AI Multimedia Lab at ImageCLEFmedical GANs 2026: Synthetic medical image generation and privacy analysis

ImageCLEF at CLEF 2026

Alexandra-Georgiana Andrei[1,*], Mihai Gabriel Constantin[1], Mihai Dogariu[1], Dan-Cristian Stanciu[1], Liviu-Daniel Ștefan[1] and Bogdan Ionescu[1]

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

Abstract

This paper presents the AI Multimedia Lab's participation in the ImageCLEFmedical GANs 2026 challenge, which addresses privacy and security risks in GAN-generated synthetic medical images across three subtasks. For Subtask 1 (training data detection), we combine k-nearest-neighbour distances in ResNet-50 feature space with pixel-level descriptors and a Gradient Boosting classifier to detect which real images were used during training. For Subtask 2 (latent vector matching), we compare four strategies of increasing complexity: supervised encoder-based GAN inversion, a dual encoder trained with a symmetric InfoNCE loss, an improved variant with hard negative mining and curriculum learning, and a training-free linear mapping in DINO feature space. For Subtask 3 (privacy-preserving CT generation), we train StyleGAN2 on the provided development set and evaluate the outputs for both image quality and privacy leakage. Direct supervised regression (Subtask 2, Method 1) is the strongest of the four strategies, reaching a matching accuracy of 0.3737, whereas our membership inference pipeline for Subtask 1 performs at chance level, a result that says as much about the difficulty of fingerprinting well-regularised GANs as it does about our pipeline. The StyleGAN2 submission for Subtask 3 achieves high image quality (FID = 1.2082) alongside only moderate privacy preservation (PPS = 0.6041), with patient re-identification standing out as the dominant leakage channel.

Keywords

ImageCLEF, GANs, generative models, medical imaging, image augmentation, privacy

1. Introduction

The fourth edition of the ImageCLEFmedical GANs task [1] builds on the previous editions [2, 3, 4] and continues to examine privacy-related concerns in GAN-generated synthetic medical images, using the platform offered by ImageCLEF [5, 6]. This paper presents the methods proposed by the AI Multimedia Lab, the task organising team, for the fourth edition of the challenge. The remainder of this paper is structured as follows: Section 2 describes the subtasks proposed in this edition, Section 3 presents the proposed methods, and the results are discussed in Section 4. Section 5 concludes the paper.

2. The 2026 ImageCLEFmedical GANs task

The 2026 edition of the ImageCLEFmedical GANs task [1] continues to explore privacy and security concerns in GAN-generated synthetic medical images. The benchmarking dataset includes both real and synthetic biomedical images. The real images consist of axial slices of 3D CT scans from approximately 8,000 lung tuberculosis patients, stored in 8-bit per-pixel PNG format at a standardised resolution of 256×256 pixels. Slices vary considerably in appearance: some are relatively normal, while others exhibit distinct lung lesions, including severe cases. This edition comprises three complementary subtasks, described below.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

alexandra.andrei@upb.ro (A. Andrei); mihai.constantin84@upb.ro (M. G. Constantin); mihai.dogariu@upb.ro (M. Dogariu); dan.stanciu1203@upb.ro (. D. Stanciu); liviu_daniel.stefan@upb.ro (L. Ștefan); bogdan.ionescu@upb.ro (B. Ionescu)

2.1. Subtask 1: Detect training data usage

This subtask has been part of the challenge since its first edition and is carried forward here with the same setup, applied to a new GAN architecture. The goal is to detect training-data "fingerprints" within synthetic outputs and to assess vulnerabilities related to memorisation. Participants must examine a set of biomedical images and decide, for each real image, whether it was used to train the generative model that produced the provided synthetic outputs, assigning a binary label (1 if used during training, 0 otherwise).

Data. The development dataset contains both real and generated images, with real images annotated according to whether they were used for training. The test dataset consists of a mixed set of real images, some of which were used during training.

Evaluation. The primary evaluation metric is Cohen's Kappa, which measures agreement between predicted and ground-truth labels while accounting for chance. Additional metrics include accuracy, sensitivity, specificity, precision, F1 score, and Matthews Correlation Coefficient.

2.2. Subtask 2: Identify corresponding latent vectors

This subtask evaluates how faithfully generative models map latent vectors to output images, and whether this mapping reveals structural or patient-specific information. The task consists in matching each generated image to its originating latent vector. The latent vectors are 128-dimensional and sampled from a standard normal distribution $\mathcal{N}(0, 1)$.

