Classification of Real and Generated Images based on Feature Similarity

Notebook for ImageCLEF Lab at CLEF 2024

Huihui Tang[1], Hancheng Wang[1,2] and Jining Chen[1,*]

[1]Guangxi Key Laboratory of Digital Infrastructure, Guangxi Zhuang Autonomous Region Information Center
[2]GUANGXI BEITOU IT INNOVATION TECHNOLOGY INVESTMENT GROUP CO.,LTD.

Abstract

Deceptive images can be shared on social network services within seconds, posing a significant risk. In the application and research of artificial intelligence on medical images, data issues have always been a challenge, including insufficient amounts of medical image data and privacy concerns. Currently, generative models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models have achieved remarkable results, generating high-quality images. Using generative models for the generation of medical images is a research focus. The essence of generative models is to learn the distribution of data, not to create something out of nothing. Therefore, we investigate whether the real data can be identified from the generated data. In this paper, based on similarity calculations, we calculate the similarity between original images and images with added perturbations. We use self-supervised Masked Autoencoders to reconstruct the images and thus achieve feature similarity calculation. By judging the similarity, we can identify the real images used for training the generative model. Our experimental results on the validation set show an accuracy of 0.743, a precision of 0.721, a recall of 0.720, and an F1 score of 0.726; the F1 score on the test set is 0.603. The experimental results indicate that generated images also pose a threat to patient privacy.

Keywords

GANs, Pre-trained Model, Feature Extraction, Similarity Calculation

1. Introduction

Currently, artificial intelligence has numerous hot topics in medical research, including studies on medical imaging, medical image classification, detection, and segmentation tasks[1, 2, 3]. However, these tasks face a common challenge: the lack of data. Initially, data augmentation[4] was used as a solution, transforming small amounts of data into larger datasets through operations such as flipping and cropping. This approach relies on the translation invariance of convolutional neural networks. However, data augmentation only addresses superficial issues, and in some tasks, the increased quantity is still insufficient. Additionally, the amount of data often determines the upper limit of the model. Large models require vast amounts of data for training, potentially exceeding tens of billions of data points. As models become larger, the demand for data also increases; otherwise, the models either overfit to noise or fail to fully utilize their capabilities.

Medical data, in particular, presents complex challenges due to privacy concerns and the difficulty of annotation, making data acquisition difficult and resulting in consistently small datasets. The advent of generative models represents a technological breakthrough, capable of creating large amounts of high-quality data. Currently, models based on Generative Adversarial Networks (GANs)[5] and Variational Autoencoders (VAEs)[6] can generate high-quality images, while the popular diffusion models not only produce high-quality images but also exhibit diversity. Therefore, generative models are an effective method for data augmentation. However, privacy concerns must be considered for medical data. Even if the generated data differs, does it still pose a privacy risk? Can the original data be identified from the generated images?

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
tanghh@gxi.gov.cn (H. Tang); 1392207107@qq.com (H. Wang); chenjn@gxi.gov.cn (J. Chen)

To address this issue, this paper employs multiple methods, including similarity calculation, enhancing image details before calculating similarity, and using deep learning and Masked Autoencoders (MAE)[7] methods to calculate the similarity of extracted image features. This approach aims to determine which real data were used for training the generative models based on the generated data.

2. Related Work

In recent years, Generative Adversarial Networks (GANs)[5] have garnered widespread attention in the medical field for various image generation and translation tasks. Numerous studies have explored the application of GANs in medical image synthesis and image-to-image translation, particularly in the recognition or detection of synthetic images. A substantial amount of work has investigated the use of GANs to generate synthetic medical images. For instance, Choi et al. proposed a method called "StarGAN"[8], a multi-domain image synthesis technique successfully applied to generate diverse and realistic brain MRI images. Similarly, the paper by Kench et al. introduced SliceGAN[9], an architecture that utilizes GANs to generate high-quality three-dimensional datasets from a single representative two-dimensional image.

