"I Forgot About You": Exploring Multi-Label Unlearning (MLU) for Responsible Facial Recognition Systems

Prommy Sultana Hossain[1] (✉), Emanuela Marasco[1], Jessica Lin[1], and Michael King[2]

[1] George Mason Univeristy, Fairfax VA 22030, USA
{phossai, emarasco, jessica}@gmu.edu
[2] Florida Institute of Technology, Melbourne FL 32901, USA michaelking@fit.edu

Abstract. The widespread adoption of machine learning and deep learning models has heightened privacy concerns, as these models can unintentionally memorize and expose personal information. Machine Unlearning (MU) has gained considerable attention for improving privacy and data control. MU addresses privacy challenges by selectively removing the influence of specific training data from deployed models. However, most current MU approaches focus on single-label classification scenarios, where each instance is assigned only one label. In contrast, Multi-Label Classification (MLC), such as those in facial recognition (facial attribute classification) systems, involve instances that can be associated with multiple, non-exclusive attribute labels. The complex interdependencies between parameters in these cases pose unique challenges when selectively removing specific knowledge. This work proposes a novel parameter space-based MU framework for MLC systems. Our data-driven generalization approach uses sparsification techniques operating directly on learned representations without retraining on the modified training data. We employ two strategies to improve state-of-the-art models for MLC unlearning: Weight Filtering, which identifies and resets critical parameters based on sensitivity and influence scores, and Weight Pruning, which strategically eliminates parameters based on their importance to the unlearned label while preserving shared representations for retained attributes. Extensive experiments demonstrate that our Weight Pruning method can achieve up to 35.5× speedup over retraining while maintaining >93% accuracy for retained labels and reducing the prediction of forgotten attributes to near zero (0.11%), a significant improvement over existing methods. The privacy analysis also confirms a substantial reduction in information leakage, which establishes a new standard for responsible facial attribute classification systems under current privacy regulations.

Keywords: Multi-label Classification · Machine Unlearning · Privacy

1 Introduction

The ubiquitous deployment of deep neural networks has created an unprecedented privacy paradox: while these systems enable remarkable capabilities in classification, they simultaneously memorize and expose sensitive personal information without explicit consent [1]. This challenge is particularly acute in facial recognition (facial attribute classification (FAC)) systems, which operate within a Multi-Label Classification (MLC) paradigm where each face simultaneously expresses multiple non-exclusive attributes—age, gender, emotion, ethnicity—encoded within shared neural representations [2][3]. Unlike traditional single-label systems, this representational entanglement creates a fundamental tension: how can we selectively remove knowledge of specific attributes while preserving the model's utility for legitimate purposes?

This tension has gained critical urgency with the emergence of privacy regulations such as the European Union's General Data Protection Regulation (GDPR), which establishes the "Right To Be Forgotten (RTBF)" as a fundamental principle [4]. Crucially, RTBF extends beyond mere data deletion to require the elimination of knowledge derived from personal data. Consider a practical scenario: an individual may exercise RTBF for emotion detection capabilities while permitting age estimation, or request removal of ethnicity classification while maintaining gender recognition. Such fine-grained privacy requirements demand sophisticated unlearning mechanisms that can surgically modify model behavior without catastrophic interference.

Machine Unlearning (MU) has emerged as the primary framework to address these demands. It offers two main paradigms: Exact Unlearning, which provides robust privacy guarantees through complete retraining but at prohibitive computational costs, and Approximate Unlearning, which achieves efficiency through direct parameter modification [5][6]. However, a critical research gap exists: existing unlearning techniques almost exclusively target Single-Label Classification (SLC) scenarios and fail catastrophically when applied to multi-label systems. When attempting to remove a single attribute from facial classification models, current methods degrade performance across all remaining attributes, rendering the system unusable [7][8][9][10].

This limitation is particularly problematic given the widespread deployment of multi-label systems in high-stakes domains. Healthcare systems must maintain diagnostic capabilities while protecting patient privacy; marketing platforms need to preserve demographic insights while respecting individual rights; and security systems require selective attribute removal without compromising legitimate functionality [11][12]. The absence of effective Multi-Label Unlearning (MLU) capabilities represents a fundamental barrier to privacy-compliant AI deployment in these critical applications.

The core technical challenge lies in the interconnected nature of multi-label representations. Unlike single-label models where each instance belongs to exactly one class, multi-label systems must handle partial label deletion (removing some but not all labels from an instance), entangled representations (shared parameters across multiple output heads), and complex label dependencies (statistical correlations between attributes). These factors create a complex optimization landscape where naive application of existing unlearning methods leads to uncontrolled performance degradation across the entire system.

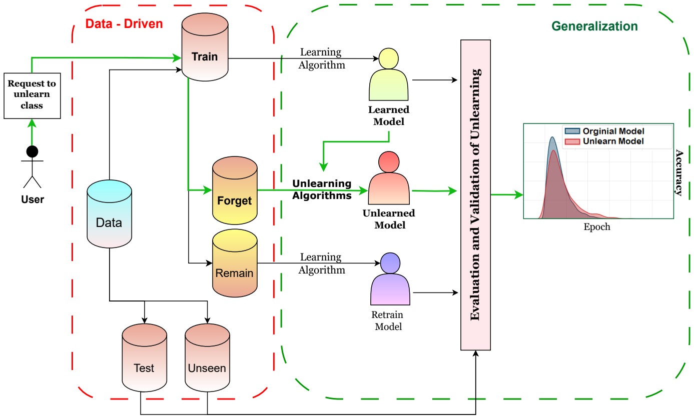

Fig. 1. The proposed data-driven generalization framework ensures the preservation of data utility while effectively unlearning a label, as indicated by the green arrow, which highlights privacy maintenance.

To address these fundamental challenges, we introduce a novel data-driven generalization framework for Multi-Label Unlearning.[3] (Figure 1). Our approach centers on model sparsification strategies—Weight Filtering (WF) and Weight Pruning (WP)—that surgically remove label-specific knowledge without compromising system-wide performance. As illustrated in the framework, when a user requests to unlearn a specific class, our system partitions the data into "Forget" and "Remain" sets, then applies targeted unlearning algorithms to produce an unlearned model that maintains utility for retained class attributes while eliminating the specified knowledge. Rather than relying on computationally expensive retraining or task-specific heuristics, our framework analyzes statistical patterns in parameter distributions to identify and neutralize parameters linked to forgotten labels while preserving the underlying model architecture [13].

