Controlling the Quality of Synthetic Medical Images Created via GANs

Notebook for the ImageCLEF Lab at CLEF 2025

Farhaan Areeb[1,†], Divyansh Vashist[1,†] and Lekshmi Kalinathan[1,*,†]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus

Abstract

The detection of training data usage in generative models is critical for preserving privacy in synthetic medical imaging. This study evaluates a set of baseline and attention-enhanced convolutional architectures to identify whether real medical images were used to train a Generative Adversarial Network (GAN). Using a Siamese network framework, we evaluated four standard backbones such as ResNet50, EfficientNetB2, DenseNet161, and Vision Transformer (ViT), along with enhanced variants that incorporate cross-attention modules. Our approach leverages an adaptive similarity metric that combines absolute difference and dot product signals to improve sensitivity to subtle "fingerprints" left by GANs. Experiments on a lung CT dataset demonstrate that cross-attention significantly improves detection performance, with ResNet50 + Cross-Attention achieving the best balance between accuracy and generalizability. These results highlight the potential of attention-guided deep networks in the forensic analysis of synthetic medical imagery and contribute to the broader effort to ensure the ethical deployment of AI in healthcare.

Keywords

Synthetic Medical Imaging, Generative Adversarial Networks (GANs), Siamese Network, Cross-Attention, Training Data Detection, Forensic Deep Learning

1. Introduction

Our team, SCOPE VIT Visioneers, participated in the ImageCLEFmedical 2025 Challenge, focusing on Subtask 1: "Detect Training Data Usage." This task aims to determine whether a specific real medical image was used to train a GAN that generated synthetic images. The challenge addresses growing concerns around data privacy and the traceability of training data in medical AI. Our work explores methods to detect such traces by analyzing the relationship between real and generated images.

Generative Adversarial Networks (GANs) have emerged as a powerful class of deep learning models capable of synthesizing highly realistic images. Their architecture, consisting of a generator that creates images and a discriminator that evaluates them, has enabled significant progress in generating synthetic content across various domains. In healthcare, GANs are particularly valuable for generating synthetic medical images to augment datasets, reduce annotation costs, and support privacy-preserving data sharing. These capabilities make GANs an attractive tool for training diagnostic models without exposing sensitive patient data. However, as synthetic images become increasingly photorealistic, concerns have grown about their misuse and the potential leakage of real training data into generated outputs.Detecting whether real medical images were used during the training of a GAN is a complex but crucial task. Modern GANs can embed subtle artifacts or statistical patterns, often referred to as 'fingerprints', that link synthetic outputs to their training data. These fingerprints may not be visible to the human eye, but can be identified through machine learning techniques that analyze image textures, frequency patterns, or deep feature representations. This challenge becomes even more critical in the medical domain, where synthetic images must maintain clinical validity while ensuring that patient

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

*Corresponding author.

†These authors contributed equally.

farhaan.areeb2021@vitstudent.ac.in (F. Areeb); divyansh.vashist2021@vitstudent.ac.in (D. Vashist); lekshmi.k@vit.ac.in (L. Kalinathan)

0000-0002-7005-742X (L. Kalinathan)

identities and attributes are not inadvertently exposed. Thus, robust forensic tools are necessary to audit synthetic datasets and guarantee the ethical use of AI-generated medical content.

To address this problem, we propose a framework based on a Siamese Neural Network (SNN) to detect training data fingerprints in synthetic medical images. It consists of twin subnetworks with shared weights that independently process a pair of images (one real and one synthetic) and extract high-dimensional feature representations. These subnetworks utilize convolutional backbones, such as ResNet50 or DenseNet161, without their final classification layers. To enhance sensitivity to subtle generative artifacts, the model incorporates a cross-attention mechanism that dynamically aligns features of real and synthetic images by computing context-aware interactions. The attended feature representations are passed to an adaptive similarity module that combines absolute differences and dot products to capture both dissimilarity and correlation. These signals are then processed through a multilayer perceptron to produce a similarity score indicating the likelihood that a real image was used during GAN training. This modular and interpretable design ensures the robustness and adaptability of the framework, even in medical imaging scenarios with limited data.

