Correlating Biomedical Image Fingerprints between Real and GAN-generated Images using a ResNet Backbone with ML-based Downstream Comparators: ImageCLEFmed GANs 2023

Haricharan Bharathi[1], Anirudh Bhaskar[1], Vishal Venkataramani[1], Karthik Desingu[1,2], and Lekshmi Kalinathan[1][0000−0002−7005−742X]

[1] Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
[2] Corresponding Author
{haricharan2010267,anirudh2010094,vishal2010271}@ssn.edu.in
{karthik19047}@cse.ssn.edu.in
{lekshmik}@ssn.edu.in

Abstract. To address the challenge of ImageCLEFmed GANs 2023, this paper proposes three different workflows based on ResNet feature extractors. Toward identifying fingerprinting patterns between images used to train a GAN and the images generated by the trained GAN, the feature vectors are evaluated using multiple approaches, each slightly more complex that the previous. First, agglomerative 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. The performance of the approaches is evaluated on a blind test set containing 200 real images and 10,000 generated images, predicting which of these images were generated using the real images as a precursor to the generative model. The experimental results demonstrate the effectiveness of the overall approach and that the relational model approach bears the most potential. A Python implementation of the experiments will be made available at: https://github.com/karthik-d/ImageCLEFmedical-gan-2023.

Keywords: Generative Adversarial Network · ImageCLEF · Support Vector Machines · Hierarchical Clustering · Machine Learning · Deep Learning · ResNet · Convolutional Neural Networks · Few-shot Learning · Relation Netwok

1 Introduction

The ImageCLEF GANs task 2023 [1] is an evaluation campaign organised by the CLEF initiative [7]. Being the first edition of the task, the primary objective is to detect similarities between synthetic biomedical images generated by GANs and the real images used for training, in order to assess the probability of specific real images being used in the training process. The main motive behind this task is to assess the potential privacy and security implications of using GAN-generated medical images. Understanding if GANs preserve identifiable information from real images can help determine the risks associated with sharing and using artificial biomedical data in real-life scenarios. The findings can inform the development of ethical guidelines and regulations for the generation and usage of synthetic medical images while safeguarding patient privacy.

With advancements in Machine Learning techniques like GANs, it has become increasingly difficult to visually distinguish between real and fake images. Generative models can produce highly realistic images that mimic the visual characteristics of real images, making it challenging for ML algorithms to identify subtle differences. Limited data for images may result in biased ML models that struggle to generalize well. This can be realised with the amount of data provided for the task. Identifying fake images may require understanding the contextual information beyond pixel-level analysis. Detecting subtle anomalies in scene composition, object interactions, or semantic consistency can be difficult for ML models that primarily rely on low-level visual features.Overcoming these difficulties requires continuous research and development of more robust and adaptable ML algorithms.

This paper provides a solution to the above challenges with the help of an ML technique known as feature extraction. Feature extraction involves the use of pre-trained deep learning models, such as convolutional neural networks (CNNs), to extract high-level features from both real and fake images. The CNNs can capture important patterns and textures that can help in distinguishing between the two. A Python implementation of the experiments will be made available at: https://github.com/karthik-d/ImageCLEFmedical-gan-2023.

2 Related Work

In recent years, Generative Adversarial Networks (GANs) have gained significant attention in the medical field for various image generation and translation tasks. Several studies have explored the use of GANs for medical image synthesis, image-to-image translation, and specifically, the identification or detection of synthetic images. Numerous works have investigated the generation of synthetic medical images using GANs. For instance, Choi et al. (2018) proposed a method called "StarGAN" [3] for multi-domain image synthesis, which was successfully applied to generating diverse and realistic brain MRI images. Meanwhile, the paper by Kench et al. presents SliceGAN [8] , an architecture that utilizes generative adversarial networks (GANs) to generate high-quality 3D datasets from a single representative 2D image.