The key innovation lies in our parameter-centric approach: by examining weight correlations across network layers, we can precisely target neurons and connections responsible for specific attribute predictions without solely relying on the original training data during the unlearning process [7]. This generalization capability allows the framework to adapt to diverse facial attribute classification tasks based on the learned parameter structure rather than domain-specific modifications [14]. Our method preserves the model's original training objec-

[3] Data and code are available in Github: `https://github.com/Promzi/unlearn_label.git`

tive while ensuring surgical modification of only the targeted label information, maintaining both utility and privacy.

Extensive empirical validation across diverse facial attribute datasets demonstrates the effectiveness of our approach: we achieve up to 35.5× computational speedup over exact retraining while maintaining >93% accuracy for retained attributes and reducing forgotten attribute prediction to near-random levels (0.11%). Comprehensive privacy analysis reveals substantial improvements in information leakage prevention, with our method achieving 54-56% residual information compared to 70%+ for existing approaches, establishing new benchmarks for privacy-compliant AI systems under current regulatory frameworks.

1.1 Main Contributions

1. **Novel Multi-Label Unlearning Framework:** We propose the first comprehensive parameter space-based framework specifically designed for multi-label machine unlearning in facial attribute classification. Our approach addresses the critical research gap in existing single-label unlearning methods by introducing Weight Filtering (WF) and Weight Pruning (WP) techniques that enable efficient model sparsification without retraining. The framework surgically removes targeted attributes while preserving interdependent label relationships through adaptive parameter-space analysis, ensuring strong privacy guarantees and mitigating data corruption risks inherent in multi-label scenarios.
2. **Comprehensive Empirical Validation:** We conduct extensive experimental validation across diverse benchmark datasets, including CelebA (adapted for multi-label settings), CIFAR-10, MNIST, and SVHN [15], and demonstrate superior performance across multiple evaluation dimensions. Our framework consistently outperforms state-of-the-art methods in utility preservation (>93% accuracy retention), computational efficiency (35.5× speedup), and output distribution integrity, establishing new performance benchmarks for multi-label unlearning in facial attribute classification and beyond.
3. **Privacy Analysis:** To analyze how parameter space modifications ensure privacy protection, we conduct experiments and show that our approach achieves significant improvement in privacy preservation (54-56% residual information vs. 70%+ for existing methods) without requiring retraining on modified datasets and provide practical privacy compliance for deployment under current regulatory frameworks.

2 Related Work

Recent advances in privacy-conscious ML have catalyzed MU development, initially through theoretical studies on convex models that provide crucial insights but face limitations with deep neural networks' non-convex optimization landscapes [16][17]. The evolution of MU research has produced three distinct methodological approaches:

Input Space: Early unlearning methods focused on alterations in the input space through data obfuscation, noise injection, label anonymization, and adversarial perturbations, which relies on direct access to the original training data during the unlearning process and introduce significant operational constraints [6]. These approaches suffer from performance degradation and privacy vulnerabilities that can be exploited by direct attacks (submitting unseen data to unlearning) and preconditioned attacks (strategically removing poisoned data) [18]. Despite defensive countermeasures including regulated algorithms and membership verification, these methods remain limited by their dependence on direct data manipulation [19].

Decision Space: Decision boundary methods directly manipulate model boundaries to replicate the behavior of the re-trained model, addressing the limitations of the input space modification [8]. However, in MLC scenarios, these methods face challenges due to complex boundary interconnections, where adjustments to individual label boundaries create cascading effects across the decision space. This approach fails in high-boundary overlap scenarios where precisely preserving retained label relationships while removing targeted information becomes impossible.

Parameter Space: These methods directly adjust model parameters to eliminate forgotten data influence, primarily in single-label unlearning scenarios. Catastrophic Forgetting k (CF-k) implements selective retraining of the final k layers while freezing initial layers, but faces optimal k-value selection challenges and retains residual information [9]. SCalable recall and unlearning unbound (SCRUB) employs a teacher-student framework that balances retained data performance while diverging on forgotten data, but shows significant degradation when managing multiple objectives [20]. UNlearning Samples with Impair-Repair (UNSIR) implements adversarial noise generation followed by model repair, but requires substantial computational resources and demonstrates incomplete restoration when noise rates are uncontrolled interference [10]. Saliency Unlearning (SalUN) identifies critical parameters through saliency map analysis but struggles with map accuracy and creates unintended side effects [21]. Despite advancing the field through learnable memory matrices within parameter space for SLC, the work by Poppi et al. [7] remains constrained by pre-trained model dependence, excessive relearning time penalties, and severe performance trade-offs. To address the limitations in parameter-space methods that expose critical deficiencies in addressing MLC unlearning challenges, our proposed *Weight Filtering* strategy advances beyond the SOTA by efficiently managing single-label unlearning and MLU scenarios while preserving classification integrity. Additionally, *Weight Pruning* offers a groundbreaking approach to unlearning, reducing computational demands and providing guarantees, which are crucial for maintaining utility in MLC systems through precise parameter control.

3 Preliminaries

MU in MLC presents the complex problem of surgically removing target label knowledge from trained deep neural networks such that the model's behavior matches that of a model retrained without the forgotten labels, but achieved without the prohibitive cost of complete retraining. Let $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ denote the training dataset, where $x_i \in \mathcal{X}$ represents an input vector in the input space $\mathcal{X} \subseteq \mathbb{R}^d$, with d denoting the input dimensionality. Each input instance x_i is associated with a label vector $y_i \in \{0, 1\}^{\mathcal{K}}$, where $\mathcal{K}$ represents the total number of possible labels (e.g., facial attributes like Arched_Eyebrows, Bald, Oval_Face). Each element in y_i is a binary indicator denoting the corresponding attribute's presence (1) or absence (0).

