Evaluating Machine Unlearning in Finger photo Presentation Attack Detection

Anudeep Vurity · Zhisheng Yan · Massimiliano Albanese

George Mason University, Fairfax, Virginia, USA

Contactless Fingerphoto Authentication

- A finger photo is captured by holding a finger in front of a smartphone camera
- Enhancement algorithms analyse ridge patterns and build a template
- The template represents the individual at verification time
- Fingerphoto-to-Fingerphoto matching has improved substantially over the last decade

A captured finger photo, and the minutiae extracted from it

Examples of Presentation Attack

DISPLAY ATTACK

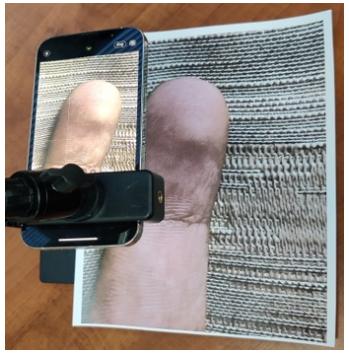
Screen replay

A high-resolution fingerphoto is rendered on a screen and presented to the camera.

Desktop monitor, tablet, smartphone

PRINT ATTACK

Paper printout



A fingerphoto is printed on paper and held in front of the camera.

Laser print, inkjet print, photo paper

OVERLAY ATTACK

Physical replica

A fabricated mould is placed over the finger to impersonate an enrolled subject.

Silicone, latex, play-doh, dragon skin

All three present an artefact to the sensor rather than a live finger. Presentation Attack Detection Algorithms detect them.

Identity Leakage in Binary PAD Models

Supervision provided: **is this a real finger or a fake?**

Representation learned: **whose finger is this?**

No identity supervision

Subject labels are withheld from PAD training. The objective is cross-entropy over two classes.

Subject signal in the input

RGB fingerphotos encode ridge flow, pore distribution and skin texture at every spatial location.

Erasure is under-specified

Deleting enrolment images does not remove subject information retained in the parameters θ.

MOTIVATION

Regulatory Basis for Model-Level Erasure

GDPR
European Union

Right to erasure covers learned model state, not just storage.

CCPA
California, USA

Equivalent deletion rights over personal information.

PIPEDA
Canada

Retention limits and secure disposal once data is no longer needed.

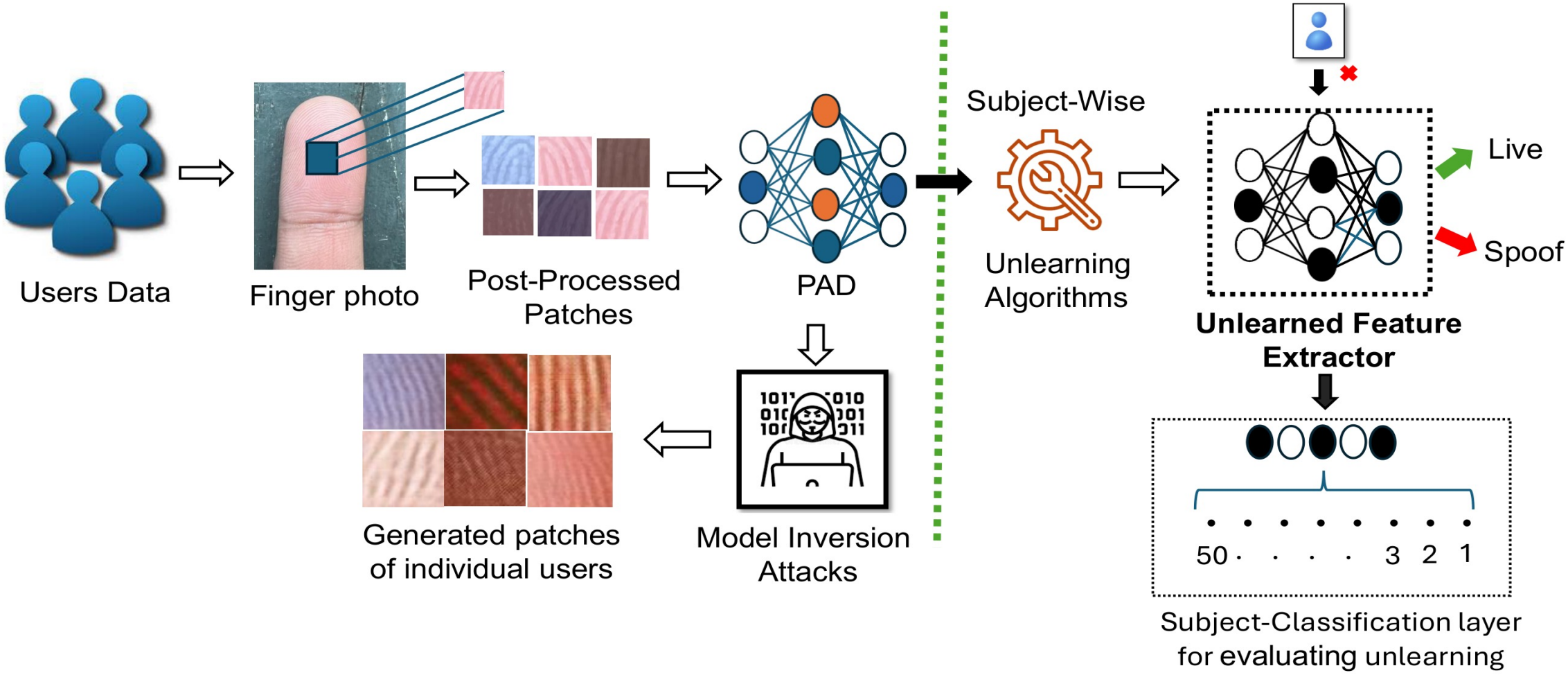
1 Do PAD models leak identity?

