# Privacy Vulnerabilities of the NIST Face Recognition Technology Evaluation

Alexander Goldberg, Giulia Fanti, and Nihar B. Shah

Carnegie Mellon University

In this document, we identify a privacy vulnerability of the NIST Face Recognition Technology Evaluation (FRTE)[1]. The FRTE benchmarks algorithms for two facial recognition tasks: (1) 1:1 face verification where an algorithm identifies if two images of faces come from the same person or not, and (2) 1:N face identification, where the algorithm takes a "gallery" of enrolled face images and a "probe" face image and returns images from the gallery that match the probe image.

We describe **a privacy attack on the 1:N benchmark** whereby a developer with access to a reasonably accurate facial recognition algorithm can detect whether one or more subjects are present in the gallery data used for benchmarking. Concretely, a developer with a face image of a given person can test whether that person's face was included in the FRTE 1:N benchmark datasets with high confidence. The attack is no harder to implement than a valid submission to the benchmark, as the developer only needs to slightly alter the implementation of the search algorithm they submit in order to leak information. Given that FRTE benchmarks algorithms on private datasets such as mugshots, visa applicants, and images taken at border crossings, this privacy vulnerability can compromise the FRTE's intended anonymity guarantees.

**Attacker Goal:** Identify whether a person's face is included in a gallery dataset of the 1:N benchmark.

**Submission:** A correctly implemented facial recognition system with slightly modified version of `Search`.

**Result:** Based on the value of the accuracy statistic released by the benchmark, the attacker learns if the person's face was included in the gallery data or not.

Figure 1: Sketch of privacy attack on FRTE 1:N benchmark.

To demonstrate the practical viability of the attack, we simulate the 1:N face recognition benchmark using the publicly available CelebA dataset, which contains faces of over 10,0000 celebrities [LLWT15]. We use an open-source, deep learning-based facial recognition model ArcFace [DGXZ19] as a baseline submission. **We find that our attack achieves nearly perfect accuracy in identifying whether subjects are present in the gallery dataset, with a True Positive Rate of 0.98 at a False Positive Rate of 0.00.**

In Section 1, we describe the proposed privacy attack on the FRTE 1:N benchmark. In Section 2, we then simulate the attack on publicly available data (a dataset of celebrity faces) using open-source facial recognition algorithms, demonstrating the practical viability of the attack.[2] Finally, in Section 3, we describe potential mitigation measures to address the privacy vulnerability.

[1] The official website for the FRTE can be found at: https://pages.nist.gov/frvt/html/frvt1N.html.

[2] Our code to simulate the attack is available at https://github.com/akgoldberg/face_recognition_privacy_attack. This repository is currently private, and we will share access with NIST on request.

# 1 Privacy Attack on FRTE 1:N

We now describe a privacy attack in which the developer submits an adversarially modified implementation of a facial recognition system to the FRTE 1:N benchmark and learns information about individuals included in the benchmark's gallery dataset based on the released accuracy statistics.

We assume that the FRTE 1:N system functions as described in API documentation. The developer implements a submission that supports the following functionality:

- `Enroll Template:` Create a "template" (e.g., an embedding) for that individual based on image(s) provided. This can be used to enroll $N$ subjects in the gallery dataset.

- `Search:` For each probe template, search the entire gallery and return a list of candidate matches and similarity scores for subjects in the gallery.

The benchmark enrolls a dataset of "gallery" images using the provided `Enroll Template` implementation, then enrolls a dataset of "probe" images using the provided `Enroll Template` implementation and finally searches for each of the probe images in the gallery using the developer's implementation of a `Search` function. Our attack assumes the developer provides a reasonably accurate implementation of these functions, modifying only the `Search` function to leak private gallery information. We describe the modified Privacy Attack Search in Algorithm 1.

