Ensemble Approaches for GAN Membership Inference and Latent Vector Prediction

ImageCLEF at CLEF 2026

Muhammad Annas Shaikh[1,*], Noor Un Nisa Shaukat[1], Hamna Inam Abro[1], Zara Masood[1] and Muhammad Atif Tahir[1]

[1]School of Mathematics and Computer Science, Institute of Business Administration (IBA), Karachi, Pakistan

Abstract

We present our system for the ImageCLEFmed GANs 2026 challenge, which comprises two complementary forensic tasks on a biomedical GAN dataset: Task 1 (membership inference, determining whether a real image was used to train a GAN) and Task 2 (latent vector matching, identifying the latent code $z \in \mathbb{R}^{128}$ corresponding to a given generated image). For Task 1, we evaluate three independent detectors: a fine-tuned ResNet-18 classifier (M1), a handcrafted pixel-statistics and GAN-fingerprint Random Forest model (M2), and a bagged Convolutional Autoencoder ensemble (M3). These are combined in four ensemble configurations and evaluated using 5-fold cross-validation. All Task 1 methods perform close to the chance baseline, with the best cross-validation accuracy of 0.5281 achieved by ResNet-18 alone and a competition test-set score of 0.4879 obtained by E3 (Pixel-RF + AE-Ensemble). These results highlight the inherent difficulty of black-box membership inference against a well-generalized GAN. For Task 2, we train three ResNet-50 regression variants using MSE, cosine, and combined losses, alongside two contrastive dual-encoder architectures (A4 and A5), and combine them in five ensemble configurations. The ensemble of all three ResNet models (E4) achieves a Cohen's κ score of 0.5312 ± 0.085 under cross-validation. Incorporating the dual encoder (E5) further improves performance to $\kappa = 0.6047 \pm 0.086$, corresponding to a matching accuracy of 60.49%. The E5 ensemble achieves a competition test-set matching accuracy of 0.3677, representing the best result among all our submitted runs. Our analysis shows that regression-based ensembles substantially outperform standalone contrastive encoders in small-batch settings and that combining complementary loss functions provides the largest performance gain among the evaluated design choices.

Keywords

GAN, membership inference, latent vector prediction, ResNet, autoencoder, contrastive learning, ensemble methods, ImageCLEF, GAN forensics

1. Introduction

Generative Adversarial Networks (GANs) [1] have achieved near-photorealistic image synthesis across a wide range of domains, including medical imaging, raising pressing questions around training-data privacy, intellectual property, and model reproducibility. In healthcare, synthetic data generated by GANs offers a promising route to augmenting limited datasets without exposing sensitive patient information; however, this benefit is contingent on the synthetic images not inadvertently memorising or leaking identifiable patterns from their training inputs. Two complementary forensic problems arise naturally in this context: *did a specific image contribute to training a GAN?* (membership inference), and *which latent code produced a given GAN output?* (latent inversion). The ImageCLEFmed GANs 2026 [2], organized as part of the Conference and Labs of the Evaluation Forum (CLEF 2026) [3], challenge operationalises both problems on a curated biomedical imaging dataset drawn from lung tuberculosis CT scans, providing a controlled benchmark that avoids the confounds of real-world deployment.

Tasks performed. This paper addresses:

1. **Task 1 — Detect Training Data Usage.** Binary classification of real images as *used* (included in

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

*Corresponding author.

m.shaikh.26919@khi.iba.edu.pk (M. A. Shaikh); n.shaukat.26971@khi.iba.edu.pk (N. U. N. Shaukat); h.abro.27113@khi.iba.edu.pk (H. I. Abro); z.masood.26928@khi.iba.edu.pk (Z. Masood); atiftahir@iba.edu.pk (M. A. Tahir)

GAN training) or *not_used*. This is a black-box membership inference attack [4]: no access to the generator, discriminator weights, or training loss is available.

2. **Task 2 — Match Latent Vector.** Given a generated image, identify its ground-truth $z \in \mathbb{R}^{128}$ from candidates via one-to-one bipartite matching.

This paper is organized as follows. Section 2 shows related work, followed by Datasets and Resources in Section 3. Section 4 presents our approaches for Task 1 (Membership Inference), while Section 5 details our methodologies for Task 2 (Latent Vector Matching). Section 6 reports the experimental results for both tasks. Section 7 provides an analysis and discussion of our findings. Implementation details are outlined in Section 8, and Section 9 concludes the paper.

