Identify Training Data Subsets 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 identifying the particular dataset of real images that contributed to 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 2 of the ImageCLEFmed GAN 2025 challenge, the problem is framed as a multi-class classification task where each generated image must be assigned a label corresponding to one of five real data subsets that contributed to its generation. The proposed approach involves fine-tuning a ResNet-18 model pretrained on ImageNet, adapted for five-class classification. Real images from five labelled subfolders and synthetic images generated from them are preprocessed using standard ImageNet normalization and resized to 224×224 pixels. The model is trained using a cross-entropy loss function and optimized with the Adam optimizer over 10 epochs. Evaluation on the official test set demonstrated strong performance, with an accuracy of 94.84%, precision of 95.04%, recall of 94.84%, F1 score of 94.87%, and specificity of 98.79%. These results highlight the effectiveness of supervised deep learning and feature-based discrimination for fine-grained attribution in synthetic medical image generation, offering a reliable method for auditing the provenance of GAN-generated data.

Keywords

GAN-generated medical images, training data attribution, ResNet-18, deep learning classification, supervised learning, multi-class image classification, dataset provenance, medical image synthesis, synthetic CT scans, real-to-synthetic image mapping, AI transparency, medical image auditing, 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 2: "Identify Training Data Subsets", which involves determining which specific subset of real images was used to generate a given synthetic image. 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 challenge is structured as a multi-class classification problem, requiring fine-grained attribution. It introduces significant technical challenges due to the complex and often subtle manner in which visual features from different data subsets are embedded and transformed within GAN-generated content.

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

Traditional methods like watermarking or forensic analysis are often intrusive or unreliable, especially with deep generative models. Manual checks are impractical, highlighting the need for scalable, automated solutions.

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 supervised classification framework for attributing synthetic medical images to the specific subset of real images used during GAN training. By learning discriminative features across known data groups, the system is able to predict the most likely training subset for each synthetic image. The approach emphasizes scalability and practical applicability, laying the groundwork for improved dataset provenance tracking and greater transparency in AI-generated medical content.

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].

Recent work has shown that GANs, especially when trained on medical datasets, can retain subtle characteristics of their training subsets, making it feasible to infer the origin subset of a generated image under certain conditions [5]. This is particularly relevant in medical imaging, where identifying the provenance of synthetic data is crucial not just for privacy assurance but also for validating model fairness and training diversity [6]. Instead of focusing on whether a single real image was seen during training—a typical membership inference task—Subtask 2 requires tracing synthetic images back to the specific subset of real data that influenced their generation.

To tackle this, we adopt a supervised feature attribution approach, leveraging the capability of deep convolutional networks to capture fine-grained distributional patterns across image subsets. A ResNet-18 model is fine-tuned to classify GAN-generated synthetic images into five known real-data subsets. This design draws inspiration from applications in medical imaging where CNNs have been successfully adapted to detect subtle structural and textural variations in CT, MRI, and other modalities [7]. The extracted deep features represent a high-dimensional embedding of the image, where subset-specific patterns are preserved and enhanced through training. The resulting classifier offers an interpretable and scalable solution for subset-level training data attribution.

Such approaches support broader goals in AI transparency, data governance, and model auditing, especially as synthetic data becomes increasingly embedded in medical workflows [8]. Compared to pixel-space or handcrafted feature methods, the deep feature route demonstrates better generalization and robustness to intra-class variability [9].

3. System Overview

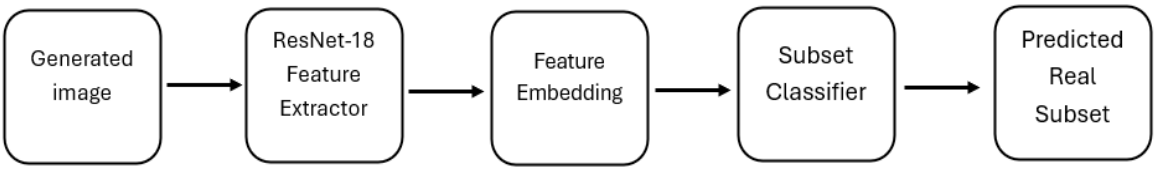

Figure 1: Overall Flow Diagram of the Proposed Model

This research addresses the task of identifying the specific subsets of real medical images that were used to generate synthetic images with generative adversarial networks (GANs). The objective is to develop a robust classification framework capable of assigning each GAN-generated image to one of several predefined subsets of real data. The approach is grounded in deep representation learning, leveraging a feature extraction and classification pipeline for effective attribution.