Synthetic images play a crucial role in the medical field as they offer several important advantages and address significant challenges. First, the generation of synthetic images allows for the augmentation of limited or insufficient datasets. In many medical imaging applications, acquiring large and diverse annotated datasets can be challenging and time-consuming. By generating synthetic images, researchers can expand the training data, thereby improving the robustness and generalization of machine learning models. Second, synthetic images enable the simulation of rare or difficult-to-obtain medical scenarios. Certain conditions or diseases may have low prevalence or be challenging to capture through traditional imaging methods. Synthetic images provide a means to create representative cases, allowing researchers and clinicians to study and understand these conditions better, develop diagnostic tools, and explore treatment strategies. Moreover, synthetic images can address privacy concerns related to patient data. Medical images often contain sensitive information, making it difficult to share or publicly release datasets. By generating synthetic images that preserve the statistical and anatomical properties of real data while removing specific patient information, privacy can be maintained, enabling more open collaboration and facilitating research advancements. In summary, synthetic images are indispensable in the medical field, serving as a valuable resource for data augmentation, rare scenario simulation and privacy preservation. Their utilization empowers researchers, clinicians, and technologists to address critical challenges, enhance diagnostic accuracy, improve patient care, and advance medical imaging technologies.

The paper by Nataraj et al. [10] proposed a novel approach for detecting GAN-generated fake images by combining co-occurrence matrices and deep learning techniques. The authors extracted the co-occurrence matrices on three color channels of the pixel domain and trained a deep convolutional neural network (CNN) model. The experimental results on two diverse GAN datasets, based on image-to-image translations and facial attributes/expressions, demonstrated the promising performance of the proposed approach, achieving over 99% classification accuracy. The approach also exhibited good generalization capabilities when trained on one dataset and tested on the other. GANs have also been utilized for image-to-image translation tasks in the medical domain. For example, The paper by Zhu et al. [14] introduces CycleGAN, an approach for translating images from one domain to another without requiring paired training data.

In summary, prior work has demonstrated the potential of GANs in generating synthetic medical images and performing image-to-image translation tasks. However, the problem of distinguishing synthetic and real medical images remains an active research area, requiring robust methodologies to ensure the reliability and integrity of generated data.

3 Task and Dataset

The objective of this task is to examine the hypothesis that Generative Adversarial Networks (GANs) produce medical images that exhibit similarities to the images used during their training. This addresses concerns regarding the security of personal medical image data in relation to the generation and utilization of artificial images in various real-life scenarios.

The task aims to identify distinctive features or "fingerprints" within synthetic biomedical image data to determine which real images were utilized during the training process to generate the produced images. The analysis involves studying test image datasets and assessing the likelihood that specific real patient images were employed for training the image generators, without assuming the ability to identify artificial images or classify image datasets as real or artificial.

3.1 Dataset

The benchmarking dataset comprises axial CT slice images, specifically focusing on lung tuberculosis patients. This dataset consists of approximately 8000 patients, encompassing a range of images that may appear relatively "normal" or exhibit various lung lesions, including severe cases. The image data is stored as 8-bit per pixel PNG images with dimensions of 256x256 pixels.

The artificial slice images are all 256x256 pixels in size and have been generated using Diffuse Neural Networks, a specific type of generative neural network. The published development dataset for this task consists of 500 artificial images, 80 real images that were not used during the training of the generative neural networks, and 80 real images extracted from the image set that was utilized for training the corresponding generative model.

For the test dataset, a similar approach was followed. However, there is a key distinction in that the two subsets of real images, i.e., those used and not used during training, are combined without disclosing the specific proportions. Consequently, the test dataset comprises a total of 10,000 generated images and 200 real images.

In summary, the benchmarking dataset encompasses a diverse range of lung tuberculosis patient images, while the development and test datasets consist of a combination of artificial and real images, with variations in the composition and undisclosed proportions of the real image subsets.

4 Methods

The following section provides a brief on the machine learning techniques and workflows that were utilized on our experiments toward meeting the objectives of this task.