Data. The dataset consists of 256×256 images paired with their corresponding z-vectors. The development set contains 10,000 image-z-vector pairs. The test data consists of a folder of GAN-generated images and a file of latent vectors that must be matched to their corresponding images.

Evaluation. Performance is measured by matching accuracy: the proportion of submitted image-z-vector pairs for which the predicted index matches the ground truth.

2.3. Subtask 3: Privacy-preserving CT slice generation

The objective of this subtask is to develop a generative model capable of producing realistic and diverse CT slices while preserving patient privacy.

Data. The dataset is released in two phases. The development dataset contains real CT slices intended for model training; all images are 2D axial slices from the same anatomical region, drawn from multiple patients, with no patient identifiers provided. The validation dataset contains additional real CT slices from patients not included in the development set, intended to support model development and hyperparameter tuning.

Evaluation. Submitted images are assessed across two dimensions: quality and privacy.
Image quality is measured using the Fréchet Inception Distance (FID), computed with a task-specific feature extractor trained on domain-relevant images.
Privacy is quantified through a Privacy Preservation Score (PPS), defined as a composite score averaging four leakage scores derived from four distinct attack types: (i) Attack 1. Nearest-Neighbour Distance (NND) - l_1: measures instance-level memorisation, i.e. whether any synthetic images are near-copies of specific training samples; (ii) Attack 2. Membership Inference Attack (MIA) - l_2: measures distributional bias, i.e. whether the synthetic distribution as a whole reveals which patients were used for training, even in the absence of individual image copies; (iii) Attack 3. Patient Re-Identification - l_3: measures patient identity leakage, i.e. whether the synthetic distribution reflects the anatomical characteristics of

individual training patients, even if no image is a direct visual copy; (iv) Attack 4. Generalised Inversion Attack - l_4: measures latent-space memorisation, i.e. whether the generator's internal representations encode training-specific anatomy in a way that allows training images to be reconstructed more faithfully than unseen held-out images. A detailed description of the privacy evaluation protocol is available in the task overview [1]. The PPS is computed as:

$$\text{PPS} = 1 - \text{mean}(l_1,\, l_2,\, l_3,\, l_4) \tag{1}$$

A higher PPS indicates stronger privacy preservation.

3. Proposed Methods

3.1. Subtask 1

Our pipeline consists of four stages: (i) feature extraction, (ii) k-NN distance computation, (iii) feature matrix assembly, and (iv) classification. While the overall structure is similar to those submitted in previous editions [7, 8, 9], this edition adopts a different approach to feature representation and classification.

Deep feature extraction. We use a ResNet-50 convolutional neural network [10] pretrained on ImageNet [11] as a fixed feature extractor. Each image is preprocessed as follows: resized to 256×256 pixels, centre-cropped to 224×224 to match ResNet-50's expected input, replicated across three channels to convert greyscale to pseudo-RGB, and normalised using the ImageNet channel mean and standard deviation. A forward hook attached to the global average pooling layer (`avgpool`) yields a 2048-dimensional embedding $\phi(x) \in \mathbb{R}^{2048}$ per image. Features are extracted for all 10,000 generated images, the 6,400 labelled real images in the development set, and the 2,140 unlabelled test images.

k-Nearest-Neighbour distance to generated set. Let $\mathcal{G} = \{\phi(g_1), \ldots, \phi(g_N)\}$ denote the set of feature vectors extracted from the $N = 10{,}000$ generated images. For each real image x, we retrieve the $k = 5$ nearest neighbours of $\phi(x)$ in $\mathcal{G}$ under Euclidean distance. From the resulting distances $d_1 \leq d_2 \leq \cdots \leq d_k$, four statistics are derived:

- $f_1 = \frac{1}{k}\sum_{i=1}^{k} d_i$ (mean distance; primary membership signal)
- $f_2 = d_1$ (distance to the single nearest generated image)
- $f_3 = \text{std}(d_1, \ldots, d_k)$ (consistency of the match across neighbours)
- $f_4 = d_k$ (distance to the furthest of the k neighbours)

A low f_1 means several generated images closely resemble the query, which is a strong indicator of training membership. A non-training image may still land near one generated image by chance, but its broader k-NN neighbourhood tends to be more dispersed – reflected in higher f_3 and f_4.