Synthetic images play a crucial role in the medical field as they offer significant advantages and address major challenges[10]. Firstly, the generation of synthetic images allows for the augmentation of limited or insufficient datasets. In many medical imaging applications, obtaining large and diverse annotated datasets can be challenging and time-consuming. By generating synthetic images, researchers can expand the training data, thereby enhancing the robustness and generalization of machine learning models. Secondly, synthetic images can simulate rare or difficult-to-obtain medical scenarios. Certain conditions or diseases may have low prevalence or be challenging to capture through traditional imaging methods, making synthetic images a valuable resource in these instances.

Synthetic images offer a novel approach to creating representative cases, allowing researchers and clinicians to study and understand these conditions better, develop more effective diagnostic tools, and explore various treatment strategies. Additionally, synthetic images address privacy concerns associated with patient data. Medical images often contain sensitive information, making data sharing and public release challenging. By generating synthetic images, it is possible to retain the statistical and anatomical characteristics of the data while removing specific patient information. This approach preserves privacy, facilitates more open collaboration, and advances research progress. In conclusion, synthetic images are indispensable in the medical field. They play a crucial role in data augmentation, simulation of rare conditions, and privacy protection. Their utilization empowers researchers, clinicians, and technologists to tackle key challenges, enhance diagnostic accuracy, improve patient care, and advance medical imaging technologies.

3. Method

3.1. Task Description and Dataset Analysis

After training the generative model, it is possible to produce new images. In order to identify the potential privacy threats of using and sharing synthetic medical data in various real-world scenarios, a new challenge (ImageCLEFmedical GANs[11]) arose as part of the medical track of the ImageCLEF Challenge 2024[12]. Our team's username is robot. This task aims to verify whether the generated images can leak information from the original training data. By analyzing the generated images, we can distinguish which images from the real dataset were used to train the generative model. The dataset used in this study is as follows:

The data is divided into two folders: development data and test data. In the development data, the dataset includes both used and not used images.

In Table 1 and Table 2, the dataset structures for Task 1 and Task 2 are shown, respectively. Although there is a significant difference in the number of real datasets between the two tasks during the

Table 1
Task1 dataset structure description.

Dataset	Generated	Real
Development	10000	100 (used) 100 (not used)
Test	5000	4000 (used and not used)

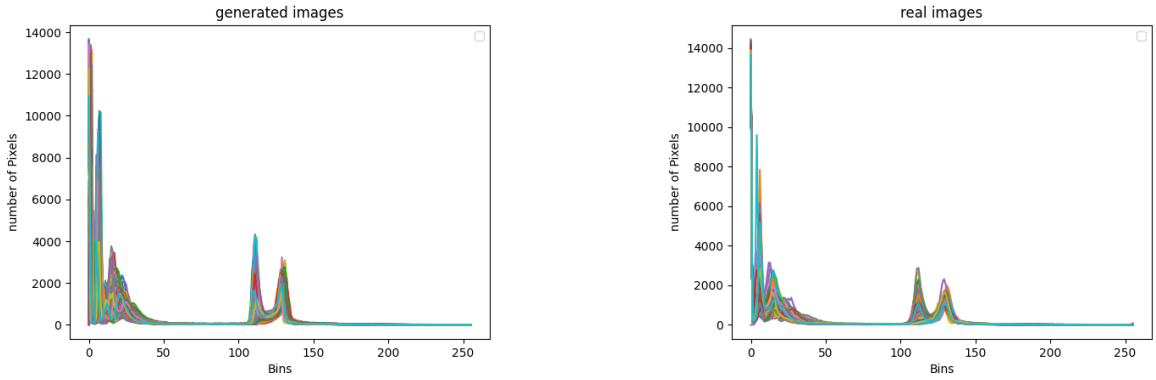
Table 2
Task2 dataset structure description.

Dataset	Generated	Real
Development	10000	3000 (used) 3000 (not used)
Test	7200	4000 (used and not used)

development phase, the approach and methods for handling both tasks are consistent.

