# Overview of ImageCLEFmedical GANs 2023 Task — Identifying Training Data "Fingerprints" in Synthetic Biomedical Images Generated by GANs for Medical Image Security

Notebook for the ImageCLEF Lab at CLEF 2023

Alexandra-Georgiana Andrei[1,*], Ahmedkhan Radzhabov[2], Ioan Coman[1], Vassili Kovalev[2], Bogdan Ionescu[1] and Henning Müller[3]

[1]*Politehnica University of Bucharest, AI Multimedia Lab, Romania*
[2]*Belarusian Academy of Sciences, Minsk, Belarus*
[3]*University of Applied Sciences Western Switzerland, Sierre, Switzerland*

**Abstract**

The 2023 ImageCLEFmedical GANs task is the first edition of this task, examining the existing hypothesis that GANs (Generative Adversarial Networks) are generating medical images that contain the "fingerprints" of the real images used for generative network training. The objective proposed to the participants is to identify the real images that were used to obtain some synthetic images using Generative Models. Overall, 23 teams registered to the task, 8 of them finalizing the task and submitting runs. A total of 40 runs were received. An analysis of the proposed methods shows a great diversity among them, ranging from texture analysis, similarity-based approaches that join inducer predictions like SVM or KNN, to deep learning approaches and even multi-stage transfer learning. This paper presents the overview of 2023 ImageCLEFmedical GANs task by describing its datasets, evaluation metrics as well as a discussion of the participants runs and results, and the future challenges.

**Keywords**

artificial intelligence and deep learning, generative models, medical synthetic data, medical imaging, ImageCLEF benchmarking lab

## 1. Introduction

ImageCLEF [1] is part of the CLEF initiative[1] and presents a set of multimedia information retrieval tasks. Medical tasks were included in the 2nd edition of ImageCLEF in 2004 and have been held every year since then. The 2023 ImageCLEFmedical GANs task is the first edition of this task, examining the existing hypothesis that GANs (Generative Adversarial Networks) are

*CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece*
*Corresponding author.
alexandra.andrei@upb.ro (A. Andrei); coman.ioan95@gmail.com (I. Coman); vassili.kovalev(at)gmail.com (V. Kovalev); bogdan.ionescu@upb.ro (B. Ionescu); henning.mueller@hevs.ch (H. Müller)

[1]https://www.clef-initiative.eu/

generating medical images that contain the "fingerprints" of the real images used for generative network training. If this hypothesis is true, it may question the very nature of the synthetic images in term of copyright issues. So far, synthetic images are considered to be totally artificial data, thus no copyright issues can occur with respect to the real images.

In recent years, the emergence of generative models in the field of Artificial Intelligence (AI) has sparked significant interest and innovation, transforming various fields and revolutionizing the way we solve complex problems. The 2023 ImageCLEFmedical GANs Task offered an environment for investigating GANs' effects on the creation of synthetic medical images by providing a benchmark to explore the impact of GANs on artificial biomedical image generation. Medical image generation plays a critical role in medical research, training healthcare professionals, and improving patient care. While real patient data can be expensive, insufficient, or ethically challenging to acquire, the ability to generate synthetic yet realistic biomedical images can bridge these gaps and empower researchers, clinicians, and educators. Thus, Generative Models have demonstrated remarkable capabilities in generating high-quality images that mimic the characteristics and patterns of real data.

In this article, we present an overview of the 2023 ImageCLEFmedical GANs Task, describing the objective, data sets, evaluation metrics and participants' solutions. The reminder of the article is organized as follows. Section 2 introduces the scope and objectives of the task. Section 3 presents the evaluation metrics and Section 4 describes and presents the methods and results obtained by each participant team. Finally, Section 5 concludes the paper and presents the conclusions.