Pixel-level descriptors. To complement the distance features, we compute four lightweight statistics directly from the normalised greyscale pixel values $p \in [0, 1]^{H \times W}$:

- $f_5 = \mathbb{E}[p]$ (mean intensity)
- $f_6 = \text{std}(p)$ (global contrast)
- $f_7 = \mathbb{E}[p] - \text{median}(p)$ (skewness proxy)
- $f_8 = -\sum_b h_b \log h_b$ (entropy proxy, where h_b is the normalised 64-bin histogram)

These descriptors capture distributional properties of the image content that may differ systematically between training and non-training images.

Classification. The eight features $\mathbf{f} = (f_1, \ldots, f_8)$ are standardised to zero mean and unit variance. We train a Gradient Boosting classifier [12] in scikit-learn [13]) with 200 estimators, a maximum tree depth of 4, a learning rate of $\eta = 0.05$, and a subsample ratio of 0.8.

To obtain a reliable estimate of generalisation performance, the classifier is evaluated using 5-fold stratified cross-validation, ensuring that no image is scored by a model that has seen it during training. The final model for test-set prediction is retrained on the full development set. As a simple interpretable baseline, we also apply a threshold directly to f_1: images whose f_1 falls below the midpoint of the two class-conditional means are assigned label 1. This threshold-based rule requires no training and provides a lower bound on achievable performance.

3.2. Subtask 2: Identify corresponding latent vectors

We explored four distinct techniques for this subtask.

Method 1: Supervised encoder-based GAN inversion. This approach frames the matching problem as a supervised regression task. A ResNet-50 backbone pretrained on ImageNet is fine-tuned on the 10K labelled (image, z) pairs by replacing the classification head with a two-layer regression head that outputs a 128-dimensional vector. The model is trained to directly predict the latent vector z from its corresponding generated image, using a combined loss of mean squared error (MSE) and cosine similarity. MSE penalises magnitude errors, while cosine loss penalises directional errors, both relevant given that z is sampled from a standard normal distribution with meaningful scale and direction. At inference time, the encoder predicts a latent vector for each test image. A cosine similarity matrix is then computed between all predicted image embeddings and all candidate z-vectors, and the optimal one-to-one assignment is found using the Hungarian algorithm.

Method 2: Dual encoder with symmetric InfoNCE loss. This approach reframes matching as a metric learning problem, drawing inspiration from contrastive vision-language models such as CLIP. Rather than regressing z directly, two separate encoders are trained to map images and z-vectors into a shared 256-dimensional embedding space on the unit hypersphere, where matched pairs are brought close together and unmatched pairs are pushed apart. The image branch uses a ResNet-50 backbone followed by a projection head; the z branch uses a three-layer MLP. Both branches are trained jointly with a symmetric InfoNCE loss: for a batch of B pairs, the loss treats diagonal entries of the $B \times B$ cosine similarity matrix as positives and all off-diagonal entries as negatives, in both the image-to-z and z-to-image directions. Inference follows the same cosine similarity matrix and Hungarian assignment procedure as Method 1.

Method 3: Improved dual encoder with hard negative mining, z-augmentation, and hybrid loss. This is a variant of Method 2 with three modifications. First, a memory bank storing EMA-updated embeddings of all 10K training samples enables hard negative mining: rather than using random in-batch negatives, the top-k most similar but incorrect z-embeddings are retrieved from the full bank and appended to the InfoNCE denominator at each step. Second, Gaussian noise augmentation is applied to z-vectors during training ($z_{\text{aug}} = z + \varepsilon, \varepsilon \sim \mathcal{N}(0, 0.15^2)$) to encourage sensitivity to small perturbations in latent space. Third, a hybrid loss combines InfoNCE with direct z-regression via a curriculum schedule: the regression weight decays linearly from 1.0 to 0.05 over the first 40 epochs, providing stable gradients during the cold-start phase when embeddings are uninformative.