2. Background Study

The rapid evolution of Generative Adversarial Networks (GANs) has led to substantial research on synthetic image generation and forensic detection. Marra et al. [1] first highlighted the presence of unique GAN 'fingerprints', similar to PRNU patterns in camera images. This foundational idea was expanded by Yu et al. [2], who demonstrated fine-grained attribution of synthetic images to specific GAN models, revealing the forensic potential of deep-generative models. Liu et al. [3] further contributed by analyzing frequency domain noise for real-versus-synthetic image detection, while Yang et al. [4] developed GFD-Net to disentangle content-independent GAN fingerprints across diverse datasets.

To accommodate the rapidly evolving landscape of generative models, Marra et al. [5] proposed an incremental learning strategy that maintains detection performance in old and new GANs. Meanwhile, efforts to secure synthetic medical images led to innovative detection pipelines. Asakawa et al. [6] employed multistage transfer learning across domains, and Ghazi et al. [7] used texture analysis to reveal whether GAN-generated medical images leak training data characteristics. Subburam et al. [8] incorporated CNN and image hashing techniques to detect synthetic medical content.

Few-shot and hybrid learning methods have shown promise in detection with limited real data. Bharathi et al. [9] used relational networks and clustering to distinguish real biomedical images from GAN-generated images, while Andrei et al. [10] provided a comprehensive review of detection methodologies from ImageCLEFmedical GANs 2023 task. Chai et al. [11] analyzed properties that consistently make fake images detectable, and Jeong et al. [12] introduced FingerprintNet - a model that leverages frequency-based signatures to detect images from previously unseen GANs.

Beyond detection, research has critically evaluated the representational fidelity and utility of GANs in medical contexts. DuMont Schütte et al. [13] examined synthetic data as a privacy-preserving alternative to patient data, providing benchmarks for quality assessment. Kelkar et al. [14] questioned the ability of GANs to capture canonical medical image statistics, while Skandarani et al. [15] conducted empirical evaluations across imaging modalities, revealing challenges in generating clinically viable synthetic images. Collectively, these studies highlight the importance of balancing innovation in image generation with robust forensic safeguards.

Recent developments in benchmarking synthetic image detection have been strongly guided by the ImageCLEFmedical challenges. The 2025 edition of the ImageCLEFmedical GANs task [16] has extended its focus toward evaluating how well detection methods generalize to unseen generative models and whether training data can be reverse-identified from generated outputs. The broader ImageCLEF 2025 overview paper [17] outlines the scope of the medical challenges, emphasizing reproducibility, cross-domain learning, and the ethical implications of synthetic medical imaging. These benchmarks have provided a structured environment for comparative evaluations and for tracking progress in forensic detection methods under real-world constraints.

3. Methodology

3.1. Dataset Description And Pre-Processing

The dataset comprises axial CT scan slices (256×256 pixels) of patients with lung tuberculosis, split into real and synthetic categories. Real images are grouped into two classes: 100 'real used' images involved in the GAN training and 100 'real not used' images excluded from it. The synthetic portion includes 5,000 training images and 2,000 testing images(as represented in Figure 1), generated using a consistent GAN architecture that mimics the visual features of real CT scans while preserving patient anonymity. This class imbalance of real images being vastly outnumbered by synthetic ones necessitated strategic sampling to ensure fair model training.

