# GAN-ISI: Generative Adversarial Networks Image Source Identification Using Texture Analysis

Notebook for the ImageCLEFmedical GANs Lab at CLEF 2023

Mehdi Mehdipour Ghazi, Mostafa Mehdipour Ghazi

Pioneer Centre for AI, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

#### Abstract

Generative adversarial networks (GANs) have emerged as powerful tools for generating realistic images in various domains, including healthcare and medicine. However, concerns surrounding the privacy and security of personal data have become prominent. This study investigates the presence of fingerprints in synthetic medical images generated by GANs, which may indicate traces of the real images used during training and raise concerns about the sharing and limitations imposed by sensitive medical data. To address this, we analyze the texture characteristics of real and synthetic images from the ImageCLEF2023 Medical GANs challenge datasets, utilizing a range of texture descriptors and analysis methods to identify discernible patterns within the synthetic image data and determine the source images employed for training. We calculate the cumulative distribution function (CDF) of texture feature maps and apply the Wasserstein distance to compare the CDFs of the query and generated images. A binary classifier is trained to predict the utilization of the query image in generating each GAN image. The obtained results demonstrate balanced performance across various evaluation metrics, with the model exhibiting good generalization to the challenge test set, achieving an accuracy of 0.54 and an F1-score above 0.5. Our findings provide valuable insights into the security and privacy considerations when generating and utilizing artificial medical images in real-life scenarios.

#### Keywords

Generative adversarial networks, source identification, texture descriptors, cumulative distribution function, Wasserstein distance

# 1. Introduction

Generative Adversarial Networks (GANs) [1] have revolutionized the field of image synthesis, enabling the generation of highly realistic and diverse images across various domains. In the context of healthcare and medicine, GANs have shown remarkable potential in generating biomedical images that capture complex patterns and characteristics [2]. However, as the use of GANs in medical imaging becomes more prevalent, concerns arise regarding the security and privacy of personal source data [3, 4].

This study aims to investigate the hypothesis that GANs generate synthetic medical images that bear discernible traces of the real images used during the training process. The presence of these fingerprints [5, 6] would raise concerns about the potential sharing and usage limitations that artificial biomedical images may inherit from real sensitive medical data. Conversely, if the

<sup>© 02023</sup> Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

hypothesis is proven incorrect, it suggests that GANs can be utilized to create vast datasets of biomedical images that are free from ethical and privacy concerns, opening up new opportunities for real-life applications.

# 2. Background

The increasing availability of large-scale medical image datasets, coupled with advances in deep learning techniques, has fueled the development and utilization of GANs for medical image synthesis. These generative models have shown remarkable success in generating realistic images that capture the complex and nuanced characteristics of medical conditions. However, with the potential integration of GANs into various applications, the concern over the source image forensics and security of the generated images becomes paramount [7].

One line of research focuses on uncovering traces of the training dataset within the generated images. Approaches based on DeepFakes Detection (DFD) have been proposed to identify manipulated or synthesized images. For example, the video-based DFD method of [8] analyzed the artifacts caused by the underlying GAN model to reveal inconsistencies between real and generated images. Similarly, the style-based metrics proposed in [9] were used to distinguish between real and GAN-generated images in the context of facial images.

Another approach involves analyzing the distributional properties and statistical characteristics of real and synthetic images. Methods such as Kernel Density Estimation (KDE) have been employed to detect deviations from the original data distribution. The effectiveness of KDE in identifying synthesized medical images was demonstrated in [10] by comparing their statistical properties with those of the training data. Additionally, approaches based on Wasserstein distance or other generative models have been explored to quantify the similarity between real and synthetic images [11].

Furthermore, advancements in deep learning interpretability have contributed to the development of techniques that visualize and understand the internal representations of GANs. These methods enable the identification of specific image regions or features that influence the generation process. The method proposed in [12] used an attribution-based approach to identify the most important regions in GAN-generated images, shedding light on the potential sources of information within the generated data. Likewise, specific characteristics associated with fake image generators were exploited in [13] to introduce a face generator representation space that allows identification of the face-image source generator model.