Method 4: Linear mapping in pretrained DINO feature space. This method sets aside learned image representations entirely. The task organisers describe the Subtask 2 generator as a BigGAN model [14], whose latent space z is known to control high-level image attributes – attributes that a self-supervised Vision Transformer such as DINO ViT-B/16 [15], trained on the same broad image distribution, tends to capture well. Building on that observation, we hypothesise and empirically test

a linear relationship between z and DINO's CLS token features.. DINO features are extracted once for all training images and cached. A ridge regression model is then fitted in both directions: $z \rightarrow$ DINO features (direction A) and DINO features $\rightarrow z$ (direction B), with the regularisation coefficient α selected via five-fold cross-validation over the range $[0.01, 1000]$. PCA whitening is applied to z-vectors before regression to normalise dimension variances. At inference, candidate z-vectors are projected into DINO feature space (direction A) and test image features are projected into z-space (direction B); the two resulting cosine similarity matrices are averaged before Hungarian assignment. This method requires no training loop, no GPU for the regression step, and no hyperparameter tuning beyond the cross-validated ridge penalty. The R^2 score of the $z \rightarrow$ feature regression on the training set serves as a direct diagnostic of how strongly z and DINO features are linearly aligned.

3.3. Subtask 3: Privacy-preserving CT slice generation

We adopted StyleGAN2 from NVIDIA [1] [16] to generate CT images, training exclusively on images from the development set. No preprocessing was applied to the training images, and the validation dataset was not used. The model was trained on a single GPU using the Adam optimiser for both the generator and discriminator ($\beta_1 = 0.0$, $\beta_2 = 0.99$, $\epsilon = 10^{-8}$).

The generator employs a non-saturating logistic loss with path length regularisation, while the discriminator uses a logistic loss with R1 gradient penalty ($\gamma = 10.0$). Training followed a progressive growing schedule, starting from an initial resolution of 8×8 and scaling up to 256×256, with resolution-dependent learning rates ($\eta = 0.0015$ at 128×128, $\eta = 0.002$ at 256×256). The minibatch size was adjusted per resolution, ranging from 128 at low resolutions down to 8 at 256×256. Horizontal mirror augmentation was applied throughout training to increase effective data diversity. At inference time, synthetic CT slices were generated by sampling latent vectors from $\mathcal{N}(0, 1)$ and passing them through the trained generator.

4. Results and discussion

4.1. Subtask 1: Detect training data usage

Table 1 reports the performance of our system on the Subtask 1 test set. The results indicate that the proposed pipeline failed to produce a meaningful membership signal on this test set. A Cohen's Kappa of -0.0056 and an MCC of -0.0056 are both consistent with random-chance performance, and all other metrics cluster closely around 0.50. The balanced accuracy of 0.4972 confirms that this failure is symmetric: the system shows no systematic tendency to favour either class.

We attribute this outcome to two factors. First, the ResNet-50 features used to compute k-NN distances were extracted from a model pretrained on natural images, without domain adaptation to CT data. The resulting embeddings likely lack the discriminative resolution needed to detect the subtle textural fingerprints that a GAN imprints on its outputs, particularly for a dataset with limited visual diversity such as axial lung CT slices at a fixed resolution. Second, the k-NN signal may simply be too weak for this specific GAN architecture: if the generator generalises well and does not strongly memorise individual training images, the distance gap between member and non-member images in feature space collapses, rendering the membership score uninformative. Together, these point to a broader lesson: membership inference against well-regularised GANs is considerably harder in practice than

[1] Source: https://github.com/nvlabs/stylegan

Table 1
Results for Subtask 1: Detect Training Data Usage.

Accuracy	Sensitivity (TPR)	Specificity (TNR)	Precision	F1 Score	Cohen's Kappa	MCC
0.4972	0.4935	0.5009	0.4972	0.4953	−0.0056	−0.0056

Table 2
Results on the test set for Subtask 2.

Run ID	Method	Matching accuracy	Correct matches	Wrong matches
965	Method 1	**0.3737**	3737.0	6263.0
968	Method 2	**0.0666**	666.0	9334.0
970	Method 4	**0.0105**	105.0	9895.0
969	Method 3	**0.0065**	65.0	9935.0

development-set cross-validation scores would suggest, a theme that resurfaces in Subtask 3, where a different GAN leaks patient identity clearly even though slice-level copying stays low.

4.2. Subtask 2: Identify corresponding latent vectors

Table 2 summarises the matching accuracy achieved by each method on the test set of 10,000 image–z-vector pairs.

