Overview of ImageCLEFMedical 2025 GANs Task: Training Data Analysis and Fingerprint Detection

Notebook for the ImageCLEF Lab at CLEF 2025

Alexandra-Georgiana Andrei[1,*], Mihai Gabriel Constantin[1], Mihai Dogariu[1], Ahmedkhan Radzhabov[2], Liviu-Daniel Ștefan[1], Yuri Prokopchuk[2], Vassili Kovalev[2], Henning Müller[3] and Bogdan Ionescu[1]

[1] *AI Multimedia Lab, National University of Science and Technology Politehnica Bucharest, Romania*

[2] *Belarusian Academy of Sciences, Minsk, Belarus*

[3] *University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland*

Abstract

The 2025 ImageCLEFmedical GANs Task - Controlling the Quality of Synthetic Medical Images created via GANs, continuing to investigate privacy and security concerns around using patient data to generate synthetic medical images. It comprises two complementary sub-tasks: the first extends prior editions by asking participants to detect which real images were used in training a Generative Adversarial Network to produce given synthetic outputs; the second builds on the 2024 findings by requiring teams to attribute each synthetic image to its specific real-image subset of origin. Ground-truth annotations and benchmark datasets of real and GAN-generated lung CT slices are provided for both tasks, and evaluation is based on Cohen's Kappa for Subtask 1 and accuracy for Subtask 2. 14 teams submitted runs for Subtask1 and 4 teams submitted runs for Subtask 2, totaling 95 submitted runs that used a variety of methods. This paper presents an overview of the task setup, datasets, and evaluation metrics, and summarizes and discusses the approaches and results of the .

Keywords

generative models, Generative Adversarial Networks, medical synthetic data, medical imaging, deep learning, ImageCLEF benchmarking lab

1. Introduction

AI systems for medical tasks like predicting, detecting, and classifying diseases rely on the availability of large and diverse datasets for training. High-quality data allow these models to learn complex patterns and improve their accuracy and reliability. However, access to real medical data is not easy due to privacy concerns. Patients are usually willing to share their medical information only for their own treatment and not for research. This makes it difficult to gather enough data to effectively train AI models, slowing progress in developing better tools for healthcare care.

One way to solve this problem is to create synthetic data: artificial data that looks like real medical data but does not come from actual patients. Generative models, such as Generative Adversarial Networks (GAN), can be used to create these datasets. Synthetic data can help researchers build and test AI systems without needing to rely on real patient data, which protects privacy and makes it easier to collect the variety of information needed for training. But there is an important challenge with synthetic data: it must not include hidden details, or "fingerprints" from the real data it was trained on. If synthetic data can somehow be traced back to the original patient data, it could risk exposing private information. Ensuring that synthetic data are completely free from such "fingerprints" is critical.

This is the third edition of the GANs task, part of ImageCLEF 2025 Lab [1], in which we continue to investigate the hypothesis that generative models generate synthetic medical images that retain

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

[*] Corresponding author.

alexandra.andrei@upb.ro (A. Andrei); mihai.constantin84@upb.ro (M. G. Constantin); mihai.dogariu@upb.ro (M. Dogariu); liviu_daniel.stefan@upb.ro (L. Ștefan); vassili.kovalev@gmail.com (V. Kovalev); henning.mueller@hevs.ch (H. Müller); bogdan.ionescu@upb.ro (B. Ionescu)

"fingerprints" from the real images used during their training based on the lessons learned from the previous two editions:

- In the first [2] and second [3] editions of this task, held at ImageCLEF 2023 [4] and 2024 [5], various generative models were analyzed within the framework of the first subtask to investigate whether synthetic images contained "fingerprints" of the real medical data used during training. The results demonstrated that the tested generative models do retain and imprint features from their training data, raising important security and privacy concerns. These findings underscore the need for robust techniques to detect and mitigate such imprints to ensure that synthetic images protect patient privacy while maintaining their utility for research and development.
- In the 2nd edition of the task [3], it was confirmed that generative models leave unique "fingerprints" on the synthetic images they produce. By analyzing images generated from various models, distinct patterns and features were identified that allowed the attribution of synthetic images to their respective generative models.

To investigate the potential privacy risks associated with synthetic medical images, the 2025 edition of the GANs Task focuses on two complementary subtasks. The first subtask challenges participants to determine whether specific real images were used during the training of a GAN to produce given synthetic outputs, thereby addressing the risk of image-level information leakage. The second subtask extends this inquiry by asking participants to attribute synthetic images - generated using a Diffusion model - to one of several predefined subsets of real training data. This focuses on whether broader, dataset-level characteristics are retained during the generative process. An overview of these two subtasks is illustrated in Figure 1.

