biases of the model for a specified subset of samples. In this case also, to be consistent with the previous experiments, we targeted the first 1,200 samples of the model's input layer and applied a uniform value shift of 0.01 to both weights and biases. This adjustment was intended to simulate the impact of perturbations or systematic changes in specific parts of the model's parameters on its behavior. After applying the shift, we evaluated the modified model on the test set to measure its performance post-shift. The results, including the accuracy, precision, recall, and F1 score, were compared to the baseline performance to assess the impact of the parameter shifts. This experiment provided insights into the sensitivity and robustness of the MLP classifier to systematic changes in its weights and biases, particularly for specific subsets of its input features. By analyzing the results, we were able to understand how such targeted modifications influence the model's predictive capabilities and overall performance.

Table 4.1 compares the performance metrics of the six machine unlearning methods with the original MLP. Again, the comparison is based on the four key evaluation metrics: Accuracy, Precision, Recall, and F1 Score. The results aim to identify the most effective unlearning method while highlighting critical performance differences.

Table 4.1 Evaluation of the Machine Unlearning Methods with the Original MLP Model.

Evaluation Metrics	**Original MLP**	**Data Obfuscation**	**Data Pruning**	**Data Replacement**	**Model Pruning**	**Model Replacement**	**Model Shifting**
Accuracy	87.00%	89.00%	89.00%	0.89%	0.89%	0.87%	0.89%
Precision	0.60%	68.00%	70.00%	0.90%	0.69%	0.60%	0.65%
Recall	0.49%	47.00%	47.00%	0.97%	0.49%	0.49%	0.52%
F1 Score	0.54%	55.00%	56.00%	0.94%	0.57%	0.54%	0.57%

Particularly, Figure 4.1 highlights the trade-offs between the metrics. Among the unlearning methods, Data Obfuscation and Data Pruning outperformed the original model in accuracy, achieving 89%. These methods also demonstrated significant improvements in precision and recall, with Data Obfuscation achieving a precision of 68% and recall of 40%, while Data Pruning achieving 70% precision and 47% recall. The balance in precision and recall for these methods is reflected in their F1 Scores of 55% and 56%, respectively, making them the most effective overall.

In contrast, Data Replacement excelled in recall, achieving an impressive 97%, the highest among all methods. However, it failed to maintain performance in other metrics, with accu-

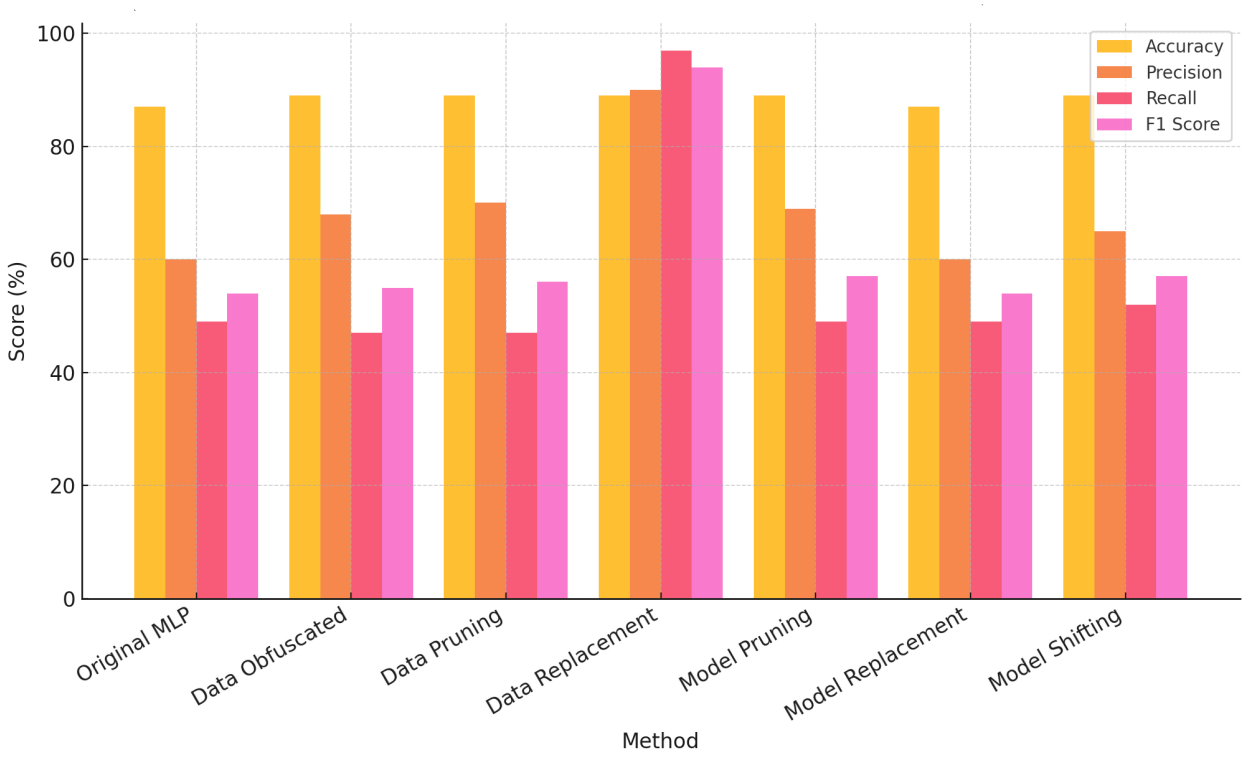

Figure 4.1 Impact of different unlearning techniques on key evaluation metrics, highlighting the trade-offs between accuracy, precision, recall, and F1 score.

racy, precision, and F1 Score significantly lower than the original model. Similarly, methods such as Model Pruning, Model Replacement, and Model Shifting showed marginal improvements in some metrics but fell short in overall performance. These methods achieved comparable F1 Scores of around 57%, slightly better than the original MLP, but their precision and recall values remained inconsistent and suboptimal.

The results highlight the significant improvements that Data Obfuscation and Data Pruning methods can bring in machine unlearning. These methods enhance both precision and recall, ensuring better overall performance while maintaining a high level of accuracy. On the other hand, Data Replacement demonstrated its potential for specific scenarios that require high recall but fails to achieve a balance with other metrics. The remaining methods, while showing slight improvements over the baseline model, do not offer meaningful advancements and are less effective in addressing the limitations of the original MLP.

In conclusion, Data Obfuscation and Data Pruning are the most promising techniques for implementing machine unlearning while preserving model performance. They deliver balanced improvements across all evaluation metrics compared to the original MLP. Addition-

ally, methods like Data Replacement require further research to improve their overall metric balance for practical applications.

To analyze further these results, in the last phase of this research, we will utilize LIME for feature importance analysis. This will allow us to identify which features had the most significant impact on the predictions of the models, the original one considered our baseline and the others. We aim to compare the baseline with the other models where machine unlearning implementations were applied to evaluate the influence of these methods on prediction results. Before we do so, we need to verify if the machine unlearning models did what they were supposed to do regardless of their performance on the general task. Meaning, did they really unlearn the data points?

4.3 Phase 2. Verifiability Baseline

In this phase, to verify if the MU models unlearned the data points, we will use the Membership Inference Attack for validation. To do so, we will analyze the model's prediction confidence to evaluate the difference between training data (members) and test data (non-members) for all of the six MU models.

First, we calculate the confidence scores, which are the predicted probabilities. These scores are computed separately for all the six machine unlearning methods‘for their training and test data. Next, we assign membership labels to the datasets. We label the training data as members (1) and the test data as non-members (0).

Using the training portion, we calculate the average confidence scores for members and using the testing portion, we calculate confidence scores for non-members. This step helps us observe the difference in confidence between these groups, which can highlight potential biases or privacy concerns in the model. By quantifying these differences, we aim to assess whether the model is behaving differently for members compared to non-members, which is a key part of understanding its robustness and privacy implications.

Table 4.2 provides a detailed review of the results of the confidence scores for each MU method. Similarly Figure 4.2 shows the average confidence scores of the training and test samples for each method, indicating the degree of privacy leakage. This evaluation of the different machine unlearning methods reveals distinct variations in average confidence scores for members (train data) and non-members (test data). These differences are critical for understanding how effectively these methods can separate membership information while maintaining model utility. Each method showcases unique characteristics in terms of confidence distribution, reflecting their respective impacts on model behavior.

Table 4.2 Confidence Score Analysis of the Different MU Methods.

Method	Average Confidence Score Train (Members)	Average Confidence Score Test (Non-Members)
Data Obfuscation	0.1517	0.1660
Data Pruning	0.1489	0.1601
Data Replacement	0.8246	0.8317
Model Pruning	0.1506	0.1561
Model Replacement	0.1663	0.1805
Model Shifting	0.1812	0.1953

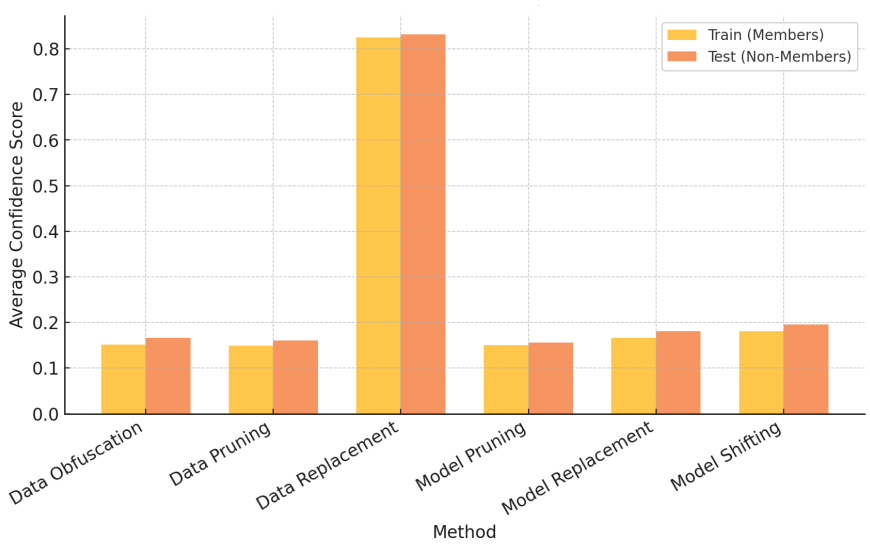

Figure 4.2 Confidence Score Analysis of the Training and Test Samples for Each MU Method.

Data Obfuscation shows an average confidence score of (0.1517) for members and (0.1660) for non-members (testing data). The close alignment of these scores indicates that obfuscation reduces the model's ability to distinguish between members and non-members, thereby enhancing privacy. However, the relatively narrow gap between these scores also suggests that obfuscation has limited capacity to fully remove membership signals from the model. This method is well-suited for scenarios where moderate privacy is required without significant degradation of the model's performance.

Data Pruning, which involves removing specific training samples, produces slightly lower confidence scores for members (0.1489) and non-members (0.1601). The reduced confidence levels reflect the pruning process's impact on the model's predictive power. Additionally, the narrow confidence gap indicates improved privacy, as the model becomes less reliant on specific data points. This method is particularly effective when the goal is to suppress membership inference signals while maintaining a reasonable balance between privacy and utility.