2. Related Work

Membership inference. Shokri *et al.* [4] introduced the shadow-model attack framework for discriminative classifiers, exploiting observable differences in a model's output confidence distributions between records it was trained on and those it was not. Hayes *et al.* [5] extended membership inference to generative models, presenting the first attacks against GANs (LOGAN). Their white-box attack leverages direct access to the GAN's discriminator, exploiting the fact that the discriminator assigns systematically higher scores to images it was trained on relative to held-out data. Their black-box attack operates without access to internal model components and relies only on generated samples, making it substantially less effective in comparison. Chen *et al.* [6] subsequently presented the first taxonomy of MIAs against deep generative models, categorising attack settings by increasing attacker knowledge: full black-box generator, partial black-box, white-box, and accessible discriminator. They also propose a generic attack model instantiable across all settings. Our Task 1 operates in their full black-box regime. Zhang *et al.* [7] proposed a generalised black-box attack that requires only the distribution of generated samples and no shadow models, demonstrating vulnerability across GANs, VAEs, and diffusion models. Matsumoto *et al.* [8] showed through experiments with DDIM and DCGAN that diffusion models exhibit comparable resistance to MIAs as GANs across both white-box and black-box settings. We build on the black-box regime studied in these works and confirm experimentally that, under strict black-box conditions with no access to the discriminator or generator loss, reliable per-sample membership inference remains an open problem.

GAN fingerprinting. Marra *et al.* [9] showed that images generated by GANs contain consistent model-specific artifacts that can be exploited for attribution tasks. They demonstrate that a simple classifier trained on noise residual features extracted from generated images can reliably distinguish outputs from different generative models, analogous in spirit to sensor pattern noise in digital imaging systems. We leverage this observation in our M2 feature set, computing GAN fingerprint correlation as one of the features fed to the Random Forest classifier.

Latent space inversion. Learning an approximate inverse mapping $G^{-1} : \mathbf{x} \to z$ was studied by Creswell and Bharath [10], who introduced an iterative optimisation-based approach that directly searches in latent space to find a z vector whose generated output best reconstructs a given image. Their method is used primarily for analysing and probing the latent structure of pretrained GANs via reconstruction quality and latent recovery behaviour. This optimisation-at-test-time paradigm is impractical for large candidate pools. We instead train feed-forward regressors and contrastive encoders that produce a single-pass embedding, enabling efficient bipartite matching via the Hungarian algorithm [11].

Contrastive representation learning. Van den Oord *et al.* [12] proposed Contrastive Predictive Coding (CPC) and the associated InfoNCE loss, a probabilistic contrastive objective that trains representations by predicting future latent states from past context using noise-contrastive estimation. CPC demonstrated strong unsupervised representation quality across speech, images, text, and reinforcement learning. Chen *et al.* [13] subsequently showed in SimCLR that a symmetric NT-Xent loss treating differently augmented views of the same image as a positive pair and all other images in the batch as

negatives substantially improves representation quality for visual recognition, while also identifying that contrastive learning requires large batch sizes to supply sufficient hard negatives. We adapt the NT-Xent formulation to image-latent pairs, pairing each generated image with its known z as the positive example and all other z vectors in the batch as negatives. Our refinement in A5 addresses the large-batch requirement of NT-Xent by freezing early backbone layers, reducing gradient noise without requiring a large batch.

3. Dataset and Resources

3.1. Task 1 — Detect Training Data

The Task 1 dataset consists of 256×256 pixel, 8-bit greyscale PNG images representing axial slices of 3D CT scans from approximately 8,000 lung tuberculosis patients. Images vary in appearance from relatively normal-looking lung tissue to scans exhibiting distinct lesions and severe pathology. The development set is divided into three folders: `generated` (10,000 GAN-synthesised images used to train the AE ensemble, M3), `real_used` (3,200 real images that were used to train the GAN), and `real_not_used` (3,200 real images that were not used to train the GAN). Within these real images, we allocate 2,240 per class for cross-validation training, 320 per class for validation, and 640 per class for held-out testing, yielding a perfectly balanced dataset (ratio 1.000). The competition test set comprises 2,140 unlabelled real images of unknown membership status. No access to the GAN's generator weights, discriminator, or training loss is provided, making this a strict black-box membership inference task. Bergen *et al.* [14] conducted a closely related study on a 3-D PET medical imaging GAN, proposing a framework for evaluating the privacy-utility trade-off and demonstrating that white-box discriminator access yields near-perfect membership inference (AUC $=$ 0.99), while noting that black-box attacks are substantially harder, motivating the difficulty of our own strict black-box setting.

3.2. Task 2 — Latent Vector Dataset

The Task 2 development dataset consists of 10,000 image-z-vector pairs generated by the **BigGAN-Deep** [15] model at 256×256 resolution. Each latent vector $z \in \mathbb{R}^{128}$ is independently sampled from a standard normal distribution $\mathcal{N}(0, I_{128})$, verified empirically (global mean $= 0.0007$, global std $=$ 1.0000, per-dimension std range: $[0.979, 1.018]$). For 5-fold cross-validation, the 10,000 development pairs are split into 8,000 training and 2000 validation samples per fold. The competition test set provides an additional 10,000 GAN-generated images with their z vectors shuffled; participants must submit a permutation mapping each z to its correct image.