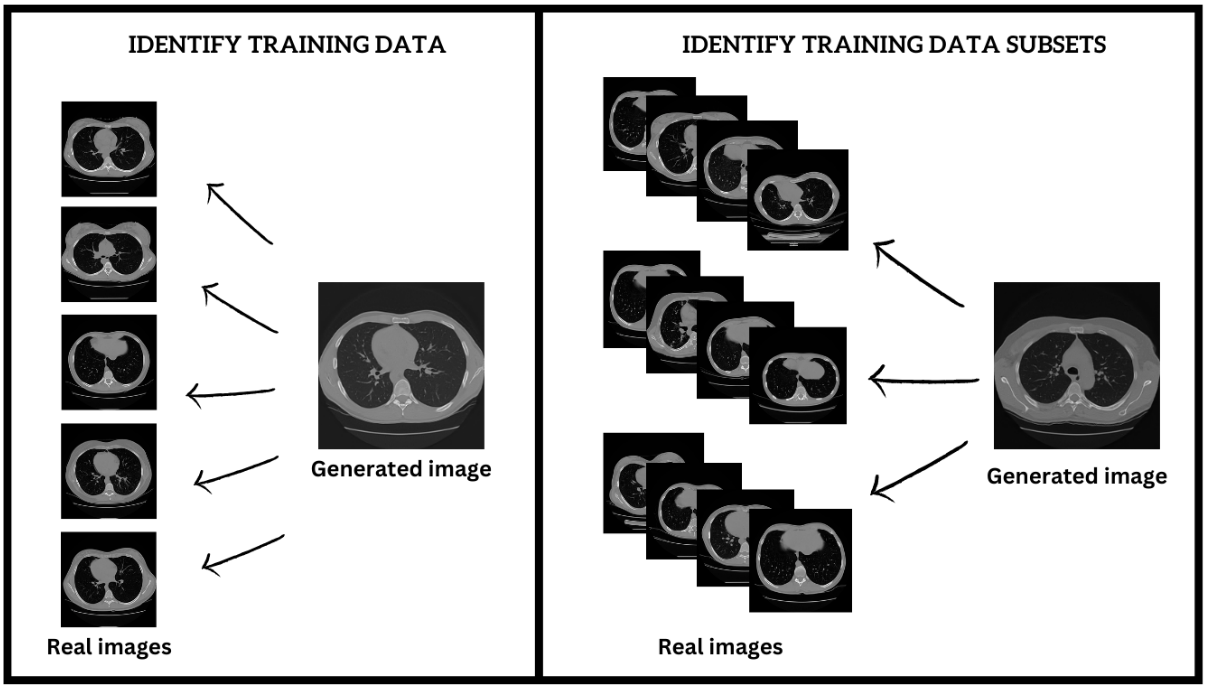


Figure 1: Overview of the two subtasks in the ImageCLEFmedical 2025 GANs Task. Left: Subtask 1 aims to determine whether specific real images were used to train a GAN that generated a given synthetic image. Right: Subtask 2 requires identifying the training subset from which the synthetic image was derived using a Diffusion model.

2. Tasks description

2.1. Sub-task 1: Detect Training Data Usage

2.1.1. Description

We continued to investigate the hypothesis that generative models are generating medical images that are in some way similar to the ones used for training. The task addresses the security and privacy concerns related to personal medical image data in the context of generating and using artificial images in different real-life scenarios. This edition continues the task proposed in both previous editions [2, 3] by investigating another GAN. The objective of the task is to detect "fingerprints" within the synthetic biomedical image data to determine which real images were used in training to produce the generated images. The task consisted in performing analysis of test image datasets and assess which images of real patients were used for training image generators and which were not. The task is formulated as follows:

- *In this subtask, participants will analyze synthetic biomedical images to determine whether specific real images were used in the training process of generative models. For each real image in the test set, participants must label it as either used (1) or not used (0) for generating the given synthetic images. This task focuses on detecting the presence of training data "fingerprints" within synthetic outputs.*

2.1.2. Data description

The benchmarking dataset includes both real and synthetic biomedical images. The real images consist of axial slices of 3D thoracic CT scans from approximately 8,000 lung tuberculosis patients. These slices vary in appearance: some may look relatively "normal", while others exhibit distinct lung lesions, including severe cases. The real images were provided in 8-bit per pixel PNG format, with dimensions of 256x256 pixels, providing a standardized resolution for analysis. The synthetic images, also sized at 256 × 256 pixels, have been generated using a GAN. By providing both real and synthetic datasets, this task enables participants to analyze and compare the characteristics of synthetic images with their real counterparts, investigating potential "fingerprints" and patterns related to the training process. Examples of real and generated images are depicted in Figure 2.

Train dataset consists of 3 folders:

- "generated" - contains 5,000 synthetic images generated using a GAN.
- "real_used" - contains 100 real images that were used to train the GAN to produce the synthetic images in the "generated" folder.

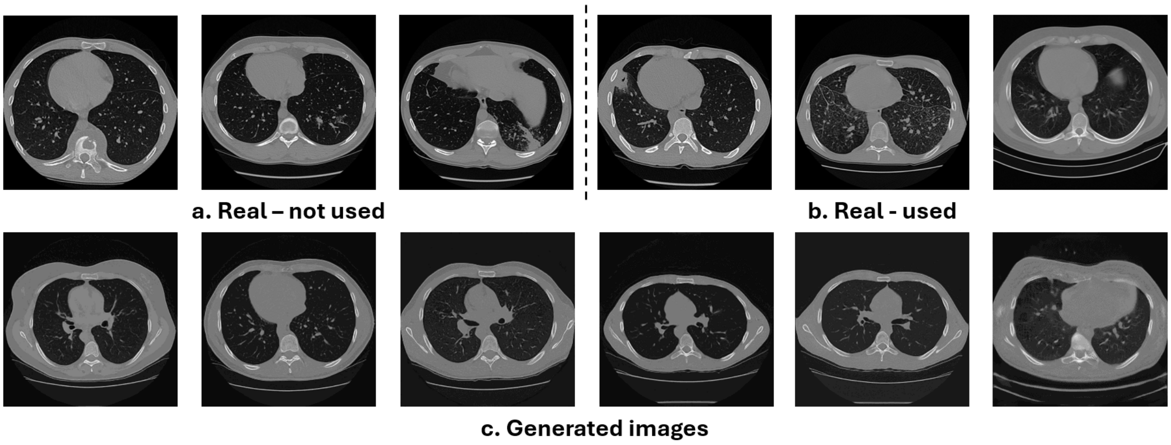


Figure 2: Examples of images provided for Subtask 1: Detect Training Usage. The first row contains real images: the first three are samples that were not used to train the generative model, followed by three real images that were used for training. The second row contains samples of generated images.

- “real_not_used”- contains 100 real images that were not used for training.

Test dataset was organized as follows:

- “generated”- contains 2,000 additional synthetic images. These images were generated using the same model trained under the same conditions as those used to create the synthetic images in the training dataset.
- “real_unknown” - contains a mix of 500 real images. Some of these images were used in training the generative model, while others were not.