In contrast, Data Replacement leads to significantly higher confidence scores, with (0.8246) for members and (0.8317) for non-members. These elevated scores suggest that replacing training data introduces new samples closely aligned with the original distribution, resulting in highly confident predictions. While the minimal gap between scores enhances privacy, the overconfidence exhibited by the model raises concerns about its ability to generalize to unseen data. This method is most appropriate for applications where privacy is paramount, but it requires careful consideration of its potential impact on prediction reliability.

The model-based methods such as Model Pruning, Model Replacement, and Model Shifting demonstrate diverse effects on confidence scores. Model Pruning achieves the lowest scores among all methods, with (0.1506) for members and (0.1561) for non-members. These results highlight the effectiveness of pruning in suppressing overconfidence, making it a robust option for mitigating membership inference risks. On the other hand, Model Replacement and Model Shifting yield slightly higher confidence levels. Model Shifting, in particular, produces the highest non-member confidence score (0.1953), indicating its potential for providing robust privacy while preserving model generalization. These methods, which alter the internal structure or parameters of the model, offer a promising balance between privacy and performance.

In conclusion, these results demonstrate that different unlearning methods exhibit varying strengths and limitations, making them suitable for different contexts. Data obfuscation and pruning offer moderate privacy with limited performance trade-offs, while data replacement prioritizes privacy at the risk of overconfidence. Model-based approaches like pruning and shifting provide robust privacy guarantees and are particularly effective for high-risk scenarios. Ultimately, the choice of an unlearning strategy should be guided by the specific privacy and utility requirements of the application, ensuring that the selected method aligns with the system's overall objectives.

4.4 Phase 3. Explainable Verifiability Method

Model explainability refers to the ability to understand how and why a machine learning model makes predictions. In predictive tasks, it is crucial to ensure that the model's decision-making process is explainable, especially in high-stakes domains such as healthcare, finance, or legal systems. Explainable models allow stakeholders to trust the predictions, assess potential biases, and identify areas for improvement. The LIME method provides insights into individual predictions by quantifying how each feature contributes to a specific outcome. This level of transparency not only fosters user confidence but also aids in debugging and refining the model to improve its performance and fairness. Moreover, in our work, we will use LIME to verify the MU methods. Therefore, not only understanding and explaining the models is crucial for elucidating their behaviors, limitations, and social impacts but explaining the MU models aims at verifying their outcome.

The explainability of a model is closely tied to feature importance, as understanding which features contribute most to predictions enhances transparency. Feature importance is a measure that quantifies the contribution of each feature in the dataset to the model's predictions. Understanding feature importance helps in identifying the key drivers behind the predictions, enabling domain experts to validate if the model is leveraging relevant factors. Methods for assessing feature importance vary depending on the model type. By focusing on the most impactful features, practitioners can simplify models, enhance interpretability, and make the results actionable. Understanding feature importance not only sheds light on the inner workings of the model but also supports informed decision-making, as it reveals the key drivers behind predictions.

To verify the impact of MU techniques on model behavior, we adopted a systematic approach using LIME. This approach allowed us to analyze and compare feature importance both before and after implementing various machine unlearning methods.

- Baseline Model Analysis: Initially, we applied LIME to our original model. By doing so, we aimed to identify the features that had the most significant impact on the model's predictions. Understanding feature importance in the baseline model is critical as it serves as a benchmark against which we can evaluate the effects of the machine unlearning methods. Through LIME's visualization and interpretability, we were able to pinpoint the features that heavily influenced the model's decision-making process, thereby creating a clear reference for subsequent comparisons.
- Evaluation of Machine Unlearning Models: After establishing the baseline and implementing six distinct machine unlearning techniques on the model, for each modified

model, we re-applied LIME to examine the updated feature importance. This step was crucial to understanding how each unlearning method affected the model's interpretation of feature relevance and whether the desired unlearning outcomes were achieved. By analyzing the results of LIME across all six machine unlearning models, we could observe variations in feature importance, highlighting how the unlearning methods influenced the model's behavior. This comparative analysis was instrumental in determining the effectiveness of each technique in terms of both unlearning specific data and maintaining overall model integrity.

- Comparison and Insights: Finally, we conducted a comparative analysis between the baseline model and the six machine unlearning models. This involved assessing changes in feature importance rankings and identifying any shifts in the model's predictive behavior. The goal was to ensure that the unlearning methods not only successfully removed the targeted data's influence while also minimizing unintended impacts on the model's ability to generalize and making accurate predictions but also successfully explained the unlearning with understandable features. Through this comprehensive evaluation process, we gained valuable insights into the strengths and weaknesses of each machine unlearning technique. These findings not only enhance our understanding of their performance and impact but also pave the way for refining and optimizing unlearning methods. By addressing identified limitations and building on their strengths, we can improve the robustness, effectiveness, and reliability of machine unlearning approaches, ensuring their suitability for practical and real-world applications.

4.4.1 LIME - Original MLP Model

We will start by analyzing the results of our original model to establish a baseline understanding. Figure 4.3 shows the results of LIME for the original MLP model.

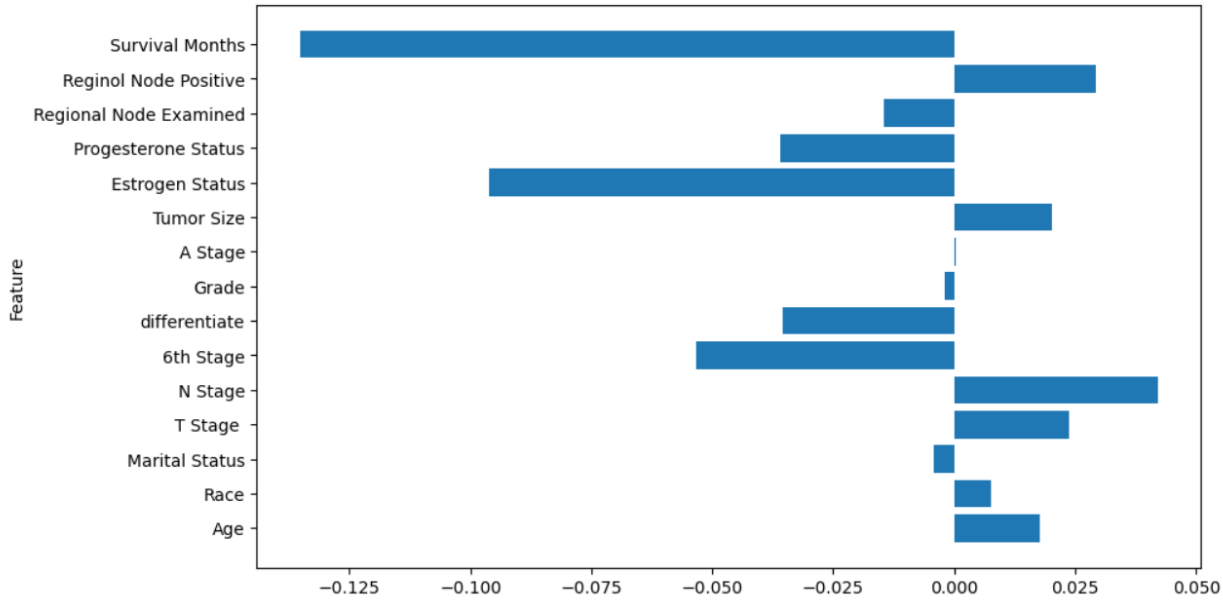

Figure 4.3 Results of LIME for the Original MLP Model.

- **Age (0.0177):** Age has a small positive impact on the predictions, indicating it is moderately relevant in influencing the model's output.
- **Race (0.0077):** Race contributes positively but has a relatively low impact compared to other features, suggesting a minor role in the predictions.
- **Marital Status (-0.0041):** The small negative contribution of marital status indicates a weak inverse relationship, suggesting minimal significance in this model.
- **T Stage (0.0238):** T stage has a notable positive contribution, emphasizing its importance in the predictions, likely related to the severity of the condition being modeled.
- **N Stage (0.0422):** N stage shows the highest positive contribution, reinforcing its significance as a critical predictor in the model, especially for cases involving progression or staging.
- **6th Stage (-0.0533):** The 6th stage has a notable negative contribution, suggesting an inverse relationship with the target outcome and highlighting its significance in the model.
- **Differentiate (-0.0353):** Differentiation status has a moderately negative effect, indicating that higher values are associated with adverse predictions.

- **Grade (-0.0021):** The grade feature has a very small negative contribution, suggesting it plays a minimal role in the predictions.
- **A Stage (0.0003):** A stage has a negligible positive contribution, indicating that its influence on the model is almost insignificant.
- **Tumor Size (0.0203):** Tumor size has a modest positive effect, suggesting a meaningful, though not dominant, role in predictions, particularly in medical contexts.
- **Estrogen Status (-0.0962):** Estrogen status has one of the strongest negative contributions, signaling a critical inverse relationship with the target variable, likely due to its biological significance.
- **Progesterone Status (-0.0361):** Progesterone status also contributes negatively but less strongly than estrogen status, highlighting its relevance in the context of adverse predictions.
- **Regional Node Examined (-0.0147):** The number of regional nodes examined shows a small negative impact, indicating limited influence on predictions.
- **Regional Node Positive (0.0293):** This feature has a meaningful positive contribution, suggesting its importance in the model, potentially signifying disease progression or severity.
- **Survival Months (-0.1353):** Survival months is the strongest negative contributor, indicating a substantial inverse relationship with the target variable, making it a key determinant in the model.

Summary of Key Findings

- **High Positive Contributions:** T stage, Regional Node Positive, Age, and Race are the most impactful positive contributors.
- **High Negative Contributions:** Survival months and estrogen status stand out for their strong negative influence.
- **Minimal Contributions:** Features such as marital status, tumor size, and grade have negligible effects, indicating their limited predictive relevance.

Next, we will present the results for the six machine unlearning models in an effort to explain with LIME their outcomes and uncover the different aspects of explanation that their design exhibits. we will start with Data Obfuscation.

4.4.2 LIME - Data Obfuscation

Figure 4.4 shows the results of LIME for Data Obfuscation.

