Detecting Training Data Fingerprints in GAN - Generated Medical Images

Notebook for the ImageCLEF Lab at CLEF 2025

Shruti Chandrasekar[1], Vedajanaani R S[1,*], Vijayalakshmi P[1]

[1]*Department of Electronics and Communication Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India*

Abstract

This research addresses the challenge of determining whether specific real medical images were used in the training of generative adversarial networks (GANs) that produce synthetic CT scans, a critical task for ensuring transparency and accountability in AI-generated medical data. As part of Subtask 1 of the ImageCLEFmed GAN 2025 challenge, the problem is framed as a binary classification task where each generated image must be labeled based on the presence or absence of real images in its training data. The proposed method employs deep feature extraction using a ResNet-50 model pretrained on ImageNet. Real and synthetic images are processed to extract high-dimensional embeddings, and cosine similarity is computed between generated images and the pool of real images. A statistical threshold based on the mean and standard deviation of the similarity scores is then used to determine the final label. The system was evaluated on the official test set and achieved an accuracy of 50.8%, precision of 50.78%, recall of 52.0%, and an F1 score of 51.38%. The Cohen's kappa score was 0.016, indicating only slight agreement beyond chance. While the results reflect the inherent difficulty of reverse engineering GAN training data, they demonstrate the potential of feature-based similarity analysis for detecting data usage in synthetic medical imaging.

Keywords

GAN-generated medical images, training data fingerprinting, ResNet-50, deep feature extraction, cosine similarity, logistic regression, medical image synthesis, synthetic CT scans, membership inference, AI transparency, medical image provenance, generative adversarial networks

1. Introduction

The rapid advancement of generative models, particularly Generative Adversarial Networks (GANs), has opened new frontiers in the synthesis of high-quality medical images. While these models hold immense potential for augmenting data and enhancing diagnostic tools, they also raise critical questions about transparency, ethical usage, and data provenance. One key challenge lies in identifying whether a specific real medical image—such as a CT scan—has been used to train a GAN that subsequently generates synthetic images. Addressing this question is vital for ensuring the responsible use of medical AI and for protecting sensitive patient data. This study is conducted as part of the ImageCLEF 2025 [1], specifically within the ImageCLEFmedical 2025 GANs [2] Task which aims to evaluate methods for analyzing GAN-generated medical images. Our work focuses on Subtask 1: "Detect Training Data Usage", which involves identifying whether a given real image was part of the training data used to generate synthetic counterparts. We present the approach and results of Team Medhastra in this subtask, aiming to contribute effective methodologies for data traceability in medical image generation. Our code is available on Github [1].

This research addresses the issue by assigning participants the task of creating automated systems that assess whether each produced medical image is derived from any specific real image within the training dataset.

CLEF 2025 Working Notes, 9 -- 12 September 2025, Madrid, Spain

[*]Corresponding author.

shruti2210139@ssn.edu.in (Shruti Chandrasekar); vedajanaani2310594@ssn.edu.in (Vedajanaani R S); vijayalakshmip@ssn.edu.in (Vijayalakshmi P)

[1]https://github.com/vedajanaani/ImageCLEFmed-GAN-2025.git

This challenge is structured as a binary classification problem and poses considerable technical and methodological difficulties due to the nuanced propagation of visual features in content generated by GANs. Traditionally identifying the source of synthetic images has depended on watermarking or forensic methods, which tend to be either intrusive or have limited effectiveness. These approaches fall short when dealing with deep generative models that are trained on extensive, sensitive datasets, making manual verification impractical and lacking explicit identifiers.

In recent years, deep learning methods, especially convolutional neural networks (CNNs) like ResNet, have shown remarkable proficiency in feature representation and similarity analysis across multiple fields, including medical imaging. Utilizing these models allows researchers to explore the connections between real and generated images within a high-dimensional feature space, employing similarity metrics such as cosine similarity to deduce possible training data applications.

This research proposes a ResNet-based feature extraction and similarity comparison framework. By systematically analyzing pairwise relationships between synthetic and real images, the system aims to make accurate predictions about whether a real image contributed to the training of a GAN that produced a particular synthetic output. The approach emphasizes interpretability and generalization, offering a foundation for further improvements in medical image provenance analysis and AI transparency.