Method 1 (supervised encoder-based GAN inversion) achieved the best performance with a matching accuracy of 0.3737 (3,737 correct matches). This confirms that directly regressing the latent vector from the image using a fine-tuned ResNet-50 backbone, combined with Hungarian assignment over the cosine similarity matrix, is a viable and effective strategy for this task. The combined MSE and cosine loss appears well-suited to the geometry of the latent space, penalising both magnitude and directional errors.

Method 2 (dual encoder with symmetric InfoNCE loss) achieved a matching accuracy of 0.0666 (666 correct matches), substantially lower than Method 1. The contrastive framing is conceptually appealing, but the limited training set size (10,000 pairs) likely constrains its effectiveness: InfoNCE-based methods generally benefit from large batch sizes and diverse negatives to learn meaningful embeddings.

Method 3 (improved dual encoder with hard negative mining, z-augmentation, and hybrid loss) performed worse still, achieving only 65 correct matches (accuracy 0.0065), despite its added complexity. This suggests that hard negative mining with a randomly initialised memory bank destabilises early training, and that the regression and contrastive objectives impose conflicting constraints on the embedding space when data is scarce.

Method 4 (linear mapping in pretrained DINO feature space) obtained 105 correct matches (accuracy 0.0105), outperforming Method 3 but falling well short of Methods 1 and 2. The low R^2 of the ridge regression likely reflects the fact that the relationship between BigGAN's latent space and DINO features, while partially linear, is insufficient to support accurate one-to-one matching at scale.

Overall, these results confirm that direct supervised regression (Method 1) outperforms both contrastive and unsupervised approaches for this task. When labelled (image, z) pairs are available, regression is a well-posed and efficient strategy; metric learning and linear approximations, by contrast, appear unable to capture the full complexity of the latent-to-image mapping under these data conditions.

4.3. Subtask 3: Privacy-preserving CT slice generation

We submitted a single run from the trained StyleGAN2, obtaining the results shown in Table 3.

Quality. The generated images achieve a very low FID score, and visual inspection of the samples in Figure 1 confirms that the model has learned to produce convincing CT slices that closely follow the real data distribution in terms of intensity, anatomy, and structural coherence.

Privacy. The privacy evaluation was conducted via four attacks of increasing specificity, applied against a held-out patient set not released to participants. A composite Privacy Preservation Score (PPS) is derived from the per-attack leakage scores, where 1.0 indicates fully private and 0.0 indicates maximal leakage.

Attack 1 - NND Audit. The NND audit measures, for each synthetic image, the distance to its nearest training image versus its nearest held-out image in InceptionV3 feature space. A ratio close to 1.0 is consistent with genuine generalisation; values well below 1.0 indicate memorisation. We obtained a

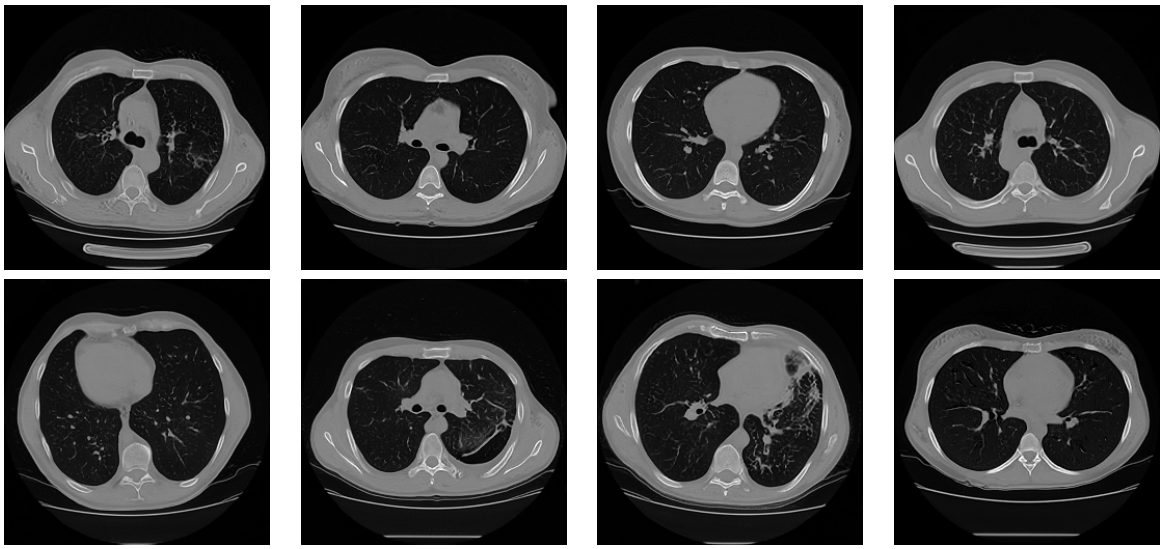

Figure 1: Samples of generated CT slices from the StyleGAN2 model.