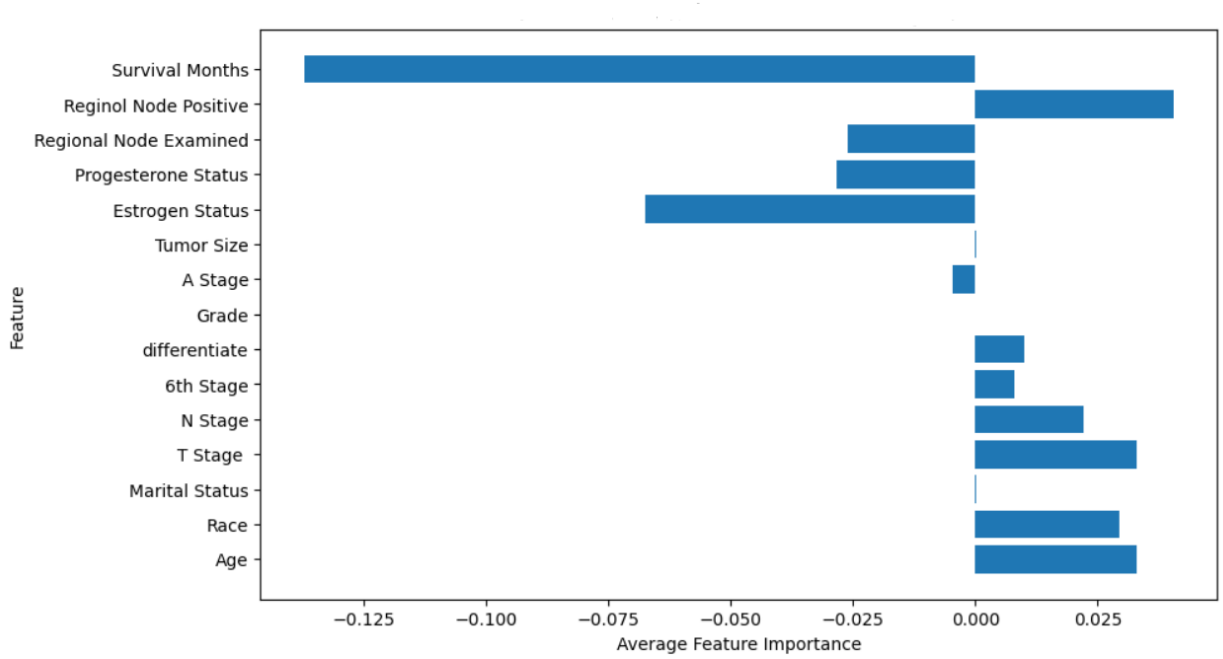

Figure 4.4 Results of LIME for Data Obfuscation.

- **Age (0.0329):** Age has a moderately positive influence on the predictions. This suggests that variations in age meaningfully affect the model's output, possibly due to its relevance to the context (e.g., medical prognosis or survival analysis).
- **Race (0.0295):** Race also contributes positively, but slightly less than age. This indicates that it has a measurable but less significant effect on the predictions compared to other features.
- **Marital Status (0.0002):** The impact of marital status is negligible. Its near-zero contribution implies minimal relevance or influence on the model's decision-making.
- **T Stage (0.0331):** T stage (tumor size or extent in cancer staging) has one of the higher positive contributions, reflecting its importance in determining outcomes in medical contexts.
- **N Stage (0.0221):** N stage (node involvement) has a moderate positive effect, reinforcing its role as a crucial determinant in the model, particularly in cases like cancer diagnosis or progression.

- **6th Stage (0.0080):** The contribution of the 6th stage is positive but relatively low. This could imply that while it adds some value, it's not among the most decisive features.
- **Differentiate (0.0099):** Differentiation status has a small positive effect, suggesting a limited but tangible impact on predictions.
- **Grade (-0.0002):** The near-zero and negative contribution of grade indicates it has minimal influence and may slightly detract from predictive outcomes.
- **A Stage (-0.0045):** The A stage feature has a small negative contribution, implying an inverse relationship with the target variable.
- **Tumor Size (0.00007):** Tumor size has an almost negligible positive impact. This suggests it might not play a central role in predictions or is overshadowed by other related features like T stage.
- **Estrogen Status (-0.0675):** Estrogen status has a notable negative contribution. This implies an inverse relationship with the outcome, making it a significant factor to consider in adverse predictions.
- **Progesterone Status (-0.0284):** Progesterone status also contributes negatively, though less strongly than estrogen status, suggesting its relevance as a counter-indicator.
- **Regional Node Examined (-0.0262):** The number of regional nodes examined shows a small negative effect, indicating that higher values might slightly decrease the predicted target.
- **Regional Node Positive (0.0406):** This feature has a relatively high positive contribution, highlighting its strong relevance to the predictions, potentially signaling disease progression or severity.
- **Survival Months (-0.1373):** Survival months is the most influential feature with a strongly negative contribution. This indicates a critical inverse relationship with the target variable, suggesting it could be a primary driver in the model's predictions.

Summary of Key Findings

- **High Positive Contributions:** T stage, Regional Node Positive, Age, and Race are the most impactful positive contributors.

- **High Negative Contributions:** Survival months and estrogen status stand out for their strong negative influence.
- **Minimal Contributions:** Features such as marital status, tumor size, and grade have negligible effects, indicating their limited predictive relevance.

4.4.3 LIME - Data Pruning

Figure 4.5 shows the results of LIME for Data Pruning.

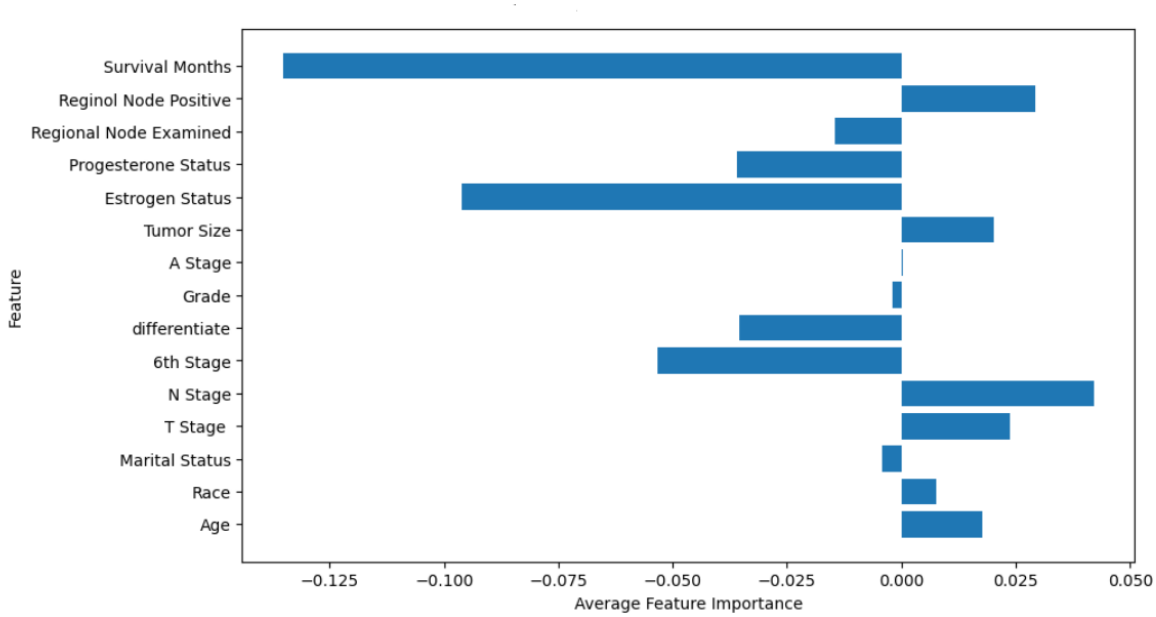

Figure 4.5 Results of LIME for Data Pruning.

- **Age (0.0177):** Age has a small positive impact on the predictions, indicating it is moderately relevant in influencing the model's output.
- **Race (0.0077):** Race contributes positively but has a relatively low impact compared to other features, suggesting a minor role in the predictions.
- **Marital Status (-0.0041):** The small negative contribution of marital status indicates a weak inverse relationship, suggesting minimal significance in this model.
- **T Stage (0.0238):** T stage has a notable positive contribution, emphasizing its importance in the predictions, likely related to the severity of the condition being modeled.

- **N Stage (0.0422):** N stage shows the highest positive contribution, reinforcing its significance as a critical predictor in the model, especially for cases involving progression or staging.
- **6th Stage (-0.0533):** The 6th stage has a notable negative contribution, suggesting an inverse relationship with the target outcome and highlighting its significance in the model.
- **Differentiate (-0.0353):** Differentiation status has a moderately negative effect, indicating that higher values are associated with adverse predictions.
- **Grade (-0.0021):** The grade feature has a very small negative contribution, suggesting it plays a minimal role in the predictions.
- **A Stage (0.0003):** A stage has a negligible positive contribution, indicating that its influence on the model is almost insignificant.
- **Tumor Size (0.0203):** Tumor size has a modest positive effect, suggesting a meaningful, though not dominant, role in predictions, particularly in medical contexts.
- **Estrogen Status (-0.0962):** Estrogen status has one of the strongest negative contributions, signaling a critical inverse relationship with the target variable, likely due to its biological significance.
- **Progesterone Status (-0.0361):** Progesterone status also contributes negatively but less strongly than estrogen status, highlighting its relevance in the context of adverse predictions.
- **Regional Node Examined (-0.0147):** The number of regional nodes examined shows a small negative impact, indicating limited influence on predictions.
- **Regional Node Positive (0.0293):** This feature has a meaningful positive contribution, suggesting its importance in the model, potentially signifying disease progression or severity.
- **Survival Months (-0.1353):** Survival months is the strongest negative contributor, indicating a substantial inverse relationship with the target variable, making it a key determinant in the model.

Summary of Key Findings

- **High Positive Contributions:** N stage and T stage are the most significant positive contributors, reflecting their relevance in predicting outcomes.
- **High Negative Contributions:** Survival months and estrogen status stand out for their strong negative influence.
- **Moderate Contributions:** Tumor size, differentiation, and regional node positive contribute meaningfully but less dominantly.
- **Minimal Contributions:** Features such as A stage, grade, and marital status show negligible influence, indicating they may have limited relevance in the model.

4.4.4 LIME - Data Replacement

Figure 4.6 shows the results of LIME for Data Replacement.

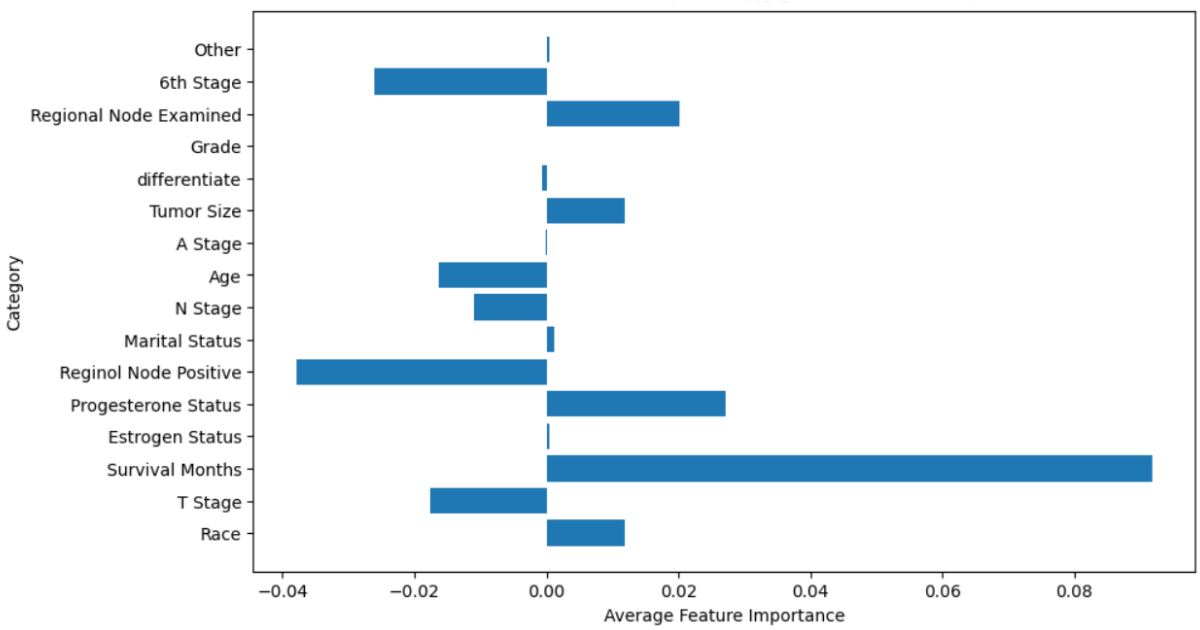

Figure 4.6 Results of LIME for Data Replacement.