The system begins with two main datasets: a labeled training dataset consisting of real images organized into five known subsets (t1 to t5), and a test dataset comprising GAN-generated images whose source subsets are unknown. A pre-trained convolutional neural network (CNN), specifically ResNet-18, is utilized to extract high-level feature embeddings from both real and generated images. The final classification layer of the network is replaced and fine-tuned to output predictions corresponding to the five real data subsets.

To improve generalizability and prevent overfitting, the backbone network is fine-tuned on the labeled dataset, while deep features from the penultimate layer are extracted and used to train a secondary classifier. A logistic regression model is employed for this purpose, offering interpretability and reliable multi-class discrimination.

All experimentation was carried out on Google Colab Pro, leveraging GPU acceleration (NVIDIA T4 / A100) for efficient feature extraction and model training. The implementation utilizes the following key Python libraries:

- PyTorch and Torchvision for model definition, transfer learning, and feature extraction.
- Scikit-learn (sklearn) for logistic regression, PCA dimensionality reduction, confusion matrix computation, and performance evaluation.
- NumPy and Pandas for numerical computation and structured data handling.
- Matplotlib and Seaborn for visualizations, including PCA plots and confusion matrices.
- TQDM, OS, and Glob for efficient file handling, preprocessing, and batch-wise data loading.

The pipeline is designed to be scalable and interpretable, enabling practical forensic evaluation of GAN-generated medical images. By examining the distribution of deep features and training a classifier on top of them, the system aims to reverse-engineer the generative process and provide evidence linking synthetic images to the real data subsets they were influenced by. This methodology contributes toward transparency, accountability, and regulatory compliance in medical image synthesis.

3.1. Dataset

The dataset used in this task is sourced from the ImageCLEFmed-GAN 2025 challenge, specifically designed to support the task of identifying the exact subgroup of real images used to generate each synthetic image. 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.

The training data is structured as follows:

- Real: This folder contains real medical images categorized into five subsets — t1, t2, t3, t4, and t5. Each subset represents a distinct group of real images.
- Generated: This folder contains five corresponding subsets — gen_t1, gen_t2, gen_t3, gen_t4, and gen_t5, each containing synthetic images generated using the respective real subset as training input.

For inference, we are provided with:

- Generated Unknown: A set of synthetic images generated by the GAN, whose corresponding real subset (t1 to t5) used during training is unknown. The task is to classify each image into one of the five real subsets.

3.2. Data Preprocessing

In this subtask the dataset consisted of synthetic medical images located in the generated_unknown folder without predefined subfolders or labels. Each image was loaded individually and processed to conform with the input requirements of the trained ResNet-18 classification model.

Images in various formats (PNG, JPG, JPEG) were read using the PIL library and converted to RGB format if needed, ensuring consistency in colour channels. Each image was resized to 224×224 pixels to match the model's expected input size. Subsequently, normalization was applied using the ImageNet mean and standard deviation values to standardize pixel intensities. The processed images were organized into batches using a DataLoader for efficient inference.

4. Methodology

4.1. Feature Extraction

The preprocessed image is then passed through the ResNet-18 network in evaluation mode, which ensures consistent inference by disabling training-specific behaviors such as dropout and batch normalization updates. We extract features from the penultimate layer (i.e., the output of the global average pooling layer), resulting in a 512-dimensional feature vector per image.

This vector, denoted as $f(I) = \text{ResNet18}_{\text{features}}(I)$, where $f(I) \in \mathbb{R}^{512}$, captures high-level semantic features of the image including shape, texture, and structure, while abstracting away pixel-level noise. These embeddings are subsequently used to compute similarities between synthetic and real subsets to infer the likely source subgroup for each generated image.

4.2. Supervised Multi-Class Classification

The problem is modelled as a supervised multi-class classification task, where each image is assigned to one of five real subsets (t1–t5). The final FC layer of ResNet-18 is replaced with a new linear layer with five outputs and trained on the real image dataset.