2.2. Sub-task 2: Identify Training Data Subsets

2.2.1. Description

In this subtask, participants will link each synthetic biomedical image to the specific subset of real data used during its generation. The goal is to identify the particular dataset of real images that contributed to the training of the generative model responsible for creating each synthetic image. This requires a more detailed attribution of synthetic images to their corresponding training subsets. The task is formulated as follows:

- *In this subtask, participants will link each synthetic biomedical image to the specific subset of real data used during its generation. The goal is to identify the particular dataset of real images that contributed to the training of the generative model responsible for creating each synthetic image. This requires a more detailed attribution of synthetic images to their corresponding training subsets.*

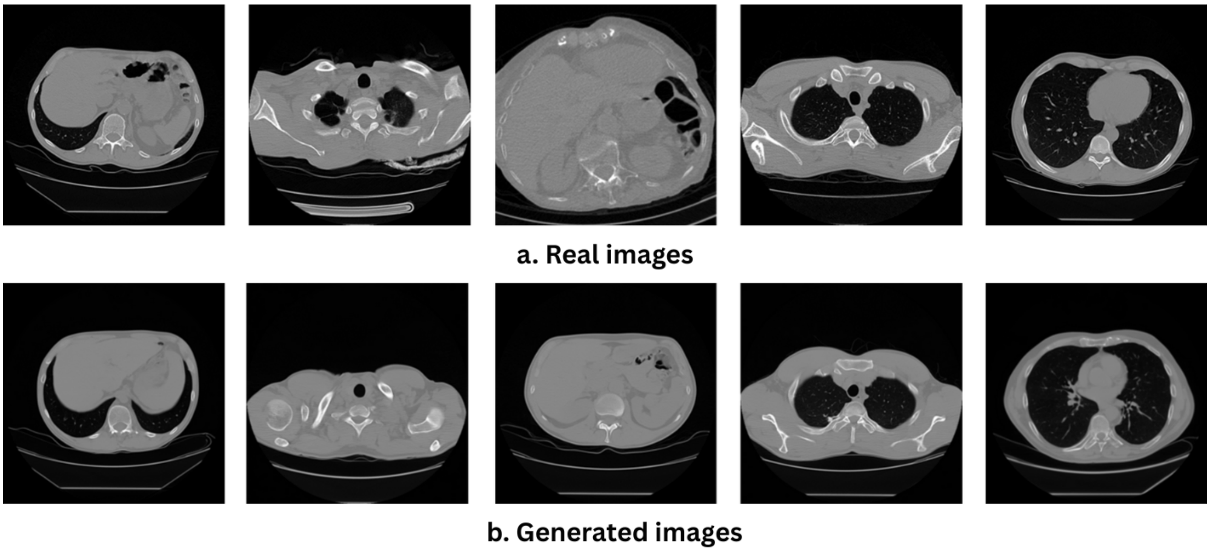
2.2.2. Data description

The benchmarking dataset for this task is similar to the dataset provided for the first subtask in terms of cohort, acquisition conditions, image dimensions (256×256) and bit-depth (8 bit per pixel); the only difference lies in the scan types. In addition to thoracic CTs, it also contains cervical and abdominal CTs, as exemplified in Figure 3. The synthetic images, also sized at 256 × 256 pixels, have been generated using a Diffusion model. By providing both real and synthetic datasets, this task enables participants to analyze and compare the characteristics of synthetic images with their real counterparts, investigating potential “fingerprints” and patterns related to the training process. Examples of generated images are shown in Figure 3. The training dataset consists of two main folders:

- “generated”- contains 5 subfolders of synthetic images. Each subset was generated using a different training dataset for the generative model.
- “real” - contains 5 subfolders, each corresponding to a specific training dataset used to train the generative model. The real images in each subfolder were used to generate the synthetic images in the corresponding "generated" subfolder.

The mapping between the real and generated images is as follows:
Folder “t1” (real images) - Used to generate synthetic images in “gen_t1”
Folder “t2” (real images) - Used to generate synthetic images in “gen_t2”
Folder “t3” (real images) - Used to generate synthetic images in "gen_t3”
Folder “t4” (real images) - Used to generate synthetic images in “gen_t4”
Folder “t5” (real images) - Used to generate synthetic images in “gen_t5”.

The test dataset contains 25,000 generated images, each derived from a real subgroup of images in the training dataset. Each image will be assigned a label consistent with those used in the training dataset.



a. Real images

b. Generated images

Figure 3: Examples of images provided for Subtask2: Identify Training Data Subsets. The first row contains samples of real images used for trianing the generative model and the second row contains samples of generated images.

3. Evaluation methodology

3.1. Sub-task 1: Detect Training Data Usage

The official evaluation metric of the task is Cohen's kappa score. In addition, the F1-score, accuracy, precision, and recall were computed to asses the task which can be considered a binary-class classification problem. Cohen's kappa [6] is a statistic technique that measures inter-annotator agreement on a classification problem and it is defined as:

$$k = (p_o - p_e)/(1 - p_e) \tag{1}$$

,where p_o is the empirical probability of agreement on the label assigned to any sample (the observed agreement ratio), and p_e is estimated using a per-annotator empirical prior over the class labels.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

where TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative.

3.2. Sub-task 2: Identify Training Data Subsets

The official metric of the task is Accuracy. In addition, Precision, Recall, F1 score and Specificity were computed.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

4. Results

4.1. Participation stats

Overall, 41 teams registered for our task. Of these, 14 teams completed the first subtask by submitting runs, and 4 teams completed the second subtask. In total, 9 teams submitted working notes papers. Notably, 2 teams participated in both subtasks, including the task organizing team. The continued interest in the first subtask, now in its third edition, highlights its ability to attract more participating teams.

Each participating team was allowed to submit up to 10 runs per task. We received a total of 95 submitted runs: 91 for Subtask 1 and 14 for Subtask 2. Table 1 presents the list of participating teams and their institutions. The rankings for Subtask 1 are shown in Table 2, and those for Subtask 2 are presented in Table 3, limited to the teams that described their methods by submitting working notes papers.