- Four model inversion attacks reconstruct training-like images from the trained detector.
- A subject-identification layer on the frozen detector measures how well the training subject is recovered.

2 Can unlearning take that identity back out?

- Ten unlearning strategies, from retraining as the gold standard through gradient ascent, relabeling, saliency and distillation.
- Seven architectures, three public datasets. The first systematic unlearning study for fingerprint models.

Experimental Pipeline



Left of the dotted line: train the detector, then attack it. Right: unlearn one subject, then test whether that subject is really gone.

PROPOSED APPROACH

Subject-wise Evaluation Protocol

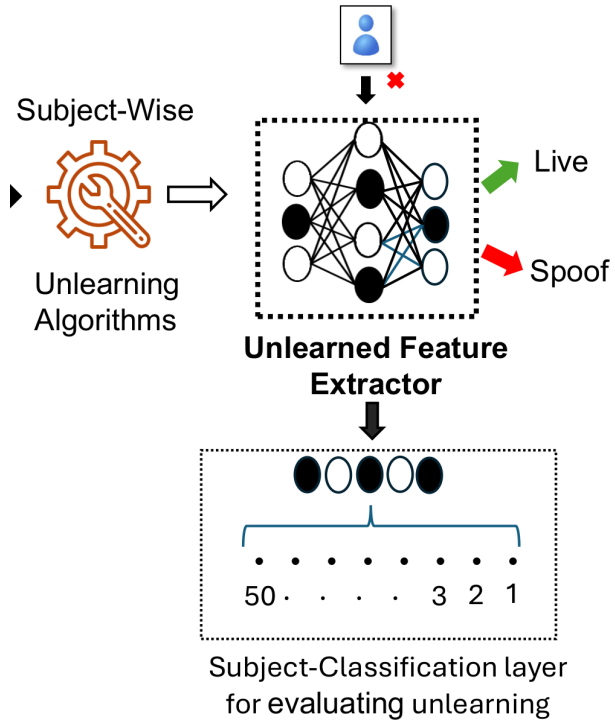
The PAD has two outputs. Neither is informative about whether a specific subject remains encoded in the backbone.

1. **Freeze the detector**
   After unlearning, nothing below the output layer moves again.

2. **Swap in a subject layer**
   Replace the real-or-fake layer with one that predicts which person the photo belongs to.

3. **Test on unseen photos**
   Score it on samples held out from both training and unlearning.

4. **Evaluate**
   Forget accuracy on the deleted person. Retain accuracy on everyone else.