3.2. Data Visualization Analysis

The quality of a generative model is evaluated based on its ability to produce high-quality and diverse images. A generative model learns the distribution of real data, and a model that can generate high-quality images indicates that it has accurately learned the data distribution of real images. Diversity reflects the creative ability of the generative model, assessing whether it can create different images based on the learned data distribution. Visualizing the data distribution allows for an intuitive understanding of the relationships between the data. In this paper, we provide histogram visualizations of the statistical data for both the generated and real images.



(a) Pixel statistics of generated images.

(b) Pixel statistics of real images.

Figure 1: Pixel statistics are performed on the generated images and real images of the two tasks respectively, and the results of the two tasks are averaged.

As illustrated in Figure 1, we have compared the pixel values of the generated images with those of the real images. It is evident that the pixel value distributions are quite similar. The horizontal axis of the histogram represents the pixel values, while the vertical axis represents the number of pixels. Based on the results of the two histogram statistics, it can be seen that the generated images and the real images are highly similar. Additionally, it is evident that the real images are also highly similar to each other, even though they include two categories: used images and not used images.

3.3. Feature Extraction

In this paper, we explore the inherent connection between generated images and used images, which were involved in the training process. As we know, in mainstream generative model methods, Generative Adversarial Networks (GANs) have both a generator and a discriminator that mutually enhance each other, ultimately generating the required image data through the generator. The generator's structure involves extracting features and then reconstructing them. The principle structure of Variational Autoencoders (VAEs) is similar, directly reconstructing the extracted features by building a reconstruction loss. Diffusion models work in the same way, adding noise and then reconstructing the noisy features. Although these generative models employ various ingenious designs when constructing loss functions, they essentially aim to achieve reconstruction loss. The calculation of reconstruction loss is typically done using the Mean Squared Error (MSE) function, which is also a method of image similarity comparison. Therefore, the overall approach adopted in this paper is reverse inference through similarity comparison.

MAE (Masked Autoencoder-Decode) is a self-supervised learning and deep learning method. We know that image similarity can be compared, and similarly, features extracted by deep learning can also be compared for similarity. The features extracted by deep learning often contain highly integrated information. Therefore, calculating the similarity of extracted features is a worthwhile approach to consider. Feature extraction networks need to be trained to accurately extract information from images. Although we can directly use models pre-trained on ImageNet for feature extraction and similarity calculation, the results are not satisfactory because there is a significant difference between ImageNet data and medical data.

To address this issue, unsupervised learning or self-supervised learning methods can be considered. As shown in Figure 2, this is the structure of our overall network model. MAE is a self-supervised learning method that treats both data and labels as inputs for model reconstruction. The uniqueness of MAE lies in its approach of masking part of the information and then reconstructing it. This is an excellent idea because, in data with high image similarity, masking part of the information forces the model to focus more on details. This increases the difficulty of reconstruction, allowing the model to learn to distinguish image details from a small amount of data. This approach ensures that the model can approximate the true image even with occlusions, thus emphasizing detailed information. Consequently, we can calculate the similarity of the extracted features.

3.4. Feature Similarity

After completing feature extraction, similarity calculation becomes a critical step for image recognition and classification. This process aims to measure how close two images are in the feature space, and it employs common similarity measures such as Euclidean distance, cosine similarity, and Manhattan distance to achieve this. Each of these measures has its own advantages and applications depending on the nature of the data and the specific requirements of the task.

The specific steps for similarity calculation using features extracted by the MAE network are as follows. First, acquire the feature vectors for each image to be compared by using a pre-trained MAE network. This involves passing the images through the encoder part of the MAE network, which compresses the input into a latent representation capturing the essential features of the image.

Next, choose an appropriate similarity measure. For instance, cosine similarity can be particularly useful in cases where the magnitude of the feature vectors is not as important as the orientation, making it ideal for assessing the angle between vectors in high-dimensional space. Alternatively, Euclidean distance might be preferred for tasks where the absolute differences in feature values are more meaningful. Once the similarity measure is selected, calculate the distance or similarity score between the feature vectors of the two images. This score quantitatively expresses how similar or different the images are based on their extracted features.