By building upon the existing research, this study aims to detect fingerprints within synthetic medical image data and determine the source images used for generation during the training process. The objective of this study does not involve the identification of artificial images or the binary classification of real-fake datasets. Our focus lies in detecting the presence of discernible features or patterns within synthetic image data using various texture descriptors, aiming to determine the real images utilized for training the GANs. The results of this study provides valuable insights into the security of personal medical image data in the context of generating and utilizing artificial images across various real-life scenarios.

# 3. Methods

To address the objective of identifying fingerprints and establishing relationships between real (source) and fake (generated) medical image datasets, this study incorporates a range of texture descriptors and analysis methods. These techniques are employed to effectively extract and analyze texture information from the images under investigation [14, 15]. Texture descriptors serve as computational representations that capture pertinent patterns from the images, facilitating the quantification of various texture aspects and enabling efficient comparisons across different images. The selection of texture descriptors used in this study for image analysis is detailed in the subsequent sections. It should however be noted that each of these approaches may have limitations in capturing robust complex textures and struggle with variations in lighting conditions and scale.

#### 3.1. Statistical features

These features capture statistical relationships between pixel intensities by computing measures such as contrast, entropy, or homogeneity [16]. In this context, we calculate the local range of pixel intensities, local standard deviation, and local entropy of the intensities in a specified neighborhood around the corresponding image pixel. We use 3×3, 3×3, and 9×9 neighborhoods for extracting the range, standard deviation, and entropy feature maps (#3), respectively.

## 3.2. Filter-based features

These features employ filter banks to capture specific spatial-frequency domain information from the images that characterize complex texture patterns at various scales and orientations. Here, we create a bank of Gabor filters at different wavelengths and orientations [17] to later capture texture information from magnitudes of the filter responses. For the construction of the Gabor filter bank, we determined the minimum and maximum lengths as  $L_{min} = 4/\sqrt{2}$  and  $L_{max} = 256/2$ , respectively. Accordingly, the wavelength parameter was discretized into  $2^{[0,1,\dots,\log_2(L_{max}/L_{min})]}L_{min}$  pixels/cycle, while the orientation parameter was regularly sampled at angles of [0, 45, 90, 135] degrees. These choices allowed us to define a comprehensive set of Gabor filters (#24) that could effectively capture a range of spatial frequencies and orientations in the images.

Besides, we use steerable filters of Gaussian derivatives [18], in which the basis filter bank is composed of separable orthogonal kernels using the first and second-order Gaussian derivatives at a specific scale ( $\sigma = 1$ ). The texture patterns can then be obtained by a linear combination of the filter responses at different orientations. In order to generate the steerable filters, we used a regular sampling at [0, 45, 90, 135] degrees for the orientation parameter. Besides, the kernel window size was set to  $2[2\sigma] + 1$ . By employing these choices, we were able to construct a set of steerable filters (#8) that would exhibit controlled directional selectivity and effectively capture various orientations within the image data.

Furthermore, the Gaussian derivatives are used as basis functions to filter the images at different scales, where the texture features are extracted from the gradient magnitude, eigenvalues of the Hessian, Laplacian of Gaussian, Gaussian curvature, and Frobenius norm (eigen magnitude) of the Hessian at each scale [19]. To construct the multiscale filters using Gaussian derivatives, we applied standard deviations  $\sigma = [0.5, 0.75, 1, 1.25, 1.5]$  to the Gaussian function. The kernel window size for each scale was determined as  $2[2\sigma] + 1$ , ensuring an appropriate spatial extent for the filter. These choices enabled us to capture texture feature maps (#30) across multiple scales, facilitating the comprehensive analysis of image content.

# 3.3. Deep features

Deep learning-based features such as convolutional neural network (CNN) representations are high-level abstract information extracted using pretrained models like ResNet [20]. These models are trained on large-scale image datasets [21] and can capture effective texture patterns from images at different scales and levels of abstraction before the output classification layer. We used the pretrained ResNet50 architecture to extract the features from the output of the 14th addition layer [22], which would result in  $7 \times 7 \times 2048$  texture maps.

## 3.4. Histogram descriptors

These descriptors capture the distribution of intensities obtained from the local statistical features or filter maps by dividing the range of intensities into bins and counting the number of pixels falling into each bin. Since the histogram bins need to be defined and adjusted for different maps, we use the empirical cumulative distribution function (CDF) that provides information about the accumulated probability of pixel intensities. It gives insights regarding dominant intensities and their spread within the texture while allowing us to compare the intensities of different maps at specified levels of probabilities.