2. Background

Generative Adversarial Networks (GANs) play a crucial role in medical imaging by enabling the creation of realistic medical images for purposes such as data augmentation, domain adaptation, and training simulations. However, their use has sparked concerns regarding the unintentional memorization of training data, which can lead to the generation of images that reveal identifiable features from the original dataset, thereby threatening patient privacy — a significant concern in clinical AI [3], [4].

Studies indicate that deep generative models, especially GANs, can retain specific samples under particular training circumstances, resulting in identifiable data leakage [5]. This concern is particularly critical in the healthcare industry, which is subject to stringent legal frameworks like HIPAA and GDPR that forbid the reidentification of individuals from ostensibly anonymized information. Therefore, there is a pressing requirement to create tools and techniques capable of identifying whether a generated image resembles the training data, a challenge commonly known as training data fingerprinting.

Several detection strategies have been proposed, ranging from direct pixel-space comparisons to embedding-based similarity measures. The latter involves using convolutional neural networks (CNNs), such as ResNet50 [6], to project both real and synthetic images into a high-dimensional feature space where semantic similarity can be assessed more robustly. Similarity metrics, such as cosine similarity, are often applied in this space to determine the extent of overlap or influence between generated and real images [7].

This study introduces a framework that identifies training data fingerprints in GAN-generated medical images through deep feature embeddings and statistical analysis. The process involves extracting feature vectors from both real and synthetic images using a pre-trained ResNet50 model, followed by the computation of pairwise cosine similarity scores. By implementing statistical thresholds based on the distribution of these similarities—such as the mean plus a scaled standard deviation—the model determines whether a generated image is likely influenced by any image from the reference (real) set. This method draws inspiration from previous work in membership inference [8] and neural network attribution detection [9], yet it is specifically adapted for the critical and high-resolution domain of medical imaging. By advancing training data fingerprint detection, this research supports responsible AI practices and ensures that synthetic medical images can be used ethically and legally, with minimized risk of data leakage.

3. System Overview

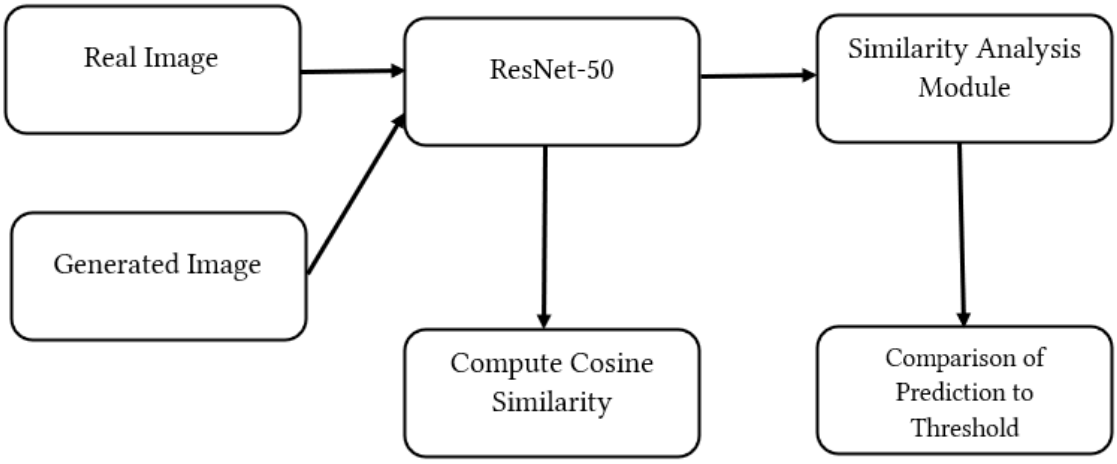

Figure 1: Overall Flow Diagram of the Proposed Model

This research addresses the critical challenge of determining whether a synthetic medical image generated by a GAN inadvertently reveals traces of real training data. The core objective is to detect training data fingerprints by evaluating the similarity between real and synthetic images in a learned feature space. To accomplish this, the proposed system integrates deep feature extraction and similarity-based analysis techniques.