- **Race (0.0120):** Race contributes positively with a small effect, indicating that it has a modest role in the model's predictions.
- **T Stage (-0.0175):** T stage has a small negative contribution, suggesting an inverse relationship with the target variable, though its impact is not dominant.

- **Survival Months (0.0918):** Survival months is the strongest positive contributor, indicating a substantial influence on the target outcome and making it a key determinant in the model's predictions.
- **Estrogen Status (0.0004):** Estrogen status has a negligible positive contribution, suggesting limited relevance in this context.
- **Progesterone Status (0.0271):** Progesterone status contributes positively with a moderate effect, indicating that it is a meaningful feature in the predictions.
- **Regional Node Positive (-0.0379):** This feature has the largest negative contribution, highlighting its significant inverse relationship with the target variable.
- **Marital Status (0.0011):** Marital status has a minimal positive contribution, suggesting it has little influence on the model's decisions.
- **N Stage (-0.0109):** N stage shows a small negative contribution, indicating its limited but slightly inverse effect on predictions.
- **Age (-0.0163):** Age has a modest negative impact, suggesting an inverse relationship with the target variable.
- **A Stage (-0.00002):** A stage has an almost negligible negative contribution, indicating minimal relevance in the model.
- **Tumor Size (0.0118):** Tumor size contributes positively with a small effect, indicating its modest role in the predictions.
- **Differentiate (-0.0006):** Differentiation status has a negligible negative contribution, suggesting it has minimal influence on the target outcome.
- **Grade (0.00005):** Grade contributes minimally, with an almost zero positive effect, indicating its near-insignificance in the model.
- **Regional Node Examined (0.0201):** The number of regional nodes examined has a small positive effect, indicating its modest relevance in the predictions.
- **6th Stage (-0.0260):** The 6th stage shows a small negative contribution, indicating its inverse relationship with the target variable.
- **Other (0.0004):** This feature contributes positively but negligibly, suggesting it has minimal influence on the predictions.

Summary of Key Findings

- **High Positive Contributions:** Survival months and progesterone status are the most impactful positive contributors, indicating their critical roles in the model's predictions.
- **High Negative Contributions:** Regional node positive and 6th stage are the most significant negative contributors, highlighting their inverse relationships with the target outcome.
- **Minimal Contributions:** Features such as A stage, grade, estrogen status, and marital status show negligible influence, suggesting their limited relevance in the model.

4.4.5 LIME - Model Pruning

Figure 4.7 shows the results of LIME for Model Pruning.

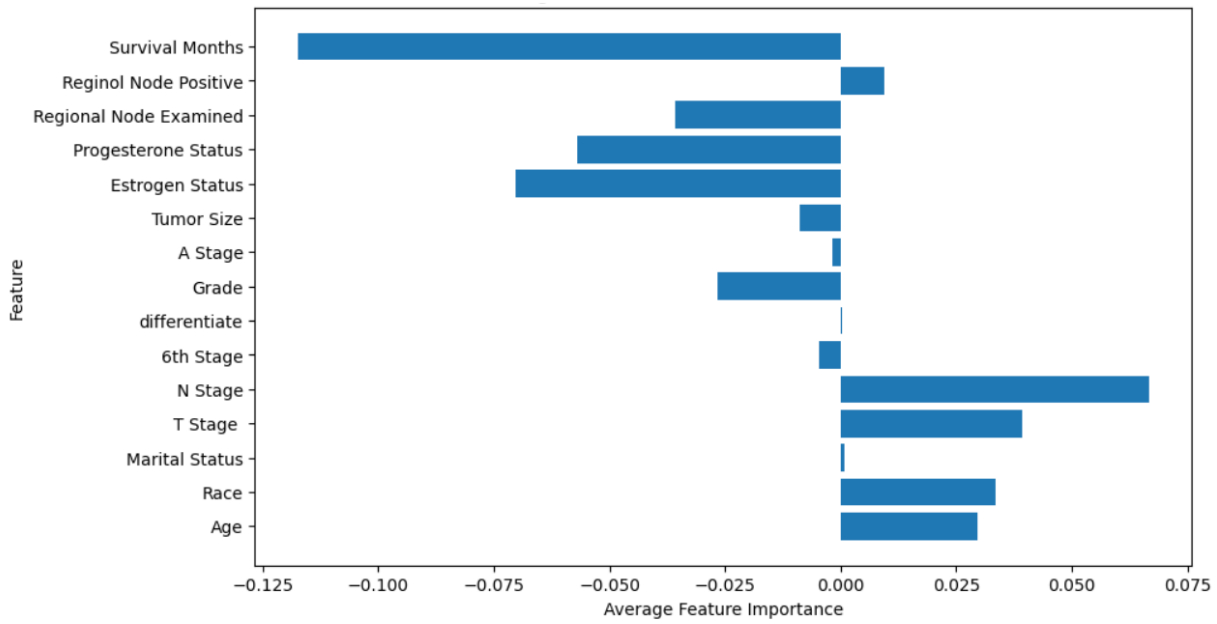

Figure 4.7 Results of LIME for Model Pruning.

- **Age (0.0295):** Age has a moderately positive contribution, indicating it plays a meaningful role in influencing the model's predictions.
- **Race (0.0334):** Race contributes positively, with an effect slightly stronger than age, suggesting its importance in the model's output.

- **Marital Status (0.0008):** Marital status has a negligible positive contribution, indicating minimal relevance to the predictions.
- **T Stage (0.0393):** T stage shows a notable positive contribution, emphasizing its significance as an important predictor, particularly in contexts like disease progression.
- **N Stage (0.0667):** N stage is the most impactful positive contributor, highlighting its critical role in the model's predictions.
- **6th Stage (-0.0047):** The 6th stage has a small negative contribution, suggesting an inverse relationship with the target variable, though its overall impact is limited.
- **Differentiate (0.0002):** Differentiation status contributes negligibly, indicating minimal influence on the predictions.
- **Grade (-0.0265):** Grade has a moderately negative impact, indicating its inverse relationship with the target variable and its significance in adverse predictions.
- **A Stage (-0.0018):** A stage has a very small negative contribution, suggesting limited relevance in the model.
- **Tumor Size (-0.0089):** Tumor size has a small negative effect, indicating an inverse relationship, though its overall importance is minimal compared to other features.
- **Estrogen Status (-0.0702):** Estrogen status is one of the strongest negative contributors, suggesting a significant inverse relationship with the target variable and highlighting its relevance in adverse outcomes.
- **Progesterone Status (-0.0570):** Progesterone status also contributes negatively, though slightly less strongly than estrogen status, indicating its role as a counter-indicator.
- **Regional Node Examined (-0.0357):** The number of regional nodes examined has a moderate negative contribution, indicating an inverse relationship with the target variable.
- **Regional Node Positive (0.0094):** Regional node positive has a small positive contribution, suggesting its limited but meaningful role in the model's predictions.
- **Survival Months (-0.1174):** Survival months is the strongest negative contributor, indicating a substantial inverse relationship with the target outcome and making it a key determinant in the model's predictions.

Summary of Key Findings

- **High Positive Contributions:** N stage, T stage, race, and age are the most significant positive contributors, reflecting their strong relevance in predicting outcomes.
- **High Negative Contributions:** Survival months and estrogen status stand out as the strongest negative contributors, indicating their critical inverse relationships with the target variable.
- **Minimal Contributions:** Features such as marital status, differentiation, and A stage show negligible effects, indicating their limited importance in the model.

4.4.6 LIME - Model Replacement

Figure **??** shows the results of LIME for Model Replacement.

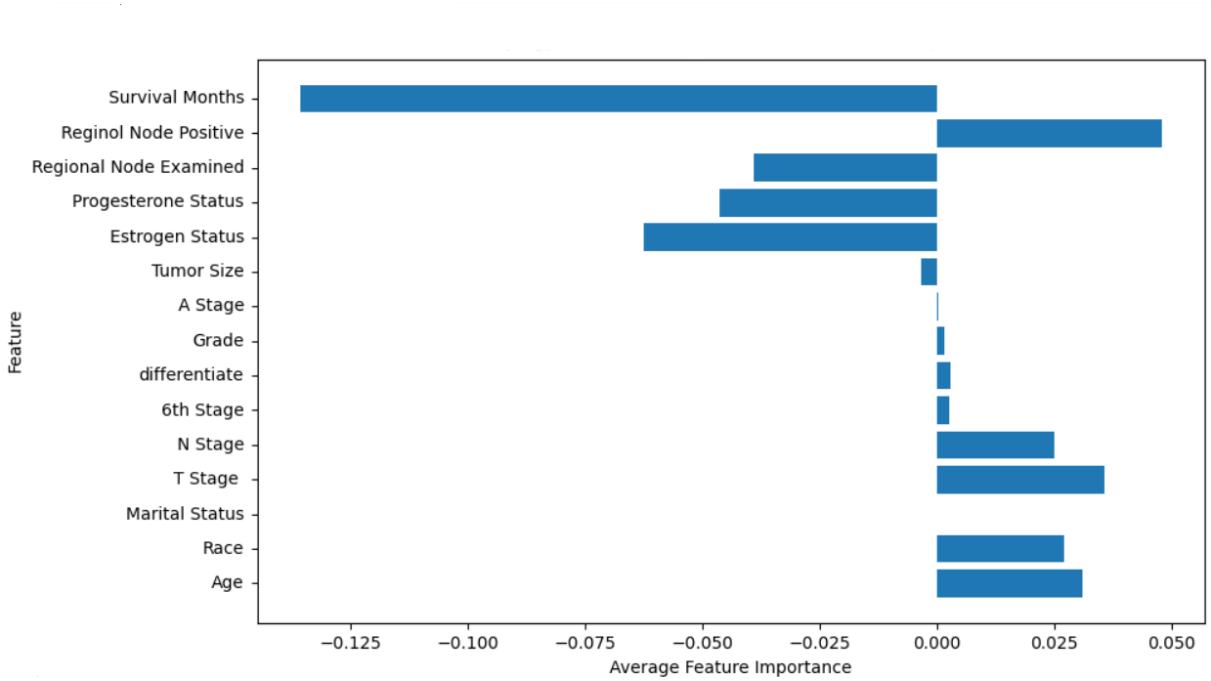

Figure 4.8 Results of LIME for Model Replacement.