4. Task 1: Membership Inference — Approach

4.1. Overview and Design Rationale

Task 1 is a black-box binary classification problem. Because no query access to the GAN's internal states is available, we cannot rely on shadow-model or loss-based attacks. We instead hypothesise that three complementary information sources may capture residual membership signals: (1) deep visual features learned by a pre-trained CNN classifier; (2) handcrafted low-level statistical and frequency-domain features correlated with GAN noise patterns; and (3) reconstruction error from autoencoders trained exclusively on GAN outputs. These three information streams are captured by methods M1, M2, and M3 respectively, and combined in four probability-averaging ensemble configurations (E1–E4).

4.2. M1: Fine-Tuned ResNet-18 Classifier

We fine-tune a ResNet-18 backbone [16] pre-trained on ImageNet with a custom classification head: $\mathrm{FC}(512 \to 256) \to \mathrm{ReLU} \to \mathrm{Dropout}(0.4) \to \mathrm{FC}(256 \to 2)$. Images are resized to 224×224 pixels

and normalised using ImageNet statistics (mean $[0.485, 0.456, 0.406]$, std $[0.229, 0.224, 0.225]$). Since the input images are greyscale, each image is replicated across three channels before the transform is applied.

Training augmentations include random horizontal flip, random vertical flip ($p = 0.3$), random affine transformation (rotation ± 15, translation 10%, scale 0.85–1.15), random colour jitter (brightness and contrast ± 0.3, $p = 0.5$), and random Gaussian blur (kernel size 5, $p = 0.2$). The model is trained for 25 epochs per fold with batch size 32 using AdamW ($\mathrm{lr} = 3 \times 10^{-4}$, weight decay 10^{-4}) and a One-Cycle learning rate scheduler (`pct_start`=0.1). Mixed-precision training (AMP) is used throughout. The best-validation-accuracy checkpoint is retained and its softmax output probabilities are used as M1's score.

4.3. M2: Pixel-Statistics + GAN Fingerprint → Random Forest

M2 extracts a rich hand-crafted feature vector from each image and passes it through a 300-tree Random Forest [17] trained on the balanced real-image set. Features are computed at 128×128 resolution and comprise the following groups:

1. **Intensity histogram** (32 bins, normalised density).
2. **Global statistics**: mean, std, skewness, kurtosis, 25th and 75th percentiles (6 values).
3. **Local Binary Patterns (LBP)**: radius 1, 8 neighbours, uniform method; 10-bin histogram (10 values).
4. **GLCM Haralick features**: contrast, correlation, homogeneity, energy, and dissimilarity computed at distance 1 and angles $\{0, \pi/4, \pi/2\}$; mean and std per property (10 values).
5. **Gradient statistics**: Sobel gradient magnitude mean, std, skewness, and kurtosis; Laplacian mean, std, and variance (7 values).
6. **DCT energy**: ratio of low-frequency ($\leq 16 \times 16$) and high-frequency ($\geq 64 \times 64$) energy to total DCT energy (2 values).
7. **FFT radial energy**: mean spectrum energy in four radial bands (0–25%, 25–50%, 50–75%, 75–100% of the Nyquist radius, 4 values).
8. **GAN fingerprint correlation:** The Pearson correlation of the image noise residual (the image minus its Gaussian-blurred version, using $\sigma = 0$ and a 3×3 kernel) with a pre-computed GAN fingerprint template (yielding a single correlation coefficient).

The GAN fingerprint template is built once before cross-validation by averaging the high-frequency noise residuals of 5,000 randomly sampled generated images (shape: 128×128; mean $= -0.0186$, std $= 0.9124$). The RF pipeline wraps a `StandardScaler` and the classifier in a `sklearn.pipeline.Pipeline`. The predicted class probabilities of the RF serve as M2's score.

4.4. M3: Bagged Autoencoder Ensemble

M3 trains an ensemble of $K = 5$ Convolutional Autoencoders (CAEs), each on a random 70% subset of the 7,000 training GAN-generated images (after splitting), for 30 epochs per fold (batch size 64, $\mathrm{lr} = 10^{-3}$, Adam). Images are resized to 64×64 pixels. Each CAE has a latent dimension of 256 and uses convolutional encoder-decoder blocks with MSE reconstruction loss. The membership score for an image x is derived from the per-ensemble reconstruction error as follows:

$$s(x) = \bar{e}(x) - 0.5\,\hat{\sigma}(x), \tag{1}$$

where $\bar{e}(x)$ and $\hat{\sigma}(x)$ are the mean and standard deviation of the MSE reconstruction errors across the K autoencoders. A logistic transformation maps this score to a probability:

$$p_{\text{used}} = \mathrm{sigmoid}\left(-\frac{s(x) - \tau}{\sigma_s}\right),$$

where σ_s is the standard deviation of the scores, and τ is a threshold calibrated by searching across score percentiles to maximise classification accuracy on a validation set. The hypothesis underlying M3 is that images containing GAN-specific textures will be reconstructed with lower error than out-of-distribution real images, providing a weak but complementary membership signal.