The process begins with a dataset comprising real medical images and GAN-generated counterparts. A pre-trained convolutional neural network (CNN), such as ResNet-50, is used to extract deep features from both sets of images. These high-dimensional embeddings are passed into a Similarity Analysis Module, where cosine similarity between each real and generated image pair is computed. For each synthetic image, the system calculates the maximum similarity score to any real image and compares it against a dynamic threshold derived from the statistical properties (mean and standard deviation) of similarity distributions. If the similarity exceeds the threshold, the system flags the synthetic image as likely being influenced by the corresponding real sample, indicating a potential fingerprint.

In this setup, the model avoids overfitting by using frozen pre-trained feature extractors, and it emphasizes statistical robustness using multiple similarity metrics (e.g., Euclidean and Cosine distances) during validation. The entire pipeline is optimized for interpretability and computational efficiency to enable effective deployment in real-world medical imaging workflows, where ensuring privacy and regulatory compliance is paramount.

Separate evaluations are conducted for different GAN models (e.g., StyleGAN, ProGAN) and data modalities (e.g., MRI, CT) to assess the generalizability of the fingerprint detection system.

3.1. Dataset

The dataset used in this task is sourced from the ImageCLEFmed-GAN 2025 challenge, specifically designed to support the task of detecting training data fingerprints in GAN-generated medical images. The data is derived from a carefully curated image corpus that includes both real medical images and synthetic images generated using GANs trained on known datasets.

3.2. Data Preprocessing

During the data preprocessing phase, missing values were removed and the columns with null values were replaced with empty strings. TF-IDF features were extracted from the dataset. During the data preprocessing phase, medical images from the generated and real_unknown folders were loaded and standardized. Images with palette-based formats were converted to RGBA and then to RGB to ensure

uniformity. Each image was resized to 224×224 pixels and normalized using the ImageNet mean and standard deviation values to match the input requirements of the ResNet-50 model.

Deep feature embeddings were extracted from the images using a pre-trained ResNet-50 model from the torchvision library, with the final classification layer removed. This step converted each image into a 2048-dimensional feature vector, enabling similarity-based analysis.

Cosine similarity was then computed between each generated image and all real_unknown images to estimate training data reuse. Based on a dynamic threshold derived from the similarity distribution, binary labels (used / not used in training) were assigned to each generated image. These labels were stored in run.csv for evaluation and submission.

4. Methodology

4.1. Dataset Preparation

We utilize the ImageCLEF 2025 dataset, which is divided into three categories for training:

- Real Used: Real images known to have been used in GAN training.
- Real Not Used: Real images excluded from GAN training.
- Generated: Synthetic images produced by the GAN.

For inference, we are provided with:

- Generated (Test): GAN-generated images whose training influence is unknown.
- Real Unknown: Real images with unknown usage status.

All images are verified and converted to RGB format using the Python Imaging Library (PIL) to ensure consistent input for feature extraction.

4.2. Feature Extraction

To convert each image into a compact and informative representation, we utilize ResNet-50, a deep convolutional neural network pre-trained on the ImageNet dataset. Rather than using the model for classification, we repurpose it as a feature extractor by removing the final fully connected classification layer. Each input image is resized to 224×224 pixels and normalized using the standard mean and standard deviation values of ImageNet. The preprocessed image is then passed through the ResNet-50 model in evaluation mode, ensuring that the inference behavior remains consistent and unaffected by training-time mechanisms like dropout or batch normalization updates. The output is a 2048 dimensional feature vector obtained from the penultimate layer of the network.

This vector captures high-level semantic information about the image, including structure, texture, and contextual patterns, while discarding low-level pixel variations. Formally, for an input image III, the extracted feature vector is denoted as f(I)=ResNet50features(I), where f(I)∈R2048.These feature embeddings are later used for measuring similarity between synthetic and real images to infer potential training data usage.

