DMK-SSN at ImageCLEF 2023 Medical: Controlling the Quality of Synthetic Medical Images Created via GANs using Machine Learning and Image Hashing Techniques[*]

Notebook for the ImageCLEFmedical GANs Lab at CLEF 2023

Dhivya Subburam[1,*,†], Shriram M SathyaNarayanan[1,†], Bhavana Anand[1,†], Kavitha Srinivasan[1,†] and Mohanavalli Subramaniam[1,†]

[1]*Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Tamil Nadu, India*

Abstract

ImageCLEF, a platform dedicated to the evaluation and advancement of visual media analysis, has conducted a novel challenge in its medical track titled Controlling the Quality of Synthetic Medical Images created via GANs. This task investigates the hypothesis whether Generative Adversarial Networks (GANs) generated synthetic images retain distinct characteristics indicative of the real images used for training. This research aims to detect these characteristics and analyze the relationship between real and artificial biomedical image datasets. The task focuses on analyzing test image datasets to identify the presence of real images in the training process. The development dataset includes both artificial and real images of lung tuberculosis patients. Our team submitted two methods for evaluation: a CNN model achieved an F1 score of 0.4804, while the SIFT-KNN approach obtained an F1 score of 0.449. In addition to the models stated above, the pHash approach was applied which enabled to map real images to their corresponding synthetic counterparts. These findings provide valuable insights into detecting real image characteristics in synthetic biomedical datasets, highlighting the importance of addressing security concerns when employing GANs for medical image generation.

Keywords

ImageCLEF2023, Lung CT axial slices, Image Hashing, Convolutional Neural Network, SIFT+KNN

1. Introduction

ImageCLEF is an evaluation forum organized annually that encompasses research tasks oriented towards image analysis and cross-language annotation. ImageCLEF 2023 [1], focussed on various challenges which aims towards the amelioration of research contributions in visual analysis, annotation, classification, and retrieval tasks. The medical based tasks has been included from the second edition of the ImageCLEF under the tag ImageCLEFMedical[2] which annually hosted several medical domain based tasks for major achievements since 2004. Amongst the four tasks proposed for the year 2023, evaluating the synthetic images generated by the Generative adversarial Networks (GANs) [3] is indeed a completely new challenge in the track. The objective mainly focussed on evaluating the prevalent theory that the GANs produce synthetic images with the fingerprints of the real images employed during the GANs training. Thus, the goal of this task is to identify whether a GAN generated image do possess the fingerprint of the real image, if so then the synthetic biomedical images are also subject to the restrictions for usage and sharing similar to the actual real data. However, if the hypothesis is incorrect, then GANs could be used for data augmentation in order to produce extensive biomedical databases that are exempted from any privacy or ethical restrictions. This document illustrates DMK-SSN's participation in the ImageCLEF medical 2023 for the task Controlling the Quality of Synthetic Medical Images created via GANs[2].

Since the discovery of x-rays, medical imaging has achieved significant advancements in medicine and fundamentally revolutionised the way diseases are identified and treated[4][5]. The most common imaging modalities used to study the internal workings of the human body are computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound scanning. Shallow and deep learning techniques [6] have advanced dramatically in discriminative tasks across the key facets of healthcare because of the development of Computer Aided Diagnosis (CADs). One of the biggest difficulties encountered while applying these algorithms for disease diagnosis is the lack of patient datasets for model training. This presents a major risk, particularly in disease diagnosis. Additionally, models struggle to learn the underlying pattern in the datasets during the training phase if there is an uneven distribution of the classes. To overcome these challenges, several techniques are employed to scale up the databases. Traditional approaches involved geometric transformations such as rotation, scaling, translation, flipping, shearing and other approaches. However, deep model-based image synthesis can be achieved using GANs and its variations. In the following, we first describe related works involved in image augmentation and their impact on classification using deep models, to compare fingerprints between two images and image hashing approaches in Section 2, followed by the description of the dataset provided for ImageCLEF medical 2023 for the task Controlling the Quality of Synthetic Medical Images created via GANs in Section 3. In Section 4, we describe the details of methods employed, and Section 5 describes the experiments and results. Section 6 elucidates the conclusion for this task.