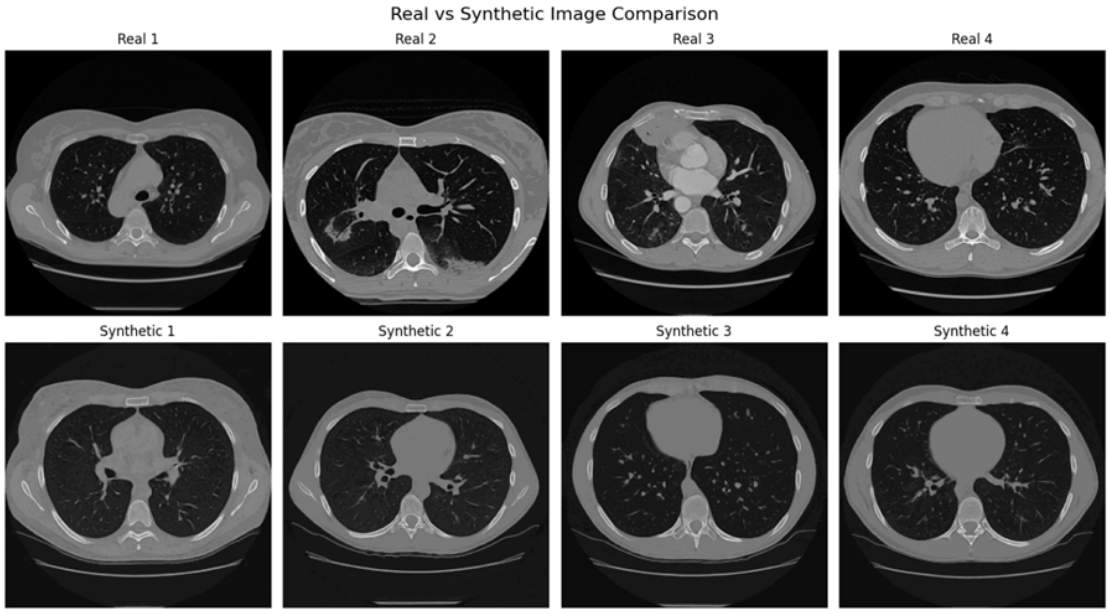

Figure 1: Real vs Generated Image Visualization.

An 80-20 data split was applied to both real and synthetic samples. Real images contributed 80 samples per class to training and 20 to testing; Synthetic images were split into 4,000 training images and 1,000 test images. To avoid distortion of subtle generative 'fingerprints', no data-augmentation techniques (e.g., flipping, rotation, noise) were applied. All images were resized to 224×224 pixels and normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) for compatibility with pretrained deep learning models. During training, a weighted random sampling strategy was used, assigning sampling weights inversely proportional to class frequencies to ensure balanced exposure to both 'used' and 'not used' images. The preprocessing pipeline, which includes resizing, normalization, and class-aware sampling, is visually summarized in Figure 2.

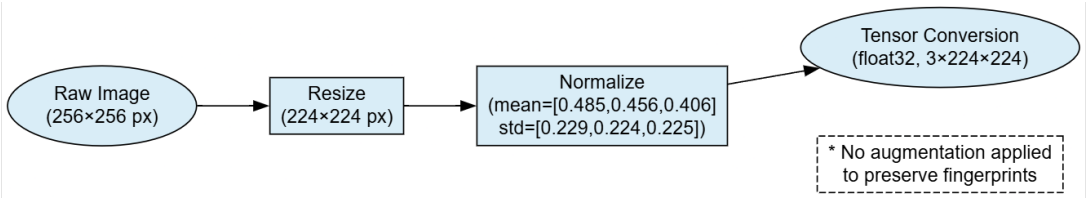

Figure 2: Data preprocessing pipeline.

3.2. Proposed Methodology

The proposed framework employs a Siamese Neural Network (SNN) to detect training data fingerprints in synthetic medical images. It consists of twin subnetworks with shared weights that independently process a pair of images; one real and one synthetic, and extract high-dimensional feature representations. These subnetworks utilize convolutional backbones, such as ResNet50 or DenseNet161, without their final classification layers. The extracted feature vectors (2048-D) encode spatial and semantic patterns critical for distinguishing real-used samples from real-not-used ones.

To enhance sensitivity to subtle generative artifacts, the model incorporates a cross-attention mechanism that dynamically aligns features of real and synthetic images. This is achieved through attention blocks that compute context-aware interactions between the paired feature maps, highlighting regions where potential fingerprints are most likely to persist. The attended feature representations are then passed to an adaptive similarity module that captures both dissimilarity (via absolute difference) and similarity (via dot product) between the inputs. These signals are concatenated and processed through a multilayer perceptron (MLP) to yield a similarity score between 0 and 1, indicating the likelihood that a real image was used during GAN training. The architecture's modular design ensures interpretability and facilitates robust generalization across diverse image types, even in low-data settings. A visual overview of the architecture is presented in Figure 3.