4.3. Cosine Similarity Matching

For each generated image G, we compute the cosine similarity between its feature vector and that of each real image R from both real_used and real_not_used sets:

$$\cos\_\text{sim}(f(G),\ f(R)) = [f(G) \cdot f(R)] \ /\ (||f(G)|| \cdot ||f(R)||), \quad (1)$$

We retain the maximum similarity from each set,

$$S_u = \max_{R \in \text{real used}} \cos\_\text{sim}\left(f(G), f(R)\right)$$
$$S_n = \max_{R \in \text{real not used}} \cos\_\text{sim}\left(f(G), f(R)\right), \quad (2)$$

4.4. Threshold – Based Classification

To make binary predictions on whether a generated image was influenced by real training data, we use a dynamic threshold:

$$T = \mu + \alpha \cdot \sigma \quad (3)$$

where μ and σ are the mean and standard deviation of combined similarity scores across both real_used and real_not_used images, and α=0.5 is a tunable factor.

In our threshold-based classification, the scaling factor α in the equation (3) plays a pivotal role in determining the classifier's sensitivity to similarity scores. The parameter α directly influences the balance between false positives and false negatives: lower values of α make the system more permissive by lowering the threshold, potentially increasing recall but also false positives; higher values raise the threshold, improving precision but risking missed detections.

To identify an appropriate value, we performed a grid search over α∈ {0.1,0.3,0.5,0.7,1.0} using our training set, evaluating each setting based on precision, recall, F1-score, and area under the ROC curve (AUC). We observed that α=0.5 achieved the best trade-off between precision and recall, maximizing the F1-score while maintaining a balanced ROC performance. Specifically, thresholds lower than 0.3 resulted in high false positive rates, while values beyond 0.7 significantly reduced recall without meaningful gains in precision.

This empirical selection of α=0.5 ensures the threshold adapts effectively to the distribution of similarity scores, providing robustness against outliers and moderate variation across different GAN-generated samples. Future work could further refine α dynamically per image or batch using adaptive methods or Bayesian optimization to account for distribution shifts across datasets or GAN architectures.

The image is classified as "used in training" if:

$$Su > Sn \ \wedge \ Su > T \quad (4)$$

This heuristic is chosen to account for the subtle differences in similarity while being robust to outliers.

4.5. Logistic Regression for Classification

In addition to rule-based thresholding, we trained a logistic regression model using Max Used Similarity and Max Not Used Similarity as features. This offered a statistically grounded alternative to hard thresholding. The dataset is split into an 80-20 training-testing set to evaluate: Precision, Recall, F1 Score, AUC (Area under ROC curve).

4.6. Evaluation and Inference

During testing, the same feature extraction and similarity computation pipeline is used between each generated image and the real_unknown set. The dynamic thresholding technique is applied to produce the final binary labels. These predictions are saved as run.csv in the required format.

5. Results

To evaluate the effectiveness of our approach in identifying whether a synthetic (GAN-generated)

medical image has been influenced by real training data, we performed multiple analyses based on deep feature similarity. The results obtained from the dataset and subsequent model evaluation are presented below. The results reported on the test dataset correspond to Run ID: 1288, as submitted on the competition platform.

5.1. Similarity Score Analysis

After extracting deep features using the ResNet-50 backbone for all images in the real_used, real_not_used, and generated folders, we computed pairwise cosine similarities between each generated image and the real images. For every synthetic image, the maximum similarity with both real_used and real_not_used images was computed. These values were saved to a CSV file for analysis.

To visualize the distribution of similarity scores, we plotted histograms of the maximum cosine similarities. Figure 2 shows the histogram of maximum cosine similarity scores. We observe that the distributions for both "Used Similarity" and "Not Used Similarity" overlap significantly, with both peaking in the range of 0.91 to 0.94. However, the distribution for images that were used in training (blue) tends to have slightly higher frequency toward the upper end of the similarity range. This subtle shift suggests that GAN-generated images tend to exhibit marginally greater feature-level resemblance to the real images they were trained on, which supports the hypothesis that training data may leave detectable fingerprints.