2. Related Works

Deep learning [7] have led the GAN has a lot of potential for medical imaging challenges because of its capacity to produce high quality, realistic, factual images. Several pioneering papers such as [8], [9] [10]have demonstrated the effectiveness of GANs in tasks such as image generation, data augmentation, and quality control. In paper [11] utilized deep convolution GANs for data augmentation in CT images of 182 liver lesions (53 cysts, 64 metastases and 65 haemangiomas). In paper [12] have employed data augmentation for an improved lymph node segmentation in the CT images. Several works have shown an improved results in the segmentation and classification of images or lesions. In general, GANs include two networks, a discriminator which discriminates a real image from the generated image and a generator network which is capable to produce synthetic images from random noise by incorporating the feedback images. Thus, if the images are produced by learning the underlying patterns, then the generated images completely mimic the real images. To identify the relationship between a real image and the synthetic image can be identified by analysing the fingerprints of the two images. This could be achieved using certain hash functions on both the image. By employing these techniques, the ideal outcome is to create a distinct fingerprint or the hash value pertaining to that image. This assists to enable a quick search for the related generated images from the dataset. Thus, the binary values which are part of the hash values indicates which real image corresponds to the synthetic images. The GAN image hash and the Real image hash are compared using the hamming distance. The Hamming distance counts how many bits in two hash values' corresponding places are different from one another[13]. Greater similarity between the images is suggested by a smaller Hamming distance, whilst greater dissimilarity is suggested by a higher Hamming distance. Instead of stating the criterion directly, we determine the pair with the smallest hamming distance and mark it as similar.

3. Datasets

For the ImageCLEF 2023, for the task Controlling the Quality of Synthetic Medical Images created via GANs , the organizers shared us the benchmarking image set collected from 8000 lung tuberculosis patients. The imaging modality is CT, and the dataset comprises of axial slices of 3D CT scans obtained from those patients. Amongst them, there are normal and as well with lung lesions, or even more serious stages. The images are of 256×256 pixels and are in PNG file format. Diffuse neural networks are employed in the generation of synthetic images which are of same resolution similar to the real images. The initial development dataset comprised 500 synthetic images, around 80 images which were not used for the training of generative models and 80 real images used for training the generative models. However, the test dataset also included images using the generative models. In the test dataset, the synthetic images are 10,000 and the real images composed of 200 images, where the images are not segregated as used for generative and not used for generative training. Figure 1 illustrates a sample image from all three classes of generated, used and not used.
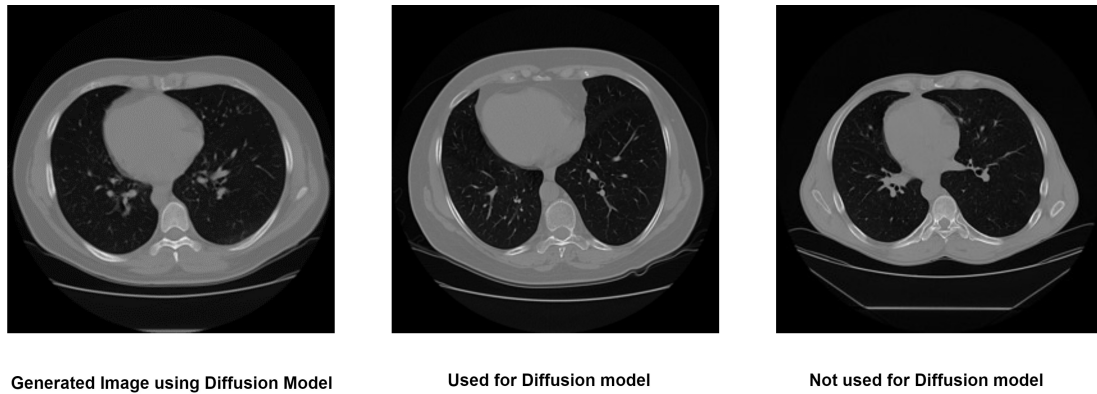
| Generated Image using Diffusion Model | Used for Diffusion model | Not used for Diffusion model |