4.5. Ensemble Configurations

The four ensemble variants combine M1, M2, and M3 by averaging their class probability vectors and taking the $\arg\max$ of the result. Table 1 lists all configurations.

Table 1
Task 1 ensemble configurations.

ID	Components	Description
E1	M1 + M2	ResNet-18 + Pixel-RF
E2	M1 + M3	ResNet-18 + AE-Ensemble
E3	M2 + M3	Pixel-RF + AE-Ensemble
E4	M1 + M2 + M3	All three detectors

5. Task 2: Latent Vector Matching — Approach

5.1. Overview and Design Rationale

Task 2 requires solving a one-to-one bipartite matching problem between N generated images and N latent vectors. Our strategy is to train models that embed images and latent vectors into a common space, compute a pairwise cosine similarity matrix $S \in \mathbb{R}^{N \times N}$, and apply the Hungarian algorithm [11] to find the minimum-cost perfect matching (equivalently, maximum-similarity assignment).

We explore two complementary paradigms: (1) *regression-based* approaches (A3 variants), which directly minimise the distance between a predicted and ground-truth z vector; and (2) *contrastive* approaches (A4 and A5), which learn aligned image and z embeddings using InfoNCE-style objectives. We hypothesise that regression and contrastive approaches capture different aspects of the image-to-latent mapping, motivating ensemble combination.

All models share the same data preprocessing: images loaded as RGB (greyscale replicated across channels), resized to 224×224 for regression models (A3) or 224×224 with an initial 256×256 resize and random crop for contrastive models (A4), and normalised with ImageNet statistics.

5.2. A3 Variants: ResNet-50 Regressors

Shared architecture. A ResNet-50 backbone [16] (IMAGENET1K_V2 weights) with the final classification layer replaced by a regression head: $\mathrm{FC}(2048 \to 512) \to \mathrm{BatchNorm} \to \mathrm{ReLU} \to \mathrm{Dropout}(0.3) \to \mathrm{FC}(512 \to 128)$. All models are trained with AdamW ($\mathrm{lr} = 3 \times 10^{-4}$, weight decay 10^{-4}) and gradient clipping at norm 1.0 using AMP. Each model runs for 20 epochs per fold with batch size 64. Training augmentations include random horizontal flip ($p = 0.5$) and colour jitter (brightness and contrast ± 0.2, saturation ± 0.1).

A3-MSE. Minimises $\mathcal{L}_{\mathrm{MSE}} = \|\hat{z} - z\|_2^2$, encouraging exact Euclidean proximity.

A3-COS. Minimises the cosine distance:

$$\mathcal{L}_{\mathrm{COS}} = 1 - \frac{\hat{z} \cdot z}{\|\hat{z}\| \, \|z\|}, \tag{2}$$

aligning predicted directions rather than magnitudes — better suited to cosine-based retrieval evaluation.

A3-COMBO. Convex combination balancing both objectives:

$$\mathcal{L}_{\mathrm{COMBO}} = 0.5 \, \mathcal{L}_{\mathrm{MSE}} + 0.5 \, \mathcal{L}_{\mathrm{COS}}. \tag{3}$$

At inference time, each regression model predicts $\hat{z}$ for every image. The pairwise similarity matrix between all predicted embeddings and all ground-truth z vectors is computed; the Hungarian algorithm then finds the globally optimal one-to-one assignment.

5.3. A4: Dual-Encoder Contrastive Model

A4 trains a dual-encoder architecture with two branches: an *image encoder* and a z *encoder*. The image encoder uses a ResNet-50 backbone with a projection head $\mathrm{FC}(2048 \to 512) \to \mathrm{BN} \to \mathrm{ReLU} \to \mathrm{FC}(512 \to 128)$ followed by L2 normalisation. The z encoder uses a three-layer MLP: $\mathrm{Linear}(128 \to 256) \to \mathrm{BN} \to \mathrm{GELU} \to \mathrm{Linear}(256 \to 256) \to \mathrm{BN} \to \mathrm{GELU} \to \mathrm{Linear}(256 \to 128)$ with L2 normalisation. Temperature is a learnable parameter initialised at 0.07 and clamped to $[0.01, 1.0]$.