Finally, based on the calculated similarity scores, perform operations such as classification, clustering, or retrieval of images. In classification tasks, images can be assigned to predefined categories based on

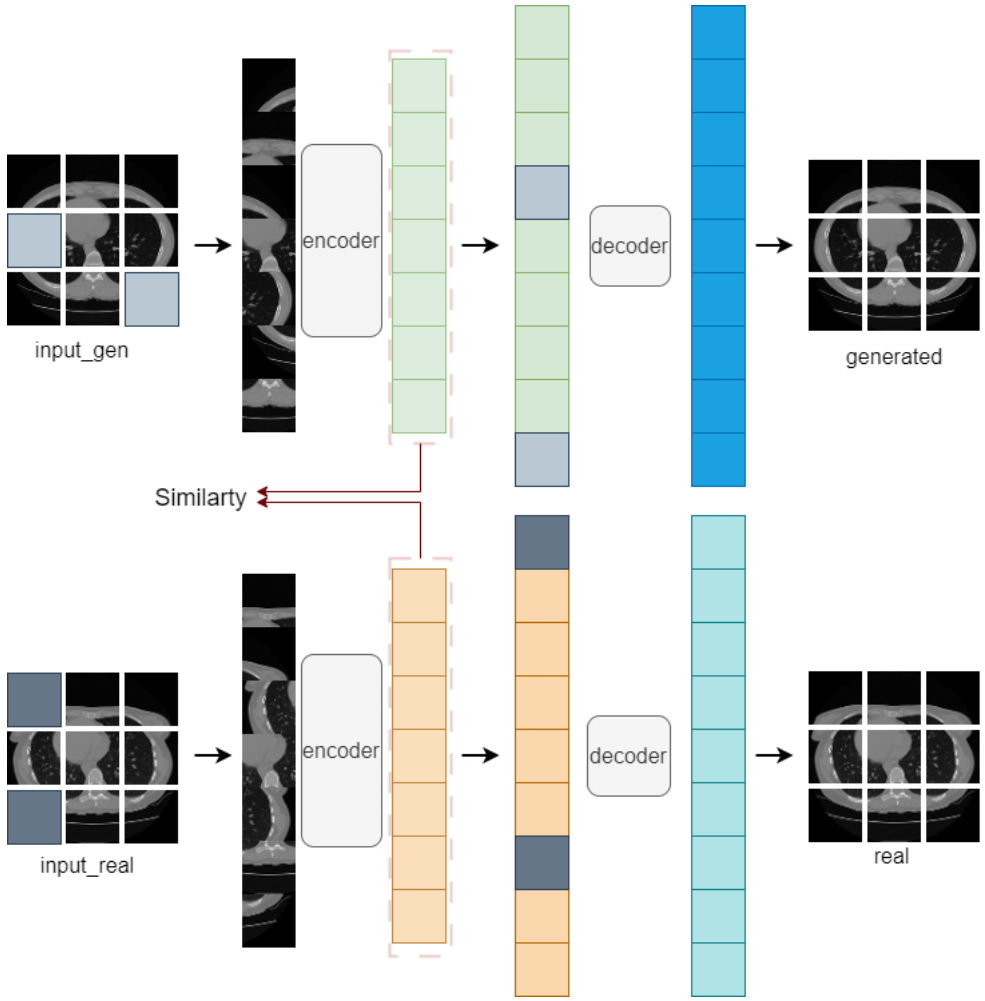


Figure 2: Comparative Learning Diagram. The generated image is obtained from the used image, so the two are similar in comparative learning. When comparing the generated image with the not used image for learning, it is dissimilar.

their similarity to representative examples. In clustering, images are grouped into clusters of similar items, which can reveal inherent structures in the data without prior labeling. For image retrieval, the similarity scores can be used to rank a database of images, retrieving those that are most similar to a given query image. This comprehensive approach ensures that the images are analyzed and utilized effectively, leveraging the power of MAE-based feature extraction and similarity calculation to enhance various image processing tasks.