In MLC setting, where each instance x_i can simultaneously belong to multiple labels, selectively forgetting an entire label $u \in \mathcal{K}$ presents unique challenges due to the interrelated nature of label representations [11][22]. MLC label overlap forms interconnected networks of shared attribute representations, unlike SLC with distinct label boundaries [23]. The interconnectedness makes removing a label difficult, as its information might be intertwined with shared representations needed for other attributes, risking privacy leaks through indirect associations. Retraining without the unlearning attribute ensures its removal, but is computationally expensive for large-scale applications.

Let f_{w_0} be the original learned model trained on $\mathcal{D}$, optimally parameterized by w_0. For any input $x \in \mathcal{D}$, the output of the MLC model $f_{w_0}(x) = [f_{w_0}^k(x)]_{k=1}^{K}$, where $f_{w_0}^k(x)$ represents the logit score for the k^{th} label. The final predictions are obtained by applying a threshold function to each logit, which allows for simultaneous attribute assignments. Based on established literature demonstrating successful unlearning through weight influence analysis [7], our study partitions the parameter space by identifying weights associated with the target unlearned label $u \in \mathcal{K}$ and allowing direct modifications to the influential parameters without relying on the original training data throughout the unlearning process.

Hence, given an unlearning request for a specific label u, we define the forget set $\mathcal{W}_f$ as $\mathcal{W}_f = \{w\{(x, y)\} \in w_0 \mid \mathcal{I}(w, u) > \epsilon\}$, where $\mathcal{I}(w, u)$ denotes the influence function that quantifies the contribution of weights towards classifying the target unlearned label, and ϵ represents the threshold determining the significant influence. The rest of the weights in the parameter space are placed in the remaining set $\mathcal{W}_r = \{w_0 \setminus \mathcal{W}_f\}$.

Next section presents the detailed methodology on measuring the influence function and subsequent weight modifications, where we aim to unlearn the information of $\mathcal{W}_f$ from f_{w_0}—without re-learning $\mathcal{W}_r$—and updating the parameters $w_0 \to w'$, where w' represents the updated parameters obtained by the unlearning methods. To validate the performance of the unlearning model $f_{w'}$, we train a model, which we call the *Retrain* model f_{w^*}, using the original learning algorithm from scratch without the targeted unlearned label u. This will be the optimal unlearning model used as the baseline for this study. In this unlearning problem, we expect the unlearning model $f_{w'}$ to be as similar to the retained model f_{w^*} as possible.

4 Methodology

We propose a parameter space-based unlearning framework that operates directly on the model's learned representations. Our approach follows a data-driven generalization framework, which identifies and partitions the parameter space based on the weights' influence on the target unlearned label u. We identify weights significantly contributing to label u classification through influence function analysis, storing them in the forget set $\mathcal{W}_f$, while retaining other weights in the remaining set $\mathcal{W}_r$. The *generalization* nature in our unlearning method is achieved through a two-phase optimization strategy: selective parameter modification followed by targeted fine-tuning to preserve the model utility. Formally, let $w'_u = \varPhi(w_0, u, \mathcal{W}_f), \;\; w' = \psi(w'_u, \mathcal{W}_r)$ where w'_u denotes the intermediate parameters after selective modification of label u, $\varPhi$ focuses exclusively on adapting parameters related to the unlearned label, and ψ refines the entire parameter space using the remaining learned parameters. Through this framework, we can modify the influencing parameters of the target unlearned label without solely relying on the original training data or eliminating any data points from $\mathcal{D}$ during the unlearning process, as this could inadvertently affect the model's performance on the remaining labels due to shared attribute representations. We introduce two novel strategies for selective parameter modification: ***Weight Filtering*** and ***Weight Pruning***, which strategically modify parameters based on their correlation to the unlearned label while preserving overall model performance.

4.1 Weight Filtering

Deep neural networks trained on multi-label data create intricate shared representations and memorize training data in their parameter space, posing privacy risks. Although existing approaches focus on data or decision boundary modifications, we observe that the original model f_{w_0} parameters show varying influence on label predictions, allowing selective parameter modification for targeted unlearning while maintaining model utility. Motivated by recent advances in influence functions and parameter sensitivity analysis [24][25], we propose weight filtering that identifies and neutralizes parameters specifically encoding information about the unlearned label[4]. For each w_{ik} associated with weight i and label $k \in \mathcal{K}$, we calculate a sensitivity score:

$$S(w_{ik}) = |\frac{\partial \mathcal{L}}{\partial w_{ik}}| \tag{1}$$

that quantifies its impact on the model's standard loss function $\mathcal{L}$. Furthermore, we compute an influence score $\mathcal{I}(x_i)$ for each training point in $\mathcal{D}$ to locate the specific influential data points that contribute most to the classification of the

[4] Data and code are available in Github: `https://github.com/Promzi/unlearn_label.git`

unlearned label u, using the formula shown in [25], without retraining on the modified training data.

$$\mathcal{I}(x_i) = -\nabla_{w_0}\mathcal{L}(x_i^{\text{pert}}, w_0) \cdot H^{-1} \cdot \nabla_{w_0}\mathcal{L}(x_i, w_0) \tag{2}$$

where x_i^{pert} represents a perturbed version of the original training example x_i, H is the Hessian matrix that captures the loss surface curvature, providing insight into how x_i affected the f_{w_0} model decision. As $\mathcal{I}(x_i)$ identified data points, $x_i' \in \mathcal{D}$, that contribute to the classification of the unlearned label u. We can then construct the forget set $\mathcal{W}_f$ in two steps; First, a sensitivity score $(S(w_{iu}))$ is calculated to analyze the influence of weights through network activation patterns for classification of u; second, a composite score $(\mathcal{S}_{i\mathcal{K}})$ examines the interaction of all attributes with weights associated with u during forward propagation [24]. These steps facilitates our comprehension of the shared representation of weights within the learned model, and assists us in establishing a threshold for filtering the weights of u.