**Algorithm 1** Privacy Attack Search

```
Require: Probe $P$, gallery set $G$, input target image $I$, implementations of Search and EnrollTemplate.
Ensure: Returns a random search result if image $I$ is in gallery set $G$ and an accurate search result otherwise.

Compute the enrollment template for $I$: $T \leftarrow$ EnrollTemplate($I$).
Search for $I$ in the gallery $G$ using Search: result $\leftarrow$ Search($T, G$).
if The person associated with $I$ is identified in the gallery from result then
    return A random image from $G$ with a random similarity score. {Run random search.}
else
    return Search($P, G$). {Run the correct search process.}
end if
```

The attack enables the developer to infer whether a target face $I$ is included in the gallery dataset. If $I$ is found in the gallery, the attack returns random search results, reducing the system's accuracy. Otherwise, the `Search` function operates correctly, yielding high accuracy. The attack's success depends on two key capabilities:

(1) The ability to create a valid submission with accuracy exceeding that of random search.

(2) Access to a clear image of the target face and the ability to match it with high accuracy against the gallery.

There are current submissions to the FRTE 1:N benchmark with very poor accuracy of $> 0.99$ False Negative Identification Rate (FNIR) on all benchmarks (e.g., the submission *intsysmsu*) but also submissions with high accuracy of $< 0.1$ FNIR on all benchmarks (e.g., *qazsmartvisionai.*) This suggests that it is possible to achieve a wide range of accuracy scores for the current set of benchmarks, meeting capability (1). Additionally, the attacker can use their high-accuracy `Enroll Template` and `Search` functions to identify the target face in the gallery, meeting capability (2). Hence, based on current submissions to the 1:N benchmark, our described attack is feasible. We simulate the attack on publicly available data in Section 2 and demonstrate its high effectiveness using open-source facial recognition models.

# 2 Simulation of the Privacy Attack

We now describe simulation of the 1:N benchmark and privacy attack using the CelebA dataset and ArcFace facial recognition model.

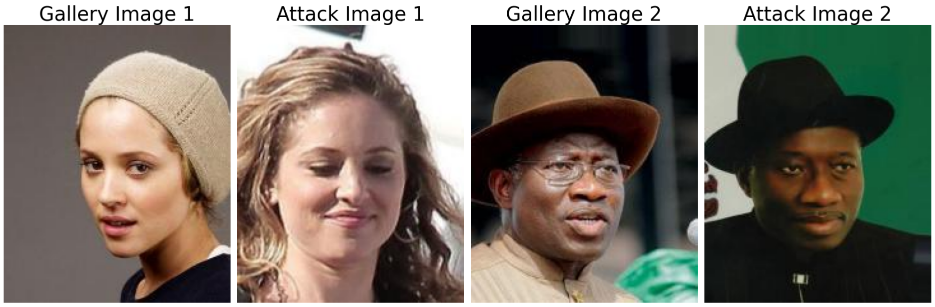

Figure 2: Example of matches by the privacy attack between gallery images and held-out attack images. The displayed attack image and gallery image are of the same person.

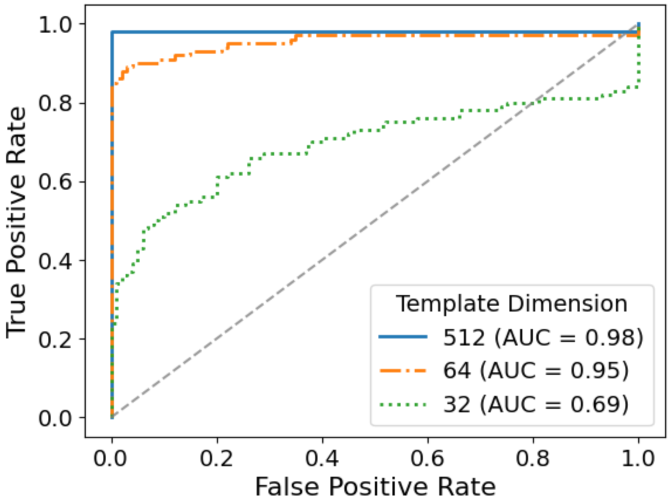

Figure 3: ROC curve of the privacy attack simulated on the CelebA dataset.