4.2. Presented methods

4.2.1. Sub-task 1: Detect Training Data Usage

SCOPE VIT Visioneers [7] team proposed a framework using a Siamese Neural Network (SNN) to detect whether real medical images were used in training. The architecture consists of twin subnetworks based on popular backbones (ResNet50, DenseNet161, EfficientNetB2, and ViT), enhanced with a cross-attention mechanism to align feature maps of real and synthetic image pairs. These attention-guided features are fed into an adaptive similarity module that combines absolute difference and dot product operations, followed by a multilayer perceptron to score the likelihood that an image contributed to GAN training. Evaluated on a lung CT dataset, the ResNet50 backbone with cross-attention achieved the best performance, demonstrating high accuracy and generalizability. All results obtained by the team are shown in Table 2 and consists in the following methods:

- Run ID 1160: Siamese network with a ResNet50 backbone augmented by cross-attention

Challengers [8] explored four different methods. The first two models applied an unsupervised framework combining ResNet50 for deep feature extraction, Principal Component Analysis (PCA) for dimensionality reduction, and various clustering algorithms (k-means, DBSCAN (A density based clustering algorithm capable of identifying arbitrarily shaped clusters and detecting outliers), Gaussian Mixture Model (GMM), and agglomerative clustering) to analyze similarities between real and synthetic images. The third model was a supervised binary classifier based on ResNet50V2, fine-tuned via transfer learning and enhanced with a custom classification head, achieving the best overall performance. Finally,

Table 1
Overview of participating teams that submitted at least one run (* task organizing team)

Team	Subtask 1	Subtask 2	Affiliation	Country
SCOPE VIT Visioneers	✓	×	Vellore Institute of Technology	India
SDVA/UCSD	×	✓	San Diego VA Health Care System	US
Challengers	✓	×	Sri Sivasubramaniya Nadar College of Engineering	India
Medhastra	✓	✓	Sri Sivasubramaniya Nadar College of Engineering	India
Neural Nexus	✓	×	Pune Institute of Technology	India
zhouyijiang1	✓	×	Yunnan University	China
taozi	✓	×	Yunnan University	China
ZOQ	✓	×	Yunnan University	China
AI Multimedia Lab*	✓	✓	National University of Science and Technology POLITEHNICA Bucharest	Romania

a fourth hybrid approach combined either deep features or Local Binary Patterns (LBP) with clustering and distance-based classification, determining training data reuse by measuring the distance between test samples and learned synthetic image centroids. All results obtained by the team are shown in Table 2 and consists in the following methods:

- Run ID 1778: Method 1 – ResNet50 + PCA + Clustering (KMeans, GMM, Agglomerative, DBSCAN);
- Run ID 1779: Method 2 – ResNet50 + PCA + Clustering;
- Run ID 1811: Method 3 – ResNet50V2 + Binary Classification (Supervised CNN Approach);
- Run ID 1776: Method 4 – Deep Feature Based Clustering and Distance Based Classification.

Team **Medhastra** [9] explored a deep feature similarity-based approach. Their method leverages a ResNet-50 model, pre-trained on ImageNet [10], to extract 2048-dimensional embeddings from both real and generated images. Cosine similarity scores are computed between each synthetic image and the real image pool, and a statistical threshold (based on mean and standard deviation) is applied to classify wheater a real image was used for training. Additionaly, a logistic regression model using similarity features was trained to support this classification. The team submitted one run (Run ID 1288) to the task, available in Table 2.

Neural Nexus [11] adopted a multi-faceted deep learning approach. Their strategy combined supervised and unsupervised techniques. They trained multiple autoencoders, including convolutional and Vision Transformer (ViT)-based variants, to learn latent representations from synthetic images. The ViT autoencoder, pretrained on a tuberculosis dataset, was used to extract features for clustering and classification tasks. These latent vectors were clustered using spectral clustering to identify patterns correlating with the use of specific real images in GAN training. In a parallel track, the team developed a ResNet-based autoencoder and supplemented the learned features with handcrafted descriptors, such as GLCM texture features, wavelet transforms, and Gabor filters. These were then input to a Random Forest classifier that also integrated Mahalanobis distance-based anomaly scores to identify potential training data influence. Finally, they explored a Dual-Contrastive Learning GAN (DCLGAN) to learn and compare attention-based feature maps from real and generated images. The overlap in discriminator attention regions was analyzed to infer training image influence, forming another signal for attribution. Each component of the pipeline was designed to capture complementary aspects of training data fingerprints. Results obtained by the team are shown in Table 2 and consists in the following methods:

- Run ID 1878: ViT-Based Encoder
- Run ID 1880: Spectral Clustering on AutoEncoder Features
- Run ID 1881: ResNet Autoencoder and Feature-Based Detection Framework

Team **zhouyijiang1** [12] explored a ViT-based two-stage detection architecture to identify potential fingerprint leakage from training data in GAN-generated medical images. Their system integrates a coarse-to-fine strategy: in the first stage, a feature approximation module using the FAISS L2 engine and cosine similarity identifies the top-50 most similar real images for each synthetic sample. Spectral clustering further refines these candidates to reduce outliers. In the second stage, a Hybrid Contrastive Attention Network (CANet) performs deep semantic matching. The model fuses features from multiple ViT layers (6, 12, 18) using a gated fusion mechanism with dynamic weights to generate a 1536-dimensional feature embedding. This is passed into a supervised contrastive module combining cross-entropy and contrastive loss to enhance discriminative power and suppress false positives. Feature encoding is enriched through dynamic data augmentation (including random masking and Gabor noise), and a custom positional encoding strategy augments anatomical awareness. The deep attention module consists of a three-stage pipeline: feature mapping, global dependency modeling via an 8-head attention mechanism, and decision refinement using adaptive temperature scaling and sigmoid classification. A final decision is made by checking if a sample has high attention-weighted similarity and confidence >0.6. The method is implemented using ViT-Base and optimized via AdamW with cosine learning rate decay. The final system balances high detection performance with low false positive rates and real-time inference. The obtained results are shown in Table 2 and consists in the following:

- Run ID 1873: Vision Transformer
- Run ID 1802:
- Run ID 1801:
- Run ID 1803:
- Run ID 1804:

Team **taozi** [13] proposed a hybrid framework. Their approach centered on contrastive learning and a Mixture of Experts (MoE) architecture that integrates three pre-trained backbones: ResNet-50, EfficientNet-B0, and ViT-B/16. Initially, feature similarity matrices were computed using these backbones to create pools of candidate positives and negatives. These were used to train a MoCo v2-style contrastive model enhanced with dynamic projection heads and adaptive sample-wise temperature scaling. The team used an online encoder with momentum updates, InfoNCE loss, and FIFO queues for negative sampling to ensure diverse and stable contrastive training. Their dynamic projection head projects 2048-dimensional features into a 256-dimensional embedding space, while generating adaptive temperature values to control learning sensitivity. A Hybrid Contrastive Loss was employed, balancing online and momentum components. For multi-model integration, the team implemented a Mixture of Experts module, which L2-normalized and concatenated outputs from all three backbones. A two-layer gating network then weighted each exper's contribution, followed by an 8-head cross-attention mechanism to refine fused features. This enabled the system to leverage complementary strengths of each backbone for better generalization across anatomical patterns. The obtained results are shown in Table 2 and consists in the following:

- Run ID 1875: ResNet50
- Run ID 1874: MoE
- Run ID 1169: EfficientNEt-B0
- Run ID 1140: Vision Transformer
- Run ID 1179: ResNet50
- Run ID 1107: ResNet50

Team ZOQ [14] explored a similarity classification approach that combines image enhancement and deep learning models. The pipeline begins with image preprocessing using four enhancement techniques: Gaussian filtering to reduce noise while preserving edges, the Laplacian operator for edge detection, the Hessian matrix to enhance curvature-based features, and bilateral filtering for edge-preserving smoothing. These transformations aim to highlight critical structural features and improve the quality of the extracted representations. Following enhancement, the team employed two deep learning architectures: CNN and ResNet50 for feature extraction. These models were used to generate high-dimensional feature vectors representing each image. To determine similarity between synthetic and real images, three different similarity metrics were evaluated: Cosine similarity (to assess directional closeness of feature vectors), Structural Similarity Index (SSIM) (capturing luminance, contrast, and structural alignment), and Jaccard similarity (measuring overlap between binarized feature regions). The final classification of an image as "used" or "not used" was determined by applying thresholds to these similarity scores, offering a multi-angle view of feature resemblance between real and generated samples. The obtained results are shown in Table 2 and consists in the following:

- Run ID 1355: CNN + SSIM
- Run ID 1427: ResNet with Jaccard

AI Multimedia Lab [15] explored a twin-branch Siamese network, where each branch received image inputs and produced feature embeddings via convolutional layers followed by dense layers with L2 normalization. The model was trained using contrastive loss to encourage similar embeddings for real-synthetic pairs originating from the same training source and dissimilar ones otherwise. Input pairs were composed of synthetic images and either real-used (positive) or real-not-used (negative) images. For inference, similarity scores between image pairs were computed using Euclidean distance, and binary predictions were derived based on learned thresholds. This architecture, previously applied in Subtask 2 for subset attribution, was adapted here to learn the nuanced similarities indicative of GAN training data usage. The obtained results are presented in Table 2 under Run ID 1696 and 1492.

4.2.2. Sub-task 2: Identify Training Data Subsets

SDVAHCS/UCSD [16] presented an ensemble-based deep learning approach to identify the origins of synthetic biomedical images by classifying them according to the specific real data subsets used in their generation. The authors employed multiple EfficientNet architectures (b0, b1, b2), leveraging a six-class classification setup to distinguish between real images and five sets of synthetic images, each linked to a distinct training subset. A max-voting ensemble strategy was used to enhance robustness, and pseudo-labeling was incorporated to utilize unlabeled test data by iteratively assigning labels and retraining models. In one configuration, a portion of the original data was sequestered to validate the generalization of the pseudo-labeling technique and to detect overfitting. The results obtained by the team are shown in Table 3 and consists in the following methods:

- Run ID 1425: A single EfficientNet-b1 model trained with a learning rate of 0.0001 and batch size 32; selected for best validation performance.

Table 2
Results of participant submissions and their results for Subtask 1: Detect Training Data Usage.

#	Participant	Run ID	Entries	Cohen's kappa	Accuracy	Precision	Recall	F1
1	Neural Nexus	1878	7	**0.148**	0.574	0.5698	0.604	0.5864
2	zhouyijiang1	1803	5	**0.136**	0.568	0.5582	0.652	0.6015
3	zhouyijiang1	1804	5	**0.136**	0.568	0.5582	0.652	0.6015
4	zhouyijiang1	1873	5	**0.136**	0.568	0.5582	0.652	0.6015
5	zhouyijiang1	1802	5	**0.132**	0.566	0.5537	0.68	0.6104
6	zhouyijiang1	1801	5	**0.128**	0.564	0.55	0.704	0.6175
7	Neural Nexus	1880	7	**0.072**	0.536	0.5542	0.368	0.4423
8	taozi	1359	8	**0.064**	0.532	0.5597	0.3	0.3906
9	taozi	1367	8	**0.044**	0.522	0.5505	0.24	0.3343
10	AIMultimediaLab	1696	2	**0.036**	0.518	0.5162	0.572	0.5427
11	taozi	1364	8	**0.032**	0.516	0.6	0.096	0.1655
12	Neural Nexus	1881	7	**0.032**	0.516	0.5222	0.376	0.4372
13	taozi	1360	8	**0.032**	0.516	0.5128	0.64	0.5694
14	Neural Nexus	1877	7	**0.028**	0.514	0.5164	0.44	0.4752
15	taozi	1366	8	**0.02**	0.51	0.5069	0.732	0.599
16	Neural Nexus	1872	7	**0.016**	0.508	0.5182	0.228	0.3167
17	Medhastra	1288	1	**0.016**	0.508	0.5078	0.52	0.5138
18	taozi	1368	8	**0.012**	0.506	0.5092	0.332	0.4019
19	Challengers	1811	5	**0.012**	0.506	0.5062	0.492	0.499
20	ZOQ	1427	5	**-0.016**	0.492	0.4905	0.412	0.4478
21	Neural Nexus	1879	7	**-0.024**	0.488	0.4732	0.212	0.2928
22	Neural Nexus	1882	7	**-0.028**	0.486	0.4646	0.184	0.2636
23	ZOQ	1355	5	**-0.032**	0.484	0.4904	0.82	0.6138
24	SCOPE VIT Visioneers	1160	1	**-0.032**	0.484	0.4831	0.456	0.4691
25	Challengers	1779	5	**-0.032**	0.484	0.4355	0.108	0.1731
26	AIMultimediaLab	1492	2	**-0.044**	0.478	0.4829	0.62	0.5429
27	ZOQ	1330	5	**-0.068**	0.466	0.4822	0.92	0.6327
28	ZOQ	1794	5	**-0.068**	0.466	0.4822	0.92	0.6327
29	taozi	1369	8	**-0.096**	0.452	0.4657	0.652	0.5433
30	Challengers	1778	5	**-0.116**	0.442	0.4461	0.48	0.4624
31	ZOQ	1356	5	**-0.132**	0.434	0.3862	0.224	0.2835
32	Challengers	1776	5	**-0.176**	0.412	0.3764	0.268	0.3131
33	Challengers	1777	5	**-0.176**	0.412	0.3764	0.268	0.3131