## 3.5. Image classification

After obtaining the texture descriptors based on the CDF of each abovementioned feature map, we measure the area between the texture CDFs of the query image and those of each generated image using the Wasserstein distance [23]. These multivariate distances are then used to train a binary classifier to predict whether the query image was utilized for the generation of each fake image using GANs or not. The overall decision is made based on the probability score fusion of the classifier outputs from all query-generated multivariate distance pairs.

# 4. Experiments and Results

## 4.1. Data

The data used in this study was sourced from the ImageCLEF2023 Medical GANs challenge datasets [24, 25], which were divided into labeled training and unlabeled test sets. The training set comprises 500 artificial images obtained by training diffusion neural networks using axial slices of 3D computed tomography (CT) scans (8-bit/pixel images of dimension 256×256 pixels) from 8000 lung tuberculosis patients, along with 80 real images that were not utilized during the training of GANs, and another 80 real images that were used for training the model. The test set consists of a total of 10,000 generated images and 200 real images, all of which are unlabeled.

### 4.2. Experimental setup

We partitioned the available training dataset into training and test sets, ensuring that the training set is exclusively used for training and inference purposes before testing the optimized models. To maintain clarity and avoid confusion between our test data split and the challenge test dataset, we will henceforth refer to them as the inner test and challenge test sets, respectively. The inner test set was generated through a stratified partitioning technique, comprising 10% of the training dataset (160×500). The input data underwent a standardization process, normalizing each feature dimension to zero mean and unit variance using the training set. The target and predicted labels were assigned values of 0 for real data instances not involved in the generation of artificial samples, and 1 for real data instances utilized in the generation of synthetic data.

We employed the support vector machine (SVM) classifiers [26] trained in a 5-fold crossvalidation fashion. To capture the non-linear relationships in the data, we utilized a radial basis function (RBF) kernel in conjunction with a logistic function, yielding membership probability scores as the output. The selection of the SVM classifiers were based on careful consideration of various classical classifiers, including discriminant analysis [27], boosted ensemble of decision trees [28], and feedforward neural networks [29]. We opted for the SVM classifiers due to their favorable inference and generalization performance in this study.

#### 4.3. Evaluation metrics

The total accuracy, precision, recall, specificity, and F1-measure were used as the evaluation metrics for the classification tasks. Specificity measures the proportion of correctly predicted negative instances out of the total actual negative instances. In addition, precision measures the proportion of correctly predicted positive instances out of the total instances predicted as positive, while recall or sensitivity quantifies the proportion of correctly predicted positive instances out of the actual positive instances. Finally, the F1-score provides a balanced assessment of both precision and recall by calculating their harmonic mean, offering a single value that reflects the overall performance of a classifier.

### 4.4. Results

Table 1 presents the cross-validated classification accuracies obtained with the proposed GAN-ISI texture analysis method for the inner test and challenge test sets. The results reveal a consistent and promising performance, demonstrating the robustness and effectiveness of the method. Notably, the proposed approach showcases a good balance between precision and recall, indicating its ability to accurately discriminate between the real source and not-used images.

Furthermore, the generalization of the model to the challenge test set underscores its capability to handle unseen data and reinforces its potential for real-world applications. The achieved accuracies, which are better than random, contribute to the understanding of the proposed method's reliability and its potential to address the security and privacy concerns associated with synthetic medical images generated by GANs.

Data	Accuracy	Precision	Recall	Specificity	F1-score
Inner test set	0.575	0.591	0.500	0.650	0.541
Challenge test set	0.535	0.540	0.470	0.600	0.503

The cross-validated classification accuracies of the proposed GAN-ISI method for different test sets.

# 5. Conclusion

Table 1

In this paper, we proposed a method for identifying the source images used in the generation of synthetic medical images using GANs. The approach leveraged texture descriptors extracted from the images through the empirical CDF of texture feature maps, along with the measurement of the Wasserstein distance between the CDFs of the query and generated images. A binary classifier was then trained using these multivariate distances to predict the usage of query images in the generation process.