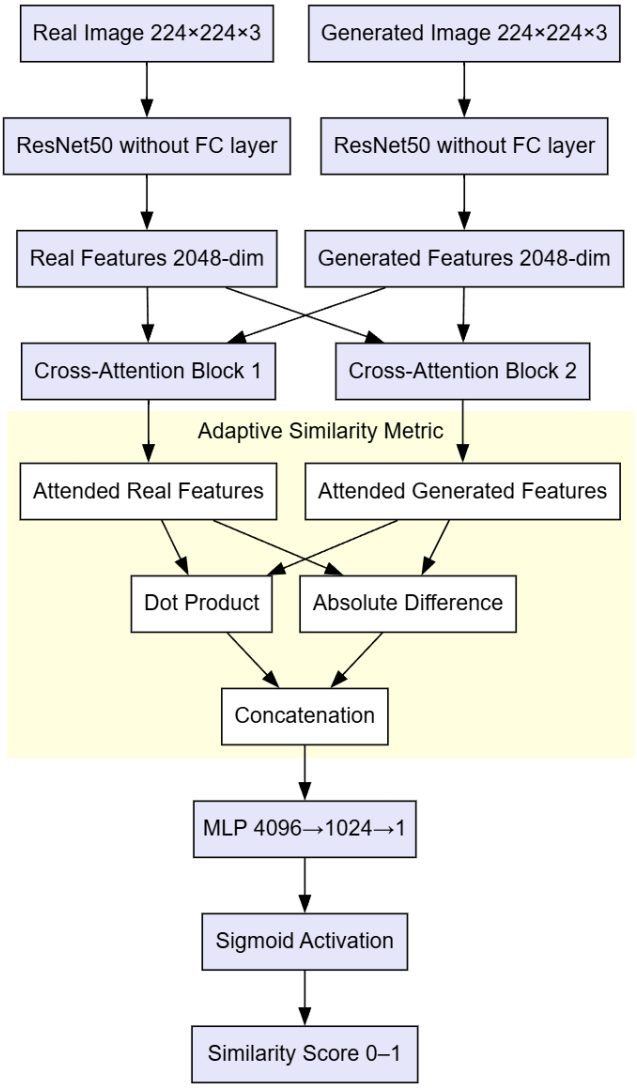

Figure 3: Visualization of the hybrid architecture for fingerprint detection.

4. Results

To evaluate fingerprint detection in synthetic medical imagery, we implemented four baseline Siamese network architectures: ResNet50, Efficient NetB2, DenseNet-161, and Vision Transformer (ViT). Among these, ResNet50 demonstrated strong performance, achieving high recall and accuracy, likely due to its deep residual connections that effectively capture spatial hierarchies in grayscale CT images. However, its relatively low precision indicated a tendency to generate false positives. EfficientNetB2, despite its success in color-based medical image analysis, underperformed on grayscale input, likely a result of its compound scaling not aligning well with domain-specific features. DenseNet161 achieved very high precision but suffered from poor recall, signaling overfitting, and limited generalization. ViT offered promising precision and F1 scores, demonstrating its capacity to suppress false positives; however, its reliance on larger data volumes and its sensitivity to class imbalance negatively affected its overall precision.

To address the limitations of the baseline models, we introduced cross-attention blocks into selected CNN backbones. This mechanism facilitated feature alignment between real and synthetic image pairs, enhancing the network's ability to detect subtle training data traces. In particular, ResNet50 with cross-attention achieved the most balanced and robust performance among all evaluated models, with an F1 score of 0.5902 and a precision of 0.9000. It improved both precision and generalization, making it the most effective architecture for fingerprint detection in this study. DenseNet161 with cross-attention demonstrated greater recall, but suffered from reduced precision, reaffirming its tendency to overfit despite improved attention alignment.The corresponding results have been summarized in Figure 4.