## 2. Task description

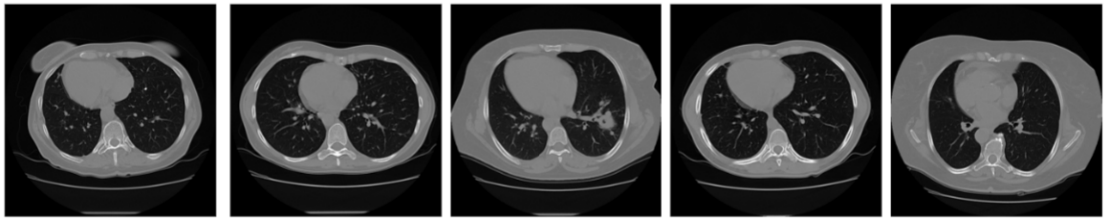
ImageCLEFmedical GANs task is a new challenge of the 2023 ImageCLEF lab [1]. The objective of the first edition of ImageCLEFmedical GANs task is to investigate the hypothesis that generative models generate medical images that exhibit resemblances to the images employed during their training. This addresses concerns surrounding the privacy and security of personal medical image data in the context of generating and utilizing artificial images in various real-world scenarios.

The task aims to identify distinctive features or "fingerprints" within synthetic biomedical image data, allowing us to determine which real images were used during the training process to generate the synthetic images. The task is formulated as following:

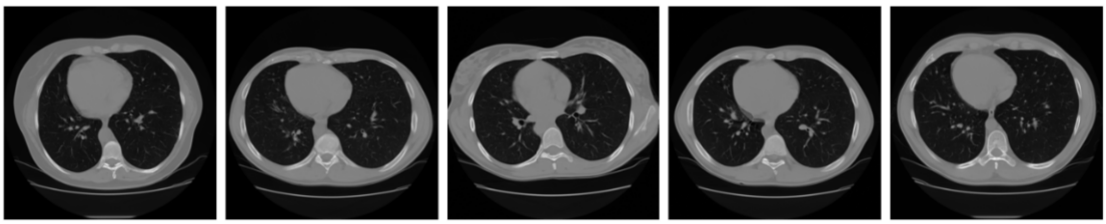
- *given a set that contains generated and real images, the participants are requested to employ machine learning and/or deep learning models to determine which of the real images were used to train the models to generate the provided synthetic images.*

### 2.1. Data Description

For the ImageCLEFmedical GANs task, we provided a data set containing axial chest CT scans of lung tuberculosis patients. This means that some of them may appear pretty "normal" whereas the others may contain certain lung lesions including the severe ones. These images are stored



a



b

**Figure 1:** Examples of images from the provided test dataset: (a) real images, (b) synthetic images generated using a Generative Model.

in the form of 8 bit/pixel PNG images with dimensions of $256 \times 256$ pixels. The artificial slice images are $256 \times 256$ pixels in size. All of them were generated using Diffuse Neural Networks.

The data is structured as following:

- Development (Train) dataset: consists of 500 artificial images and 160 real images annotated according to their use in the training of the generative network. Out of the real images, 80 were used during training.
- Test (Evaluation) dataset: it was created in similar way. The only difference is that the two subsets of real images are mixed and no proportion of non-used and used ones has been disclosed. Thus, a total of 10,000 generated and 200 real images are provided. Examples of real and generated images are shown in Fig. 1.

## 3. Evaluation Methodology

The task was evaluated as a binary-class classification problem and the evaluation was carried out by measuring the F1-score, accuracy, precision, recall and specificity metrics. The official evaluation metric of this year's edition is the F1-score. The metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

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

| Group name | Main institution | Country |
|---|---|---|
| Clef—CSE—GAN—Team | Sri Sivasubramaniya Nadar College of Engineering | India |
| DMKS—SSN | Sri Sivasubramiya Nadar College of Engineering | India |
| GAN—ISI | University of Copenhagen | Denmark |
| KDE—lab | Toyohashi University of Technology | Japan |
| one five one zero | Yunnan University in Kunming | China |
| PicusLabMed | University of Naples Federico II | Italy |
| VCMI | INESC TEC | Portugal |
| AIMultimediaLab* | University Politehnica of Bucharest | Romania |

$$F1 - score = \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

## 4. Participant Runs

Each participating team could submit up to 10 runs in total. Table 1 presents the list of participants and their institutions. The ranking is presented in Table 2 and it was according to the F1-score. A total of 40 runs were received from eight teams.