If a probe can still pick you out of the frozen features, you are still in the model.

DATABASE

Datasets and Evaluation Scope

MFPAD-i-22

112 subjects

iPhone 13 Pro

Display and print spoofs, indoor and outdoor.

MFPAD-g-23

100 subjects

Google Pixel 3 XL

Display and print spoofs, varying illumination.

IIIT-D

64 subjects

iPhone 5

The oldest and least challenging of the three.

Seven PAD architectures

ResNet18 · ResNet50 · VGG16 · EfficientNet-B0
EfficientNet-B5 · Swin Transformer · ViT

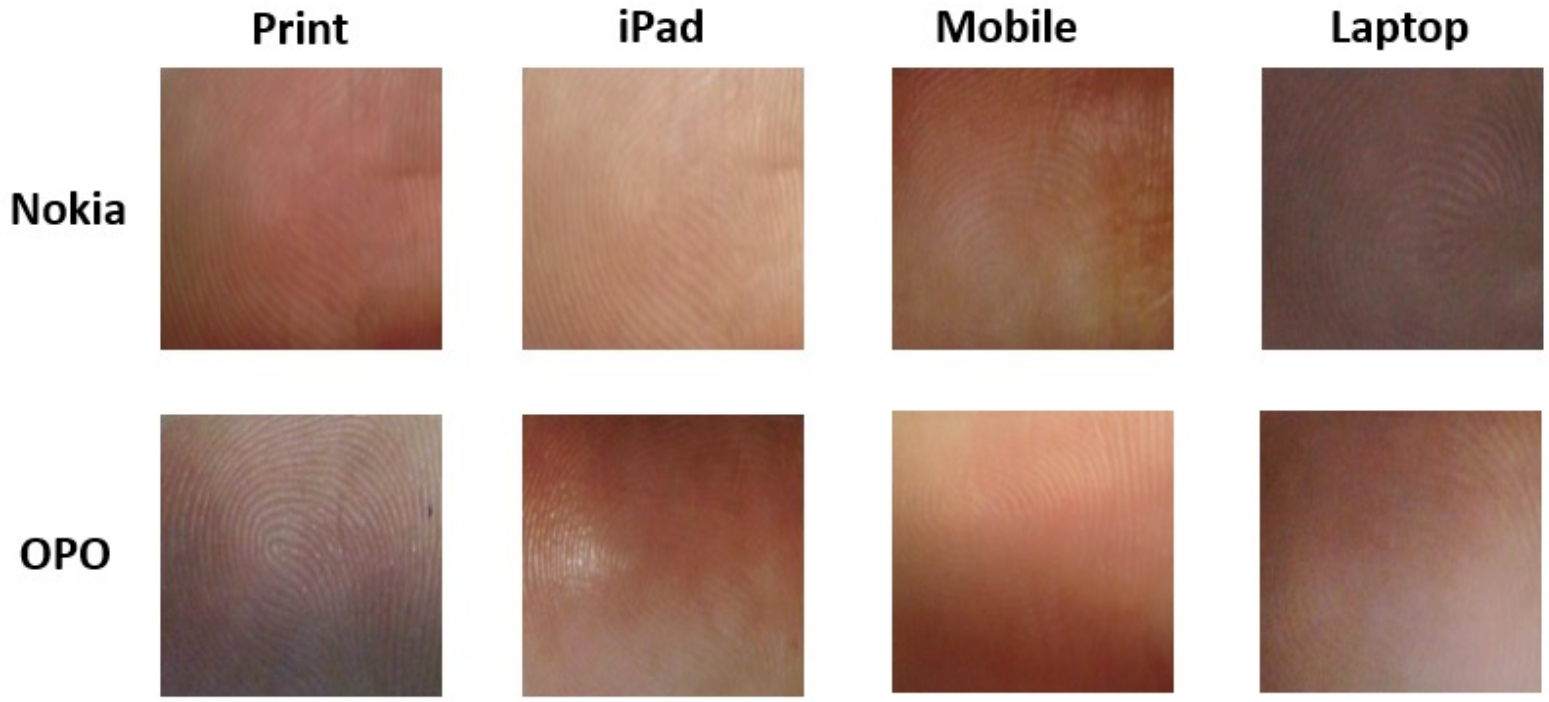
All ImageNet-initialized, then fine-tuned identically.