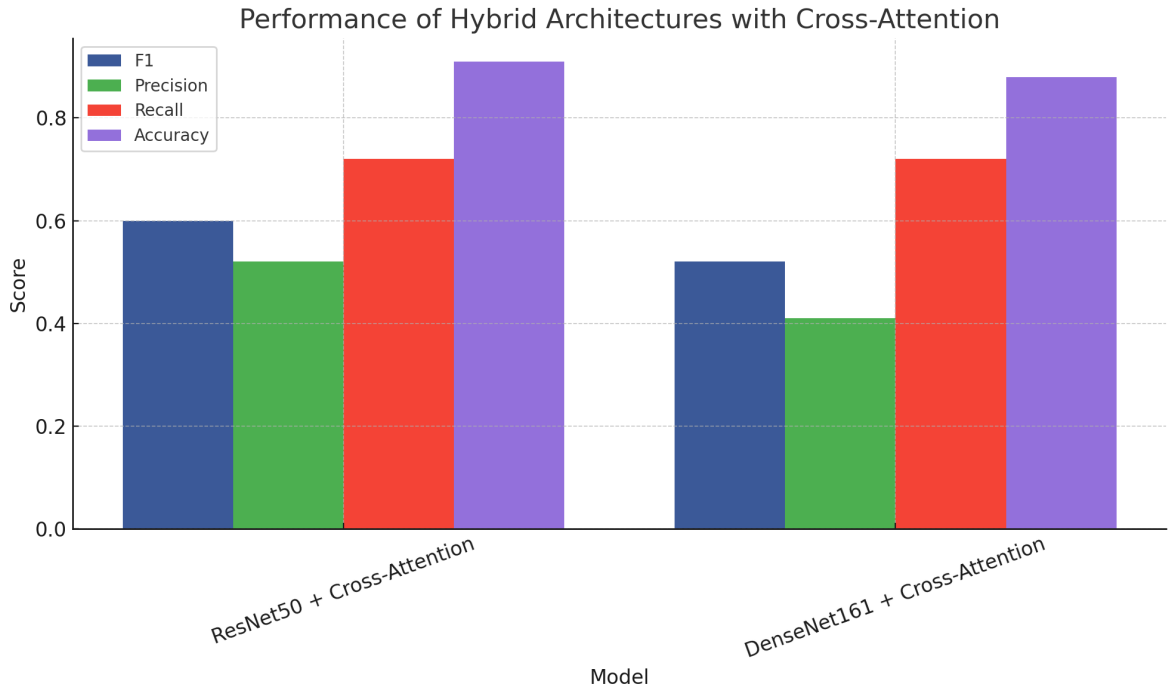

Figure 4: Performance of ResNet and DenseNet models.

The improvements introduced by cross-attention were further supported by ablation insights. The cross-attention module proved critical in dynamically focusing the model's attention on discriminative regions likely to carry generative fingerprints. In parallel, the adaptive similarity metric, which is based on the combination of absolute difference and dot product, enabled the model to assess both semantic alignment and divergence between paired features. Together, these components consistently outperformed their baseline counterparts, highlighting the value of contextual attention and informed similarity scoring for forensic analysis of synthetic medical images.

Table 1
Performance metrics comparison of all implemented models

Model	F1	Precision	Recall	Accuracy
ResNet50 (Baseline)	0.5627	0.4256	0.8300	0.8710
EfficientNetB2 (Baseline)	0.4106	0.2707	0.8500	0.7560
DenseNet161 (Baseline)	0.5714	0.9700	0.4000	0.7000
ViT (Baseline)	0.4299	0.9104	0.7625	0.7500
ResNet50 + Cross-Attention	0.5902	0.5000	0.7200	0.9000
DenseNet161 + Cross-Attention	0.5196	0.4033	0.7300	0.8650

In addition to the model evaluations presented above, our final submission to the ImageCLEFmedical 2025 challenge, registered under submission ID #1160, achieved the following performance metrics: Kappa score of -0.032, accuracy of 48.4%, precision of 48.3%, recall of 45.6%, and an F1-score of 46.9%. These results reflect the real-world performance of our framework in detecting training data usage in synthetic medical images and underscore the practical challenges in achieving high generalizability across diverse generative models.

5. Conclusion And Future Work

This study presented a robust Siamese network-based framework to detect the usage of training data in synthetic medical images generated by GANs. By evaluating multiple convolutional and transformer-based backbones and enhancing select architectures with cross-attention mechanisms and an adaptive similarity metric, the model demonstrated strong performance in identifying subtle generative fingerprints. Among the configurations tested, the Siamese network with a ResNet50 backbone augmented by cross-attention achieved the best balance of accuracy and generalization, underscoring the effectiveness of learning pairwise relationships between real and synthetic images in a shared embedding space.

To support reproducibility and further exploration, the implementation notebooks used in this study have been made publicly available at Github [18]

The results highlight the potential of attention-guided Siamese architectures for forensic analysis in synthetic medical imaging, offering a scalable and interpretable approach to ensuring data integrity and privacy. Future research could explore the extension of this framework to 3D medical volumes such as MRI and full-body CT by adapting the Siamese structure to volumetric data. Further robustness can be achieved through adversarial training against evolving generative techniques. In addition, integrating explainability modules, such as attention heat maps, could enhance transparency in clinical settings. Finally, deploying this system in a federated learning setup and optimizing for real-time performance on edge devices would broaden its applicability across diverse healthcare environments while preserving data privacy.