- Loss Function: Cross-entropy
- Optimizer: Adam, learning rate = 1e-4
- Batch Size: 64
- Epochs: 10
- Device: GPU

To evaluate generalization, the model is validated using the synthetic generated data (gen_t1 to gen_t5), which mimics the distribution of the test set (generated_unknown).

4.3. Fine-Tuning of ResNet-18

Fine-tuning began by loading the pre-trained ResNet-18 weights and modifying the final classification layer to fit our five-class task. The model was then trained on the labeled real images while validating on synthetic images to monitor performance on data closer to the inference distribution.

Standard image preprocessing was consistently applied to both training and validation datasets to ensure uniformity in input. The training loop employed batch-wise gradient descent with backpropagation to optimize model weights by minimizing cross-entropy loss.

Throughout the ten epochs of training, we tracked the loss and validation accuracy to detect signs of overfitting or underfitting and adjusted training parameters accordingly. For inference and evaluation phases, the model was set to evaluation mode to deactivate dropout and batch normalization updates, ensuring consistent predictions. Upon completion of training, the fine-tuned model was saved for subsequent testing and deployment.

4.4. Inference on Unlabeled Synthetic Data

During inference, the trained ResNet-18 model was loaded in evaluation mode and used to predict the subgroup labels of the unlabelled synthetic images located in the generated_unknown directory. The synthetic images underwent the same preprocessing pipeline as during training.

Predictions were generated by forwarding batches of images through the model on the available compute device (GPU). For each image, the class with the highest output score was selected as the predicted subgroup. These predictions were then collated with their corresponding image filenames.

Finally, the results were formatted into a CSV file without headers, adhering to the challenge's submission specifications. This file enabled automatic evaluation of the model's ability to infer the training data subgroup influence on synthetic medical images.

5. Results

To evaluate our approach for Subtask 2, we trained a ResNet-18 classifier on real image data labeled from subgroups t1 to t5 and tested it on GAN-generated images to predict the corresponding real subset used during generation. We also conducted evaluations on the training dataset to validate the model's learning behavior using both quantitative and visual techniques.

5.1. Classification Performance on Training Data

The trained ResNet-18 model was evaluated on the training dataset itself to inspect class-wise prediction performance. The model achieved strong validation accuracy of 94.89% on the real image dataset, indicating successful learning of discriminative features between the five subgroups. To further understand class-wise prediction behavior, we visualized the confusion matrix on the training data.