**Benchmark:** The CelebA is a dataset of celebrity faces containing 10,177 subjects, that we split into a gallery of 9,119 subjects and 1,058 held-out subjects used as “non-mated” proxies. Following the FRTE 1:N benchmark’s ”Identification” task, we report FNIR at FPIR 0.05, where FNIR is calculated on proxies in the gallery and FPIR on non-mated proxies.

**Face Recognition:** We use ArcFace, an open-source deep learning model, as a baseline submission. ArcFace generates 512-dimensional embeddings for face images. For subjects with multiple images, we use the mean of the image embeddings as a gallery template. Gallery-proxy searches are performed by calculating cosine similarity between embeddings and selecting the highest similarity. This baseline achieves an FNIR of 0.066 at FPIR 0.05. Since the gallery size of our benchmark is smaller than the FRTE gallery size, it may be more difficult to achieve high accuracy on the FRTE than in our simulations. To test our attack with lower accuracy submissions, we degrade the accuracy of the ArcFace model by using reduced embedding dimensions of 64 and 32.

**Privacy Attack:** We implement the privacy attack by modifying the `Search` function as in Algorithm 1. The attack infers that a target image is in the gallery if the benchmark reports FNIR $> 0.5$. We select target attack images by holding out additional images of faces included in the gallery, using these faces exclusively for the attack. Figure 2 shows successful matches between held-out attack and gallery images. We evaluate the attack using 100 gallery-member images and 100 non-member images, calculating true positive and false positive rates by varying the threshold to match an attack target image to a gallery image.

In Figure 3, we show the ROC curve of the attack. Using 512-dimensional embeddings, the attack achieves a TPR of 0.98 at an FPR of 0.00, successfully identifying 98 out of 100 gallery members while avoiding false matches. Even with 64- or 32-dimensional embeddings, the attack remains effective, achieving high AUC scores. This result highlights that the attack is feasible even on a larger FRTE 1:N benchmark using simple, open-source facial recognition models.

# 3 Potential Mitigation

Our privacy attack exploits the release of accuracy statistics by the benchmark directly, rather than relying on a side channel or an implementation bug. As a result, we believe it is unlikely that this vulnerability can be completely eliminated. Instead, we propose three potential mitigation measures to reduce the associated risks:

1. **Additional screening of developers:** The current FRTE benchmark allows submissions from a wide range of developers, with over 150 participants worldwide. The only screening requirement is an organizational email address, which offers minimal oversight. To reduce the risk of ill-intentioned submissions, NIST could implement stricter developer screening procedures, such as verifying credentials, affiliations, or prior work.

2. **Auditing submissions to limit the amount of information leakage:** As described in Section 1, our attack reveals the data of a single individual by exploiting a single released accuracy statistic. However, NIST releases multiple accuracy metrics (e.g., FNIR values across thresholds) along with other measurements like timing and storage statistics. *An attacker could extend the algorithm to reveal membership for multiple individuals by manipulating these additional metrics.*

   To mitigate this, FRTE could audit submissions to test their dependency on individual data points. This audit would split the gallery dataset into two subsets and evaluate the submission's performance on each subset. If results vary significantly between the two subsets, the submission could be flagged as potentially malicious and reported as a failure. This approach rejects submissions that show widely varying performance based on subsets of data, which is indicative of poor robustness in addition to potential privacy risks.

   While this auditing process introduces privacy leakage, it limits information exposure to a binary "pass/fail" outcome, significantly reducing the potential for malicious data extraction. Additionally, developers face an increased risk of being caught if they attempt an attack.

3. **Obfuscating results by addition of random noise:** There is extensive research in Computer Science on releasing statistical information while protecting individual privacy, particularly in the field of Differential Privacy [DR14]. These methods add random noise to statistics to obscure the presence or absence of specific individuals in the dataset. We anticipate that applying these techniques to the FRTE 1:N benchmark without substantially degrading its utility would be challenging. Preventing our proposed attack would require adding enough noise to make accurate submissions indistinguishable from random ones. This level of obfuscation would hinder legitimate developers from reliably assessing whether their submissions outperform random guesses, compromising the benchmark's effectiveness. *Hence, we view this as the least promising of the proposed mitigations.*

# References

[DGXZ19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, aug 2014.

[LLWT15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.