Table 3
Results for Subtask 3: Privacy-preserving CT slice generation.

Run ID	FID	PPS	l_1	l_2	l_3	l_4
973	1.2082	**0.6041**	0.1283	0.3514	0.9060	0.1964

mean NND to training of 7.4454 versus 8.5416 to the held-out set, yielding a ratio of 0.8717 and a leakage score of 0.1283. This moderate bias indicates that synthetic images are noticeably closer to training images than to unseen patients, suggesting partial absorption of training-specific patterns. That said, the ratio remains far from the extremes, and 1359 images fell below the memorisation threshold (6.8986), implying that outright slice-level copying is limited. The model generalises to some extent, but there is a detectable distributional pull toward the training set.

Attack 2 - MIA. The MIA assigns each real CT image (training = member, held-out = non-member) a score based on its proximity to the nearest synthetic neighbour, and uses the AUC-ROC of this binary classifier to quantify how well the synthetic distribution reveals training membership. An AUC of 0.5 indicates no leakage; 1.0 indicates perfect membership disclosure. We obtained an AUC of 0.6757 and a leakage score of 0.3514, indicating that a distance-based adversary can achieve meaningful separation between members and non-members. At a 5% false positive rate, 22.5% of training images are correctly identified (TPR@5%FPR), confirming that the synthetic distribution carries detectable statistical fingerprints of the training cohort even though the leakage remains at a moderate rather than severe level.

Attack 3 - Patient Re-ID. This attack constructs a per-patient anatomical centroid from training slices and assigns each synthetic image to its top-ranked patient by cosine similarity. Rank-k coverage measures the fraction of training patients appearing as the top-k match for at least one synthetic image. We obtained a Rank-1 patient coverage of 0.9060 and a Rank-5 coverage of 0.9866, with a leakage score of 0.9060. These figures indicate near-complete patient-level coverage: 90.6% of the 149 training patients appear as the top-ranked match for at least one synthetic image with 135$\times$ above the chance level of 0.0067. The normalised assignment entropy of 0.8579 further suggests that synthetic images are spread fairly broadly across patient identities, rather than collapsing onto a small number of individuals. This result is particularly notable: even without verbatim image copying, as indicated by the relatively low NND leakage, the generator has internalised the full identity space of the training cohort. The synthetic dataset effectively functions as an anatomical index of the training patients, allowing an adversary with knowledge of those patients to link synthetic images back to specific identities.

Attack 4 - GAN Inversion. This white-box attack uses gradient-based optimisation on the submitted generator checkpoint to find the latent vector that best reconstructs each target image. If training images are reconstructed with substantially lower error than held-out images, this confirms that the generator's latent space has memorised training-specific structure. We obtained a mean reconstruction error of 8.2847 for training images versus 9.1932 for held-out images, yielding a delta of 0.9085 and a leakage score of 0.1964. While the positive delta confirms marginally greater fidelity for training images, the magnitude falls in the low-leakage range, suggesting that latent-space bias toward the training set, though detectable, is not severe.

The overall PPS of 0.6041 reflects moderate privacy preservation, with risk concentrated in a single attack surface. The dominant leakage originates from Attack 3 (re-identification, $l_3 = 0.9060$), indicating very high patient-level identity exposure. Attacks 1 (NND, $l_1 = 0.1283$) and 4 (GAN inversion, $l_4 = 0.1964$) both fall in the low-leakage range, confirming that the model neither copies individual slices nor exhibits severe latent-space memorisation. Attack 2 (membership inference, $l_2 = 0.3514$) sits at a moderate level, indicating detectable but limited distributional bias.

This pattern is characteristic of models that suppress output-level copying without formal privacy guarantees: individual slice reproduction is avoided and latent-space overfitting remains bounded, yet the generator has nonetheless internalised the full distribution of patient anatomies. Patient identity is thus recoverable at the distributional level even in the absence of verbatim memorisation, underscoring the fundamental gap between incidental privacy and formal privacy guarantees.