Training uses the symmetric NT-Xent (InfoNCE) loss [12, 13]:

$$\mathcal{L}_{\text{NT-Xent}} = \frac{1}{2B} \sum_{i=1}^{B} \Big[\ell(\mathbf{v}_i, \mathbf{u}_i;\, \{\mathbf{u}_j\}_{j \neq i}) + \ell(\mathbf{u}_i, \mathbf{v}_i;\, \{\mathbf{v}_j\}_{j \neq i}) \Big], \tag{4}$$

where $\mathbf{v}_i$ is the image embedding, $\mathbf{u}_i$ is the z embedding, $B = 256$ is the batch size, and $\ell(\cdot)$ is the cross-entropy loss over in-batch negatives. The model is trained for 40 epochs per fold with AdamW ($\mathrm{lr} = 10^{-4}$, weight decay 10^{-4}). At inference, the cosine similarity matrix $S_{ij} = \mathbf{v}_i \cdot \mathbf{u}_j$ is computed for all image-z pairs and the Hungarian algorithm finds the optimal assignment.

5.4. A5: Refined Contrastive Dual-Encoder with Partial Freeze

A5 is an independently trained variant of the dual-encoder architecture in which the ResNet-50 backbone is partially frozen to stabilize training. Layers `conv1`, `bn1`, `relu`, `maxpool`, `layer1`, `layer2`, and `layer3` remain frozen, while `layer4` and `avgpool` are fine-tuned. The projection head is $\mathrm{Linear}(2048 \to 256) \to \mathrm{BN} \to \mathrm{ReLU} \to \mathrm{Linear}(256 \to 128) \to$ L2-norm. The z encoder is $\mathrm{Linear}(128 \to 256) \to \mathrm{BN} \to \mathrm{ReLU} \to \mathrm{Linear}(256 \to 256) \to \mathrm{BN} \to \mathrm{ReLU} \to \mathrm{Linear}(256 \to 128) \to$ L2-norm.

Training employs symmetric InfoNCE loss with fixed temperature $\tau = 0.2$. We use AdamW optimization with differential learning rates: $\mathrm{lr}_{\text{backbone}} = 3 \times 10^{-5}$ and $\mathrm{lr}_{\text{head}} = 3 \times 10^{-4}$ (weight decay 10^{-4}), paired with cosine annealing ($\eta_{\min} = 10^{-7}$), batch size 64, and gradient clipping (norm 1.0). Input images are resized to 128×128 and augmented with random horizontal flip ($p = 0.5$), color jitter (brightness and contrast ± 0.15), and Gaussian blur ($p = 0.2$).

To assess the effect of training duration, we evaluate A5 under two epoch budgets: 60 epochs per fold and 100 epochs per fold during 5-fold cross-validation.

Test image and latent-vector embeddings are matched using the Hungarian algorithm on their cosine similarity matrix.

5.5. Ensemble Configurations

Task 2 ensembles combine the z-space embeddings (or predicted vectors) of their component models by averaging the similarity matrices prior to Hungarian assignment. Table 2 lists all configurations.

Table 2
Task 2 ensemble configurations.

ID	Components	Description
E1	A4 + A3-MSE	Contrastive + MSE regressor
E2	A4 + A3-COS	Contrastive + Cosine regressor
E3	A4 + A3-COMBO	Contrastive + Combined regressor
E4	A3-MSE + A3-COS + A3-COMBO	All three ResNets (no A4)
E5	E4 + A4	All three ResNets + A4

6. Results

6.1. Task 1: 5-Fold Cross-Validation Analysis from Training Data

Table 3 reports the 5-fold cross-validation results for all Task 1 methods. All individual methods and ensembles achieve accuracies near the chance baseline of 0.50, with small but positive Cohen's κ values indicating a marginal above-chance signal. The best individual cross-validated accuracy is achieved by ResNet-18 (M1: 0.5281 ± 0.0120), while the AE ensemble (M3: 0.5127 ± 0.0095) shows the lowest variance, suggesting it captures a more stable but weaker signal. The pixel-RF (M2) fails to exceed chance on average ($\kappa = -0.0103$).

Ensembling degrades rather than improves accuracy in every configuration: E1 (0.5145) through E4 (0.5016) all underperform the best individual method. This indicates that the three detectors introduce noise when combined, likely because M2 and M3 are below chance in some folds.

Table 3
Task 1 5-fold cross-validation results (mean $\pm$ std across folds).