**CLEF—CSE—GAN**. The best performing run from the CLEF—CSE–GAN team achieved an F1-score of 0.614. This team proposed three different workflows based on ResNet feature extractors [2]. First, aglomerative clustering is used to group similar images together based on generated features. By identifying clusters predominantly composed of real images, the authors enhance the ability to distinguish between real and artificial images effectively. Second, an SVM is implemented as a classifier that discerns real from artificial images, this time based on a one-dimensional flattened concatenation of the features from corresponding images pairs. The SVM model is trained using the combined feature representations obtained from the real and artificial images. And finally, a relation network based out of few-shot learning is used to fine-tune the backbone to learn fingerprints, learn a custom similarity comparison metric, and preserve spatial context by concatenating features as two-dimensional representations. All results obtained by the team are shown in Table 2 and consists in the following methods:

- Submission #1: relational model that used ResNet-101 as the backbone model for feature extraction.
- Submission #2: Hierarchical clustering approach.
- Submission #3: SVM.

**DMK—SSM**. The best performing run from DMK—SSM team achieved an F1-score of 0.480. They used Perceptual Hashing (pHash) algorithm [3]. It was applied on the generated images, which produces a hash, a fingerprint of the image. Similarly, hashes for the used and unused real images were generated and Hamming distance was computed between the used and the generated images. By doing this, pair of source image and generated images were identified and a threshold value of 50 was selected for the hamming distance. Two models were used: Convolutional

**Table 2**
Summary on the participant submissions and their results.

| Group rank | Group name | Submission # | F1-score |
|---|---|---|---|
| #1 | VCMI | submission 2 | 0.802 |
| #2 | VCMI | submission 1 | 0.731 |
| #3 | VCMI | submission 3 | 0.707 |
| #4 | PicusLabMed | submission 8 | 0.666 |
| #5 | VCMI | submission 4 | 0.654 |
| #6 | AIMultimediaLab | submission 1 | 0.626 |
| #7 | PicusLabMed | submission 6 | 0.624 |
| #8 | VCMI | submission 5 | 0.621 |
| #9 | Clef-CSE-GAN-Team | submission 1 | 0.614 |
| #10 | VCMI | submission 7 | 0.613 |
| #11 | VCMI | submission 6 | 0.605 |
| #12 | VCMI | submission 10 | 0.594 |
| #13 | AIMultimediaLab | submission 2 | 0.585 |
| #14 | one five one zero | submission 2 | 0.563 |
| #15 | PicusLabMed | submission 9 | 0.562 |
| #16 | PicusLabMed | submission 4 | 0.552 |
| #17 | KDE lab | submission 5 | 0.548 |
| #18 | one five one zero | submission 3 | 0.522 |
| #19 | Clef-CSE-GAN-Team | submission 2 | 0.521 |
| #20 | VCMI | submission 9 | 0.514 |
| #21 | one five one zero | submission 1 | 0.507 |
| #22 | GAN-ISI | submission 5 | 0.502 |
| #23 | GAN-ISI | submission 2 | 0.489 |
| #24 | PicusLabMed | submission 10 | 0.487 |
| #25 | GAN-ISI | submission 3 | 0.486 |
| #26 | GAN-ISI | submission 4 | 0.483 |
| #27 | DMK | submission 1 | 0.480 |
| #28 | PicusLabMed | submission 2 | 0.470 |
| #29 | KDE lab | submission 2 | 0.469 |
| #30 | GAN-ISI | submission 1 | 0.469 |
| #31 | KDE lab | submission 1 | 0.465 |
| #32 | KDE lab | submission 4 | 0.457 |
| #33 | DMK | submission 2 | 0.449 |
| #34 | VCMI | submission 8 | 0.448 |
| #35 | PicusLabMed | submission 1 | 0.434 |
| #36 | Clef-CSE-GAN-Team | submission 3 | 0.431 |
| #37 | PicusLabMed | submission 3 | 0.419 |
| #38 | PicusLabMed | submission 5 | 0.417 |
| #39 | KDE lab | submission 3 | 0.407 |
| #40 | PicusLabMed | submission 7 | 0.093 |

Neural Network (CNN) and Scale-Invariant Feature Transformation (SIFT) algorithm with the K-Nearest Neighbors (KNN) classifier. All results obtained by the team are shown in Table 2

and consists in the following methods:

- Submission #1: 3 layers CNN model.
- Submission #2: SIFT-KNN model.