5. Conclusions

This paper described the participation of the AI Multimedia Lab in all three subtasks of the ImageCLEFmedical GANs 2026 challenge, covering membership inference, latent vector matching, and privacy-preserving image generation for synthetic lung CT data.

Subtask 1 turned out to be the hardest of the three: our pipeline did not beat chance on the test set (Cohen's Kappa $= -0.0056$), which suggests that when a GAN generalises well rather than memorising, the distance gap between member and non-member images in a generic feature space collapses far enough to defeat indirect, distance-based attacks. Domain-adapted embeddings, discriminator-based confidence scores, or GAN-inversion reconstruction error all seem like more promising directions than the approach we took here.

Subtask 2 gave a much clearer picture: direct supervised regression (Method 1, matching accuracy 0.3737) comfortably outperformed the contrastive approach (Method 2, 0.0666), the training-free linear mapping (Method 4, 0.0105), and the hard-negative-mining variant (Method 3, 0.0065). With labelled (image, z) pairs available, regression is simply a better-posed problem than metric learning at this data scale, where contrastive methods are constrained by both the size of the training set and the complexity of the latent-to-image mapping.

Subtask 3 delivered the paper's most striking finding. StyleGAN2 produces visually convincing CT slices (FID = 1.2082) that closely track the real data distribution, and neither instance-level copying ($l_1 = 0.1283$) nor latent-space memorisation ($l_4 = 0.1964$) is severe. Yet patient re-identification ($l_3 = 0.9060$) is near-complete: the generator has internalised the identity space of essentially all 149 training patients without ever reproducing an individual slice. In other words, a model can look private by every naive quality and copying check while remaining highly vulnerable to identity-based attacks – exactly the gap between incidental privacy and formal privacy guarantees that motivates mechanisms like DP-SGD [17] as a next step for this line of work.

Acknowledgments

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to polish specific sections of the manuscript. All scientific content and references were provided by the authors, who reviewed and verified all GPT-edited text and take full responsibility for the content of the article.

References

[1] A. Andrei, M. Constantin, M. Dogariu, D. Karpenka, Y. Prokopchuk, A. Radzhabov, D. Stanciu, L. Ştefan, V. Kovalev, H. Müller, B. Ionescu, Overview of the 2026 ImageCLEFmedical GANs task: Fingerprint detection, latent space analysis, and privacy-preserving medical image synthesis, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[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.-G. Andrei, A. Radzhabov, D. Karpenka, Y. Prokopchuk, V. Kovalev, B. Ionescu, H. Müller, Overview of the 2024 imageclefmedical gans task-investigating generative models' impact on biomedical synthetic images., in: CLEF (Working Notes), 2024, pp. 1412–1428.

[4] A.-G. Andrei, M. G. Constantin, M. Dogariu, A. Radzhabov, L. Ştefan, 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, CEUR-WS. org, Madrid, Spain, 2025.

[5] B. Ionescu, H. Müller, D.-C. Stanciu, A. Radzhabov, A. G. S. de Herrera, A.-G. Andrei, A. Băicoianu, A. Neacşu, A. Storås, A. B. Abacha, et al., Imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: European Conference on Information Retrieval, Springer, 2026, pp. 336–344.

[6] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[7] A.-G. Andrei, B. Ionescu, Aimultimedialab at imageclefmedical gans 2023: Determining" fingerprints" of training data in generated synthetic images., in: CLEF (Working Notes), 2023, pp. 1379–1386.

[8] A.-G. Andrei, M. G. Constantin, M. Dogariu, B. Ionescu, Ai multimedia lab at imageclefmedical gans 2024: Deep learning approaches for analyzing synthetic medical images., in: CLEF (Working Notes), 2024, pp. 1480–1489.

[9] 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, CLEF2025 Working Notes, Madrid, Spain (2025).

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[11] 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, Ieee, 2009, pp. 248–255.
[12] J. H. Friedman, Stochastic gradient boosting, Computational statistics & data analysis 38 (2002) 367–378.
[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.
[14] A. Brock, J. Donahue, K. Simonyan, Large scale gan training for high fidelity natural image synthesis, arXiv preprint arXiv:1809.11096 (2018).
[15] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660.
[16] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.
[17] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, 2016, pp. 308–318.