Feature Unlearning: Theoretical Foundation and Practical Applications with Shuffling

Yue Yang[1,2], Jinhao Li[2], Hao Wang[2]

[1] Maincode
[2] Monash University

Background – Machine Unlearning and Feature Unlearning

MONASH University

What is machine unlearning?

- Remove specific information from trained models
- Forget certain data points or features upon request
- Be more effective compared to retraining from scratch

Why is machine unlearning important?

- Privacy concerns
- Ethical use of data
- Regulatory framework – the right to be forgotten
  - EU General Data Protection Regulation[1]

Why feature unlearning?

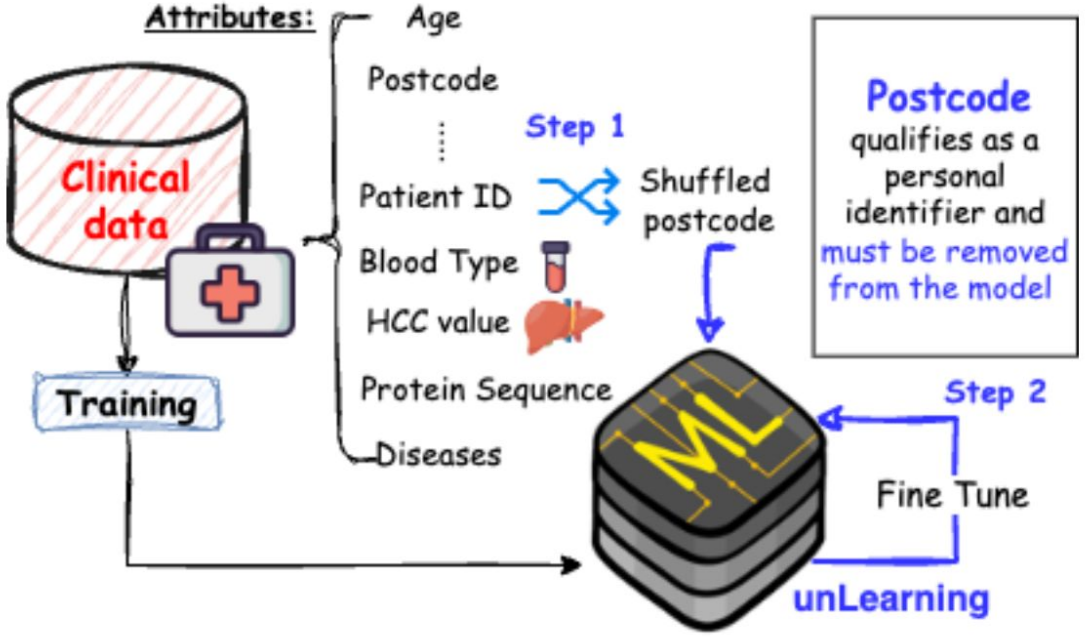
- Most machine unlearning focuses on instance-based unlearning
  - Remove the impacts of a certain subset of data samples
- Unique challenges in feature unlearning
  - Cannot disrupt the influence of others
  - Avoid affecting the learned relationships and decision boundaries associated with other features
- Literature review[2]-[4]
  - **Performance degradation** as the number of removed feature increases
  - No **theoretical guarantees** to fully negate the influence of removed features while maintaining the integrity of the remaining features

[1] Voigt et al., The EU general data protection regulation (GDPR), 2017
[2] Warnecke et al.,Machine unlearning of features and labels, NDSS, 2025
[3] Guo et al., Efficient attribute unlearning: Towards selective remove of input attributes from feature representations, arXiv, 2022
[4] Moon et al., Feature unlearning for pre-trained GANs and VAEs, AAAI, 2024

Shuffle-based Feature Unlearning Approach

Contributions

- Rigorous theoretical guarantees
  - The impact of the specified features diminishes to an insignificant level
  - Further validation under the concept of Shapley value and mutual information
- Comprehensive empirical evaluations
  - Eight metrics, four neural models
  - Both tabular and visual datasets (seven in total)
  - Effectiveness of single-feature and multi-feature unlearning



Example of our unlearning method

Approach Overview

1. Draw a random permutation
2. Construct the shuffled dataset by applying the derived permutation on the unlearned feature
3. Fine-tune trained model with the original loss function