Figure 1: Sample images from the dataset

4. Methods

With the development dataset shared as a part of the ImageCLEF Medical task we analysed the fingerprints between the two images, i.e generated image and the real images. The unique fingerprints were identified using the Perceptual Hashing algorithm illustrated in Figure 2. The pHash algorithm [13] was applied on the generated images, which produces a hash, a fingerprint of the image. Similarly, hashes for the Used and Unused images were generated. Note, the Used and Unused images are a part of the training dataset. Upon finding the hashing, hamming distance was computed between the used images and the GAN images. Once the hamming distance was computed, we check the closest pair and declare it to form a pair of source image and generated image. This observation was used to check the pattern of hamming distance between source and generated images. It was evident that the hamming distance values circled around 50. With this observation, we perform the same algorithm on the test datasets that consists of the real and generated images. Here upon calculating the minimum distance, we pass a threshold of 50 as observed. The source image will be found for anything below the threshold and the real image will be declared Unused if the minimum distance is more than the threshold. This approach is used as a proof of concept to check whether the images are appropriately classified as used or unused. The primary models in this section include a Convolutional Neural Network and a Scale-Invariant Feature Transform (SIFT) algorithm with the K-Nearest Neighbors (KNN) classifier. The hashing technique is a means to verify the results and acknowledge the outputs from these classifiers.

4.1. Model 1 : Convolutional Neural Network

The first model devised for this task is based on a Convolutional Neural Network (CNN) architecture. CNNs are widely used for image classification tasks due to their ability to effectively capture spatial features from images. The model begins by loading and preprocessing the images. Resizing the images to a uniform size ensures that all images have the same dimensions, which is essential for further processing. The images are resized to a specific width and height, in this case, 64×64 pixels. By resizing the images, we eliminate any variations in size that could hinder
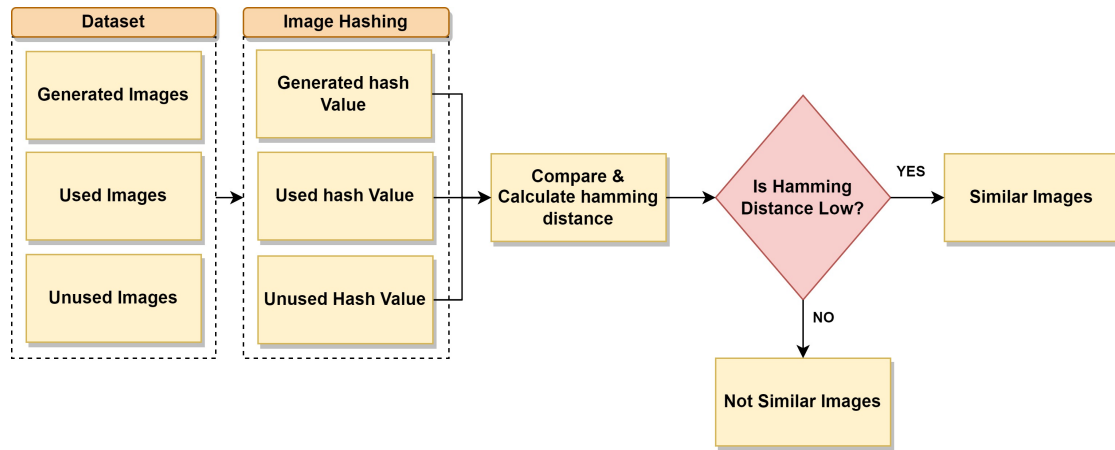
Figure 2: Image hashing approach on development dataset