4.1 Relational Model Approach

Inspired by relation networks used in few-shot learning tasks, a relational model was employed in this workflow. Backbone networks based on ResNet were used as feature extractors. Modeled as a siamese network, it takes a pair of input images comprising a real image and a generated image. The generated feature vectors are concatenated spatially and passed to a relational module that outputs a similarity score between the two images.

4.1.1 Deep Learning Architectures The family of ResNet [13] mdodels were considered as backbones, ResNet-10, ResNet-12, ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152

ResNet101, a widely adopted convolutional neural network model introduced in 2015, addresses the degradation problem encountered in deep networks. This problem refers to the phenomenon where increasing network depth leads to saturated accuracy followed by a rapid decline. To overcome this, ResNet101 incorporates shortcut connections that bypass one or more layers. These connections, inspired by the Highway network, utilize gated shortcuts to regulate the flow of information. By employing these mechanisms, ResNet101 effectively mitigates the degradation problem and improves the overall accuracy of deep neural networks.

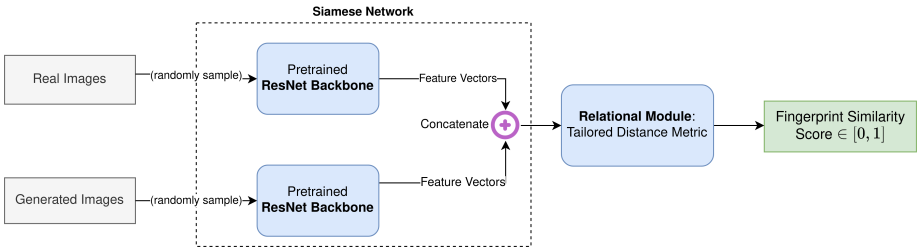

Fig. 1. Workflow described by the relational model where image feature vectors are concatenated and suitable distance metric employed.

4.1.2 Few-shot Learning Approach Few-shot learning [12] is a machine learning approach that deals with the challenge of learning new concepts or classes with limited labeled training data. In traditional machine learning, a substantial amount of labeled data is typically required to train models effectively. Few-shot learning aims to recognise novel visual categories from very few labelled examples. The availability of only one or very few examples challenges the standard 'fine-tuning' practice in deep learning. However, few-shot learning aims to enable learning from a small number of labeled examples, often referred to as the support set or the few-shot training set. The key idea behind few-shot

learning is to leverage the knowledge gained from a larger set of base classes or categories to learn new classes with limited labeled data. This is achieved by exploiting the similarities and transferable knowledge between the base classes and the novel classes. Few-shot learning can be achieved through various techniques, including metric-based approaches, model-based methods, or generative models. These methods often utilize techniques such as siamese networks, meta-learning, or data augmentation strategies to enable effective learning with limited labeled data. [12] deals with the implementation of relation network for few-shot learning.

4.2 Hierarchical Clustering Approach

In order to understand the patterns underlying in the data provided, other methods were implemented. One such method which gave success was using an Unsupervised mode of learning, in particular, forming clusters of the real and generated images. Over the several algorithms of clustering available, the one used is Hierarchical Agglomerative Clustering. Hierarchical Agglomerative clustering is a clustering algorithm that aims to group similar data points into clusters based on their pairwise distances or similarities. It starts with each data point as an individual cluster and iteratively merges the most similar clusters until a termination criterion is met. This merging process continues until all data points are part of a single cluster. Hierarchical clustering has the ability to capture the nested structure of data. It is a bottom up approach which builds clusters from the bottom by merging smaller clusters into larger ones.

The choice of similarity or distance metric is crucial in hierarchical clustering. Commonly used metrics include Euclidean distance, Manhattan distance, cosine similarity, or correlation coefficients, depending on the nature of the data being clustered. The linkage criterion determines how the distances or similarities between clusters are calculated. Some popular linkage criteria are Single Linkage, Complete Linkage, Average Linkage, Ward's Linkage. The type of linkage used in the model is Ward's linkage where the distance between two clusters is determined by the increase in the total within-cluster sum of squares that would result from merging the clusters. Although it is computationally expensive, it can perform on the given dataset with ease.