Method	Accuracy	Cohen's κ	AUC-ROC	F1-macro
Individual Methods				
M1: ResNet-18	0.5281 ± 0.0120	0.0563 ± 0.0240	0.5260 ± 0.0186	0.5133 ± 0.0220
M2: Pixel-RF	0.4949 ± 0.0141	-0.0103 ± 0.0283	0.5033 ± 0.0185	0.4947 ± 0.0141
M3: AE-Ensemble	0.5127 ± 0.0095	0.0254 ± 0.0191	0.4919 ± 0.0082	0.4800 ± 0.0322
Ensemble Methods				
E1: M1 + M2	0.5145 ± 0.0140	0.0290 ± 0.0281	0.5162 ± 0.0233	0.5110 ± 0.0162
E2: M1 + M3	0.5062 ± 0.0131	0.0125 ± 0.0262	0.4990 ± 0.0133	0.4690 ± 0.0341
E3: M2 + M3	0.5029 ± 0.0092	0.0058 ± 0.0183	0.4930 ± 0.0108	0.4713 ± 0.0315
E4: M1 + M2 + M3	0.5016 ± 0.0116	0.0031 ± 0.0231	0.4997 ± 0.0151	0.4661 ± 0.0412

6.2. Task 1: Official Competition Submissions

The official Task 1 submission method was selected on a separate held-out validation split, which was used as an additional estimate of out-of-sample generalization beyond cross-validation. The run submitted to the competition leaderboard was E3 (Pixel-RF + AE-Ensemble). Table 4 details the official scores obtained by this submission on the 2,140-image real_unknown test set.

Table 4
Task 1 official submission scores for E3 on the real_unknown test set. (RunID: 1170)

Metric	Score
Accuracy	0.4879
Balanced Accuracy	0.4879
Sensitivity	0.1037
Specificity	0.8720
Precision	0.4476
F1 Score	0.1684
Cohen's κ	-0.0243
MCC	-0.0380

6.3. Task 2: 5-Fold Cross-Validation Analysis from Training Data

Table 5 reports the 5-fold cross-validation results for all Task 2 methods. Among the regression models, the A3-COMBO regressor is the strongest individual model ($\kappa = 0.3619 \pm 0.1764$, 36.22% accuracy), outperforming A3-MSE (25.62%) and A3-COS (17.39%). The standalone A4 contrastive encoder initially

underperforms all regression variants ($2.12 \pm 0.64\%$ accuracy, $\kappa = 0.0207$). This poor A4 performance is attributed to the small effective batch size relative to NT-Xent's requirement for many negative samples.

To address this, the A5 refined dual-encoder was evaluated with extended training budgets. A5-V2 (100 epochs) achieves $50.50 \pm 1.71\%$ accuracy and $\kappa = 0.5048 \pm 0.0172$, substantially outperforming A5-V1 (60 epochs: $38.24 \pm 1.20\%$). The per-fold best epoch for V2 ranged from 90–100 (mean 96), with the best fold reaching 52.65% and the worst fold at 48.15% (a spread of 4.5 percentage points). The InfoNCE loss at epoch 100 (V2) reached 0.6698, versus 0.8896 for V1 at epoch 57, indicating substantially more training progress with the extended budget.

Ultimately, ensembling consistently improves over individual models. E4 (combining all three A3 ResNet variants, excluding A4) achieves $\kappa = 0.5312 \pm 0.0850$ and 53.14% accuracy. E5 (all three ResNets plus the baseline A4) further improves to $\kappa = 0.6047 \pm 0.0861$ and 60.49% accuracy, confirming that the contrastive model contributes a complementary signal to the regression ensemble even when weak in isolation.

Table 5
Task 2 5-fold cross-validation results (mean ± std) for individual models and ensembles.

Method	Cohen's κ	Accuracy (%)	Matched CosSim	Matched MSE
A3: Regression				
A3-MSE	0.2558 ± 0.1323	25.62 ± 13.23	0.2939 ± 0.1284	1.4088 ± 0.2581
A3-COS	0.1735 ± 0.0547	17.39 ± 5.46	0.2148 ± 0.0553	1.5689 ± 0.1108
A3-COMBO	0.3619 ± 0.1764	36.22 ± 17.64	0.3953 ± 0.1716	1.2082 ± 0.3451
Contrastive Dual-Encoders (A4 & A5)				
A4	0.0207 ± 0.0064	2.12 ± 0.64	0.0432 ± 0.0091	1.9147 ± 0.0201
A5-V1 (60 ep)	0.3821 ± 0.0120	38.24 ± 1.20	0.3947 ± 0.0109	1.2104 ± 0.0231
A5-V2 (100 ep)	0.5048 ± 0.0172	50.50 ± 1.71	0.5128 ± 0.0179	0.9763 ± 0.0361
Ensembles				
E1: A4 + A3-MSE	0.2798 ± 0.1130	28.02 ± 11.30	0.3146 ± 0.1111	1.3691 ± 0.2224
E2: A4 + A3-COS	0.2178 ± 0.0449	21.82 ± 4.49	0.2534 ± 0.0447	1.4929 ± 0.0880
E3: A4 + A3-COMBO	0.3722 ± 0.1560	37.25 ± 15.59	0.4026 ± 0.1512	1.1935 ± 0.3047
E4: A3-MSE + COS + COMBO	0.5312 ± 0.0850	53.14 ± 8.49	0.5582 ± 0.0798	0.8827 ± 0.1607
E5: E4 + A4	$\mathbf{0.6047 \pm 0.0861}$	$\mathbf{60.49 \pm 8.60}$	$\mathbf{0.6261 \pm 0.0818}$	$\mathbf{0.7468 \pm 0.1652}$