Roadmap to Theoretical Unlearning Guarantees

Objective: Prove the model after unlearning is $(\epsilon - \delta)$-close to the model retrained from scratch

First, prove the independency between X_j^π and $(\mathbf{X}_{-j}, Y)$

- The former variable: Unlearned feature
- The latter variables: Remaining features and label

Unlearned feature after shuffling is independent with both remaining features and label

Theorem 4.4. *If X_j^π is independent of $(Y, \mathbf{X}_{-j})$, then using the same loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ and the same optimizer, the model $f_{\theta'}$ trained on the shuffled dataset $\hat{\mathcal{D}}_j^\pi$ and the model g_ϕ trained from scratch on $\hat{\mathcal{D}}_{-j}$ are (ϵ, δ)-close.*

If loss is *k*-strongly convex, for any $f_{\theta'}$ and g_ϕ, they are $(\epsilon - \delta)$-close

If loss is not strongly convex, under mild assumptions for any $f_{\theta'}$, there must exist a g_ϕ such that they are $(\epsilon - \delta)$-close

Insights from Mutual Information and Shapley Values:

Mutual information between unlearned feature and *both label and other features* **vanishes to zero**

Shapley values of non-shuffled features are ***almost surely identical*** between unlearned model and retrained model

Experimental Settings

Dataset

- **Tabular** — Six datasets from OpenML[1]
- **CV** — CelebA[2] Shuffle specific visual features like nose

Evaluation Metrics

- **TRI**[1] — Compare accuracy of unlearned model and retrained model
- **EI**[2] — Assess efficiency of unlearning compared to retraining
- **RASI**[3] — Shuffle unlearned feature and observe the impact on predictions
- **SRI**[4], **SDI**[5] — SRI: Calculate the relative importance of unlearned feature after unlearning; SDI: Calculate the absolute deviation of the SHAP value from zero
- **MIA**[6] **Accuracy** — Privacy metric. Assess whether unlearned model is indistinguishable from retrained model
- **DP**[7], **EO**[8] — Fairness metrics

1: Test retention index
2: Efficiency index
3: Unlearning Robustness against shuffling index
4: SHAP retention index
5: SHAP distance-to-zero index
6: Membership inference attack
7: Demographic parity
8. Equalized odds