Acknowledgments

The authors would like to express their sincere gratitude to Vellore Institute of Technology (VIT), Chennai, for their continuous support and encouragement in carrying out this research work. Although the computational experiments reported in this paper were executed on local resources (Lenovo Thinkstation P348 with Intel Core i7-11700 @ 2.5 GHz, 64 GB RAM, 2 TB storage, and a 12 GB NVIDIA GPU), the institutional support provided by VIT was instrumental in the successful completion of this study.

Declaration on Generative AI

During the preparation of this work, the author(s) used GAN-based methods to generate synthetic medical images as part of the research methodology, as described in the paper. The author(s) did not use any Generative AI tools for writing, editing, or creating figures beyond this methodological purpose. All textual content was written and verified by the authors, who take full responsibility for the publication's content.

References

[1] F. Marra, D. Gragnaniello, L. Verdoliva, G. Poggi, Do gans leave artificial fingerprints?, in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2019, pp. 506–511.

[2] N. Yu, L. S. Davis, M. Fritz, Attributing fake images to gans: Learning and analyzing gan fingerprints, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, IEEE, 2019, pp. 7556–7566.

[3] B. Liu, F. Yang, X. Bi, B. Xiao, W. Li, X. Gao, Detecting generated images by real images, in: European Conference on Computer Vision, Springer Nature Switzerland, Cham, 2022, pp. 95–110.

[4] T. Yang, J. Cao, Q. Z. Sheng, L. Li, J. Ji, X. Li, S. Tang, Learning to disentangle gan fingerprint for fake image attribution, arXiv preprint (2021).

[5] F. Marra, C. Saltori, G. Boato, L. Verdoliva, Incremental learning for the detection and classification of gan-generated images, in: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2019, pp. 1–6.

[6] T. Asakawa, H. Shinoda, T. Togawa, K. Shimizu, M. Aono, Real and generated image classification using multi-stage transfer learning, in: CLEF (Working Notes), 2023, pp. 1396–1402.

[7] M. M. Ghazi, M. M. Ghazi, Gan-isi: Generative adversarial networks image source identification using texture analysis, in: CLEF (Working Notes), 2023, pp. 1588–1595.

[8] D. Subburam, S. M. SathyaNarayanan, B. Anand, K. Srinivasan, M. Subramaniam, Dmk-ssn at imageclef 2023 medical: Controlling the quality of synthetic medical images created via gans using machine learning and image hashing techniques, in: CLEF (Working Notes), 2023, pp. 1702–1710.

[9] H. Bharathi, A. Bhaskar, V. Venkataramani, K. Desingu, L. Kalinathan, Correlating biomedical image fingerprints between gan-generated and real images using a resnet backbone with ml-based downstream comparators and clustering: Imageclefmed gans, 2023, in: CLEF (Working Notes), 2023, pp. 1415–1422.

[10] 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.

[11] L. Chai, D. Bau, S.-N. Lim, P. Isola, What makes fake images detectable? understanding properties that generalize, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI, Springer International Publishing, Cham, 2020, pp. 103–120.

[12] Y. Jeong, D. Kim, Y. M. Ro, P. Kim, J. Choi, Fingerprintnet: Synthesized fingerprints for generated image detection, in: European Conference on Computer Vision, Springer Nature Switzerland, Cham, 2022, pp. 76–94.

[13] A. D. Schütte, J. Hetzel, S. Gatidis, T. Hepp, B. Dietz, S. Bauer, P. Schwab, Overcoming barriers to data sharing with medical image generation: A comprehensive evaluation, NPJ Digital Medicine 4 (2021) 141.

[14] V. A. Kelkar, D. S. Gotsis, F. J. Brooks, P. Kc, K. J. Myers, R. Zeng, M. A. Anastasio, Assessing the ability of generative adversarial networks to learn canonical medical image statistics, IEEE Transactions on Medical Imaging 42 (2023) 1799–1808.

[15] Y. Skandarani, P.-M. Jodoin, A. Lalande, Gans for medical image synthesis: An empirical study, Journal of Imaging 9 (2023) 69.
[16] A.-G. Andrei, M. G. Constantin, M. Dogariu, A. Radzhabov, L.-D. 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.
[17] 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.
[18] F. Areeb, D. Vashist, L. Kalinathan, Github for codebase of imageclef medicalgan 2025 challenge submission, https://github.com/ADwar616/ImageCLEF_MedicalGAN, 2025.