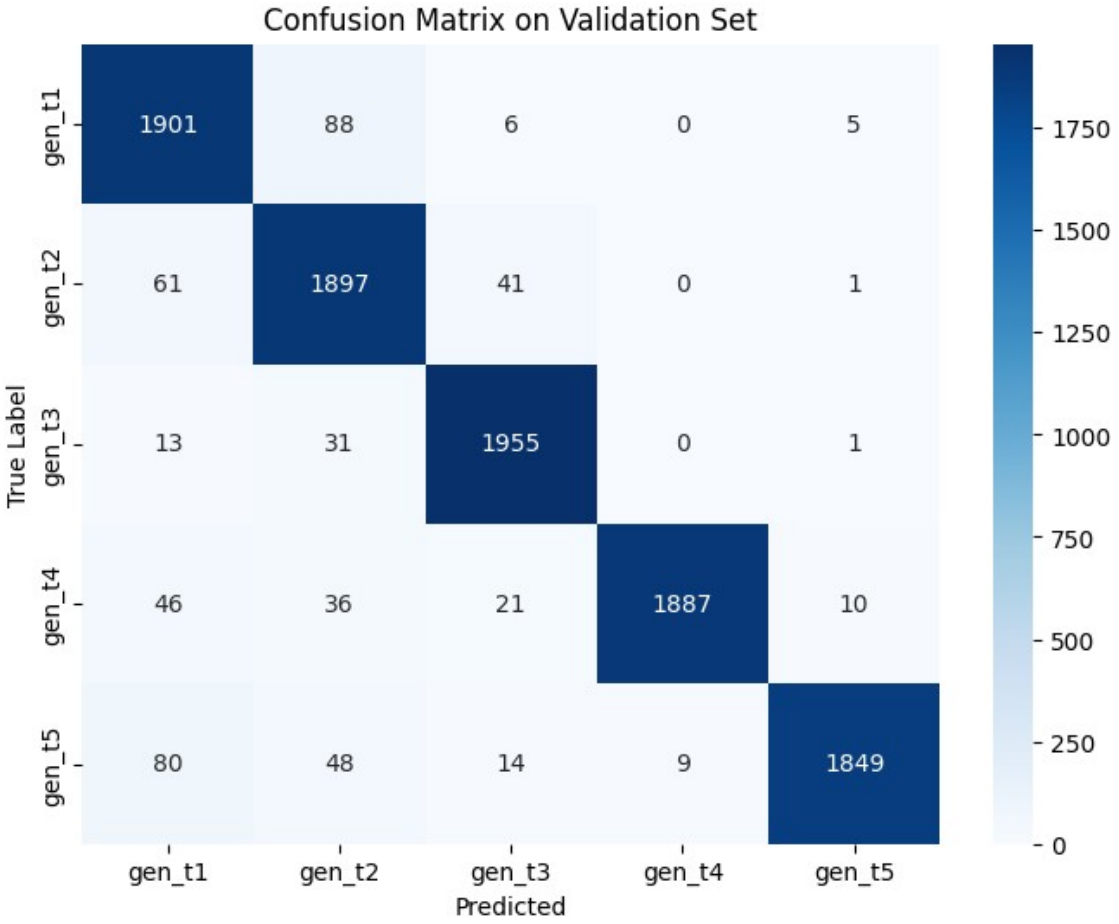

Figure 2: Confusion matrix of ResNet-18 model on the training dataset

The confusion matrix reveals that the model achieves high accuracy in distinguishing between most subgroups, with minor confusion observed between visually similar classes. This suggests that the learned feature representations are largely discriminative across the five classes.

5.2. Feature Representation Analysis using PCA

To further investigate how well the model differentiates between subgroups, we applied Principal Component Analysis (PCA) to the extracted deep features of real training images. The two-dimensional PCA plot below shows how the features cluster in reduced space.