**GAN—ISI**. The best performing run from GAN—ISI team achieved an F1-score of 0.489. The authors used texture analysis to study characteristics of real and synthetic images [4]. A range of texture descriptors and analysis methods were used to identify discernible patterns within the synthetic image data and determine the source images employed for training. The cumulative distribution function (CDF) of texture feature maps was calculated and the Wasserstein distance was applied to compare the CDFs of the query and generated images. A binary classifier was trained to predict the utilization of the query image in generating each GAN image. Five different runs using the same method were submitted and the best results were obtained for submission #5 with an F1-score of 0.502.

**KDE—lab**. The best performing run from the KDE—lab team achieved an F1-score of 0.548. The team proposed a fine-tuning deep neural network model that uses multi-stage transfer learning [5]. The first stage transfer learning uses casia dataset [6], the second stage transfer learning uses COVID-19 dataset [7], the third stage transfer learning uses the development dataset provided with the task, while the fourth stage transfer learning uses test dataset. Several methods were used for predicting the used/not used label of the real images. The best results of an F1-score of 0.548 was obtained using ViT B/32 using multi-stage transfer learning. The results obtained by the team are shown in Table 2 (there was no reference in team's working notes to the method used to obtain the results provided in the Submission file #2) and consists of the following methods:

- Submission 1 - Conv model using multi-stage transfer learning.
- Submission 3 - ResNet18 using multi-stage transfer learning.
- Submission 4 - VGG11 using multi-stage transfer learning.
- Submission 5- ViT B/32 using multi-stage transfer learning.

**One five one zero**. The best performing run from the one five one zero team achieved an F1-score of 0.507. The authors used a contrastive learning architecture, combined with transfer learning [8]. They used different pre-trained feature extraction modules such as Inception V3, ResNet and EfficientNet to find the target with large response value through the similarity calculation method in order to find the original image. Euclidian distance was used to determine the distance between the target, positive and negative examples, and use it was used an input to the loss function. All results obtained by the team are shown in Table 2 and consists in the following methods:

- Submission #1: Inception V3.
- Submission #2: ResNet50.
- Submission #3: EfficientNet.

**PicusLabMed**. The best performing run from the PicusLabMed team achieved an F1-score of 0.666. The team studied the ability of Deep-Learning models to provide a representation

of the input data, relying on CNN, to extract the features from the real and generated images [9]. These features were analysed using a ML model for the identification of the samples used during the development of the generative model among all the real instances. The authors proposed two variants for the features extraction step, introducing Vector-Net, a convolutional network that learns how to map in the input image in an efficient representation, and leveraging a Deforming Autoencoder (DAE), that provides a latent vector in an unsupervised manner. All results obtained by the team are shown in Table 2 and consists in the following methods:

- Submission #1: Vector-Net applied to the images that were not used for training and to the images that were used for training (1) and Linear-SVM classifier.
- Submission #2: Vector-Net (1) and SVM-2 classifier.
- Submission #3: Vector-Net (1,2) and Linear-SVM classifier.
- Submission #4: Vector-Net (1,2) and SVM-2 classifier.
- Submission #5: Vector-Net (1,2,3) and Linear-SVM classifier.
- Submission #6: Vector-Net (1,2,3) and SVM-2 classifier.
- Submission #7: DAE applied to the generated images and the images used for training and SVM-Linear Classifier.
- Submission #8: DAE applied to the generated images and the images that were not used for training and SVM-Linear Classifier.
- Submission #9: DAE applied to the generated images, the images that were not used for training and to the images that were used for training and SVM-Linear Classifier.
- Submission #10: voting strategy among the other results.

**VCMI**. The best performing run from the VCMI team achieved an F1-score of 0.802. The authors used similarity-based approaches such as: auto-encoders (AE) to classify the images through outlier detection techniques and patch-based methods that operate on patches extracted from real and generated images to measure their similarity [10].