Ten unlearning methods

Retrain · Finetune · Gradient Ascent · Bad Teacher
SalUn · Random Relabel · L2UL · Boundary Shrink · Fisher · DELETE

Every method edits weights only. No architectural change.

IIIT-D Dataset Patches



Subject-Level Partitioning Protocol

1

Localize and crop

Finger regions are found with a Faster R-CNN detector. Samples with unreliable region extraction are discarded, and every experiment runs on cropped patches.

2

Split subjects in half

Training and test subjects are disjoint, roughly half each. Test subjects are never touched by any experiment.

3

Hold out 20% per subject

For every training subject, a fifth of their photos are set aside. Those are excluded from training and from unlearning, and are what the probe is scored on.

4

Delete one subject at a time

All training samples of one subject form the forget set; everything else is the retain set. One at a time isolates subject-specific features.

METRICS

Metric Definitions and Directions

PAD performance

ACC ↑
Spoof-detection accuracy. Higher is better.

EER ↓
Equal error rate, where false accepts equal false rejects.

APCER / BPCER ↓
ISO/IEC 30107 error rates, read from ROC curves.

Privacy leakage

FID ↓
Distance between reconstructions and real fingerprints. Lower means the inversion attack worked better.

Dist-L / Dist-S ↓
Feature-space distance to real live and spoof samples.

Unlearning

Forget accuracy ↓
Subject-probe accuracy on the deleted person. Zero means gone.

Retain accuracy ↑
Subject-probe accuracy on everyone kept.

Careful with one thing: PAD accuracy and retain accuracy are different measurements. One is spoof detection, the other is a subject probe on frozen features.

RESULTS

Baseline PAD Performance

Model	MFPAD-g-23		MFPAD-i-22		IIIT-D	
	EER %	ACC %	EER %	ACC %	EER %	ACC %
ResNet18	5.78	96.89	7.94	94.17	1.20	98.54
VGG16	6.48	96.86	7.75	94.21	0.41	98.44
ResNet50	5.41	96.21	7.23	94.64	0.18	98.48
Swin	**4.98**	**97.56**	**4.71**	**96.89**	**0.001**	**99.86**
EffiB0	6.41	96.16	8.23	94.22	0.55	99.18
EffiB5	5.35	96.42	8.11	94.51	0.11	99.59
ViT	6.11	94.34	6.84	95.77	0.41	99.45

Observations

- All seven architectures exceed 94% accuracy on every dataset.
- Swin Transformer attains the lowest EER and highest accuracy on all three.
- IIIT-D is near-saturated: EER below 1.2% for every architecture.

RESULTS

Privacy Leakage under Model Inversion

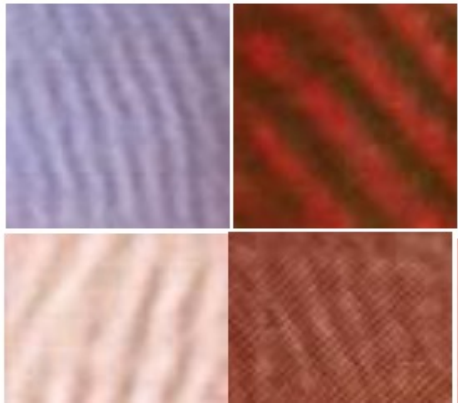
Dataset	Model	PLGMI	KEDMI	GMI	LOKT
MFPAD-i-22	IR152	347.56	**168.82**	**156.78**	387.37
	VGG16	358.13	**174.14**	**148.68**	383.34
	IR152-pretrained	331.82	**161.21**	**151.45**	388.48
	VGG16-pretrained	346.18	**168.82**	**148.68**	389.45
MFPAD-g-23	IR152	282.17	**142.38**	**145.02**	357.61
	VGG16	294.33	**156.14**	**160.66**	384.22
	IR152-pretrained	297.28	361.09	**141.13**	361.09
	VGG16-pretrained	263.05	355.29	**158.44**	355.29

FID measures distributional distance between reconstructions and real fingerphotos. Lower values indicate stronger inversion. Shading marks the strongest attack per row.