the training process. After resizing, the training images are divided into two categories: used images and unused images. These categories are represented by the labels 1 and 0, respectively and are finally concatenated as train labels. This prepares the data for training the CNN model. The CNN model architecture consists of three convolutional layers, each followed by a max pooling layer. Convolutional layers apply filters to the input images, capturing local patterns and features. The max pooling layers downsample the output of the convolutional layers, reducing the spatial dimensions and retaining the most important features. This helps in reducing the computational complexity of the model. After the convolutional and max pooling layers, the output is flattened and passed through two dense layers. The flattened output is fed into the dense layers, which are fully connected layers. The activation function used in the convolutional layers and the first dense layer is the Rectified Linear Unit (ReLU) function, which introduces non-linearity and helps in capturing complex relationships in the data. The final dense layer uses a sigmoid activation function, which squashes the output to a value between 0 and 1, representing the probability of the image belonging to the used category. During training, the model is fine-tuned by attempting to minimize the binary cross entropy loss. This loss function measures the dissimilarity between the predicted probabilities and the true labels. The optimizer used for minimizing the loss is ADAM, which is an efficient optimization algorithm commonly used for training deep neural networks. In the testing phase, a threshold of 0.5 is used to convert the predicted probabilities into binary predictions. If the predicted probability is greater than or equal to 0.5, the image is classified as used (1), otherwise, it is classified as unused (0). These predictions are then printed out. The use of a CNN model along with appropriate preprocessing techniques and activation functions allows for the effective classification of images into used and unused categories. By training the model on a labeled dataset and fine-tuning it with optimization techniques, the model can learn to distinguish between the features of used and unused images. The resulting predictions provide insights into whether an unknown image was used to generate GAN images or not. The proposed models are depicted in Figure 3.
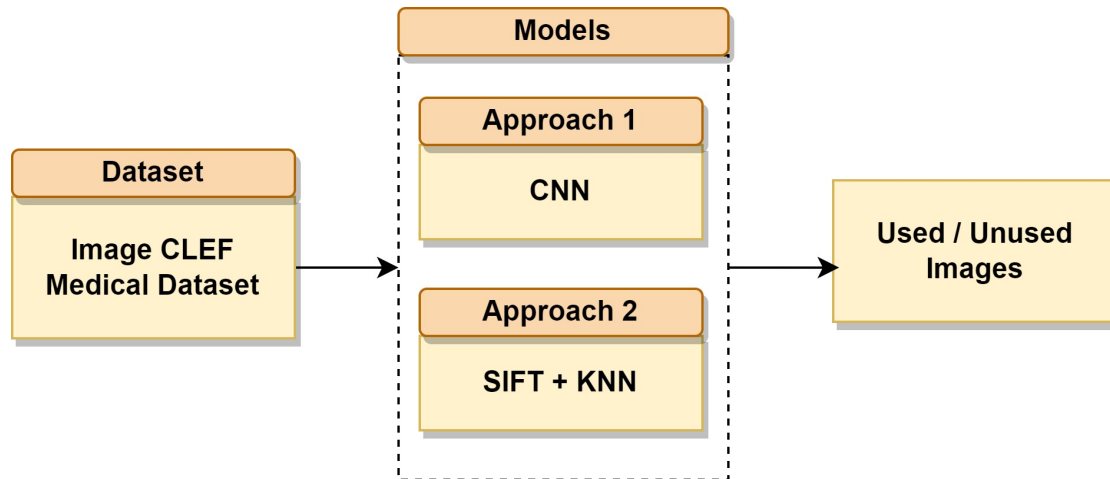
Figure 3: Image hashing approach on the development dataset