4. Experiments

4.1. Experimental Design

In our experiment, we used the NVIDIA GeForce RTX 3090 graphics card to complete two tasks. The detailed process and time for each experiment are as follows:

Task 1: To ensure the accuracy and efficiency of the experiment, we loaded all the data onto the RTX 3090 graphics card for processing. Throughout the experiment, we leveraged its powerful computational capabilities and efficient parallel processing features, significantly enhancing the data processing speed. After multiple iterations and optimizations, the total experiment time for Task 1 was approximately 15 hours. During this period, the graphics card operated efficiently, ensuring the integrity of the data and the reliability of the experimental results.

Task 2: Following the completion of Task 1, we continued to utilize the RTX 3090 graphics card

for the second experiment. Similar to Task 1, we performed multiple data loading and processing operations, fully exploiting the card's advantages in deep learning and large-scale data processing. Through repeated experiments and optimizations, we successfully completed Task 2 in approximately 16 hours. Throughout this period, the graphics card maintained high efficiency, ensuring the continuity of the experiment and the consistency of the results.

4.2. Evaluation Metrics

This task was approached as a binary classification problem, and its evaluation involved several key performance metrics: F1-score, accuracy, precision, recall, and specificity. Among these, the F1-score has been designated as the primary metric for this year's evaluation. The definitions of these metrics are as follows:

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

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

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

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

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

4.3. Experimental Results

To conduct a more comprehensive and detailed experimental analysis, we divided the validation dataset into two parts: one as the validation set and the other as the test set. This division allows the validation set to be used for parameter tuning and performance evaluation of the model, ensuring that adjustments made during training are effective. Meanwhile, the test set is used for the final performance evaluation to assess the model's generalization ability on unseen data. This approach helps us to more accurately measure the actual performance and stability of the model, thereby obtaining more reliable and representative experimental results. The experimental results are shown in the table 3.

4.3.1. Ablation Study

Table 3
The average of the experimental results of task1 and task2 on the development dataset. We used different feature extraction models for ablation experiments.

Model	Accuracy	Precision	Specificity	Recall	F1-score
VGG[13]	0.655	0.652	0.661	0.706	0.676
InceptionNet[14]	0.616	0.637	0.590	0.850	0.677
Resnet50[15]	0.632	0.640	0.648	0.652	0.651
Resnet101[15]	0.650	0.873	0.812	0.484	0.587
Mobilenetv2[16]	0.721	0.732	0.748	0.715	0.720
Mobilenetv3[17]	0.722	0.717	0.744	0.694	0.708
EfficientNet[18]	0.648	0.752	0.815	0.487	0.644
MAE[7]	0.743	0.721	0.733	0.720	0.726

Table 4
Scores of the eight different submitted results.

Submission	Model	F1	Acc	Prec	Recall	F1	Acc	Prec Recall	F1
VGG	0.312	0.583	0.812	0.216	0.341	0.559	0.761	0.174	0.283
InceptionNet	0.314	0.583	0.812	0.216	0.341	0.559	0.747	0.178	0.287
ResNet50	0.503	0.503	0.504	0.409	0.451	0.504	0.503	0.619	0.555
Mobilenetv2	0.350	0.615	0.600	0.688	0.641	0.504	0.579	0.031	0.058
Mobilenetv3	0.429	0.503	0.504	0.409	0.451	0.593	0.751	0.279	0.407
EfficientNet	0.524	0.615	0.600	0.688	0.641	0.593	0.751	0.279	0.407
MAE	0.603	0.711	0.824	0.538	0.651	0.504	0.503	0.619	0.555

We conducted an ablation study on different feature extraction modules to evaluate their performance in image classification tasks. Specifically, we used VGG, InceptionNet, ResNet50, ResNet101, MobileNetV2, MobileNetV3, EfficientNet, and MAE pre-trained models for feature extraction. Subsequently, we performed similarity calculations to classify the images based on these features.