- Run ID 1426: An ensemble of EfficientNet-b0, b1, and b2 models, using max-voting. Only models with a validation accuracy higher than 98% were included.
- Run ID 1782: An ensemble of 26 EfficientNet models (b0-b2) trained with various hyperparameters. It incorporated pseudo-labeled test data and used a sequestered set for overfitting checks.
- Run ID 1871: An ensemble of EfficientNet-b1 models trained on all available data (including pseudo-labeled test images), without using a sequestered validation set.
- Run ID 1883

Team Medhastra [17] employed a ResNet-18-based deep learning pipeline to classify. The approach involved two main components: feature extraction and supervised multi-class classification. For feature extraction, ResNet-18 (pretrained on ImageNet [10]) was repurposed by removing its classification layer, transforming each image into a 512-dimensional vector capturing high-level semantic features. These embeddings were used to measure similarity between synthetic and real images and support subgroup inference. For classification, the team fine-tuned ResNet-18 to perform five-way classification. The model was trained on labeled real images and validated using synthetic images to better simulate inference-time conditions. During inference, the trained model predicted subgroup labels for unlabeled synthetic images in the test set. The results of the team are available Table 2 under Run ID 1287.

AI Multimedia Lab [15] employed two distinct methods: (i) Method 1 –Feature Clustering, their previous feature-based clustering pipeline [18] was adaptedto the 2025 subset attribution task. The approach begins with feature extraction using four pretrained convolutional models: MobileNetV2, ResNet50, EfficientNet, and DenseNet, all originally trained on ImageNet [10]. These models generate high-dimensional embeddings from the input images to capture their semantic content. To infer which subset (T1-T5) was used to generate each synthetic image, the team experimented with three classification strategies: k-means clustering, hierarchical clustering, and Support Vector Machines. The first two are unsupervised techniques that group similar images based on feature similarity - k-means via centroid minimization and hierarchical clustering via distance-based tree structures while SVM serves as a supervised baseline trained on labeled subset data. The models were fine-tuned on the training data, with 10% of the set held out for validation to evaluate clustering and classification performance. This method investigates whether deep features from pretrained networks can effectively capture patterns indicative of the GAN training subsets, using both supervised and unsupervised grouping strategies. ii) A Siamese network was trained using contrastive loss to learn whether a real-synthetic image pair came from the same training subset. Each synthetic image was compared with real images from all five subsets, and average embedding distances were calculated per subset. The subset with the smallest average distance was selected as the predicted origin. This approach focuses on learning visual similarity through a relational, pairwise metric. The results of the team are available in Table 3:

- Run ID 1396: Siamese Neural Network
- Run ID 1271: Densenet + SVM
- Run ID 1269: EfficientNet + SVM
- Run ID 1268: ResNet + SVM
- Run ID 1267: MobileNetV2 + SVM.

5. Discussion

A wide range of strategies was deployed for Subtask 1, spanning from contrastive learning and Siamese architectures to hybrid clustering approaches, vision transformers, and handcrafted feature modeling.

SCOPE VIT Visioneers adopted a SNN enhanced with a cross-attention module and an adaptive similarity head, using backbones such as ResNet50 and ViT. Despite the architectural sophistication, their best-performing submission (Run ID 1160) achieved a Cohen's kappa of -0.032, indicating performance worse than chance, despite achieving a moderate accuracy of 0.484. Similarly, Medhastra employed

Table 3
Results of participant submissions and their results for Subtask 2: Identify Training Data Subsets.

#	Participant	Run ID	Entries	Accuracy	Precision	Recall	F1	Specificity
1	AIMultimediaLab	1396	5	**0.9904**	0.9904	0.9904	0.9904	0.9972
2	SDVAHCS/UCSD	1782	5	**0.988**	0.9882	0.988	0.9881	0.9969
3	SDVAHCS/UCSD	1871	5	**0.988**	0.9882	0.988	0.9881	0.9969
4	SDVAHCS/UCSD	1883	5	**0.988**	0.9882	0.988	0.9881	0.9969
5	SDVAHCS/UCSD	1426	5	**0.9878**	0.9881	0.9878	0.988	0.9969
6	SDVAHCS/UCSD	1425	5	**0.9708**	0.9716	0.9708	0.9711	0.9931
7	Medhastra	1287	1	**0.9484**	0.9504	0.9484	0.9487	0.9879
8	AIMultimediaLab	1268	5	**0.5236**	0.5982	0.5236	0.5327	0.8799
9	AIMultimediaLab	1269	5	**0.4913**	0.5822	0.4913	0.4934	0.8744
10	AIMultimediaLab	1271	5	**0.4904**	0.5691	0.4904	0.4832	0.8753
11	AIMultimediaLab	1267	5	**0.4112**	0.4645	0.4112	0.3945	0.8547