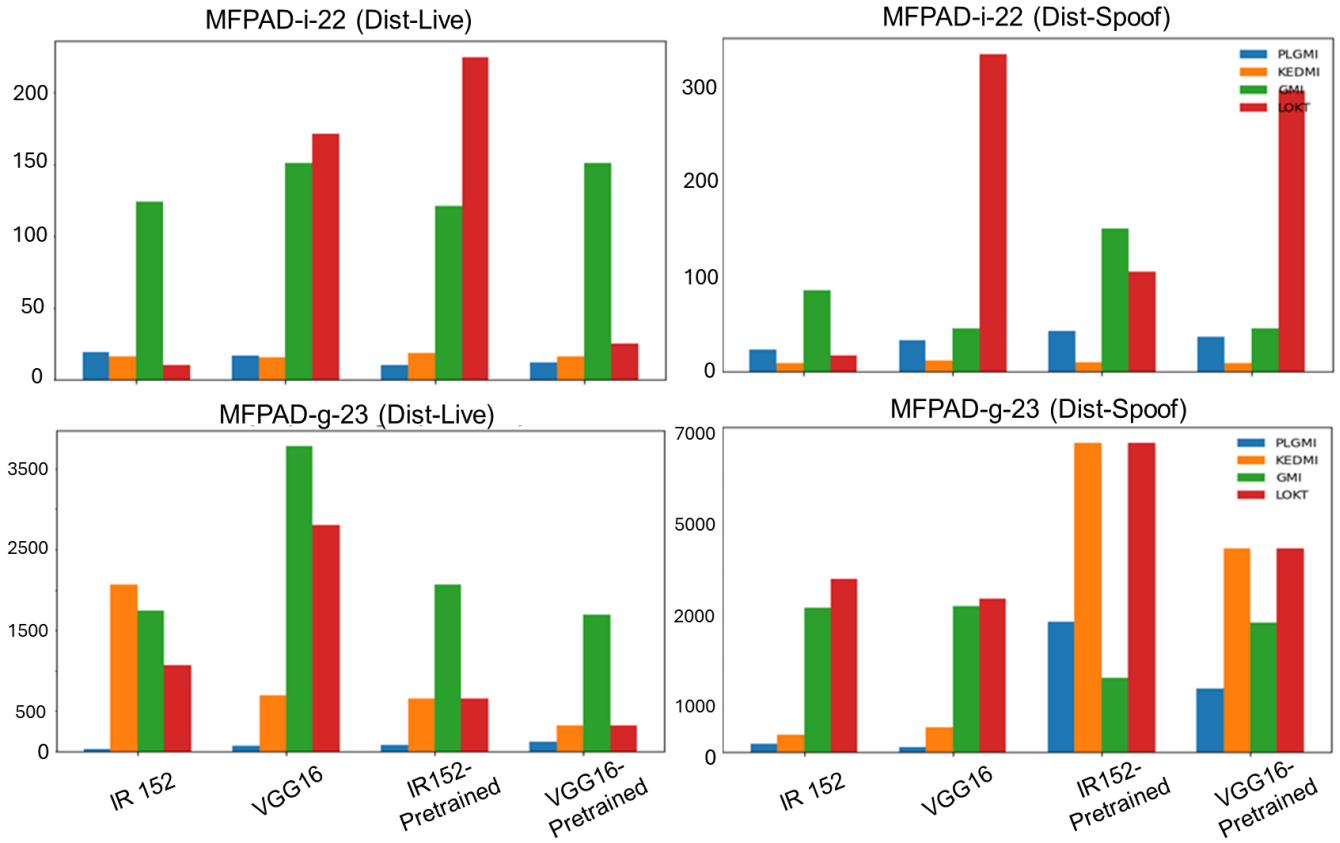
Observations

- GMI and KEDMI attain the lowest FID on both datasets and both backbones.
- Attack ranking is invariant to dataset and backbone, indicating systematic extraction rather than stochastic reconstruction.

Reconstructions from a model supervised only on bona fide versus spoof labels.



Reconstruction Distances



GMI and KEDMI stay below the others on both classes, matching the FID ranking. Leakage is not confined to one class.

These models discriminative structure for bona fide and spoof samples alike.