Structural Similarity Index Measure (SSIM) between real and generated images was studied and different methods were applied as described in the following: (i) Threshold approach was used to find and classify as "used" real images whose similarity to their most similar generated image is higher than a threshold. The threshold was calculated based on the similarity between real images; (ii) Retrieval approach was used to find a set of real images that are the most similar to at least one generated image, and those images were classified as "used". All retrieved images that are, therefore, the most similar to at least one of the generated images, were classified as "used". Real images that are not retrieved were classified as "not used"; (iii) Ranking approach was used to classify real images based on a ranking that defines how similar they are to the generated images. The method starts by calculating a threshold that represents the average rank of similarity of a real image when compared with other real images. Finally, if this average rank is higher than the threshold, then the image is classified as "used", as it shows high similarity with respect to the generated images. Otherwise, the image is classified as "not used"; (iv)(4) Clustering approach was used to find outliers in the data, to classify them as "not used". First, the method maps both generated and real images into a common space. Then, it uses the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to form clusters, and to find outliers. Outliers identified in the subset of real images were classified as

"not used", while the remaining images were classified as "used"; (v) Ensemble method was used to merge the results of different methods presented by the team; (vi) AE based methods using two types of auto-encoders (Basic AE and ResNet AE) were used to classify images by two ways: Computing the similarity between the images based on their latent representations, enabling the direct application of the techniques defined in the previous section; Applying outlier detection techniques to identify data points from the real data that do not follow the probability distribution of the generated data; (vii) Patch-based method were used to extract patches from images and perform different operations such as : matching patches using Triplet loss and replacing patches from real images with patches extracted from the generated images.

The best results were obtained with a similarity-based approach that uses Structural Similarity SSIM to compute the similarity between real and generated images, achieving an F1-score of 0.802. All results obtained by the team are shown in Table 2 and consists in the following methods:

- Submission #1: Ranking/ Ensemble method using as a similarity SSIM metric.
- Submission #2: Threshold (MAX) method using SSIM as a similarity metric.
- Submission #3: Retrieval method using SSIM as a similarity metric.
- Submission #4: Simple AE (AVG).
- Submission #5: Ranking /Ensemble with Simple AE trained on all 10,000 generated images and 200 real images.
- Submission #6: Ranking /Ensemble with Simple AE trained on 600 generated images and 200 real images.
- Submission #7: Ranking /Ensemble with ResNet AE.
- Submission #8: Ranking /Ensemble with ResNet.
- Submission #9: Matching Patches.
- Submission #10: Replacing Patches.

**AIMultimediaLab**. The best performing run from the AIMultimediaLab achieved an F1-score of 0.626. The team proposed two approaches for addressing the task [11]. Both approaches start by generating synthetic images from the real unused images provided in the development dataset. Subsequently, distinct descriptors/features are extracted and utilized to train a binary SVM classifier that was further used for identifying which of the 200 provided real images were used for generating the 10,000 artificial images from the test dataset. The analyzed features were extracted using two methods: a hand-crafted feature extraction technique called Local Binary Pattern (LBP) to capture the local spatial patterns and the gray scale contrast of the images and a deep-learning approach utilizing a pre-trained VGG-16 convolutional network. All results obtained by the team are shown in Table 2 and consists in the following methods:

- Submission #1: Hand-crafted feature extraction method and an radial SVM classifier.
- Submission #2: Deep-learning feature extraction method and an radial SVM classifier.

Fig. 2 shows the corresponding confusion matrix for each team's best run.

VCMI team achieved the best F1-score of 0.802 for the experiment in which they used Threshold (MAX) method using SSIM as a similarity metric to check whether the distance

between each real image and its closest generated image is higher than a threshold. The threshold approach finds real images whose similarity to their most similar generated image is higher than a threshold, classifying them as "used". The threshold is calculated based on the similarity between real images. The MAX threshold was considered the one with the maximum similarity between two images from the real data.