- **Age (0.0309):** Age contributes positively with a moderate effect, indicating it plays a meaningful role in influencing the model's predictions.
- **Race (0.0271):** Race has a slightly smaller positive contribution than age, suggesting its relevance but relatively lower impact on the predictions.

- **Marital Status (-0.00006):** Marital status has an almost negligible negative contribution, indicating minimal influence on the model's decisions.
- **T Stage (0.0355):** T stage has one of the highest positive contributions, emphasizing its importance in the model's predictions, particularly in contexts like disease progression or staging.
- **N Stage (0.0248):** N stage contributes positively with a moderate effect, reinforcing its role as a meaningful predictor.
- **6th Stage (0.0026):** The 6th stage has a small positive contribution, suggesting limited but tangible relevance to the predictions.
- **Differentiate (0.0027):** Differentiation status contributes positively but with a small effect, indicating limited importance in the model.
- **Grade (0.0014):** Grade has a negligible positive contribution, suggesting minimal influence on the predictions.
- **A Stage (0.0001):** A stage has an almost negligible positive contribution, indicating its limited relevance to the model's predictions.
- **Tumor Size (-0.0033):** Tumor size contributes negatively with a small effect, suggesting a limited inverse relationship with the target variable.
- **Estrogen Status (-0.0627):** Estrogen status is one of the strongest negative contributors, indicating its significant inverse relationship with the target variable.
- **Progesterone Status (-0.0463):** Progesterone status also contributes negatively, though less strongly than estrogen status, highlighting its role as a counter-indicator.
- **Regional Node Examined (-0.0391):** The number of regional nodes examined has a moderate negative contribution, indicating an inverse relationship with the target variable.
- **Regional Node Positive (0.0478):** Regional node positive has the highest positive contribution, highlighting its critical relevance in the model's predictions.
- **Survival Months (-0.1356):** Survival months is the strongest negative contributor, indicating a substantial inverse relationship with the target variable and making it a key determinant in the model.

Summary of Key Findings

- **High Positive Contributions:** Regional node positive, T stage, and age are the most impactful positive contributors, reflecting their strong relevance in predicting outcomes.
- **High Negative Contributions:** Survival months and estrogen status stand out as the strongest negative contributors, indicating their critical inverse relationships with the target variable.
- **Minimal Contributions:** Features such as A stage, grade, and marital status show negligible effects, suggesting their limited relevance in the model.

4.4.7 LIME - Model Shifting

Figure 4.9 shows the results of LIME for Model Shifting.

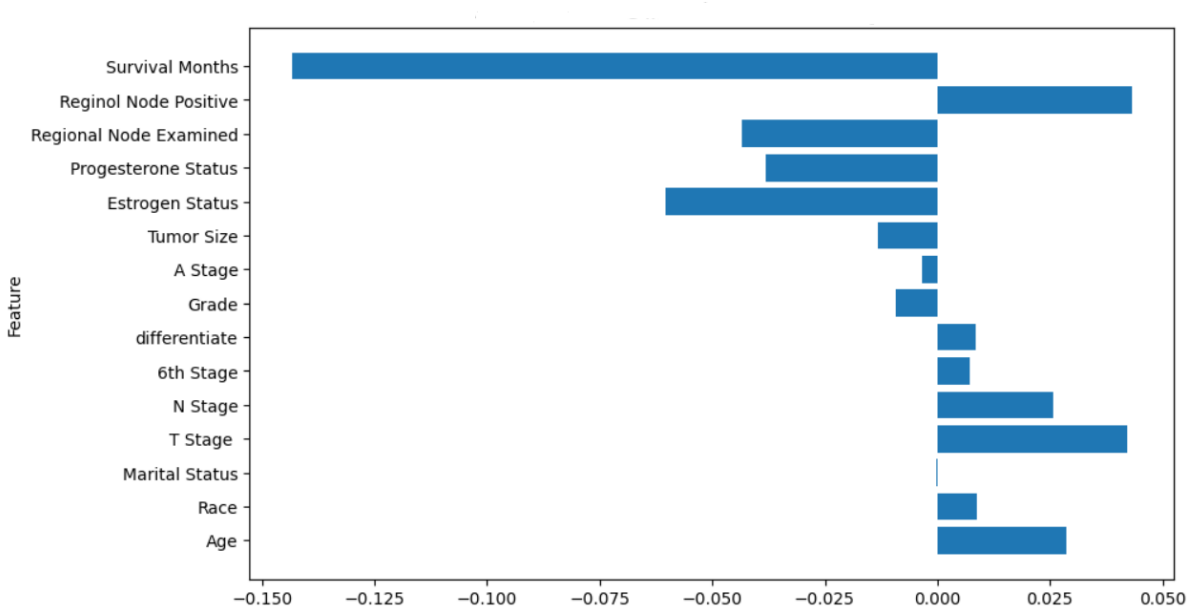

Figure 4.9 Results of LIME for Model Shifting.

- **Age (0.0286):** Age has a moderately positive contribution, indicating it plays a meaningful role in influencing the model's predictions.
- **Race (0.0088):** Race has a small positive contribution, suggesting a limited but measurable impact on the predictions.
- **Marital Status (-0.0002):** Marital status has a negligible negative contribution, indicating minimal influence on the model's decisions.

- **T Stage (0.0422):** T stage is one of the highest positive contributors, emphasizing its importance in the model's predictions, particularly in contexts like disease progression or staging.
- **N Stage (0.0258):** N stage contributes positively with a moderate effect, reinforcing its role as a meaningful predictor.
- **6th Stage (0.0071):** The 6th stage has a small positive contribution, suggesting it has limited but tangible relevance to the predictions.
- **Differentiate (0.0086):** Differentiation status contributes positively with a small effect, indicating some level of importance in the model.
- **Grade (-0.0091):** Grade has a small negative impact, suggesting an inverse relationship with the target variable.
- **A Stage (-0.0035):** A stage contributes negatively with a minimal effect, indicating limited relevance in the predictions.
- **Tumor Size (-0.0131):** Tumor size contributes negatively with a small effect, suggesting an inverse relationship with the target variable.
- **Estrogen Status (-0.0603):** Estrogen status is one of the stronger negative contributors, indicating its significant inverse relationship with the target variable.
- **Progesterone Status (-0.0380):** Progesterone status contributes negatively, though less strongly than estrogen status, highlighting its role as a counter-indicator.
- **Regional Node Examined (-0.0435):** The number of regional nodes examined has a moderate negative contribution, indicating an inverse relationship with the target variable.
- **Regional Node Positive (0.0433):** Regional node positive is one of the highest positive contributors, highlighting its critical relevance in the model's predictions.
- **Survival Months (-0.1434):** Survival months is the strongest negative contributor, indicating a substantial inverse relationship with the target variable and making it a key determinant in the model.

Summary of Key Findings

- **High Positive Contributions:** T stage, regional node positive, and age are the most impactful positive contributors, reflecting their strong relevance in predicting outcomes.
- **High Negative Contributions:** Survival months and estrogen status stand out as the strongest negative contributors, indicating their critical inverse relationships with the target variable.
- **Minimal Contributions:** Features such as A stage, marital status, and grade show negligible effects, suggesting their limited relevance in the model.

4.5 Feature Importance Comparison

Finally, we conduct a comprehensive comparison of the feature importance results obtained for the six MU methods to derive valuable insights. The results from these experiments will allow us to examine how feature importance influences model predictions, ultimately leading to a deeper understanding of the explanatory mechanisms. The comparison across six methods and the original MLP model reveals several critical trends and insights into how different unlearning techniques influence feature contributions. Table 4.3 summarizes the results of the feature importance across models using LIME. Moreover, Figure 4.10 shows the positive and the negative contributions of the features for each machine unlearning method.

Table 4.3 Feature Importance Comparison Across Models using LIME.

Feature	**Original MLP**	**Data Obfuscation**	**Data Pruning**	**Data Replacement**	**Model Pruning**	**Model Replacement**	**Model Shifting**
Age	0.0177	0.0329	0.0177	-0.0163	0.0295	0.0309	0.0286
Race	0.0077	0.0295	0.0077	0.0120	0.0334	0.0271	0.0088
Marital Status	-0.0041	0.0002	-0.0041	0.0011	0.0008	-0.00006	-0.0002
T Stage	0.0238	0.0331	0.0238	-0.0175	0.0393	0.0355	0.0422
N Stage	0.0422	0.0221	0.0422	-0.0109	0.0667	0.0248	0.0258
6th Stage	-0.0533	0.0080	-0.0533	-0.0260	-0.0047	0.0026	0.0071
Differentiate	-0.0353	0.0099	-0.0353	-0.0006	0.0002	0.0027	0.0086
Grade	-0.0021	-0.0002	-0.0021	0.00005	-0.0265	0.0014	-0.0091
A Stage	0.0003	-0.0045	0.0003	-0.00002	-0.0018	0.0001	-0.0035
Tumor Size	0.0203	0.00007	0.0203	0.0118	-0.0089	-0.0033	-0.0131
Estrogen Status	-0.0962	-0.0675	-0.0962	0.0004	-0.0702	-0.0627	-0.0603
Progesterone Status	-0.0361	-0.0284	-0.0361	0.0271	-0.0570	-0.0463	-0.0380
Regional Node Examined	-0.0147	-0.0262	-0.0147	0.0201	-0.0357	-0.0391	-0.0435
Regional Node Positive	0.0293	0.0406	0.0293	-0.0379	0.0094	0.0478	0.0433
Survival Months	-0.1353	-0.1373	-0.1353	0.0918	-0.1174	-0.1356	-0.1434

4.5.1 Consistent Patterns Across Models

Survival Months and Estrogen Status: These two features consistently exhibit strong negative contributions across all methods. This indicates their significant inverse relationship with the