Neural Models

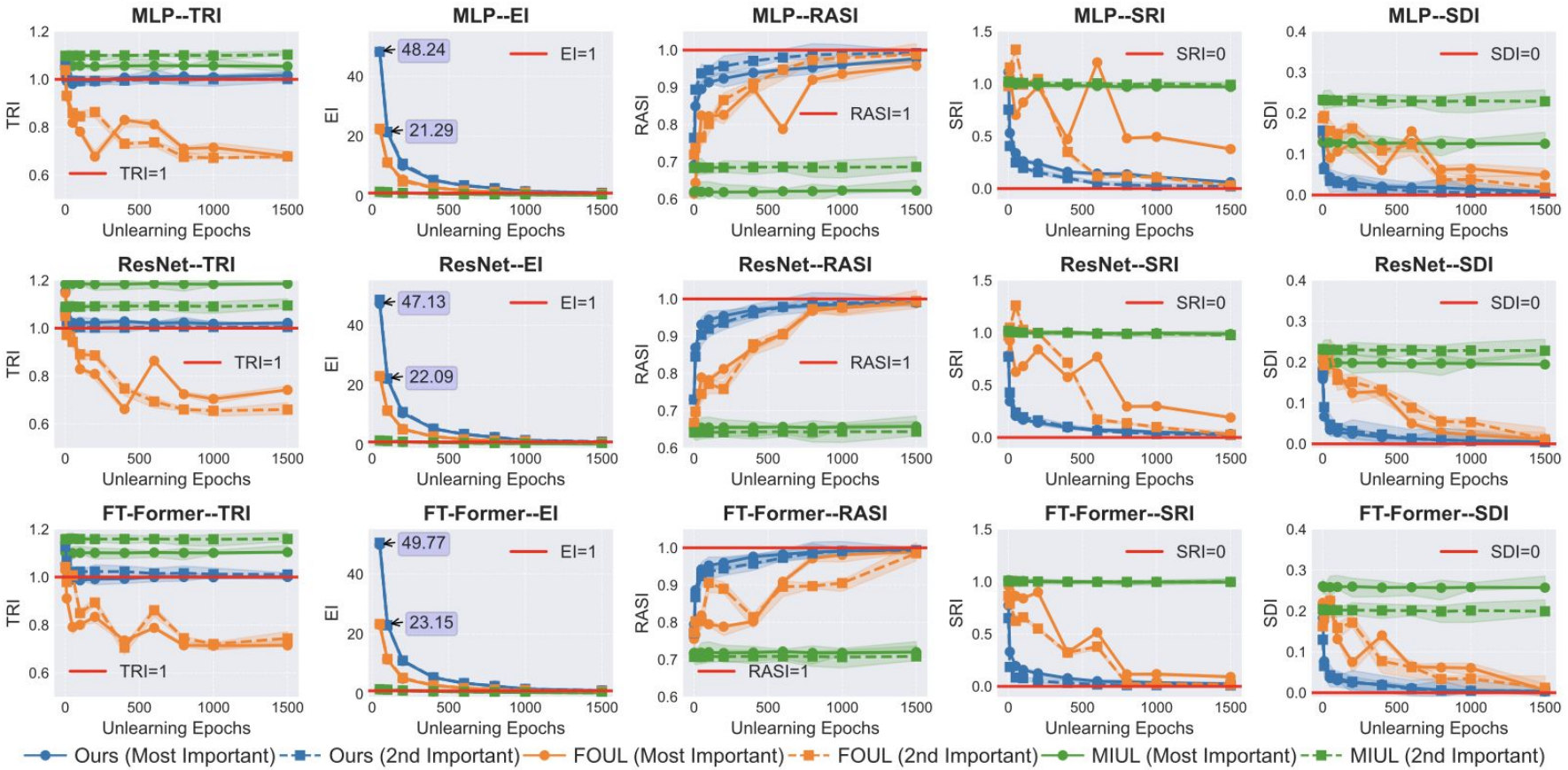
- **Tabular** — MLP; ResNet; FT-Transformer (FT-Former)
- **CV** — Vision Transformer (ViT)

[1] Feurer et al., OpenML-Python: An extensible Python API for OpenML, Journal of Machine Learning Research, 2021
[2] Liu et al., Deep learning face attributes in the wild, IEEE International Conference on Computer Vision (ICCV), 2015

Experimental Results – Tabular Dataset

Baselines FOUL[1]; MIUL[2]

Unlearn the most and second most important features



- ***TRI:*** **close to 1** across all datasets for neural models and features
- ***EI:*** **10 times** faster than retraining from scratch
- ***RASI:*** **close to 1** – predictions not impacted by unlearned feature
- ***SRI&SDI:*** close to zero – feature importance diminished to nearly zero.

[1] Warnecke et al.,Machine unlearning of features and labels, NDSS, 2025
[2] Guo et al., Efficient attribute unlearning: Towards selective remove of input attributes from feature representations, arXiv, 2022

Experimental Results – Tabular Dataset

MIA Accuracy

Method \ Epoch	1	10	50	100	200	400	600	800	1000	1500
Ours	0.544	0.520	0.496	0.509	0.525	0.528	0.525	0.520	0.512	0.511
FOUL	0.574	0.620	0.603	0.588	0.569	0.651	0.751	0.763	0.763	0.763
MIUL	0.651	0.596	0.597	0.597	0.605	0.603	0.595	0.612	0.594	0.614

DP Difference

Method \ Epoch	1	10	50	100	200	400	600	800	1000	1500
Ours	0.0774	0.0330	0.0336	0.0127	0.0271	0.0147	0.0132	0.0206	0.0243	0.0097
FOUL	0.3791	0.2174	0.1654	0.3151	0.2065	0.0417	0.0598	0.1111	0.1054	0.2345
MIUL	0.1869	0.1319	0.0440	0.0954	0.0371	0.0424	0.0895	0.1026	0.0778	0.0291