The sensitivity score $S(w_{iu})$ for each weight in w_0 to the unlearned label u is calculated using the following equation, $S(w_{iu}) = |\frac{\partial \mathcal{L}}{\partial w_{iu}}|$, while the composite score, $\mathcal{S}_{i\mathcal{K}} = (S(w_{ik}))^{\mathcal{T}} \cdot S(w_{iu})$, comprehensively measures each weight i in w_0 for each $k \in \mathcal{K}$ to understand its association of k to u. The resulting representation matrix of $\mathcal{S}_{i\mathcal{K}}$ contains the association of retained labels with unlearned labels for each $i \in w_0$. Hence the dimensions of $\mathcal{S}_{i\mathcal{K}}$ is $\mathcal{K} \times |u|$ for each i, as the dimension of $(S(w_{ik}))^{\mathcal{T}}$ is $\mathcal{K} \times |w_0|$ and the dimension of $S(w_{iu})$ is $|w_0| \times |u|$. Therefore, the forget set $\mathcal{W}_f$ now will contain parameters that require modification to unlearn the knowledge of the label u, which can be written as $\mathcal{W}_f \leftarrow \{w\{x_i'\} \mid \exists w_{iu} : (w_{iu} \text{ influences } f_{w_0}^u(x_i))\}$. The sensitivity score of the u label then guides the selective modification of the parameters of $w(x_i') \in \mathcal{W}_f$ according to:

$$w_{ik}' = \begin{cases} 0 \text{ or } \mathcal{N}(0, \sigma^2) & \text{if } S_{iu} < \phi \\ w_{ik} & \text{otherwise} \end{cases}$$

where ϕ is an adaptive threshold that determines parameter modification to balancing unlearning effectiveness to model performance. This adaptive threshold dynamically adjusts based on the loss landscape curvature during fine-tuning, preventing over-filtering or under-filtering as the model converges. The relationship between the unlearned label u and the retained label k influences the hierarchy of attribute importance, with higher overlap requiring a more conservative threshold adaptation to preserve shared representations [26]. This targeted strategy preserves crucial parameters and maintains representations of remaining attributes while modifying only those below the threshold for the unlearned label. To verify complete unlearning, we employ a secondary verification process to confirm the removal of explicit and implicit label representations using attribute inference attacks and membership inference attacks with shadow datasets (data points not involved in learning or unlearning) [19]. The process concludes with a fine-tuning phase, which is discussed in later sections. Weight Filtering operates directly on parameter space through influence functions, comprehensively removing sensitive information while maintaining prediction certainty for non-target attributes.

4.2 Weight Pruning

We develop weight pruning as a more efficient alternative to address the computational challenges of weight filtering, which scales cubically due to full Hessian matrix calculations. This method achieves quicker unlearning with substantially lower computational cost by utilizing only diagonal Hessian elements and first-order gradients, resulting in linear time complexity. [5] Weight pruning determines the importance of the parameters through a composite metric combining sensitivity analysis and second-order derivatives.

$$I(w_{ik}) = \alpha S(w_{ik}) + \beta(\frac{1}{2}H_{ii}w_{ik}^2), \tag{3}$$

where $S(w_{ik})$ is the gradient magnitude calculated as in equation (eq. 1), and the second term represents local curvature using only the diagonal Hessian element H_{ii}. Hyperparameters α and β balance gradient-based sensitivity and curvature information, optimized through cross-validation. The construction of the forget set $\mathcal{W}_f$ is then performed similarly to the weight filtering method. However, unlike weight filtering's binary threshold approach, weight pruning establishes three thresholds—ϕ_l, ϕ_m, and ϕ_h—set at the 25th, 50th, and 75th percentiles of the importance score distribution. This enables hierarchical parameter modification:

- Parameters with $I(w_{ik}) < \phi_l$ are set to zero.
- Parameters with $\phi_l \leq I(w_{ik}) < \phi_h$ are scaled by $\exp(-\lambda I(w_{ik}))$ where λ controls the decay rate.
- Parameters with $I(w_{ik}) \geq \phi_h$ undergo fine-tuning with reduced learning rate α_r.

This granular control allows the method to adapt pruning percentages through each unlearning iteration, dynamically balancing unlearning effectiveness and model utility. This iterative approach makes weight pruning particularly suitable for large-scale models where full Hessian computation would be prohibitive.

4.3 Generalization

Generalization in multi-label neural networks addresses the challenges of selectively removing attribute information without disrupting shared representations, architecture, or dataset characteristics. Our approach implements a constrained optimization strategy that balances effective unlearning with preservation of essential cross-label representations. The fine-tuning phase , optimizes model parameters while preserving unlearning effects through two key mechanisms:

1. Parameter updates using gradients computed exclusively from the remaining set $\mathcal{W}_r$.
2. Constrained updates for filtered weights $\mathcal{W}_u$ associated with the unlearned label: $w'_{ik} \leftarrow \min(\max(w'_{ik}, w_{iu} - \epsilon), w_{iu} + \epsilon)$

[5] Data and code are available in Github: `https://github.com/Promzi/unlearn_label.git`

This constraint ensures filtered weights remain within an ϵ distance of their modified values while allowing sufficient flexibility for utility preservation.

5 Limitation

Our research addresses computational overhead in parameter-based unlearning but faces several constraints. While Weight Filtering method shows strong utility preservation and privacy guarantees, it incurs $O(n^3 + md)$ time complexity for networks with n parameters, m samples, and d label dimensions due to complete Hessian computation. Weight Pruning method reduces this to $O(n + md)$ using diagonal Hessian elements while maintaining comparable effectiveness. Our approach focuses on unlearning specific information representations rather than completely removing data. Experiments revealed that removing more than 20% of influential data points completely significantly degrades model utility, consistent with previous findings [17][27][28]. Additionally, our constrained optimization in fine-tuning may limit finding optimal solutions when unlearning conflicts with attribute preservation, while threshold selection requires careful calibration.

6 Experimental Settings

Datasets We conducted extensive experiments across multiple facial attribute classification datasets (CelebA [29], MUFAC [15], Vggface2 [30], and benchmark vision datasets (CIFAR-10, MNIST, and SVHN) to evaluate our unlearning methods' performance under diverse conditions.

Baselines We implemented several SOTA parameter-space unlearning techniques as benchmarks: *Retrain* (baseline), *CF-k* [9], *SCRUB* [20], *UNSIR* [10], and *SalUN* [21].