4.2. Model 2 : Scale-Invariant Feature Transform (SIFT) + K-Nearest Neighbors (KNN)

The SIFT + KNN model is a powerful technique for image recognition that combines the Scale-Invariant Feature Transform (SIFT) algorithm with the K-Nearest Neighbors (KNN) classifier. This model leverages the strengths of both algorithms to achieve accurate and efficient image classification. The SIFT algorithm plays a crucial role in the model by extracting robust and distinctive features from images. It identifies key points that are invariant to scale, rotation, and affine transformations. These keypoints are characterized by their scale, orientation, and local descriptors that capture the image's appearance in their local neighborhoods. The SIFT algorithm provides robustness to various challenges such as changes in lighting conditions, partial occlusions, and viewpoint variations. The KNN classifier, on the other hand, serves as the classification mechanism in the model. Given a set of labeled training samples with their corresponding feature vectors extracted using SIFT, the KNN classifier assigns labels to new, unseen samples based on their proximity to the training samples in the feature space. It employs the majority voting scheme among the K nearest neighbors to determine the label of the input sample. In the training phase, the SIFT algorithm was first applied to the GAN images, used images and the unused images. The features were clustered accordingly to create a vocabulary of visual words. Similarly SIFT features were extracted for the test images. This was then passed through our classifier (KNN) to perform the required classification.

4.3. Model 3 : Hashing techniques for verification

An improvement in the given tasks can be made by finding the source image from the set of real images for the GAN images provided. This can be done using hashing techniques, one such being perceptual hashing. This model starts by loading in the given test images, both real and GAN. Perceptual hashing is a technique where Discrete Cosine Transform is used to convert the image into a frequency domain. The low frequency components are then used to

construct hash values. This approach makes pHash robust to changes in scale, rotations and minor modifications in the image. This ideally aims at generating a unique fingerprint (hash value) for the images which allows efficient searching and matching of similar images from our dataset. The hash values generated by pHash are binary strings that represent the perceptual content of the images. Once the hash values are computed and stored, hamming distance between the GAN image hash and Real image hash is performed. The Hamming distance measures the number of bit positions at which the corresponding bits in two hash values differ. A lower Hamming distance indicates a higher similarity between the images, while a higher Hamming distance suggests greater dissimilarity. As mentioned earlier in the paper, a threshold of 50 was used to check if it is a used or unused image. Upon identification the source images are found for the Used images. This approach allows for studying the connections between the generated and real images, shedding light on the effectiveness of the SIFT+KNN and CNN models in determining the source images

5. Experiments and Results

In this section, the proposed models are implemented, and the corresponding performance metrics are discussed.

5.1. System specification

The proposed model is executed on a workstation with a six core Intel processor of 3.9GHZ, 32GB DDR4 RAM and NVIDIA GEFORCE RTX 11GB GPU.

5.2. Results of the proposed methods

The obtained results are illustrated in Table 1. The submission 1 is based on the CNN model, which involves a 3 layered CNN to classify the used and unused images. The model is trained for 10 epochs using the Adam optimizer, and the loss function used is binary cross entropy.. With this model, we achieved an accuracy of 0.535 and a precision of 0.544. The specificity and recall are obtained as 0.64 and 0.43. For submission 2, the SIFT + KNN model was used and this model achieved an accuracy of 0.61, precision of 0.512, specificity of 0.6 and a recall of 0.4 as their performance metrics. Apart from the above models, as mentioned earlier in the paper, pHash technique was used. This technique helped us with the mapping of source images to their generated counterparts as in Figure 4. Below are some of the outputs from the pHash model. The first one was the output when the used set of images from the training dataset was used along with the GAN images. The second one shows the same content for the real images. The images with "None" value show that they were unused in the generation of images illustrated in Figure 5.