To comprehensively assess the effectiveness of these feature extraction modules, we evaluated multiple metrics including accuracy, precision, specificity, recall, and F1-score. By calculating and analyzing these metrics, we were able to compare the strengths and weaknesses of each model in similarity computation and image classification tasks.

The data in the table 3 indicates that the MAE pre-trained model achieved the best performance. This model excelled across all the evaluated metrics, demonstrating its robust capability in feature extraction and image classification. These results suggest that the MAE pre-trained model not only captures detailed features of images but also effectively performs classification tasks, providing strong support for future research and applications.

4.3.2. Submission

We submitted results for the IMAGECLEFmed GANS 2024: Identify Training Data Fingerprints competition. Each submission file had to contain predictions (1 - used, 0 - not used) for all 4,000 images generated by each model. The evaluation method primarily used the F1-score as the evaluation metric, while accuracy was used as the secondary metric. In total, we submitted eight results, with our best score being an F1-score of 0.711.

As shown in Table 4, the submitted results exhibited significant score differences, which may be due to the selection of less optimal features. When classifying based on the similarity scores, the resulting score differences were considerable. However, the overall experimental results indicate that our method is capable of identifying which real images were used to train the image generation model from the generated images.

The results obtained from the development dataset differed somewhat from those of the test dataset, likely due to dimensional variations between the datasets. Despite these differences, we successfully developed methods that achieved high F1-scores and accuracy in identifying images used across both datasets. These findings reinforce the hypothesis that synthetic images generated by deep generative models can potentially expose patient identities.

5. Conclusion

We performed feature extraction on the images, using advanced deep learning models to obtain high-dimensional feature representations. Subsequently, we calculated the similarity of these extracted features, using cosine similarity to evaluate the similarity scores between each pair of image features. Based on these similarity scores, we accomplished the binary classification task, categorizing the images as either used or unused. This method allows us to effectively identify and classify images, providing a solid foundation for subsequent image processing and analysis.

In conclusion, this paper and the ImageCLEFmed GANS challenge contribute to raising awareness about the potential privacy risks associated with the use and sharing of synthetic medical data in real-world applications. We underscore the importance of implementing privacy protection techniques when developing deep generative models using sensitive medical data.

6. Acknowledgements

This work is supported by Open Project Program of Guangxi Key Laboratory of Digital Infrastructure No.GXDINB2024001.

References

[1] L. Cai, J. Gao, D. Zhao, A review of the application of deep learning in medical image classification and segmentation, Annals of translational medicine 8 (2020).

[2] D. Wang, Y. Zhang, K. Zhang, L. Wang, Focalmix: Semi-supervised learning for 3d medical image detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3951–3960.

[3] K. Ramesh, G. K. Kumar, K. Swapna, D. Datta, S. S. Rajest, A review of medical image segmentation algorithms, EAI Endorsed Transactions on Pervasive Health and Technology 7 (2021) e6–e6.

[4] N. Salem, H. Malik, A. Shams, Medical image enhancement based on histogram algorithms, Procedia Computer Science 163 (2019) 300–311.

[5] 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.

[6] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16000–16009.

[8] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8789–8797.

[9] H. Chung, J. C. Ye, Feature disentanglement in generating three-dimensional structure from two-dimensional slice with slicegan, arXiv preprint arXiv:2105.00194 (2021).

[10] J. T. Guibas, T. S. Virdi, P. S. Li, Synthetic medical images from dual generative adversarial networks, arXiv preprint arXiv:1709.01872 (2017).

[11] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. Garcıa Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[12] A. Andrei, A. Radzhabov, D. Karpenka, Y. Prokopchuk, V. Kovalev, B. Ionescu, H. Müller, Overview of 2024 ImageCLEFmedical GANs Task – Investigating Generative Models' Impact on Biomedical Synthetic Images, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich,

Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[15] 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.
[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.
[17] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1314–1324.
[18] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.