Implementation The research was conducted using an NVIDIA GeForce RTX 4060 GPU, Intel Core i9-12900K CPU, 64GB DDR5 RAM, with CUDA 11.8, PyTorch 2.0.1, Python 3.12.4, on Ubuntu 22.04 LTS. For facial attribute classification (FAC), we fine-tuned pre-trained ResNet-18 and ResNet-50 models by replacing the final fully connected layer and applying multi-label sigmoid activation. ResNet-50 was trained from scratch with appropriate input normalization and softmax activation for standard image datasets for single-label classification. Dataset configurations were organized with 65% for training ($\mathcal{D}$), 25% validation ($\mathcal{D}_v$) and 10% test ($\mathcal{D}_t$) data, with verified integrity to ensure no overlap between sets. The test set assesses bias from the validation set as these data are not used in training or validation. Forget ($\mathcal{W}_f$) and remaining ($\mathcal{W}_r$) sets are established based on weight contributions to the unlearned label u, ensuring $\mathcal{W}_f \cap \mathcal{W}_r = \varnothing$. Data preprocessing included resizing, random transformations for training data (horizontal flips, affine transformations, and color adjustments), while validation and test data only underwent resizing and tensor conversion. The training procedure employed a Stochastic Gradient Descent optimizer with 0.9 momentum, a constant learning rate of 0.01, a weight decay of 5e-4, a batch size of 64, and 50 epochs. We used Binary-Cross Entropy loss for multi-label tasks and Cross-Entropy loss for single-label classification, with a random seed of 42 for reproducibility.

Metrics For utility guarantees, we measure the model's ability to maintain performance on preserved attribute while reducing accuracy on unlearned attribute, using three accuracy metics on: $\mathcal{D}$, $\mathcal{D}_v$ and $\mathcal{D}_t$ [15][31]. We also evaluate the efficacy of shared representation by examining the correlation between weight importance and attribute performance. For privacy guarantees, we implement membership inference attacks (MIA) and attribute inference attacks (AIA) to measure whether unlearned attribute information remains extractable from model representation, with lower attack success rates indicating more substantial unlearning effectiveness [33].

Hyperparameter Sensitivity We assess how sensitive our Weight Filtering technique is to its key hyperparameters: the forgetting strength ϵ and the convergence threshold ϕ. Where Table 1 reports representative results for varying ϵ (with fixed ϕ) and varying ϕ (with fixed ϵ). We observe that ϵ has a dominant effect on performance. Smaller ϵ (stronger forgetting) consistently increases the forgetting score but at a cost to accuracy, whereas larger ϵ preserves accuracy but weakens forgetting. By contrast, changing ϕ produces only modest changes in both accuracy and forgetting. For instance, reducing ϵ from 1.0 to 0.1 (with $\phi = 1.0$) might drop accuracy from 90.0% to 85.0% while boosting the forgetting score from 70.0% to 95.0%. Varying ϕ between 0.01 and 1.0 (with $\epsilon = 0.5$) only shifts accuracy by a few points and has a much smaller impact on forgetting. These trends indicate that ϵ primarily governs the trade-off between utility and forgetting, whereas ϕ mainly fine-tunes the unlearning update.

Table 1. Impact of varying ϵ and ϕ on model accuracy on predicting the attribute "Brown_Hair" and forgetting effectiveness of attribute "Gender". (First three rows fix $\phi = 1.0$ and vary ϵ; last two rows fix $\epsilon = 0.5$ and vary ϕ.

ϵ	ϕ	**Accuracy (%)**	**Forgetting (%)**
0.1	1.0	85.0	95.0
0.5	1.0	88.0	85.0
1.0	1.0	90.0	70.0
0.5	0.01	87.0	86.0
0.5	0.10	88.0	85.0

We tune hyperparameters ϵ and ϕ, calibrate ϵ for utility-forgetting trade-offs, and set ϕ roughly. Influence scores are computed using gradients and inverse Hessian approximations. Parameters above the threshold are zeroed, while those below are pruned based on the threshold index.

7 Performance Evaluation

7.1 Utility Guarantee

An efficient unlearning method should minimize knowledge of the unlearned attribute while preserving performance on the retained attribute [18][19]. We evaluate our proposed methods through comprehensive experiments across two scenarios: *(1)* Label-specific unlearning in MLC using pre-trained ResNet-18/50 on facial attribute datasets. *(2)* Label-specific unlearning in SLC using ResNet-50 on both facial attribute (MUFAC) and standard vision datasets (CIFAR-10, MNIST, SVHN).

Unlearning in Multi-Label Classification The deployment of facial attribute classification has raised significant privacy concerns, particularly regarding sensitive attributes such as gender and age in automated decision-making. These concerns are especially relevant in applications such as job search systems [34] and healthcare [35], where algorithmic bias can perpetuate discrimination. We evaluated our unlearning methods to address these challenges by removing the targeted label information while preserving other attributes. From the complete set of attributes available in the datasets, we selected a representative subset of 10 diverse facial attributes ($\mathcal{K}$ = {Arched_Eyebrows, Bald, Big_Lips, Brown_Hair, Double_Chin, Gender, No_Beard, Oval_Face, Pointy_Nose, Young_Old}) to demonstrate our approach, as showing results for all attributes would be impractical. After fine-tuning pre-trained models to classify these attributes with ≈ 98% accuracy, we focused on unlearning specific-label classification while maintaining performance on the other attributes. $\mathcal{W}_f = \{\forall\ w(x' \in \mathcal{D})\}$ contains parameters of data influencing u label classification, while the remaining parameters are set to $\mathcal{W}_r$. Table 2 shows the variation in the performance of attribute classification between different unlearning methods ($f_{w'}$). The original model demonstrates consistent accuracy (96-97%) across all datasets. Baseline methods show different levels of performance degradation: Retrain experiences minor generalization loss (3-4% drop), CF-3 performs poorly (37-50% accuracy), while SCRUB, UNSIR, and SalUN show progressive improvements (81-89% range). Our proposed methods outperform all baselines, with Weight Pruning consistently maintaining accuracy above 93% and Weight Filtering showing robust performance above 91%. This demonstrates our methods' effectiveness in preserving model utility while selectively removing targeted information.