Subject-Wise Unlearning: MFPAD-i-22

Retain accuracy / forget accuracy (%). MFPAD-i-22 is the harder of the two mobile datasets.

Method	ResNet18	VGG16	ResNet50	Swin	EffiB0	EffiB5	ViT
Original	92.21 / 88.51	93.74 / 89.12	90.67 / 88.47	88.46 / 82.91	91.17 / 92.31	92.47 / 91.88	92.31 / 91.44
Retrain	90.11 / 0.00	93.11 / 0.00	87.42 / 0.00	86.71 / 0.00	90.41 / 0.00	91.66 / 0.00	90.14 / 0.00
Finetune	88.41 / 0.00	91.11 / 0.00	85.74 / 0.00	87.57 / 0.00	88.48 / 2.15	81.40 / 4.11	85.57 / 6.45
GA	87.12 / 6.42	86.51 / 5.44	84.11 / 6.54	84.14 / 1.56	85.48 / 5.48	**74.55 / 54.69**	88.74 / 0.00
BT	82.15 / 7.84	82.44 / 7.99	83.45 / 7.44	84.47 / 10.45	**81.15 / 70.14**	77.48 / 0.00	87.74 / 0.00
SalUn	90.15 / 12.48	90.41 / 8.77	78.46 / 12.41	82.15 / 17.44	**96.15 / 87.17**	**90.22 / 55.45**	87.87 / 0.00
RL	87.44 / 12.45	82.14 / 14.44	85.55 / 5.48	86.41 / 3.88	79.48 / 2.44	84.74 / 0.00	87.33 / 13.47
L2UL	87.41 / 4.58	88.41 / 10.40	80.17 / 14.41	82.41 / 12.44	**83.54 / 37.48**	51.24 / 0.00	86.22 / 0.00
BS	86.77 / 0.00	80.77 / 0.00	79.54 / 0.00	80.74 / 0.00	79.48 / 25.41	84.54 / 0.00	81.47 / 0.00
Fisher	84.56 / 0.00	89.47 / 0.00	83.41 / 0.00	74.41 / 0.00	79.44 / 4.85	81.87 / 4.51	82.44 / 7.88
Delete	90.85 / 0.00	90.41 / 0.00	86.42 / 0.00	85.14 / 0.00	85.77 / 4.44	87.12 / 0.00	89.55 / 0.00

SalUn leaves 87.17% on EfficientNet-B0, Bad Teacher leaves 70.14% on the same backbone, gradient ascent leaves 54.69% on B5. Retrain is exactly zero everywhere.

RESULTS

Subject-Wise Unlearning: MFPAD-g-23

Each cell is retain accuracy / forget accuracy (%). Shading marks forget accuracy above 5, 15 and 30%.

Method	ResNet18	VGG16	ResNet50	Swin	EffiB0	EffiB5	ViT
Original	95.17 / 97.41	95.11 / 96.14	94.47 / 97.42	93.22 / 79.41	95.12 / 96.42	95.41 / 87.14	94.61 / 65.62
Retrain	94.45 / 0.00	94.07 / 0.00	94.21 / 0.00	92.84 / 0.15	93.41 / 0.00	94.21 / 0.00	93.11 / 0.00
Finetune	91.17 / 0.00	91.14 / 5.14	92.01 / 0.00	92.14 / 1.45	94.84 / 9.74	93.14 / 0.00	97.74 / 6.74
GA	90.79 / 0.00	90.78 / 0.00	90.74 / 0.00	87.10 / 0.00	93.41 / 5.41	90.92 / 0.00	90.71 / 0.00
BT	93.21 / 5.15	90.78 / 0.00	92.47 / 0.00	56.14 / 0.00	92.41 / 21.42	**85.56 / 41.58**	90.85 / 0.00
SalUn	90.14 / 12.54	89.02 / 12.45	88.44 / 21.74	76.48 / 14.58	48.74 / 27.14	46.42 / 12.48	89.74 / 12.44
RL	**94.11 / 50.10**	69.37 / 8.45	100.00 / 9.14	85.41 / 2.15	58.24 / 0.00	64.48 / 0.00	69.78 / 8.77
L2UL	90.44 / 4.51	90.14 / 24.48	90.47 / 0.00	87.47 / 0.00	**93.74 / 37.41**	**84.57 / 38.48**	90.77 / 0.00
BS	89.74 / 0.00	91.48 / 0.00	92.31 / 22.14	88.74 / 7.11	91.74 / 11.58	91.48 / 6.47	91.58 / 0.00
Fisher	74.41 / 0.00	25.41 / 0.00	71.45 / 0.00	77.30 / 4.45	64.52 / 0.97	67.78 / 8.58	77.07 / 5.25
Delete	93.46 / 0.00	92.11 / 0.00	93.47 / 0.00	92.47 / 0.00	92.37 / 0.00	93.74 / 0.00	92.65 / 0.00