6.4. Task 2: Official Competition Submissions

Three runs were submitted for Task 2, whose official scores are reported in Table 6. The E5 ensemble (all three ResNets plus A4 dual-encoder) achieves the best competition result with 3,677 correct matches out of 10,000 (matching accuracy 0.3677). A5-V2 achieves 2,135 correct matches (0.2135), while A5-V1, underperforms on the competition test set with only 1,709 correct matches (0.1709), suggesting overfitting to the cross-validation splits or a distribution gap in the competition test set.

Table 6
Task 2 official competition test-set submission results.

Run ID	Submitted Run	Correct Matches	Wrong Matches	Matching Accuracy
1169	E5 (All-3-ResNets + A4)	3,677	6,323	**0.3677**
1179	A5-V2 (100 ep)	2,135	7,865	0.2135
1178	A5-V1 (60 ep)	1,709	8,291	0.1709

7. Analysis and Discussion

Task 1: Fundamental hardness of black-box membership inference. All Task 1 approaches achieve accuracies that hover within ± 3 percentage points of the chance baseline of 0.50, with Cohen's κ values near zero. This is consistent with the theoretical expectation that well-generalised GANs leave no reliable signature in their real training inputs that can be distinguished from non-training images without access to the generator's internal states [5]. The disconnect between cross-validation performance (ResNet-18 best at 0.5281) and competition test-set performance (E3: 0.4879) indicates that small differences between methods are within the noise floor, reflecting the difficulty of generalising across dataset splits when all methods operate near chance.

The highly imbalanced specificity/sensitivity of the official submission (specificity 0.872 vs. sensitivity 0.1037) reveals that the ensemble is biased toward predicting "not used", consistent with a model that has not learned a genuine membership signal and defaults to the majority or lower-confidence class.

Task 2: Regression ensembles vs. contrastive encoders. The most striking finding for Task 2 is the large performance gap between regression and contrastive approaches in isolation. The A4 contrastive dual-encoder achieves only 2.12% accuracy in cross-validation. We attribute this to the difficulty of learning a meaningful latent-space correspondence with 40 training epochs when the InfoNCE objective requires a large number of effective negatives to distinguish each (x_i, z_i) pair from all other in-batch pairs. The A5 refinement addresses this by freezing seven of eight residual layers, concentrating gradient updates on the final convolutional stage and projection head; this produces substantially better stability (38.24% at 60 epochs, 50.50% at 100 epochs with variance ± 1.71 versus A4's ± 0.64).

Ensemble E4 (all three regression objectives, no contrastive) achieves a dramatic uplift to 53.14%, demonstrating that the three loss functions capture complementary aspects of the latent space geometry. A3-COS aligns directional similarity, A3-MSE aligns magnitude, and A3-COMBO balances both; their predicted $\hat{z}$ embeddings disagree on different examples, and averaging the similarity matrices before Hungarian matching reduces the effective error rate. E5 adds A4 to this ensemble, yielding a further improvement to 60.49% ($\kappa = 0.6047$), confirming the weak but complementary signal of the contrastive encoder even when individually underperforming.

Cross-validation vs. competition test set. A notable discrepancy exists between cross-validation and competition performance: A5-V2 achieves 50.50% in CV but only 21.35% on the competition test set, while E5 achieves 60.49% CV and 36.77% competition accuracy. This generalisation gap likely stems from (i) the Hungarian algorithm solving a 10000×10000 assignment problem on the full test set, where embedding errors compound over a much larger candidate pool than the 2000-sample validation folds; and (ii) potential domain shift in the competition test set's z distribution or image characteristics. The regression ensemble (E5) degrades more gracefully than the contrastive-only A5-V2, as its multi-objective training produces more robust z embeddings.

8. Implementation Details

Reproducibility. All experiments fix global random seeds (`SEED=42`) across `random`, `numpy`, and `torch` libraries with `deterministic=True` and `benchmark=False` in cuDNN. Stratified k-fold (Task 1) and standard k-fold (Task 2) use the same random state across all methods within a sweep.

Task 1 hyperparameters. ResNet-18: image size 224×224, batch 32, 25 epochs per fold, AdamW $\texttt{lr} = 3 \times 10^{-4}$, weight decay 10^{-4}, One-Cycle LR (`pct_start`$= 0.1$). AE Ensemble: image size 64×64, batch 64, 30 epochs, Adam $\texttt{lr} = 10^{-3}$, $K = 5$ autoencoders, each trained on 70% of generated images. Random Forest: 300 trees, `StandardScaler` in pipeline.