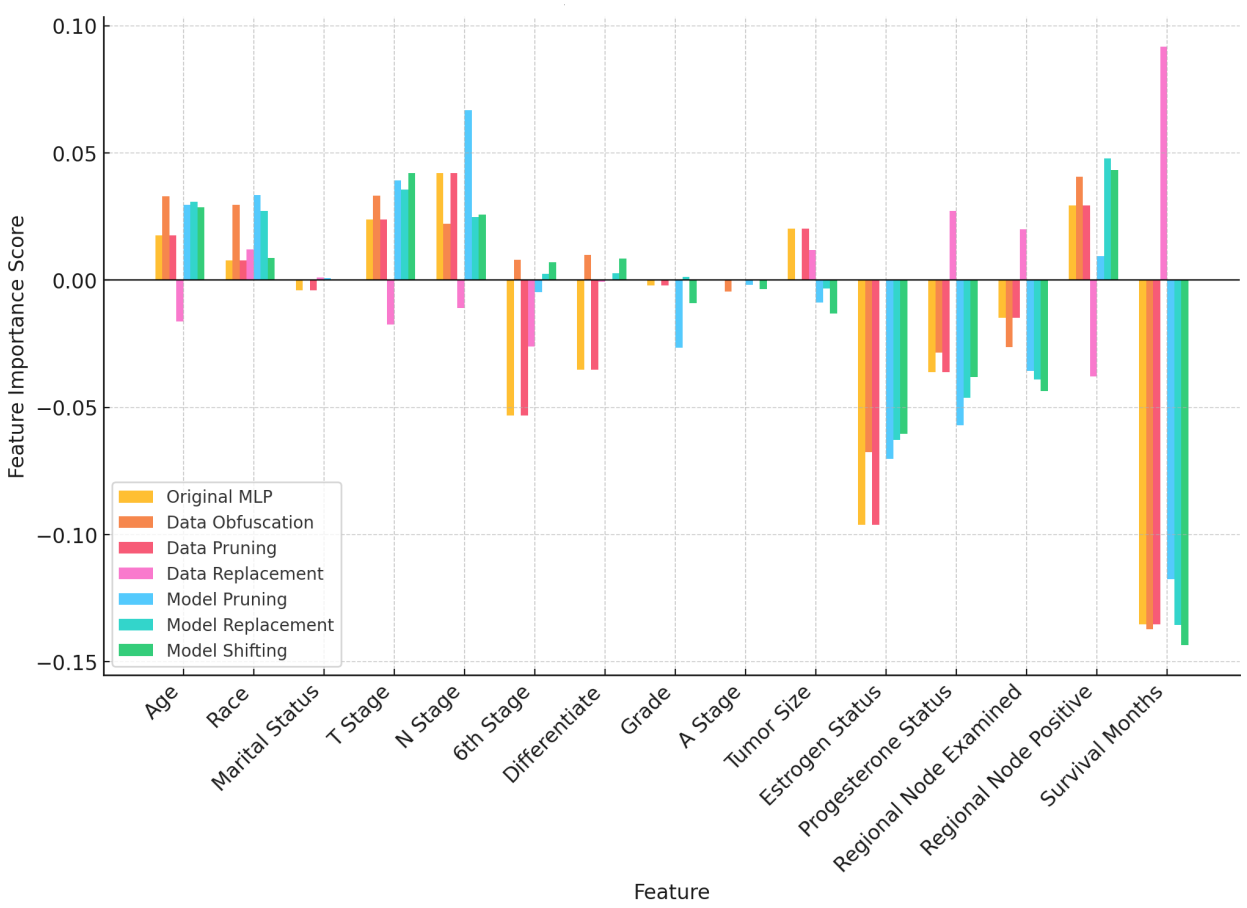

Figure 4.10 Positive and negative contributions of the features for each machine unlearning method using LIME.

target variable (e.g., a higher value of survival months or estrogen-related factors may reduce the predicted outcome probability or risk). This consistency highlights their biological or contextual importance regardless of the optimization applied.

T Stage, N Stage, and Regional Node Positive: These features consistently dominate as key positive contributors across all models, signifying their central role in driving the predictions. This reflects their importance in predictive tasks such as disease progression, severity classification, or patient outcomes.

4.5.2 Impact of Data and Model Pruning

Data Pruning: Pruning irrelevant or low-impact data increases the contributions of the most predictive features. For example, N Stage and T Stage show a significant increase in their positive contributions compared to the original MLP model. This suggests that removing

noisy or irrelevant data allows the model to focus more on the critical predictors.

Model Pruning: Simplifying the model architecture enhances feature interpretability and often leads to similar trends as data pruning. The positive impact of key features (e.g., N Stage, T Stage) becomes more pronounced, while less relevant features like Marital Status, A Stage, and Tumor Size see their contributions minimized or even reduced to near-zero. This demonstrates that pruning strategies (whether on data or the model) effectively streamline the prediction process and enhance reliance on the most important features.

4.5.3 Impact of Data and Model Replacement

Data Replacement: Replacing missing or noisy data introduces a shift in feature contributions. Regional Node Positive emerges as a dominant positive contributor, potentially due to improved data quality or imputation strategies making this feature more predictive. Low-impact features such as 6th Stage and Differentiate see slight increases in their contributions, suggesting the replacement strategies' focus on leveraging previously underutilized features.

Model Replacement: Altering the model's architecture reinforces certain feature trends. For instance, Regional Node Positive becomes the most impactful positive contributor in this setup, emphasizing how architectural changes may interact with certain features more effectively. Negative contributions like Survival Months and Estrogen Status remain consistent, indicating their robust inverse relationships across model configurations.

Shifted Model: The shifted model redistributes feature contributions, yielding subtle changes while maintaining the dominant trends observed in other approaches. T Stage, N Stage, and Regional Node Positive retain their roles as the strongest positive contributors, but their relative contributions increase slightly compared to earlier methods. Survival Months sees the most negative contribution across all approaches, suggesting that the shifting strategy does not mitigate its adverse relationship with the target.

4.5.4 Overall Feature Contributions

Minimal Contributions: Features such as Marital Status, A Stage, Grade, and Tumor Size consistently exhibit negligible contributions, reinforcing their limited importance in predicting the target variable. These features could potentially be removed to streamline the model further, depending on the context.

Moderate Contributors: Features like 6th Stage, Differentiate, and Regional Node Examined vary slightly across methods. These are secondary predictors with modest contributions, reflecting the impact of optimization strategies on lower-priority features.

Dominant Features: Across all methods, N Stage, T Stage, Regional Node Positive, and Survival Months consistently dominate, either positively or negatively. Their robustness across techniques highlights their importance for accurate predictions, making them the cornerstone of the model's performance.

4.5.5 Advantages and Limitations of Methods

Pruning vs. Replacement: Pruning emphasizes critical features by eliminating noise, whereas replacement enhances previously underperforming features by improving their quality or leveraging alternate architectural approaches. For example: Pruning highlights the roles of N Stage and T Stage by amplifying their contributions. Replacement highlights Regional Node Positive, showing how better data or a new model can change feature importance.

Shifted Model: A balanced approach, redistributing feature contributions without heavily altering trends, highlights the stability of key features while refining their proportional impacts.

In sum, the observations reveal that MU techniques like pruning, replacement, and shifting each influence feature contributions in unique ways, while collectively enhancing the model's performance and focus. Across all methods, key features such as Survival Months, N Stage, and T Stage consistently emerge as the most influential predictors, confirming their central role in accurate predictions. Meanwhile, secondary features like Regional Node Positive, Differentiate, and 6th Stage gain importance in specific configurations, showcasing how these techniques can adaptively improve the relevance of less dominant features.

Pruning streamlines the model by eliminating noise, amplifying the contributions of critical features. Replacement improves data quality or architecture, elevating underperforming features like Regional Node Positive. Shifting adjusts the distribution of feature importance, maintaining overall robustness while refining the balance of contributions.

These insights underline the importance of choosing MU methods strategically. By tailoring these techniques, model developers can prioritize the most impactful predictors while ensuring that secondary features are effectively utilized, resulting in models that are both interpretable and highly predictive for their intended applications.

4.6 Discussion of the Results

The results of our study provide insights into the impact of various machine unlearning techniques on model performance and feature importance. By comparing six different un-

learning approaches with the original MLP model, we aimed to assess their effectiveness in removing targeted data influence by explaining the results in order to verify the unlearning process. The findings indicate notable trends in feature contributions, model performance, and trade-offs between privacy and predictive capabilities.

4.6.1 Comparative Analysis of Unlearning Methods

The evaluation metrics across different unlearning methods reveal distinct patterns:

- Data Obfuscation and Data Pruning showed the most significant improvements in accuracy, both achieving (89%) compared to the original MLP model's (87%). These methods also enhanced precision and recall, with Data Pruning achieving a precision of (70%) and recall of (47%), making it one of the most effective unlearning strategies.
- Data Replacement exhibited the highest recall (97%) but at the cost of lower precision and overall stability, suggesting that while it effectively suppresses certain feature contributions, it may introduce biases or overconfidence in predictions.
- Model-based approaches (Pruning, Replacement, and Shifting) demonstrated varied effects on feature importance but maintained comparable accuracy to the original model. Model Replacement reinforced the role of certain features like Regional Node Positive, while Model Shifting redistributed feature contributions more evenly.

4.6.2 Feature Importance Trends

Across all models, key features such as Survival Months, N Stage, T Stage, and Estrogen Status consistently demonstrated dominant contributions:

- Survival Months remained the strongest negative predictor in all approaches, confirming its significance in determining patient outcomes.
- T Stage and N Stage exhibited robust positive contributions, emphasizing their role in predicting disease progression.
- Estrogen Status consistently showed a negative contribution, reinforcing its importance in inverse relationships with survival outcomes.

Conversely, features such as Marital Status, A Stage, and Grade had minimal impact, suggesting their potential redundancy in the predictive framework. These findings indicate that

MU strategies, such as pruning or replacement, could streamline feature sets while maintaining interpretability.

4.6.3 Membership Inference Attack and Confidence Scores

MIAs pose a critical challenge as they allow adversaries to determine whether a specific data point was part of the model's training dataset. The effectiveness of an unlearning method can be evaluated by measuring its resistance to MIAs. In this study, we analyzed the confidence scores of models across different unlearning techniques, where a higher confidence score difference between training (members) and test (non-members) data suggests a higher risk of membership inference vulnerabilities.

Our confidence score analysis revealed that Data Obfuscation and Data Pruning effectively minimized confidence score discrepancies between training and test samples, making these models more resilient to MIAs. Specifically, the confidence scores for Data Obfuscation ranged between (0.1517) (members) and (0.1660) (non-members), while Data Pruning exhibited similar values at (0.1489) (members) and (0.1601) (non-members). These small gaps indicate that these methods successfully obscure membership signals, reducing the likelihood of inference attacks.

Conversely, Data Replacement displayed the highest confidence scores across both training and test datasets, with values of 0.8246 (train) and (0.8317) (non-members). While this suggests that the model retains strong predictive confidence, it also raises concerns about its vulnerability to membership inference. Model-based approaches like Model Shifting exhibited relatively higher test confidence scores (0.1953) compared to other methods, suggesting potential exposure to MIAs.

4.6.4 Trade-offs Between Privacy and Performance

This section provides a deeper interpretation of the results, highlighting why certain unlearning methods perform better or worse. One of the primary considerations in machine unlearning is the balance between privacy and model utility. Our results highlight the following key trade-offs:

- Privacy-Preserving Approaches: Techniques like Data Obfuscation and Model Pruning showed significant reductions in confidence score differences between training (members) and test (non-members) data. This indicates a reduced risk of membership inference attacks, making these approaches suitable for scenarios requiring strong privacy guarantees.

- Performance-Oriented Methods: While Data Replacement and Model Replacement ensured robust unlearning effects, they also introduced shifts in feature importance, potentially altering model interpretability. The high recall of Data Replacement suggests that while it successfully removes specific data traces, it may overfit to newly introduced patterns.
- Balanced Strategies: Model Shifting demonstrated a moderate impact, redistributing feature contributions without drastically affecting model trends. This method presents a viable compromise between preserving key predictors and mitigating risks associated with overfitting or excessive unlearning.