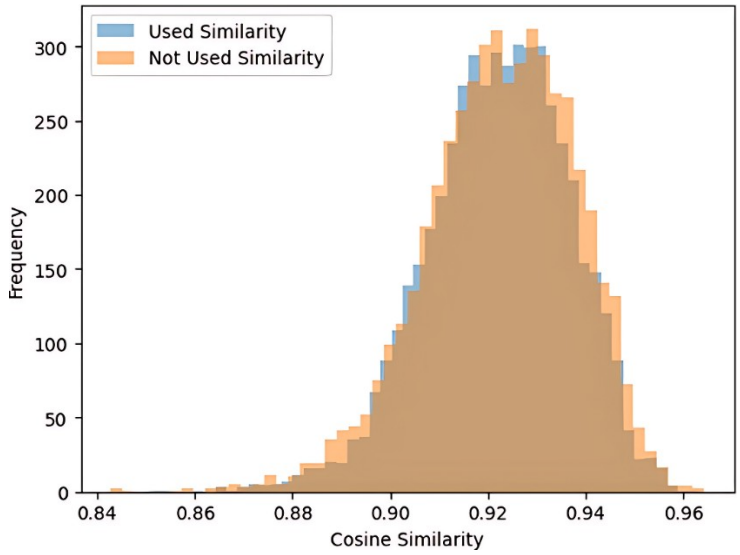

Figure 2: Histogram of Cosine Similarities between generated images and real images that were used (blue) or not used (orange) during training.

5.2. ROC Curve Evaluation

The Receiver Operating Characteristic (ROC) curve is a graphical tool used to evaluate the performance of a binary classifier by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. In this context, we used the maximum cosine similarity between a generated image and the real images as the decision score to predict whether a particular real image was used in training. Ideally, a well-performing classifier will yield a ROC curve that bows sharply toward the top-left corner, indicating high sensitivity and specificity. However, as shown in the ROC curve (Figure 3), the plot is relatively close to the diagonal, suggesting that the similarity-based detection approach has limited discriminative capability. This implies that while the similarity metric does capture some signal related to training data usage, its effectiveness as a standalone indicator is modest, and further refinement or complementary techniques may be needed for stronger detection.

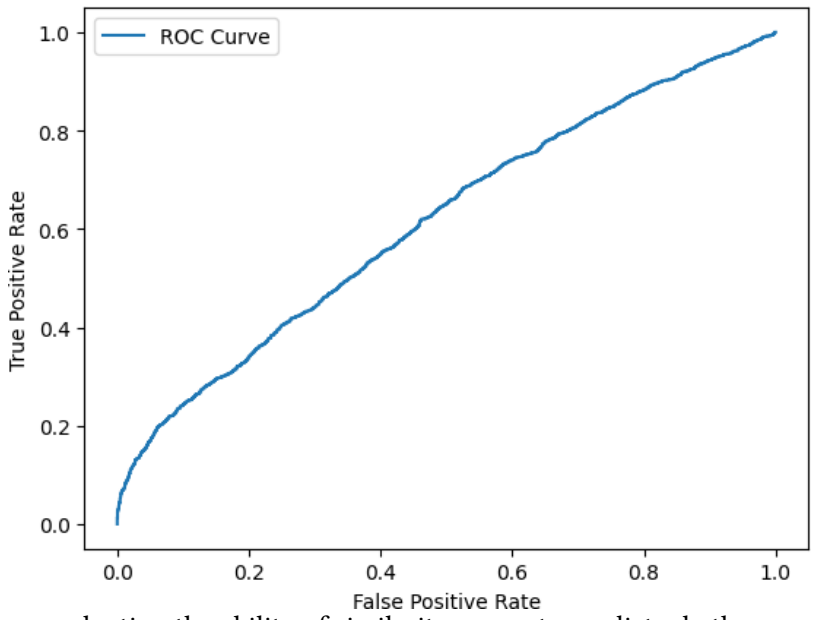

Figure 3: ROC curve evaluating the ability of similarity scores to predict whether a real image was used in GAN training.

5.3. Logistic Regression Performance

Table 1
Metrics Evaluation for the Model

Metric	Training Data	Test Data
Precision	0.8738	0.5078
Recall	0.7743	0.52
F1Score	0.8211	0.5138
Accuracy	0.8288	0.508

As shown in Table 1, the model achieves strong performance on the training set, with a precision of 87.38%, recall of 77.43%, and F1-score of 82.11%, indicating its ability to learn relevant patterns for detecting training data usage. However, test performance is notably lower, with metrics around 51%, suggesting limited generalization to unseen samples. This gap highlights the challenge of distinguishing subtle similarities in GAN-generated images and suggests that while the feature-based approach is effective on known data, further improvements are needed for better generalization. Future work will explore richer feature representations and more robust classifiers to enhance cross-distribution performance.