Task 2 hyperparameters (A3/A4). Image size 224×224. A3: batch 64, 20 epochs/fold, AdamW $\texttt{lr} = 3 \times 10^{-4}$, weight decay 10^{-4}, AMP, gradient clip 1.0; final full-data training for 20 epochs. A4: batch 256, 40 epochs/fold, AdamW $\texttt{lr} = 10^{-4}$, weight decay 10^{-4}, learnable temperature initialised at 0.07.

Task 2 hyperparameters (A5). Image size 128×128. Backbone LR 3×10^{-5}, head LR 3×10^{-4}, weight decay 10^{-4}, CosineAnnealingLR ($T_{\max} = 100$, $\eta_{\min} = 10^{-7}$), batch 64, gradient clip 1.0. V1: 60 epochs/fold (mean best epoch 52). V2: 100 epochs/fold (mean best epoch 96), early stopping patience 10 (checked every 5 epochs).
Hungarian matching. Applied at inference time using `scipy.optimize._sum_assignment` on the negative cosine similarity matrix.

9. Conclusions

In this paper, we presented our ensemble approaches for the ImageCLEFmed GANs 2026 challenge, tackling two critical forensic problems in biomedical generative modelling: membership inference and latent vector prediction.

For Task 1, our comprehensive evaluation of deep classifiers, handcrafted statistical features, and autoencoder ensembles demonstrated the fundamental hardness of strict black-box membership inference. All evaluated models, including our submitted E3 ensemble, performed near the random-chance baseline. This empirical result suggests that a well-generalised GAN does not leak easily detectable membership fingerprints in the absence of generator or discriminator access, thereby offering a robust degree of privacy protection for sensitive medical training data.

Conversely, for Task 2, we showed that the latent space of a GAN can be effectively navigated and partially inverted using multi-objective embedding strategies. By framing latent vector matching as a bipartite assignment problem, we found that combining distinct regression objectives (MSE and Cosine similarity) with contrastive representation learning yields highly complementary signals. Our combined ensemble (E5) successfully mitigated the small-batch limitations of standalone contrastive encoders, achieving the best competition test-set matching accuracy of 0.3677 across $10,000$ candidates.

Acknowledgments

Experiments were conducted on Kaggle Notebooks with NVIDIA T4 GPU access. The authors thank the ImageCLEFmed GANs 2026 organisers for providing the benchmark and evaluation infrastructure.

Declaration on Generative AI

During the preparation of this work, the authors employed generative AI tools to assist with paragraphing, improving grammatical flow, and creating tables. All AI-assisted content was thoroughly verified and checked by the authors, who assume full responsibility for the integrity and accuracy of the final manuscript.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

[2] 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.

[3] 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.

[4] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE symposium on security and privacy (SP), IEEE, 2017, pp. 3–18.

[5] J. Hayes, L. Melis, G. Danezis, E. De Cristofaro, Logan: Membership inference attacks against generative models, arXiv preprint arXiv:1705.07663 (2017).

[6] D. Chen, N. Yu, Y. Zhang, M. Fritz, Gan-leaks: A taxonomy of membership inference attacks against generative models, in: Proceedings of the 2020 ACM SIGSAC conference on computer and communications security, 2020, pp. 343–362.

[7] M. Zhang, N. Yu, R. Wen, M. Backes, Y. Zhang, Generated distributions are all you need for membership inference attacks against generative models, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 4839–4849.

[8] T. Matsumoto, T. Miura, N. Yanai, Membership inference attacks against diffusion models, in: 2023 IEEE Security and Privacy Workshops (SPW), IEEE, 2023, pp. 77–83.

[9] F. Marra, D. Gragnaniello, D. Cozzolino, L. Verdoliva, Detection of gan-generated fake images over social networks, in: 2018 IEEE conference on multimedia information processing and retrieval (MIPR), IEEE, 2018, pp. 384–389.

[10] A. Creswell, A. A. Bharath, Inverting the generator of a generative adversarial network, IEEE transactions on neural networks and learning systems 30 (2018) 1967–1974.

[11] H. W. Kuhn, The hungarian method for the assignment problem, Naval research logistics quarterly 2 (1955) 83–97.

[12] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[13] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PmLR, 2020, pp. 1597–1607.

[14] R. V. Bergen, J.-F. Rajotte, F. Yousefirizi, A. Rahmim, R. T. Ng, Assessing privacy leakage in synthetic 3-d pet imaging using transversal gan, Computer Methods and Programs in Biomedicine 243 (2024) 107910.

[15] A. Brock, J. Donahue, K. Simonyan, Large scale gan training for high fidelity natural image synthesis, arXiv preprint arXiv:1809.11096 (2018).

[16] 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.

[17] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.