6. Conclusion

In this paper, we proposed two approaches using CNN and SIFT+KNN to classify the used and not used images for the generation of synthetic images. The F1 score of the first approach is

```
The source image for generated_0478.png is used_0029.png with 52
The source image for generated_0479.png is used_0044.png with 60
The source image for generated_0480.png is used_0027.png with 34
The source image for generated_0481.png is used_0012.png with 52
The source image for generated_0482.png is used_0067.png with 48
The source image for generated_0483.png is used_0009.png with 56
The source image for generated_0484.png is used_0051.png with 44
The source image for generated_0485.png is used_0023.png with 54
The source image for generated_0486.png is used_0022.png with 62
```

Figure 4: Mapping of source image for the given generated image

```
The source image for generated_00001.png is None with 116
The source image for generated_00002.png is real_unknown_0022.png with 48
The source image for generated_00003.png is None with 116
The source image for generated_00004.png is real_unknown_0064.png with 48
The source image for generated_00005.png is real_unknown_0141.png with 32
The source image for generated_00006.png is None with 116
The source image for generated_00007.png is None with 116
The source image for generated_00008.png is None with 116
The source image for generated_00009.png is None with 116
The source image for generated_00010.png is real_unknown_0014.png with 44
The source image for generated_00011.png is real_unknown_0166.png with 44
```

Figure 5: Mapping the generated images to the given test dataset

Table 1
Results of various loss functions for proposed model

Team	Submission #	accuracy	precision	specificity	recall	f1_score	tp	tn	fp	tn
DMK	submission 1	0.535	0.544	0.64	0.43	0.4804	43	64	36	57
	submission 2	0.61	0.512	0.6	0.4	0.449	40	62	38	60

0.4804 while that of the second approach is 0.449. In terms of accuracy we find that the SIFT and KNN approach has an value of 0.61 while that of the CNN model is 0.535. Along with that, an image hashing technique, which identifies the fingerprint of the images through hash values is discussed. The hamming distance is calculated and based on the value, the generated images can be mapped to their appropriate real images. This assists us to segregate the used and unused images for the training of the diffusion network. In future work, these images can be experimented with using one-shot or few-shot learning, which assists in achieving faster

and more efficient classification of generated images using different imaging modalities.

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, D. J. A. A. A. R. I. C. V. K. A. S. G. I. Nikolaos Papachrysos, Johanna Schöler, 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] A. 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, CEUR-WS.org, Thessaloniki, Greece, 2023.

[3] I. Goodfellow, Nips 2016 tutorial: Generative adversarial networks, arXiv preprint arXiv:1701.00160 (2016).

[4] O. H. Karatas, E. Toy, Three-dimensional imaging techniques: A literature review, European journal of dentistry 8 (2014) 132.

[5] K. Aruleba, G. Obaido, B. Ogbuokiri, A. O. Fadaka, A. Klein, T. A. Adekiya, R. T. Aruleba, Applications of computational methods in biomedical breast cancer imaging diagnostics: A review, Journal of Imaging 6 (2020) 105.

[6] S. Dhivya, R. Jenifer Anjali, S. Mohanavalli, N. Sripriya, K. Srinivasan, Investigations of shallow and deep learning algorithms for tumor detection, in: 2020 IEEE-HYDCON, 2020, pp. 1–5. doi:10.1109/HYDCON48903.2020.9242888.

[7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (2015) 436–444.

[8] S. Dhivya, S. Mohanavalli, S. Karthika, S. Shivani, R. Mageswari, Gan based data augmentation for enhanced tumor classification, in: 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP), IEEE, 2020, pp. 1–5.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, nature 542 (2017) 115–118.

[10] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, Q. Liao, Deep learning for single image super-resolution: A brief review, IEEE Transactions on Multimedia 21 (2019) 3106–3121.

[11] 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.

[12] Y.-X. Tang, Y.-B. Tang, M. Han, J. Xiao, R. M. Summers, Abnormal chest x-ray identification with generative adversarial one-class classifier, in: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), IEEE, 2019, pp. 1358–1361.

[13] A. S. Shaik, R. K. Karsh, M. Islam, R. H. Laskar, A review of hashing based image authentication techniques, Multimedia Tools and Applications (2022) 1–28.