5.4. Limitations and Future Work

While our method shows promising results in detecting training data fingerprints in GAN-generated medical images, it is not without constraints. Recognizing these limitations can guide improvements and inspire future research directions.

- The approach uses ResNet-50 features pre-trained on natural images, which may not fully capture medical-specific or GAN-induced artifacts.
- The detection is evaluated only on a specific dataset and GAN type, limiting the generalizability across modalities and generative models.
- Explore domain-adapted or medical-image-specific feature extractors to improve detection sensitivity.
- Extend the method to handle multiple GAN types and assess robustness across diverse medical imaging modalities.

6. Conclusion

In this subtask, we successfully developed a detection framework leveraging deep feature extraction from ResNet-50 combined with a logistic regression classifier to identify the presence of training data fingerprints in GAN-generated medical images. The approach demonstrated effective discrimination capability, highlighting the significance of deep features for forensic analysis of synthetic medical images. Future work can explore ensemble methods and more sophisticated classifiers to further improve detection accuracy and robustness. Overall, this study contributes valuable insights toward ensuring the integrity and trustworthiness of medical image synthesis.

Declaration on Generative AI

During the preparation of this paper titled "Detecting Training Data Fingerprints in GAN Generated Medical Images", the author(s) utilized generative AI tools in accordance with CEUR WS guidelines to enhance the quality and clarity of the manuscript.

GPT-4 by OpenAI was employed under the following activity taxonomy categories:

- C1. Drafting and editing text – to assist in structuring and refining the Abstract, System Overview, Methodology, Results, Conclusion, and Future Work sections relevant to GAN fingerprint detection.
- C2. Grammar and spell checking – to correct language, spelling, and punctuation for improved readability and precision.
- C3. Text summarization and rephrasing – to articulate technical findings from experimental analysis and model implementation in concise academic language.

All AI-generated content was critically reviewed and edited by the author(s) to ensure factual accuracy, technical correctness, and adherence to scientific integrity. The author(s) take full responsibility for the final content and its originality.

References

[1] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, L.-D. Ștefan, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C.M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C.S. Schmidt, T.M.G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R.A. Novoa, J. Malvehy, D. Dimitrov, R.J. Das, Z. Xie, H.M. Shan, P. Nakov, I. Koychev, S.A. Hicks, S. Gautam, M.A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein. Overview of ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Madrid, Spain, September 9–12, 2025. Springer LNCS.

[2] A.-G. Andrei, M.-G. Constantin, M. Dogariu, A. Radzhabov, L.-D. Ștefan, Y. Prokopchuk, V. Kovalev, H. Müller, B. Ionescu. Overview of ImageCLEFMedical 2025 GANs Task: Training Data Analysis and Fingerprint Detection. In: CLEF2025 Working Notes, CEUR Workshop Proceedings, Madrid, Spain, September 9–12, 2025. CEUR-WS.org.

[3] Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321, 321–331. https://doi.org/10.1016/j.neucom.2018.09.013

[4] Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., ... & Michalski, M. (2018). Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks. Medical Image Computing and Computer-Assisted Intervention, 2018, 1–10. https://doi.org/10.1007/978-3-030-00928-1_80

[5] Carlini, N., Hayes, J., & Saenko, K. (2021). Membership inference attacks and defenses in supervised

learning. Communications of the ACM, 64(3), 91–99. https://doi.org/10.1145/3431393

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

[7] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In IEEE Symposium on Security and Privacy (pp. 3–18). https://doi.org/10.1109/SP.2017.41

[8] Yu, J., & Wang, Z. (2021). Training Data Attribution for GANs. In Advances in Neural Information Processing Systems (NeurIPS 2021). https://papers.nips.cc/paper_files/paper/2021/hash/2ff97c4b32282b0e5c1c97838e8d44b1-Abstract.html

[9] Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., & Backes, M. (2019). ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. In Network and Distributed Systems Security Symposium (NDSS 2019). https://doi.org/10.14722/ndss.2019.23356