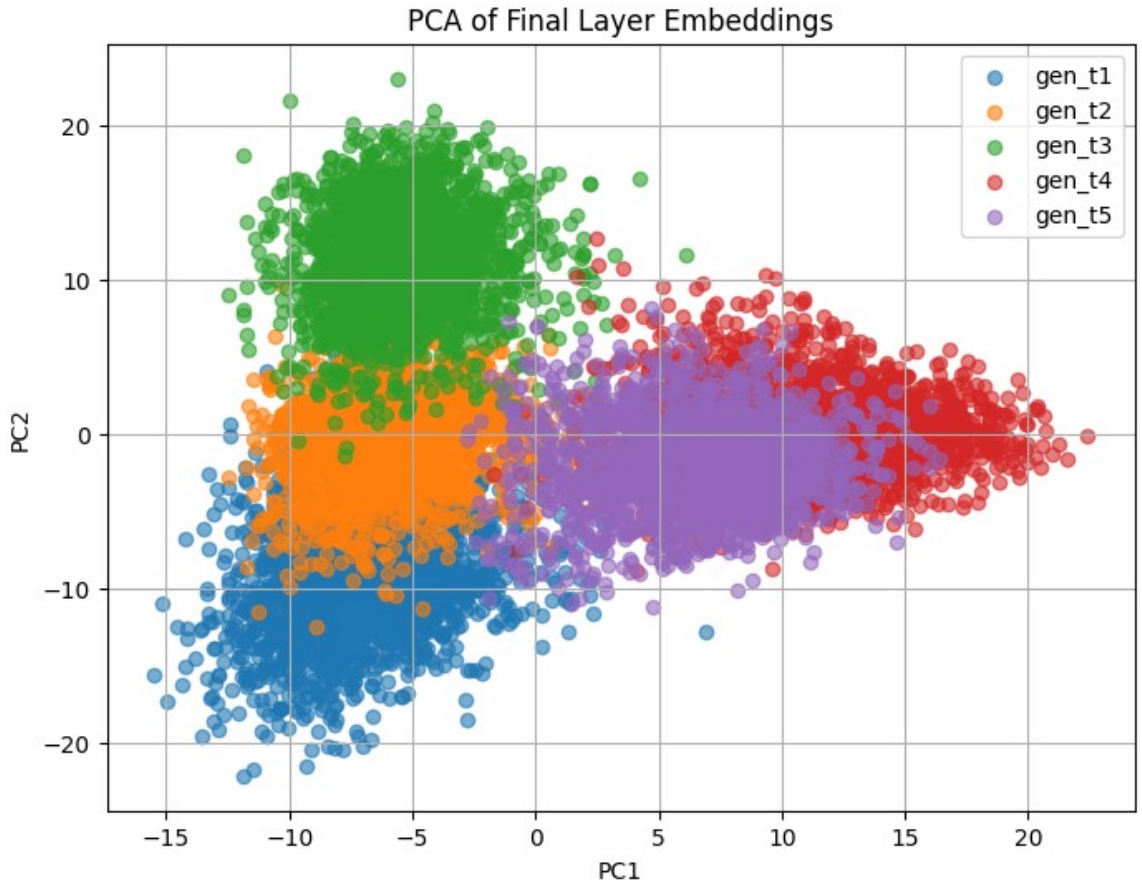

Figure 3: PCA projection of feature vectors extracted from the penultimate layer of ResNet-18 trained on real image subgroups

From the PCA visualization, it is evident that the five subgroups form relatively well-separated clusters, indicating that the network has learned subgroup-specific features. However, some overlaps do occur, particularly between subgroups that may have shared visual traits or structural similarities.

5.3. Test-Time Prediction Performance

Following model training and validation, the trained ResNet-18 model was applied to the generated_unknown test dataset containing 25,000 GAN-generated images. The model predicted the subgroup label (1 to 5) for each image.
The results reported on the test dataset correspond to Run ID: 1287, as submitted on the competition platform.
The process was efficient, completing inference in under 25 minutes and generating a prediction file (run.csv) compatible with the submission requirements. Evaluation of this output yielded the following performance metrics:

- Accuracy: 94.84%
- Precision: 95.04%
- Recall: 94.84%
- F1-score: 94.87%
- Specificity: 98.79%

These scores demonstrate that the model generalizes well to unseen GAN-generated images and is capable of reliably identifying the real image subset from which a synthetic image was generated. The high specificity further confirms that false positive rates were low across all classes.

5.4. Limitations and Future Work

While the ResNet-18-based classifier captures domain-relevant features, its generalization is limited by visual similarity between certain subgroups. Further improvements could involve:

- Using deeper models (e.g., ResNet-50 or EfficientNet).
- Augmenting training data with real-synthetic hybrids to improve subgroup discrimination.
- Employing metric learning or contrastive loss to better cluster similar image classes.
- Adding explainability methods (e.g., Grad-CAM) to visualize which regions influence the prediction most.

6. Conclusion

In this subtask, we proposed a supervised learning approach to trace the origin of GAN-generated medical images by classifying them based on the real data subsets used during their generation. Using a ResNet-18 model trained on real biomedical image subgroups, we achieved high classification performance on both validation and test datasets, with an overall accuracy of **94.84%** on synthetic test images. The results indicate that deep convolutional networks are capable of learning meaningful and discriminative representations that persist in the generated images.

Visualizations using PCA confirmed the separability of feature embeddings, and a detailed confusion matrix analysis revealed strong class-wise prediction reliability.

This framework provides a promising direction for developing automated tools to assess data provenance and transparency in medical image synthesis. As a next step, we plan to explore more advanced models, uncertainty quantification, and feature attribution techniques to further improve interpretability and robustness across diverse GAN architectures.

Declaration on Generative AI

During the preparation of this paper titled "Identifying Training Data Subsets 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.
- 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] Hilprecht, B., Härterich, M., & Günther, M. (2019). Monte Carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(4), 232–249.

[6] Kaissis, G., Makowski, M., Rückert, D., & Braren, R. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6), 305–311. https://doi.org/10.1038/s42256-020-0186-1

[7] Shin, H.-C., et al. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5), 1285–1298. https://doi.org/10.1109/TMI.2016.2528162

[8] Ouyang, C., et al. (2023). Synthetic data in radiological imaging: current state and future outlook. *British Journal of Radiology Artificial Intelligence*, 1(1), ubae007. https://academic.oup.com/bjrai/article/1/1/ubae007/7679083

[9] Zhang, J., et al. (2020). The Secret Revealer: Generative Model-Informed Membership Inference Attacks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 253–261. https://doi.org/10.1109/CVPR42600.2020.00034