Several binary classifiers and texture descriptors were evaluated, and the experimental results revealed the superior performance of the SVMs and stacked classical texture descriptors compared to other methods, including deep learning-based approaches. To ensure robustness and avoid overfitting, cross-validation techniques were employed and applied to an inner test set, resulting in a reasonably consistent performance that generalized well to the challenge test set. The achieved performance surpassed random guessing, indicating the effectiveness of the proposed method.

These findings offer valuable insights into the security and privacy aspects of personal medical image data when generating and utilizing synthetic images in real-life scenarios. By accurately identifying the source images, our method contributes to addressing concerns related to the potential sharing and usage limitations inherited from sensitive medical data. The use of classical texture descriptors and the balanced performance obtained demonstrate the potential of our approach in practical applications involving the generation and analysis of artificial medical images.

## References

- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (2020) 139–144.
- [2] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: A review, Medical Image Analysis 58 (2019) 101552.
- [3] N. Clarke, G. Vale, E. P. Reeves, M. Kirwan, D. Smith, M. Farrell, G. Hurl, N. G. McElvaney, GDPR: An impediment to research?, Irish Journal of Medical Science 188 (2019) 1129–1135.
- [4] Z. Chen, T. Zhu, P. Xiong, C. Wang, W. Ren, Privacy preservation for image data: A GAN-based method, International Journal of Intelligent Systems 36 (2021) 1668–1685.
- [5] N. Yu, L. S. Davis, M. Fritz, Attributing fake images to GANs: Learning and analyzing

GAN fingerprints, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7556–7566.

- [6] Y. Ding, N. Thakur, B. Li, Does a GAN leave distinct model-specific fingerprints, in: Proceedings of the BMVC, 2021.
- [7] P. Yang, D. Baracchi, R. Ni, Y. Zhao, F. Argenti, A. Piva, A survey of deep learning-based source image forensics, Journal of Imaging 6 (2020) 9.
- [8] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, Interfaces (GUI) 3 (2019) 80–87.
- [9] A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427–436.
- [10] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, Neurocomputing 321 (2018) 321–331.
- [11] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 214–223.
- [12] Y. Shen, J. Gu, X. Tang, B. Zhou, Interpreting the latent space of GANs for semantic face editing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 9243–9252.
- [13] M. Salama, H. Hel-Or, Face-image source generator identification, in: Computer Vision– ECCV 2020 Workshops: Part V, Springer, 2020, pp. 511–527.
- [14] R. M. Haralick, K. Shanmugam, I. H. Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man, and Cybernetics (1973) 610–621.
- [15] T. Ojala, M. Pietikäinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognition 29 (1996) 51–59.
- [16] T. Randen, J. H. Husoy, Filtering for texture classification: A comparative study, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (1999) 291–310.
- [17] A. K. Jain, F. Farrokhnia, Unsupervised texture segmentation using Gabor filters, Pattern Recognition 24 (1991) 1167–1186.
- [18] W. T. Freeman, E. H. Adelson, et al., The design and use of steerable filters, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (1991) 891–906.
- [19] L. Sørensen, C. Igel, N. Liv Hansen, M. Osler, M. Lauritzen, E. Rostrup, M. Nielsen, Early detection of Alzheimer's disease using MRI hippocampal texture, Human Brain Mapping 37 (2016) 1148–1161.
- [20] 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.
- [21] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, Communications of the ACM 60 (2017) 84–90.
- [22] N. Ahmed, H. M. S. Asif, Perceptual quality assessment of digital images using deep features., Computing & Informatics 39 (2020).
- [23] M. De Angelis, A. Gray, Why the 1-Wasserstein distance is the area between the two marginal CDFs, arXiv preprint arXiv:2111.03570 (2021).

- [24] B. Ionescu, H. Müller, A.-M. Drăgulinescu, W.-w. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. G. Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A.-G. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L.-D. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, social media and recommender systems applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association, Springer LNCS, 2023.
- [25] 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: CLEF2023 Working Notes, CEUR Workshop Proceedings, 2023.
- [26] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273-297.
- [27] Y. Guo, T. Hastie, R. Tibshirani, Regularized linear discriminant analysis and its application in microarrays, Biostatistics 8 (2007) 86–100.
- [28] M. K. Warmuth, J. Liao, G. Rätsch, Totally corrective boosting algorithms that maximize the margin, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 1001–1008.
- [29] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.