Unlearning in Single-Label Classification We evaluated the efficacy of our methods in SLC scenarios to validate them beyond multi-label settings. This capability addresses critical privacy concerns in FAC systems, particularly for selectively removing demographic information that could enable discriminatory practices, such as dating apps charging higher prices for older users [36][37]. We use the MUFAC dataset that classifies East Asian facial images into one of five

Table 2. Performance comparison of unlearning methods on multi-label FAC. Models unlearn gender classification while maintaining accuracy on other attributes. Results show the attribute classification accuracy (%) without the unlearned attribute on training ($\mathcal{D}$), validation ($\mathcal{D}_v$) and test ($\mathcal{D}_t$) data. **Bold** and *italic* values indicate the best and second-best performance on the CelebA and VggFace2 datasets for each model.

Model	CelebA [29]						VggFace2 [30]					
	ResNet-18			ResNet-50			ResNet-18			ResNet-50		
	$\mathcal{D}$	$\mathcal{D}_v$	$\mathcal{D}_t$	$\mathcal{D}$	$\mathcal{D}_v$	$\mathcal{D}_t$	$\mathcal{D}$	$\mathcal{D}_v$	$\mathcal{D}_t$	$\mathcal{D}$	$\mathcal{D}_v$	$\mathcal{D}_t$
Original	96.87	95.45	96.32	97.12	96.89	96.74	96.43	95.87	96.22	96.78	96.12	96.45
Retrain	92.34	93.21	91.78	93.67	94.45	93.12	91.98	92.65	93.23	93.45	93.87	92.98
CF-3	42.54	48.23	45.67	37.89	42.11	39.76	49.87	50.12	47.34	44.12	45.23	43.67
SCRUB	82.56	81.45	83.21	84.23	83.89	83.67	81.98	82.12	80.87	83.45	82.87	84.23
UNSIR	88.78	87.45	89.21	86.98	88.23	87.65	88.12	87.89	86.45	89.12	88.45	87.98
SalUN	89.34	88.76	88.98	87.89	88.65	88.23	89.76	88.45	87.98	88.12	89.23	88.67
WF	91.78	92.12	93.21	92.87	91.45	92.34	93.45	**94.12**	92.67	**94.23**	92.98	*93.78*
WP	*93.67*	*94.23*	**95.12**	**94.12**	**94.78**	*93.89*	92.45	*93.76*	**94.87**	*93.98*	**94.12**	93.56

Note: WF = Weight Filtering and WP = Weight Pruning

Table 3. Performance comparison of unlearning methods on single-label age classification after removing label $u = \{31 - 45\}$. Results show classification accuracy (%) on training ($\mathcal{D}$), validation ($\mathcal{D}_v$), and test ($\mathcal{D}_t$) data using MUFAC dataset with ResNet-50. The original model achieved 96% accuracy before unlearning.

Models	Retrain	CF-3	SCRUB	UNSIR	SalUN	**WF**	**WP**
Acc on $\mathcal{D}$	92.34	38.45	65.78	82.67	78.89	91.45	93.12
Acc on $\mathcal{D}_v$	93.12	37.89	63.21	81.34	79.23	92.34	92.87
Acc on $\mathcal{D}_t$	91.87	34.67	67.54	83.21	80.45	90.78	94.12

age groups $\mathcal{K} = \{$0-6, 13-16, 20-30, 31-45, 46-60$\}$, with a pre-trained ResNet-50 model and unlearning $u = \{$31-45$\}$ age label. Hence, $\mathcal{W}_f$ consists of parameters with large influence on label u. After unlearning this specific experiment, we implement distance-based heuristics to reassign instances from the unlearned label to neighboring retained labels based on decision boundary , as demonstrated in [8]. Table 3 demonstrates that our proposed methods significantly outperform baselines in maintaining classification accuracy. Weight pruning achieved consistently high performance (92-94%) across all evaluation sets, with Weight Filtering showing similar efficiency (90-92%). In contrast, baseline methods struggled with precise parameter adjustments needed for specific label unlearning in the MUFAC dataset, with CF-3 showing severe degradation (34-38%), and SCRUB (63-67%), UNSIR (81-83%) and SalUN (78-80%) demonstrating moderate performance.

7.2 Privacy Guarantee

Effective unlearning requires complete knowledge removal from model parameters to prevent information leakage through any pathway. We evaluate label-

Table 4. Success rates (%) for Attribute Inference Attack (AIA) and Membership Inference Attack (MIA) after unlearning on the CelebA dataset. Lower scores indicate better privacy; 50% denotes ideal unlearning.

Attack Type	Retrain	CF-3	SCRUB	UNSIR	SalUN	WF	**WP**
AIA	50.13	72.00	81.00	78.50	84.00	65.00	**46.00**
MIA	50.06	68.50	76.30	74.20	79.10	62.40	**50.06**

level unlearning using two complementary frameworks: Attribute Inference Attack (AIA) and Membership Inference Attack (MIA), which assess whether label-specific information remains discoverable after unlearning [39][40][38]. We present results for the CelebA dataset (experimental setting as section 7.1) as it contains rich demographic attributes that are particularly challenging to unlearn due to their entangled representations in the model's parameter space. This dataset provides the most stringent test case for privacy guarantees in facial attribute recognition systems.

Table 4 consolidates the observed success rates of AIA and MIA across all evaluated methods, highlighting the superior privacy performance of our Weight Pruning approach relative to both retraining and competitive baselines. As for AIA, the Retrain baseline achieved near-random prediction rates (50.13%), indicating optimal attribute removal. Among the baselines, CF-3 showed moderate information leakage (72%), while SCRUB, UNSIR, and SalUN demonstrated substantial retained knowledge (75-85%). Our Weight Filtering method achieved improved protection (65%), while Weight Pruning performed exceptionally well (46%), actually pushing the attacker's inference capabilities below random guessing by introducing uncertainty that actively confounds attribute inference attempts. Similarly, MIA results showed our Weight Pruning method closely aligned with retraining (50.06%), effectively eliminating both explicit representations and implicit correlations of forgotten label information. Our method achieves near-minimal privacy leakage by minimizing the KL-divergence between confidence distributions of in-label and out-label samples. Our parameter space-based unlearning framework ensures strong privacy with theoretical limits on information leakage, as confirmed by empirical results against advanced inference attacks. For AIA and MIA, a score near 50% denotes optimal unlearning.