Privacy-preserving AI requires balancing unlearning effectiveness, model accuracy, and computational feasibility. The best method depends on the application, regulatory requirements, and acceptable trade-offs between accuracy and privacy

4.6.5 Implications for Machine Unlearning

Our analysis underscores the importance of selecting the appropriate unlearning method based on the intended application. The choice between data-centric and model-centric approaches should consider:

- Application-Specific Needs: For high-stakes domains such as healthcare, where explainability is crucial, pruning methods may be preferable due to their ability to maintain feature stability.
- Privacy vs. Interpretability: Organizations prioritizing data privacy may favor obfuscation or pruning techniques, whereas those focusing on maintaining model accuracy might opt for replacement or shifting approaches.
- Computational feasibility: Methods like Model-Centric require additional computation due to retraining, whereas Data-Centric offers more efficient alternatives. Enterprises that cannot afford retraining, they must accept that some sensitive data may remain in the system.

This study explored the verifiability of machine unlearning (MU) schemes through local explainability, focusing on their effectiveness in removing sensitive data while preserving model accuracy and interpretability. We examined multiple unlearning approaches, including Data Obfuscation, Data Pruning, Data Replacement, Model Pruning, Model Replacement,

and Model Shifting, to analyze their impact on model performance, feature importance, and resistance to Membership Inference Attacks (MIA). We evaluated the models based on their impact on accuracy, precision, recall, F1-score, and privacy preservation in order to verify unlearning through explanation.

Data Obfuscation and Data Pruning (Data-Centric Methods) achieved the best balance between privacy and accuracy. These methods modify or remove data at the input level, effectively reducing the model's ability to memorize sensitive data without drastically altering learned decision boundaries.

Data Replacement had the highest recall, but lower precision, meaning it could identify most positive cases but also generated more false positives. Higher privacy comes at the cost of a slight reduction in precision. Data Replacement is useful in applications like fraud detection, where false positives are preferable to false negatives, however, high recall may lead to oversensitivity, making the model less reliable in tasks where precision matters (e.g., medical diagnosis).

In terms of Model-Centric Methods, Model Pruning and Model Replacement provided moderate improvements in privacy but did not significantly outperform data-centric methods. These methods alter model weights post-training rather than removing sensitive information at the input level. This can lead to residual memory of the unlearned data.

Model Shifting showed the weakest privacy improvement, indicating that shifting weights slightly does not fully erase the influence of removed data.

Our findings demonstrate that while traditional MU techniques successfully eliminate target data, their effectiveness varies in balancing privacy, model utility, and explainability. Data-centric methods, such as Data Obfuscation and Data Pruning, provided the most promising results in preserving accuracy while mitigating MIA risks by reducing confidence score disparities. Model-based methods, such as Model Pruning and Model Shifting, demonstrated their adaptability in redistributing feature importance, but their effectiveness varied depending on application constraints and computational feasibility.

The explainability of unlearning processes remains a critical aspect in verifying the success of data removal. The use of Local Interpretable Model-Agnostic Explanations (LIME) provided valuable insights into feature contributions, allowing us to assess the changes in model behavior post-unlearning. The results highlighted that models undergoing successful unlearning should exhibit a redistribution of feature importance rather than merely suppressing confidence scores.

Furthermore, privacy considerations in unlearning are crucial for compliance with legal frame-

works, including GDPR and the Right to Be Forgotten. Our results suggest that methods that prioritize membership indistinguishability, such as pruning and obfuscation, align well with privacy regulations while maintaining practical feasibility. However, certain techniques, such as Data Replacement, may introduce risks by significantly altering confidence distributions, which could impact the reliability of unlearning claims.

In sum, in this work, we provide local evidence of successful machine unlearning using LIME-based explanations. Since LIME offers a local approximation of model behavior, our analysis focuses on how individual predictions change following the unlearning process. Specifically, we examine the redistribution of feature importance at the local level — where dominant, moderate, and minimal contributing features are identified for each sample. This redistribution indicates that the model no longer relies on the same features to make decisions for the unlearned data points. Such behavioral shifts offer evidence of successful unlearning than merely comparing prediction confidence score of the model. our study underscores the importance of choosing the appropriate MU technique based on the specific use case requirements — whether prioritizing privacy, interpretability, or computational efficiency. By integrating explainability-driven verification methods, such as LIME, we enhance trust and transparency in MU claims, contributing to the advancement of responsible AI deployment in sensitive domains such as healthcare, finance, and cybersecurity. Future research should focus on refining hybrid approaches that maximize privacy protection while preserving model integrity and performance.

CHAPTER 5 CONCLUSION

5.1 Contributions

This study makes several key contributions to the field of machine unlearning by exploring verification methodologies through local explainability techniques and using MIA as a baseline for validation. In the course of our thesis research, we make the following contributions:

- Verification Framework for Machine Unlearning: One of the primary contributions of this study is the introduction of a verification framework that leverages local explainability methods, specifically Local Interpretable Model-Agnostic Explanations (LIME), to evaluate the effectiveness of unlearning techniques. This approach provides an interpretable and transparent means of assessing whether sensitive data has been successfully removed from a machine learning model without compromising its performance.
- Comparative Evaluation of Unlearning Methods: We conducted an extensive analysis of multiple machine unlearning techniques, including Data Obfuscation, Data Pruning, Data Replacement, Model Pruning, Model Replacement, and Model Shifting. By comparing these methods, we offer insights into their effectiveness in ensuring data removal while maintaining model utility. This comparative assessment allows practitioners to select the most suitable unlearning approach based on application specific needs.
- Membership Inference Attack (MIA) Resistance Analysis: This research contributes to the security and privacy landscape by evaluating the susceptibility of different machine unlearning methods to Membership Inference Attacks (MIA) for the purpose of verification of unlearning. Our results show that techniques like Data Obfuscation and Data Pruning offer improved resistance to MIA, reducing privacy risks for users who request their data to be unlearned. This contribution is significant for privacy-preserving machine learning, particularly in contexts requiring compliance with legal regulations such as the GDPR and the Right to Be Forgotten.
- Impact on Feature Importance and Model Interpretability: By utilizing explainability tools, we demonstrate how various unlearning methods affect feature importance and model decision-making processes. This study highlights that successful unlearning should not only reduce confidence scores but also redistribute feature importance in a meaningful way. This insight is crucial for ensuring that unlearning methods do not accidentally introduce biases or significantly alter model interpretability.

- Implications for AI Trustworthiness: The findings of this study emphasize the need for verifiable and interpretable unlearning mechanisms to enhance trust in AI systems. By bridging the gap between machine unlearning and explainability, our research provides a practical roadmap for organizations implementing unlearning solutions in sensitive applications such as healthcare, finance, and cybersecurity. The adoption of our verification framework can increase transparency, build user trust, and improve regulatory compliance in AI-driven decision-making systems.

In summary, this study advances the field of machine unlearning by introducing an explainability-driven verification framework, providing a better understanding of evaluation of unlearning techniques, and analyzing their privacy and interpretability implications. These contributions collectively strengthen the foundation for developing trustworthy, privacy-preserving AI systems that align with ethical and legal standards. Overall, the results suggest that analyzing local explanations can offer meaningful insights into the effects of unlearning. Our findings demonstrate that successful unlearning can be evidenced by the redistribution of feature importance at the local level, as approximated using LIME. While this approach provides valuable localized verification, it also highlights limitations in scaling to global model behavior. Future work should aim to strengthen verification by expanding the evidence of feature importance redistribution across broader regions of the model's decision space. The observed redistribution of feature importance at the local level supports our hypothesis that successful unlearning should lead to a meaningful change in the model's decision behavior around the affected data points.

5.2 Limitations

- Scalability Issues: Machine unlearning becomes significantly harder as datasets and models grow larger. For example, unlearning data from Large Language Models (LLMs) with billions of parameters is extremely resource-intensive and computationally complex. These scalability challenges make applying unlearning to real-world, large-scale systems a tough problem.

- Incomplete Data Removal: Even after applying unlearning techniques, traces of the removed data might still exist within the model's parameters or decision-making processes. This is particularly problematic in complex models like neural networks, where overlapping patterns in the data make it hard to ensure the exact complete removal.

- Performance Aspect: Removing data from a model can negatively impact its performance. If the data being unlearned contributed significantly to the model's decisions,

the accuracy and reliability of the model might drop. Balancing effective unlearning with maintaining model performance remains a key challenge.

- High Computational Costs: Many unlearning techniques require retraining or fine-tuning the model, which can be expensive in terms of time and resources. For organizations that need to unlearn data frequently or in real-time, this computational overhead can make unlearning impractical.

- Privacy and Security Gaps: The gaps highlight a critical limitation in the current scope of unlearning techniques, as they may not fully mitigate the broad spectrum of privacy threats. Ensuring robust protection requires unlearning methods that are resilient against diverse attack vectors not just the MIAs while maintaining their ability to safeguard sensitive data comprehensively. These residual vulnerabilities need more comprehensive security evaluations of unlearning methods.

- Difficulty in Verification: Proving that data has been fully unlearned is not straightforward. Most current methods lack robust ways to confirm that the influence of the removed data is completely eliminated, leaving room for uncertainty even when the explanation is provided.

- Limited General Applicability: Not all unlearning techniques work across all types of models making it hard to standardize unlearning practices.

- Regulatory and Ethical Challenges: Ensuring that unlearning methods align with laws like GDPR's "right to be forgotten" is a complex process. Uncertainties in legal definitions and technical implementations create challenges in proving compliance and building user trust.

- Lack of Standardized Frameworks: There is no universal framework or guideline for implementing machine unlearning. This lack of standardization leads to inconsistencies in methods and limits the adoption of unlearning in practical applications.

- Data Inter dependencies: In datasets where samples are interconnected or overlap in features, removing one piece of data can unintentionally affect the integrity of related data. This can introduce biases or reduce the generalizability of the model, complicating the unlearning process further.

5.3 Ethical and Practical Considerations in Machine Unlearning

While this research primarily focuses on privacy and verification of machine unlearning schemes, ethical concerns and real-world deployment challenges must also be addressed. The effectiveness of MU techniques in ensuring data privacy does not automatically translate to ethical compliance or practical feasibility. This section explores the potential ethical dilemmas, risks of misuse, and real-world challenges in implementing machine unlearning.