cosine similarity thresholds and logistic regression on ResNet50 embeddings, but their submission (Run ID 1288) also yielded a negative kappa score of 0.016, further confirming the difficulty of the task. AI Multimedia Lab, which previously applied a Siamese approach to Subtask 2, adapted the same architecture here, but their runs (IDs 1696 and 1492) achieved low kappa scores of 0.036 and -0.044, respectively. These results highlight that feature-level similarity, even when trained with contrastive loss, may be insufficient to reliably detect GAN training inclusion at the image level. Challengers investigated both unsupervised and supervised pipelines, including PCA-reduced clustering and a fine-tuned ResNet50V2 classifier. While their best supervised model (Run ID 1811) performed relatively well in terms of accuracy (0.506), it still produced a very low kappa score of 0.012, suggesting that even supervised classifiers struggle with consistent predictions beyond chance. Neural Nexus introduced an ambitious multi-track pipeline combining ViT-based autoencoders, handcrafted features, and GAN attention analysis. Their best run (ID 1878) achieved the highest Cohen's kappa of 0.148, still modest but indicating some ability to detect data usage patterns, supported by a higher F1-score of 0.5864. Team zhouyijiang1 implemented a ViT-based architecture featuring coarse-to-fine filtering, spectral clustering, and a hybrid contrastive attention network. Their suite of runs (IDs 1801-1804, 1873) consistently achieved kappa values in the 0.128-0.136 range, among the best overall, demonstrating slight but stable agreement beyond chance. Taozi pursued a contrastive learning framework enriched with a Mixture of Experts (MoE) module integrating ResNet, EfficientNet, and ViT backbones. Despite their architectural complexity, none of their runs exceeded a kappa of 0.108, and several dropped below zero, reflecting the difficulty of leveraging contrastive embeddings for precise image-level attribution. ZOQ applied deep similarity classification using image enhancement (e.g., Gaussian, Laplacian, Hessian filtering) and various similarity metrics (cosine, SSIM, Jaccard). Their submissions, including Run IDs 1355 and 1427, achieved Cohen's kappa values ranging from -0.016 to -0.132, reinforcing that even enriched structural representations could not capture consistent training reuse signals.

Cohen's kappa scores across all submissions were notably low, with only a few teams surpassing the 0.1 threshold. This statistical measure, which ranges from -1 (systematic disagreement) to 1 (perfect agreement), reflects the agreement between predicted labels and ground truth while correcting for chance. The prevalence of negative or near-zero kappa values indicates that most models performed at or below chance level, implying methods' inability to detect the source images used for training. This also could reflect model's ability to generate synthetic images that do not contain "fingerprints" that can be traced back to the source of training.

In Subtask 2, participants were tasked with linking each synthetic biomedical image to the specific subset of real data used during its generation. This required detailed attribution of synthetic images to one of five real training subsets (T1-T5), posing a significant challenge in capturing subtle dataset-specific features embedded by the generative model. The top performance, excluding the organizing team was obtained by SDVAHCS/UCSD team (agentili), submitting multiple ensemble-based runs. Their most notable configurations (Run IDs 1782, 1871, and 1883) each achieved an accuracy of 98.8%. These submissions relied on ensembles of EfficientNet models (b0 - b2) and employed a max-voting strategy for robust decision-making. Pseudo-labeling was introduced to incorporate unlabeled test data by iteratively assigning labels and retraining the models. Notably, Run ID 1782 included a sequestered subset of the training data to monitor and prevent overfitting during pseudo-label-based augmentation. This strategy enabled the team to verify that the pseudo-labeling process did not lead to data leakage or artificial performance inflation, enhancing model generalization and credibility. A simpler configuration (Run ID 1425) using a single EfficientNet-b1 model also performed strongly with 97.08% accuracy.

The organizing team, AI Multimedia Lab, whose submission (Run ID 1396) achieved the highest accuracy of 99.04% was based on a SNN trained with contrastive loss to learn similarity relationships between synthetic and real images. During inference, the model compared each synthetic image to examples from each subset and attributed the image to the subset with the smallest average embedding distance. This pairwise metric-learning approach effectively captured fine-grained inter-image similarities, yielding highly discriminative performance across all five subsets. Team Medhastra (Run ID 1287) adopted a more streamlined approach using a ResNet-18 backbone. Their pipeline combined deep feature extraction and supervised multi-class classification. Feature embeddings were generated from ResNet-18, and the model was fine-tuned to distinguish between the five subsets using real training images. During inference, predictions for synthetic images were generated directly via softmax classification. Despite its relative simplicity and lack of ensembling, this method achieved a commendable 94.84% accuracy, demonstrating the efficacy of targeted fine-tuning on well-designed architectures.

In contrast, AI Multimedia Lab's second method, which extended their prior year's pipeline [18], yielded lower performance. This approach (Run IDs 1267 - 1271) relied on deep feature clustering and classification using models such as MobileNetV2, ResNet50, EfficientNet, and DenseNet. The extracted embeddings were processed via various clustering or classification techniques, including k-means, hierarchical clustering, and SVMs. While methodologically diverse and interpretable, these models achieved accuracies ranging from 41.12% to 52.36%, indicating that static pretrained features lacked the discriminative capacity required for this fine-grained attribution task. The relatively poor performance suggests that such representations may be insufficient for capturing the generative nuances that distinguish training subsets in GAN outputs.

Overall, the results indicate that contrastive learning frameworks and ensemble deep learning models, particularly those using EfficientNet architectures, are highly effective for subset attribution. Techniques such as pseudo-labeling, max-voting, and overfitting control through sequestered validation further enhance performance. Meanwhile, feature clustering methods, though appealing for their simplicity and generality, may require significant enhancement or task-specific tuning to match the performance of supervised or metric-learning-based approaches.

Lessons learned

The 2025 edition of the ImageCLEFmedical GANs task provided valuable insights into the challenges of detecting training data exposure in synthetic medical images, with each subtask offering lessons from a different generative modeling approach.

Subtask 1 focused on detecting whether individual real images had been used to train GAN. Despite the use of a wide array of methods – including Siamese networks, contrastive learning, and attention-based models –participants achieved only modest performance, with the highest Cohen's kappa score reaching 0.148. This suggests that detecting image-level membership in GAN training data remains a highly challenging problem. It may also indicate that, under the given task conditions, the GAN was effective at generating images that do not exhibit easily detectable traces of specific training samples. This is a potentially encouraging result from a privacy standpoint, but it simultaneously underscores the limitations of current detection frameworks and the need for continued exploration of more sensitive or robust analysis techniques.

Subtask 2, in contrast, explored attribution at a broader scale: identifying the training subset used to generate synthetic images from a diffusion model. Here, performance was significantly higher, with multiple teams exceeding 98% accuracy. While this task did not involve image-level membership inference, it revealed that diffusion-generated images can still reflect statistical characteristics of the training data subsets. The success of various supervised and semi-supervised classification techniques suggests that diffusion models, while effective at preserving image realism, may still encode latent information about their source data distributions. This raises important considerations about the use of synthetic data for data sharing or augmentation, particularly when the provenance or diversity of training data must remain confidential.