7.3 Runtime Analysis

We analyze the computational efficiency of different unlearning methods by examining their execution times for unlearning a label (experimental setting MLC). All methods demonstrate significantly reduced computational costs compared to complete retraining as shown in Figure 2. For CelebA dataset, Weight Filtering and Weight Pruning require only 34 and 12 seconds, respectively, representing a speed-up factor of approximately 9.15x and 27.45x compared to retraining. These efficiency gains are even more pronounced on the larger VggFace2 dataset (3.31 million images), where our methods achieve remarkable speed-up factors

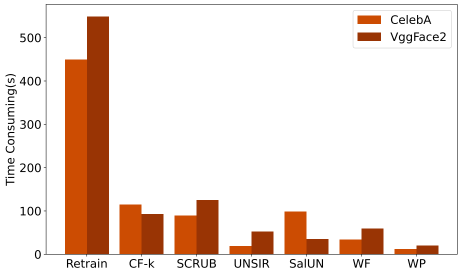

Fig. 2. The time it takes to run each unlearning method to unlearn a class u in MLC experiemnts section 7.1. The *"Retrain"* time represents the time it takes to learn from scratch.

of 13.2x and 35.5x. Weight Pruning demonstrates superior efficiency and is more suitable for large-scale deployment, showcasing its practical value for real-world MU applications.

7.4 Distribution of Entropy of Model Output

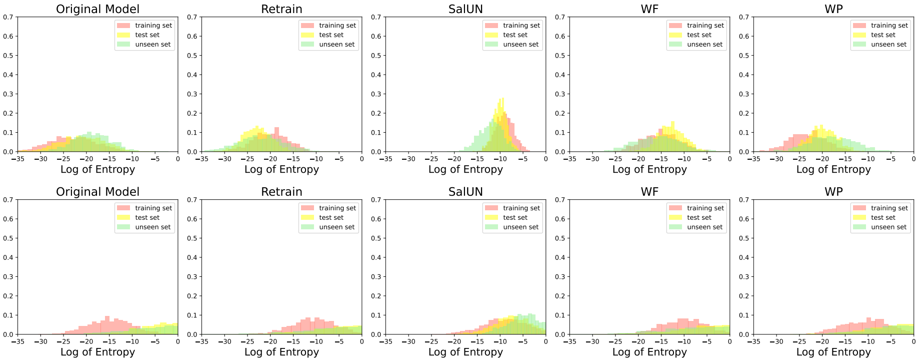

Fig. 3. Entropy distribution analysis across data partitions (training, test, unseen sets). **First row:** CelebA dataset with MLC attributes $y = \{u, k\}$ (Section 7.1). **Second row:** MUFAC dataset classifying single-label $y = \{k\}$ while unlearning u label (Section 7.1). Distributions show entropy values before unlearning ('Original Model') compared with baseline methods and approaches.

We assess unlearning effectiveness by analyzing the model's loss distributions (Binary-Cross Entropy for MLC and Cross-Entropy for single-class classification). Effective unlearning yields entropy patterns like those of a *Retrain* model; deviations suggest incomplete unlearning or leakage (Streisand effect)

[41]. Figure 3 shows entropy distributions for CelebA (MLC) and MUFAC (SLC) datasets. The original model has low entropy across all sets, with the Retrain model slightly increasing entropy across training ($\mathcal{D}$), validation ($\mathcal{D}_v$), and test ($\mathcal{D}_t$) sets. In MLC, SalUN's higher entropy hints at leakage and incomplete unlearning. Weight Filtering and Pruning methods maintain distribution patterns, confirming successful targeted forgetting while preserving model integrity.

8 Conclusion

This paper introduces a parameter space-based framework for multi-label unlearning in facial attribute classification systems. Our Weight Filtering and Weight Pruning methods selectively remove specific attribute knowledge while preserving shared representations essential for retained attributes, without solely relying on the original training data. Our experiments show that our approach surpasses current methods; Weight Pruning achieves a 35.5× speedup over retraining, keeping retained label accuracy above 93% and lowering forgotten attribute predictions to 0.11%. Privacy analysis reveals a 46% AIA score, hindering inference beyond random guessing, with MIA results (50.06%) comparable to full retraining. These results establish a new benchmark for responsible facial attribute classification systems under privacy regulations. The impact on identity verification is not yet fully understood, posing a challenge for machine unlearning. We suggest a pilot study to ensure accuracy when users withdraw consent, though we currently make no broad identity claims. Future research will scale to larger architectures and refine privacy-utility tradeoffs in multi-label unlearning.

References

1. B. Attard-Frost, A. De los Ríos, D. R. Walters, "The ethics of AI business practices: a review of 47 AI ethics guidelines," *AI and Ethics*, vol. 3, no. 2, pp. 389–406, 2023.
2. E. Gündoğdu, A. Unal, G. Unal, "A Study Regarding Machine Unlearning on Facial Attribute Data," *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1-5, 2024. doi: 10.1109/FG59268.2024.10581972.
3. S. Zhang, Y. Feng, N. Sadeh, "Facial recognition: Understanding privacy concerns and attitudes across increasingly diverse deployment scenarios," *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pp. 243–262, 2021
4. "Regulation(EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/ec (General Data Protection Regulation)," *OJ*, vol. L 119, pp. 1-88, 2016.
5. L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, N. Papernot, "Machine unlearning," *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2021.
6. H. Xu, T. Zhu, L. Zhang, W. Zhou, P. S. Yu, "Machine Unlearning: A Survey," *ACM Comput. Surv.*, vol. 56, no. 1, Article 9, August 2023.
7. S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, "Multi-Class Unlearning for Image Classification via Weight Filtering," *IEEE Intelligent Systems*, pp. 1-8, 2024.