5.3.1 Bias and Fairness in Unlearning Decisions

- Who Decides What to Unlearn? If an individual or company requests unlearning, what criteria determine whether the request is legitimate or an attempt to erase unfavorable truths?
- Regulatory bodies need to define clear guidelines for who is allowed to request unlearning and under what conditions.
- Disproportionate Impact on Minority Groups: If unlearning is not applied equally across datasets, it could introduce bias in decision-making. For example: In healthcare AI, if certain demographic data is over-unlearned (e.g., minority groups' medical records are disproportionately removed), the AI could become less effective for those populations.
- Ethical Transparency and Explainability: Users should have the right to know when and how their data is unlearned. Explainability tools (like LIME) should be required to verify that unlearning was conducted ethically and did not distort the model's fairness.
- Practical Implications of Deploying MU in Real-World Applications: Machine unlearning is closely tied to data privacy laws such as: GDPR and CCPA which require that consumers can request data deletion, but do not define technical standards for MU. The main challenges is that without standardized unlearning verification, companies may claim compliance without actually removing the data's influence.
- Computational Cost and Feasibility: Full model retraining is computationally expensive, making approximate unlearning techniques (e.g., Data Pruning, Model Pruning) more attractive. Enterprises may prefer cheaper methods over privacy-preserving ones, leading to a trade-off between efficiency and data security. Large-scale AI systems like Google's search ranking algorithms or healthcare predictive models would struggle to apply exact unlearning due to the high retraining costs. Research should optimize

computationally efficient MU methods that provide privacy guarantees without full retraining.

Finally, MU can be misused for unethical purposes, such as erasing evidence or reinforcing biases. There is no clear global regulation on who can request unlearning and under what conditions. Compliance with laws like GDPR and CCPA requires strict verification mechanisms. Computational costs make some MU methods infeasible for large-scale AI applications. In addition, over-unlearning may degrade AI performance, impacting trust in AI-driven decisions. Machine unlearning is not just a technical challenge it is an ethical and societal issue that requires regulatory oversight, transparency, and careful implementation.

5.4 Indications for future research

The findings of this study open several avenues for future research in the field of machine unlearning and verifiability. Below, we highlight key future directions that can advance the field further.

- Scalability: One of the main challenges for Machine Unlearning is scaling these techniques to larger and more complex datasets, especially for models like Large Language Models (LLMs) that are trained on vast and diverse data. Future research should focus on designing explainable scalable algorithms that can handle the complexity of these models, ensuring effective unlearning without compromising their performance.
- Technique Exploration: The diversity of existing MU techniques makes it challenging to identify the most effective solution. Each method comes with its own strengths and trade-offs, and it's not always clear which is best for a given scenario. Future work should systematically compare these methods through explainability and explore combining their strengths to develop hybrid approaches that can deliver better results.
- Efficiency Metrics: While current MU techniques focus on accuracy and privacy, their computational efficiency is often overlooked. Many methods are not optimized for time or resource usage, making them impractical for real-world applications. Future efforts should prioritize creating lightweight and efficient unlearning algorithms that are fast, resource-friendly, and suitable for deployment at scale.
- Extended Security Analysis: This research primarily evaluated privacy risks using Membership Inference Attacks, but other threats like Backdoor Attacks and Model Inversion remain to be used to validate the explainable approach. These advanced attacks could

expose vulnerabilities in unlearning techniques. Future research should expand security testing to include these threats, ensuring that MU methods are robust and comprehensive in protecting data privacy.

- Legal and Ethical Considerations in Machine Unlearning: ML models are trained on vast amounts of user-generated data, raising concerns about privacy and compliance with regulations like GDPR and CCPA. Ensuring that user-specific data can be effectively unlearned without leaving residual traces is a complex but essential task. Future work should not only aim to develop scalable algorithms but also ensure that these methods align with ethical standards, transparently prioritize user privacy, and address the broader societal implications of data removal in AI systems.

- Innovative Approaches: There's a real opportunity to invent new methods specifically designed for unlearning. Current techniques often adapt existing tools, but future work should focus on creating novel algorithms that are efficient, scalable, and tailored to unlearning. Innovations like real-time unlearning and ideas from differential privacy or federated learning could lead to breakthroughs in this field.

- Developing Automated Unlearning Verification Pipelines: A significant step forward would be the automation of unlearning verification through standardized evaluation frameworks. By integrating explainability-driven verification techniques into automated pipelines, organizations can ensure seamless compliance with privacy regulations like GDPR while minimizing human intervention. This would facilitate real-time monitoring and auditing of unlearning requests.

- Investigating Hybrid Unlearning Approaches: A promising direction for future research is the combination of data-centric and model-centric unlearning techniques to develop hybrid approaches. By blending techniques like data perturbation with model pruning or parameter shifting, researchers may find more effective methods that balance privacy, interpretability, and computational feasibility. The proposed explainable verification method should then be tested for these approaches.

By addressing these challenges, the field can move toward more trustworthy, scalable, and legally compliant machine unlearning frameworks, ensuring greater privacy and security in AI-driven applications.

My hope is simple: To pave the way for AI systems that respect our rights while serving humanity. Sometimes, forgetting is the smartest thing we can do and remember: even machines deserve a second chance to forget!

REFERENCES

[1] H. Xu, T. Zhu, L. Zhang, W. Zhou, and P. S. Yu, "Machine unlearning: A survey," *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2306.03558

[2] Y. Qu, X. Yuan, M. Ding, W. Ni, T. Rakotoarivelo, and D. Smith, "Learn to unlearn: Insights into machine unlearning," *Computer*, vol. 57, no. 3, pp. 79–90, 2024.

[3] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–34, 2023.

[4] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, p. 100211, 2024.

[5] D. Liu, M. Yang, X. Qu, P. Zhou, Y. Cheng, and W. Hu, "A survey of attacks on large vision-language models: Resources, advances, and future trends," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2407.07403

[6] General Data Protection Regulation (GDPR). (2018) General data protection regulation (gdpr). Online; Retrieved in March 20, 2022. [Online]. Available: https://data.stats.gov.cn

[7] California Consumer Privacy Act (CCPA). (2018) California consumer privacy act (ccpa). Online; Retrieved in March 19, 2022. [Online]. Available: https://oag.ca.gov/privacy/ccpa

[8] Japan Data Protection Overview (JDPO). (2019) Japan data protection overview (jdpo). Online; Retrieved in March 19, 2022. [Online]. Available: https://www.dataguidance.com/notes/japan-data-protection-overview

[9] Consumer Privacy Protection Act (CPPA). (2022) Consumer privacy protection act (cppa). Online; Retrieved in March 19, 2022. [Online]. Available: https://blog.didomi.io/en-us/canada-data-privacy-law

[10] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE Symposium on Security and Privacy*. IEEE, 2015.

[11] Y. Xu, "Machine unlearning for traditional models and large language models: A short survey," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2404.01206

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

[13] D. Garreau and U. Luxburg, "Explaining the explainer: A first theoretical analysis of lime," in *International Conference on Artificial Intelligence and Statistics*, 2020.

[14] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: A systematic review on the application of lime and shap in alzheimer's disease detection," *Brain Informatics*, 2024.

[15] N. Li, C. Zhou, Y. Gao, H. Chen, A. Fu, Z. Zhang, and Y. Shui, "Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2403.08254

[16] W. Wang, Z. Tian, and S. Yu, "Machine unlearning: A comprehensive survey," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2405.07406

[17] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021.

[18] T. Zhu, D. Ye, W. Wang, W. Zhou, and S. Y. Philip, "More than privacy: Applying differential privacy in key areas of artificial intelligence," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[19] K. Wei, J. Li, C. Ma, M. Ding, C. Chen, S. Jin, Z. Han, and H. V. Poor, "Low-latency federated learning over wireless channels with differential privacy," *IEEE Journal on Selected Areas in Communications*, 2021.

[20] N. Xiang, X. Zhang, Y. Dou, X. Xu, K. Yang, and Y. Tan, "High-end equipment data desensitization method based on improved stackelberg gan," *Expert Systems with Applications*, 2021.

[21] M. Alomari, F. Li, D. C. Hogg, and A. G. Cohn, "Online perceptual learning and natural language acquisition for autonomous robots," *Artificial Intelligence*, 2022.

[22] T. Chen, D. Ye, H. Zhu, L. Zhang, and W. Wang, "Overcoming catastrophic forgetting by bayesian generative regularization," in *International Conference on Machine Learning*, 2021.

[23] H. Liu, Y. Yang, and X. Wang, "Overcoming catastrophic forgetting in graph neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 2021.

[24] A. G. Chowdhury, M. M. Islam, V. Kumar, F. H. Shezan, V. Jain, and A. Chadha, "Breaking down the defenses: A comparative survey of attacks on large language models," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2403.04786

[25] S. Huang, S. Mamidanna, S. Jangam, Y. Zhou, and L. H. Gilpin, "Can large language models explain themselves? a study of llm-generated self-explanations," *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2310.11207

[26] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, and M. Du, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–38, 2024.

[27] B. Zhang, Z. Chen, C. Shen, and J. Li, "Verification of machine unlearning is fragile," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2408.00929

[28] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, and N. Papernot, "Machine unlearning," in *Proceedings of the IEEE Symposium on Security and Privacy (SP)*. IEEE, May 2021, pp. 141–159.

[29] P. Zhang, G. Bai, Z. Huang, and X.-S. Xu, "Machine unlearning for image retrieval: A generative scrubbing approach," in *Proceedings of the 30th ACM International Conference on Multimedia (MM)*, J. Magalhães, A. Del Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, and L. Toni, Eds. ACM, October 2022, pp. 237–245.

[30] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *Proceedings of the IEEE Symposium on Security and Privacy (SP)*. IEEE, May 2015, pp. 463–480.

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

[32] S. Schelter, S. Grafberger, and T. Dunning, "Hedgecut: Maintaining randomised trees for low-latency machine unlearning," in *Proceedings of the International Conference on Management of Data (SIGMOD)*, G. Li, Z. Li, S. Idreos, and D. Srivastava, Eds. ACM, June 2021, pp. 1545–1557.

[33] T. Baumhauer, P. Schöttle, and M. Zeppelzauer, "Machine unlearning: Linear filtration for logit-based classifiers," *CoRR*, vol. abs/2002.02730, 2020.

[34] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, "Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 357–372, 2022.

[35] D. M. Sommer, L. Song, S. Wagh, and P. Mittal, "Towards probabilistic verification of machine unlearning," *CoRR*, vol. abs/2003.04247, 2020.

[36] H. Zhang, Y. Li, Y. Huang, Y. Wen, J. Yin, and K. Guan, "Mlmodelci: An automatic cloud platform for efficient mlaas," in *Proceedings of the 28th ACM International Conference on Multimedia (MM)*. ACM, October 2020, pp. 4453–4456.

[37] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. S. Kankanhalli, "Fast yet effective machine unlearning," *CoRR*, vol. abs/2111.08947, 2021.

[38] C. Guo, T. Goldstein, A. Y. Hannun, and L. van der Maaten, "Certified data removal from machine learning models," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, July 2020, pp. 3832–3842.

[39] A. Warnecke, L. Pirch, C. Wressnegger, and K. Rieck, "Machine unlearning of features and labels," *CoRR*, vol. abs/2108.11577, 2021.

[40] S. Krishna, J. Ma, and H. Lakkaraju, "Towards bridging the gaps between the right to explanation and the right to be forgotten," in *International Conference on Machine Learning*. PMLR, July 2023, pp. 17 808–17 826.

[41] T. Shaik, X. Tao, H. Xie, L. Li, X. Zhu, and Q. Li, "Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.