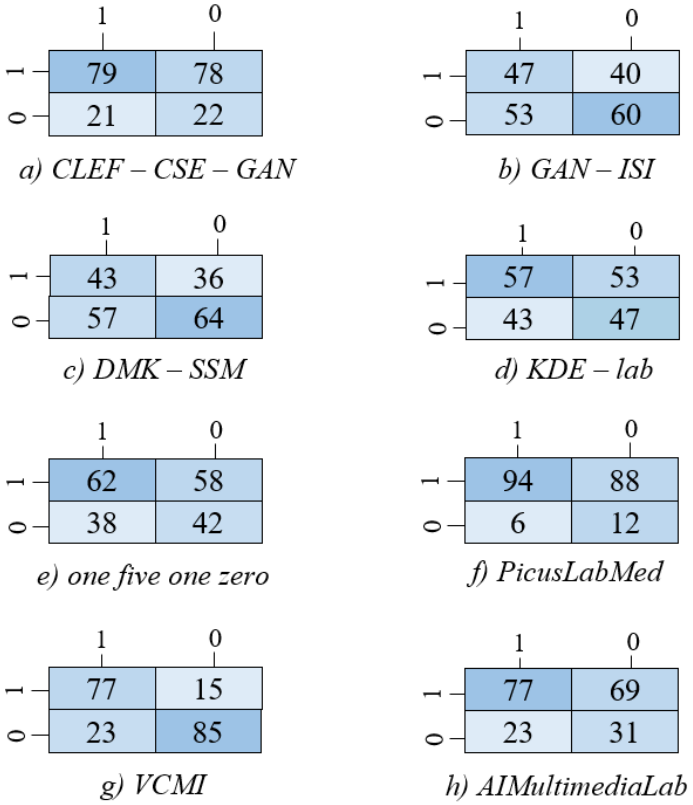



**Figure 2:** Confusion matrices for each team's best run. Vertical axes —- true label, horizontal axes —- predicted label, "1" — images used for training, "0"—images not used for training.

## 5. Conclusions

The first edition of the ImageCLEF medical GANs task attracted a total of 8 teams that submitted runs, with all of them completing their submissions by creating a working notes paper. One task was proposed to the participants, a prediction-based task that uses real and generated CT images. All the participant teams show interesting methods and results. The best result for the task is an F1-score of 0.802 obtained by VCMI team followed by PicusLabMed with an F1-score of 0.666 and AiMultimediaLab with an F1-score of 0.626. Regarding the identification methods proposed by the participants, we are happy to report a high degree of diversity among them. Proposed methods include multi-stage transfer learning, analysis of similarity of different features, patch

extraction methods, threshold methods, Perceptual Hashing algorithm and different deep-learning feature extraction methods that were further classified using both traditional (SVM, kNN) and deep learning models for prediction. We find this to be truly motivating, and we are looking forward to development of this tas in the future editions of ImageCLEF.

Future editions of this task will expand the study areas of synthetic medical data, varying different aspects such as datasets and generation methods. Also, we plan to add other tasks based on different aspects of the privacy and security of the generated data.

## Acknowledgments

The contribution of Alexandra Andrei, Bogdan Ionescu and Henning Müller to this task is supported under project AI4Media, A European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant #951911.

## References

[1] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. Garcıa 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. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, 2023.

[2] H. Bharathi, A. Bhaskar, V. Venkataramani, K. Desingu, L. Kalinathan, CLEF-Correlating Biomedical Image Fingerprints between Real and GAN-generated Images using a ResNet Backbone with ML-based Downstream Comparators: ImageCLEFmed GANs 2023, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023.

[3] D. S, S. M. S, B. A, Kavitha, M. S, Dmk-ssn at imageclef 2023 medical: Controlling the quality of synthetic medical images created via gans using machine learning and image hashing techniques, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023.

[4] M. Mehdipour-Ghazi, M. Mehdipour-Ghazi, Gan-isi: Generative adversarial networks image source identification using texture analysis, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023.

[5] T. Asakawa1, H. Shinoda1, T. Togawa, K. Shimizu, M. Aono, Real and generated image classification using multi-stage transfer learning, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023.

[6] P. Sovathna, casia dataset, kaggle, 2018. URL: https://www.kaggle.com/datasets/sophatvathana/casia-dataset.

[7] M. Rahimzadeh, A. Attar, S. M. Sakhaei, A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset, Biomedical Signal Processing and Control 68 (2021) 102588.

[8] S. Cao, X. Zhou, Finding the source images from the generated images with contrastive learning methods, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023.

[9] M. Gravina, S. Marrone, C. Sansone, Analyzing the similarity between artificial and training images in generative models: The picuslabmed contribution, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023.

[10] H. Montenegro, P. Neto, C. Patrício, I. Rio-Torto, T. Gonçalves1, L. F. Teixeira1, Evaluating privacy on synthetic images generated using gans: Contributions of the vcmi team to imageclefmedical gans 2023, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023.

[11] A.-G. Andrei, B. Ionescu, Aimultimedialab at ImageCLEFmedical GANs 2023:determining "fingerprints" of training data in generated synthetic images, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023.