8. M. Chen, W. Gao, G. Liu, K. Peng, C. Wang, "Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7766–7775, 2023.
9. S. Goel, A. Prabhu, A. Sanyal, S. Lim, P. Torr, P. Kumaraguru, "Towards adversarial evaluations for inexact machine unlearning," *arXiv preprint arXiv:2201.06640*, 2022.
10. A. K. Tarun, V. S. Chundawat, M. Mandal, M. Kankanhalli, "Fast Yet Effective Machine Unlearning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-10, 2023.
11. M. Priyadharshini, A. F. Banu, B. Sharma, S. Chowdhury, K. Rabie, T. Shongwe, "Hybrid Multi-Label Classification Model for Medical Applications Based on Adaptive Synthetic Data and Ensemble Learning," *Sensors*, 2023. doi: 10.3390/s23156836
12. C. Gérardin et al., "Multilabel classification of medical concepts for patient clinical profile identification," *Artificial Intelligence in Medicine*, 2022. doi: 10.1016/j.artmed.2022.102311
13. S. Sai, U. Mittal, V. Chamola, K. Huang, I. Spinelli, S. Scardapane, Z. Tan, A. Hussain, "Machine un-learning: an overview of techniques, applications, and future directions," *Cognitive Computation*, vol. 16, no. 2, pp. 482–506, 2024.
14. D. Choi, S. Choi, E. Lee, J. Seo, D. Na, "Towards Efficient Machine Unlearning with Data Augmentation: Guided Loss-Increasing (GLI) to Prevent the Catastrophic Model Utility Drop," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 93-102, June 2024.
15. D. Choi, D. Na, "Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems," *arXiv preprint arXiv:2311.02240*, 2023.
16. H. Asi, J. Duchi, A. Fallah, O. Javidbakht, K. Talwar, "Private adaptive gradient methods for convex optimization," *International Conference on Machine Learning*, pp. 383–392, 2021.
17. C. Dwork, "Differential privacy: A survey of results," *International conference on theory and applications of models of computation*, pp. 1–19, 2008.
18. Z. Liu, H. Ye, C. Chen, Y. Zheng, K. Lam, "Threats, attacks, and defenses in machine unlearning: A survey," *arXiv preprint arXiv:2403.13682*, 2024.
19. J. Xu, Z. Wu, C. Wang, X. Jia, "Machine Unlearning: Solutions and Challenges," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 3, pp. 2150-2168, 2024.
20. M. Kurmanji, P. Triantafillou, J. Hayes, E. Triantafillou, "Towards unbounded machine unlearning," *Advances in neural information processing systems*, vol. 36, 2024.
21. C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, S. Liu, "Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation," *arXiv preprint arXiv:2310.12508*, 2023.
22. X. Liu et al., "Emotion classification for short texts: an improved multi-label method," *Humanities and Social Sciences Communications*, 2023. doi: 10.1057/s41599-023-01816-6.
23. H. Fallah, E. Bruno, P. Bellot, E. Murisasco, "Exploiting Label Dependencies for Multi-Label Document Classification Using Transformers," *Proceedings of the ACM Symposium on Document Engineering 2023*, pp. 1–4, Aug. 2023. doi: 10.1145/3573128.3609356
24. A. Warnecke, L. Pirch, C. Wressnegger, K. Rieck, "Machine unlearning of features and labels," *arXiv preprint arXiv:2108.11577*, 2021.

25. R. Chen, J. Yang, H. Xiong, J. Bai, T. Hu, J. Hao, Y. Feng, J. T. Zhou, J. Wu, Z. Liu, "Fast model debias with machine unlearning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
26. A. Chan, A. Gujarati, K. Pattabiraman, S. Gopalakrishnan, "Hierarchical Unlearning Framework for Multi-Class Classification," *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*, 2024.
27. C. Fan, J. Liu, A. Hero, S. Liu, "Challenging forgets: Unveiling the worst-case forget sets in machine unlearning," *arXiv preprint arXiv:2403.07362*, 2024.
28. W. Chang, T. Zhu, H. Xu, W. Liu, W. Zhou, "Class Machine Unlearning for Complex Data via Concepts Inference and Data Poisoning," *arXiv preprint arXiv:2405.15662*, 2024.
29. Z. Liu, P. Luo, X. Wang, X. Tang, "Deep Learning Face Attributes in the Wild," *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
30. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74, 2018.
31. A. Sekhari, J. Acharya, G. Kamath, A. T. Suresh, "Remember what you want to forget: Algorithms for machine unlearning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18075–18086, 2021.
32. Huang, S., Hu, W., Lu, B., Fan, Q., Xu, X., Zhou, X., & Yan, H. (2024). "Application of Label Correlation in Multi-Label Classification: A Survey". Applied Sciences, 14(19), 9034.
33. E. Triantafillou et al., "NeurIPS 2023 - Machine Unlearning," Kaggle, 2023. [Online]. Available: https://kaggle.com/competitions/neurips-2023-machine-unlearning.
34. E. Kubiak, M. I. Efremova, S. Baron, K. J. Frasca, "Gender equity in hiring: examining the effectiveness of a personality-based algorithm," *Frontiers in psychology*, vol. 14, 2023.
35. C. Y. Johnson, *Book Chapter: Racial Bias in a Medical Algorithm Favors White Patients over Sicker Black Patients (1st Ed.)*. Auerbach Publications, 2022, ISBN: 9781003278290.
36. A. Rosales, J. Linares-Lanzman, "Yes, dating apps discriminate against older users," *COMeIN [online]*, no. 142, April 2024.
37. M. C. Kaufmann, F. Krings, L. A. Zebrowitz, S. Sczesny, "Age Bias in Selection Decisions: The Role of Facial Appearance and Fitness Impressions," *Frontiers in psychology*, vol. 8, 2017.
38. R. Shokri, M. Stronati, C. Song, V. Shmatikov, "Membership inference attacks against machine learning models," *IEEE symposium on security and privacy (SP)*, pp. 3–18, 2017.
39. Jia, J., & Gong, N. Z. (2018). AttriGuard: A practical defense against attribute inference attacks via adversarial machine learning. In 27th *USENIX Security Symposium* (USENIX Security 18) (pp. 513-529).
40. Lu, Z., Liang, H., Zhao, M., Lv, Q., Liang, T., & Wang, Y. (2022). Label-only membership inference attacks on machine unlearning without dependence of posteriors. *International Journal of Intelligent Systems*, 37(11), 9424-9441.
41. J. Hagenbach, F. Koessler, "The Streisand effect: Signaling and partial sophistication," *Journal of Economic Behavior & Organization*, vol. 143, pp. 1–8, 2017.