Collectively, the lessons from both subtasks emphasize that generative model outputs can leak different types of information, depending on both the model architecture and the nature of the detection task. Continued benchmarking and task diversification will be essential for building a more comprehensive understanding of privacy risks in generative medical imaging and for guiding the development of responsible and secure data generation practices.

6. Conclusions

The 2025 edition of the ImageCLEFmedical GANs Task further investigated the privacy and security implications of using GAN-generated synthetic medical images by addressing two key challenges: determining whether specific real images were used during GAN training (Subtask 1), and attributing synthetic images to their original training data subsets (Subtask 2).

In Subtask 1, despite the broad range of approaches: from SNN and contrastive learning to attention-based ViTs and handcrafted similarity metrics, overall performance remained low. The highest Cohen's kappa score reached only 0.148, suggesting that current methodologies struggle to reliably detect whether individual real images were used to generate synthetic outputs. These results may indicate the capability of the employed generative model to produce realistic images that do not easily leak training data, thereby better preserving patient privacy. However, the challenge also highlights a pressing need for more refined evaluation techniques and robust detection frameworks capable of identifying subtle "fingerprints" in high-fidelity synthetic medical imagery. This remains an open area for further investigation.

In contrast, Subtask 2 achieved considerably better results, with multiple teams surpassing 98% classification accuracy. The most successful approaches utilized supervised classification with EfficientNet ensembles and contrastive Siamese networks, some enhanced with pseudo-labeling and max-voting strategies. These findings suggest that while identifying whether an individual image was used in GAN training remains difficult, it is more feasible to capture coarse-grained, dataset-level attributions embedded in synthetic images generated by Diffusion models. This supports the hypothesis that Diffusion models may retain latent features reflecting the broader statistical properties of their training subsets.

Overall, the task underscores the need for ongoing research into privacy-preserving generative modeling, especially in sensitive domains such as healthcare. Benchmarking challenges like this remain crucial for understanding the evolving capabilities and risks associated with synthetic medical data. We will continue investigating these issues in future editions of ImageCLEF, where we plan to introduce new tasks aimed at further exploring generative model privacy, attribution, and generalization across modalities and model types.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4o in order to: Grammar and spelling check and improve writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

7. Acknowledgments

The work of Alexandra Andrei was supported by a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-IV-P6-6.3-SOL-2024-0049, within PNCDI IV. The work of Mihai Gabriel Constantin was supported by a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-IV-P6-6.3-SOL-2024-0060, within PNCDI IV.

References

[1] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, 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), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[2] A.-G. Andrei, A. Radzhabov, I. Coman, V. Kovalev, B. Ionescu, H. Müller, Overview of imageclefmedical gans 2023 task: identifying training data “fingerprints” in synthetic biomedical images generated by gans for medical image security, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 3497, 2023.

[3] A. Andrei, A. Radzhabov, D. Karpenka, Y. Prokopchuk, V. Kovalev, B. Ionescu, H. Müller, Overview of 2024 imageclefmedical gans task–investigating generative models’ impact on biomedical synthetic images, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Grenoble, France, 2024.

[4] B. Ionescu, H. Müller, A.-M. Drăgulinescu, W.-W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. García Seco de Herrera, et al., Overview of the imageclef 2023: multimedia retrieval in medical, social media and internet applications, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 370–396.

[5] B. Ionescu, H. Müller, A.-M. Drăgulinescu, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, et al., Overview of the imageclef 2024: Multimedia retrieval in medical applications, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 140–164.

[6] J. Cohen, A coefficient of agreement for nominal scales, Educational and psychological measurement 20 (1960) 37–46.

[7] F. Areeb, D. Vashist, L. Kalinathan, Controlling the quality of synthetic medical images created via gans, in: CLEF2025 Working Notes, Madrid, Spain, 2025.

[8] K. S. S, M. S, A. Narayanan, B. P, K. Ayyamperumal, Detecting training data usage in synthetic images using machine learning techniques, in: CLEF2025 Working Notes, Madrid, Spain, 2025.

[9] S. Chandrasekar, V. R. S, V. P, Detecting training data fingerprints in gan-generated medical images, in: CLEF2025 Working Notes, Madrid, Spain, 2025.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[11] M. Barve, S. Nambiar, N. Bhedasgaonkar, A. Date, I. Shah, D. G. V. Kale, Reverse engineering generative fingerprints in medical images: A deep learning approach to training data attribution, in: CLEF2025 Working Notes, Madrid, Spain, 2025.

[12] Y. Zhou, H. Ding, Vit-based generative model fingerprinting, in: CLEF2025 Working Notes, Madrid, Spain, 2025.

[13] D. Zhang, X. Yang, Evaluation of the privacy of images generated by imageclefmedical gans 2025 based on pre-trained model feature extraction methods, in: CLEF2025 Working Notes, Madrid, Spain, 2025.

[14] H. Zuo, X. Zhou, Evaluating of the privacy of images generated by imageclefmedical gan 2025 using similarity classification method based on image enhancement and deep learning mode, in: CLEF2025 Working Notes, Madrid, Spain, 2025.

[15] A. Andrei, M. G. Constantin, M. Dogariu, L. Ştefan, B. Ionescu, Ai multimedia lab at imageclefmedical gans 2025: Identifying real-image usage in generated medical images, in: CLEF2025 Working Notes, Madrid, Spain, 2025.

[16] A. Gentili, Identifying the origins of synthetic biomedical images: Anensemble approach with pseudo-labeling, in: CLEF2025 Working Notes, Madrid, Spain, 2025.

[17] S. Chandrasekar, V. R. S, V. P, Identify training data subsets in gan-generated medical images, in: CLEF2025 Working Notes, Madrid, Spain, 2025.

[18] A. Andrei, M. G. Constantin, M. Dogariu, B. Ionescu, Ai multimedia lab at imageclefmedical gans 2024: Deep learning approaches for analyzing synthetic medical images, in: CLEF2024 Working Notes,CEUR Workshop Proceedings, Grenoble, France, 2024.