Before unlearning the model identifies the deleted subject up to 97.4% of the time. Retrain and DELETE take it to zero. Random relabeling leaves 50.1% on ResNet18, Bad Teacher leaves 41.6% on EfficientNet-B5.

Subject-Wise Unlearning: IIIT-D

Retain accuracy / forget accuracy (%). IIIT-D is the oldest and easiest dataset, so the exceptions matter more.

Method	ResNet18	VGG16	ResNet50	Swin	EffiB0	EffiB5	ViT
Original	90.91 / 62.41	92.88 / 66.48	92.88 / 70.55	87.31 / 78.33	94.88 / 83.14	70.94 / 47.11	92.24 / 79.40
Retrain	89.61 / 0.00	90.91 / 0.00	90.94 / 0.00	84.97 / 0.00	95.48 / 0.33	69.88 / 0.00	90.87 / 0.00
Finetune	86.95 / 0.00	86.94 / 0.00	86.95 / 0.00	87.89 / 0.00	96.92 / 0.00	69.94 / 0.00	90.88 / 0.00
GA	87.25 / 0.00	84.05 / 0.00	64.35 / 24.55	77.21 / 14.54	61.92 / 0.51	61.91 / 0.78	83.95 / 0.00
BT	80.69 / 0.00	81.02 / 0.00	77.36 / 0.00	74.24 / 0.00	64.91 / 0.54	61.24 / 0.44	66.94 / 0.12
SalUn	86.87 / 0.00	87.06 / 0.00	76.13 / 0.00	77.94 / 14.27	90.91 / 4.15	67.88 / 4.55	64.24 / 24.74
RL	80.95 / 0.00	80.92 / 0.00	78.05 / 0.00	90.58 / 0.00	86.80 / 0.00	57.91 / 0.00	77.35 / 0.00
L2UL	78.02 / 0.00	**70.91 / 41.55**	60.91 / 4.74	60.91 / 0.51	73.88 / 0.11	60.94 / 0.00	66.94 / 0.00
BS	69.95 / 0.00	60.91 / 2.14	77.05 / 0.00	79.41 / 0.15	**74.91 / 45.72**	60.95 / 0.00	60.94 / 0.00
Fisher	73.91 / 0.00	78.02 / 0.85	78.21 / 0.36	76.91 / 0.00	88.91 / 0.00	64.24 / 0.00	83.91 / 0.00
Delete	88.13 / 0.00	84.95 / 0.00	84.05 / 0.00	83.98 / 0.00	**63.59 / 37.55**	64.94 / 0.00	86.95 / 0.00

Boundary shrink leaves 45.72% on EfficientNet-B0 and L2UL leaves 41.55% on VGG16. DELETE leaves 37.55% here after scoring zero on every architecture in MFPAD-g-23.

CONCLUSION

Findings and Implications

Findings

- MFPAD-i-22 leaks most consistently: 82.9 to 92.3% across architectures.
- Newer capture hardware records richer ridge and texture detail.
- Model inversion reconstructs recognizable ridge structure, FID as low as 141.
- Retraining erases reliably: 0.00 to 0.33% forget accuracy.

Implications

- PAD accuracy is **not a privacy guarantee**.
- Only **retraining currently satisfies** subject-level erasure.
- **Method selection must** be validated per architecture and per dataset.

Compliant biometric PAD requires architecture-aware unlearning

Thank you

Questions welcome

Anudeep Vurity
George Mason University
avurity@gmu.edu