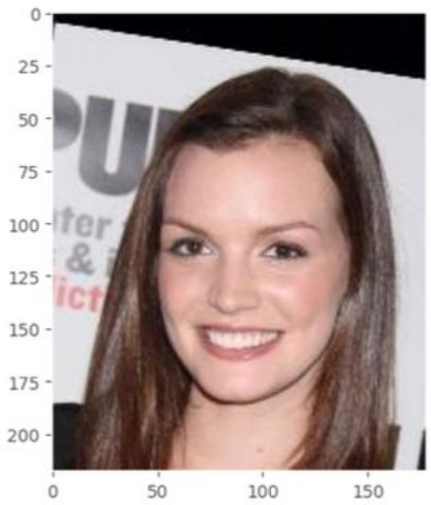
EO Difference

Method \ Epoch	1	10	50	100	200	400	600	800	1000	1500
Ours	0.0394	0.0281	0.0310	0.0158	0.0288	0.0234	0.0175	0.0222	0.0192	0.0192
FOUL	0.1728	0.1463	0.2902	0.1616	0.2665	0.7684	0.5679	0.3727	0.1668	0.1290
MIUL	0.1841	0.2069	0.1477	0.0747	0.0719	0.0601	0.0574	0.0683	0.0958	0.1170

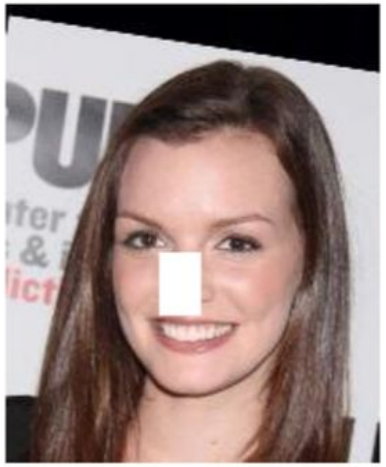
- ***MIA:* close to 0.5** – attack cannot do better than random guessing
- ***DP&EO:* close to zero** – successful unlearning from the fairness perspective

Experimental Results – CV Dataset

Example of shuffled visual features



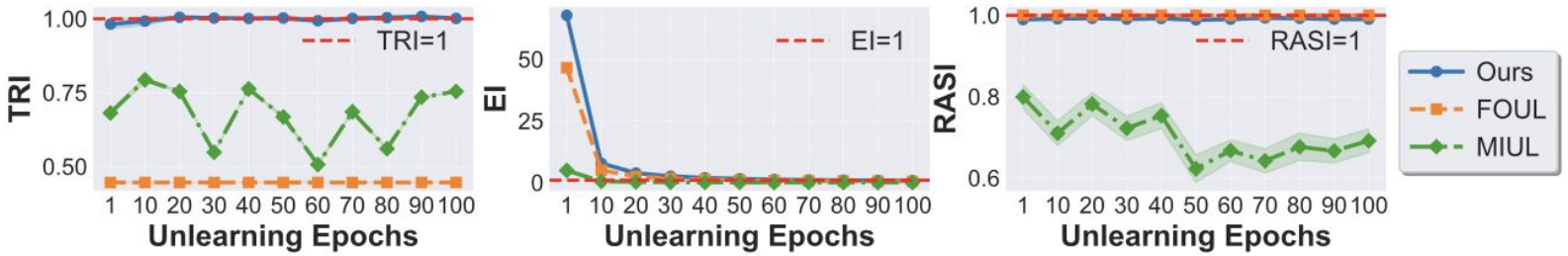
Original Image



Processed Image for Retrain-from-Scratch

Processed Image with Shuffled Feature

Evaluation results of unlearning Nose and Eyes for the label of gender



Our model attains an average ***TRI of 97.31% after 20 epochs and 99.12% after 100 epochs***

Take-Aways

1. We developed a straightforward yet effective method for feature unlearning via shuffling
2. We provided theoretical guarantees that the unlearning outcomes are comparable to retraining a model from scratch
3. We presented comprehensive empirical evaluations over both tabular and visual datasets with SOTA performance

Contact Information

Yue Yang (Maincode; Monash University) yue@maincode.com

Jinhao Li (Monash University): jinhao.li@monash.edu

Hao Wang (Monash University): hao.wang2@monash.edu

https://research.monash.edu/en/persons/hao-wang