4.3 Support Vector Machine Approach

Another method used encompasses the Supervised mode of Learning with the usage of Support Vector Machines. SVM works by finding an optimal hyperplane that maximally separates the data points of the two classes in the feature space. The hyperplane is defined by a subset of training examples called support vectors. SVMs aim to achieve a balance between maximizing the margin, i.e., the distance between the hyperplane and the closest data points of each class, and minimizing classification errors. It performs well even in high-dimensional feature spaces, where the number of features is much larger than the number of training examples. SVM can handle non-linear classification problems by employing

kernel functions. Kernel functions transform the data into a higher-dimensional space where a linear hyperplane can be used for separation. Common kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid. The RBF kernel used in the model allows SVM to effectively handle complex, non-linear decision boundaries. It can capture intricate relationships between features and adapt to non-linear patterns in the data. The RBF kernel implicitly maps the input data into a higher-dimensional space, avoiding the need for explicit feature engineering or manual transformation. This makes it suitable for cases where the underlying data structure is not well-defined or the relationship between features is non-linear. SVM is utilized in medical research for disease diagnosis, prognosis, and prediction. It can classify medical data such as patient records, genetic data, or medical images to aid in decision-making and support clinical diagnosis.

5 Experiments

Corresponding to the three approaches described, the following experiments were conducted to evaluate the proposed techniques.

5.1 Relational Model Approach

An abstract base class for a few-shot learning model was introduced where the backbone model ResNet-101 used for feature extraction [9] were instantiated taking a support set as an input. Feature maps were extracted from both images using separate backbone networks with identical architcture. These feature maps were then concatenated to form a comparison candidate tensor. Subsequently, the comparison candidate tensor was passed through the relation module, which is a convolutional neural network that produces relation scores indicating the relationship or similarity between the two images.

The relation module is defined as a simple convolutional neural network with two convolutional blocks, each comprising a 2D convolution, a batch normalization, and a max-pooling layer. A rectified linear unit activation is used after each batch normalization step. The final output is passed through sigmoid activation to limit the output between 0 and 1.

The network is trained end-to-end. While training, the similarity score is set to 1 when a "used" real image is passed as input along with a generated image, and 0, in the case of an "unused" real image. Cross-entropy loss is used to back-propagate. Over the training phase, the relation network learns a task-specific feature comparison metric, while the backbones learn to extract features that are likely to "fingerprint" the real images used to prepare the underlying diffusion network model. During the testing phase, the inputs to the siamese network is again a pair of images, the first a real image, and the second a generated image. The generated similarity score from the relational module is thresholded at 0.5, predicting a match for over the threshold, and a mismatch for under.

5.2 Hierarchical Clustering Approach

The ResNet50 model pre-trained on the ImageNet dataset is loaded. The model is configured to exclude the top dense layer and use global average pooling for feature extraction. Feature extraction is performed by normalization of the pixel values of the images from 0 to 1 and feeding them to the ResNet50 model. Feature agglomeration is used to reduce the dimensions of the feature vectors. A label of 0 is assigned to the 80 not used images, a label of 1 is assigned to the 80 used images. Finally, Hierarchical Agglomerative clustering [11] with the ward linkage method is implemented on the feature vectors of the training images. The labels for the 200 real test images are predicted by applying the trained clustering model to their feature vectors. A Dendrogram [2] is plotted which computes the Cophenetic correlation coefficient [5] to evaluate the hierarchical clustering model's quality.

5.3 Support Vector Machines

A ResNet101 model is employed without its top dense layer to extract feature vectors from each image. The train images are rescaled to lie between 0 and 1 and their features are extracted using the employed model. Additional feature vectors are created by concatenating the feature vectors from the 80 used and 80 not used training images with the feature vectors from the 500 generated images. These concatenated feature vectors are used for training a support vector machine [4] (SVM) classifier. The SVM classifier is trained using the concatenated feature vectors and their corresponding labels (1 for used images, 0 for not used images) with the help of rbf kernel [6].The 10,000 generated images along with 200 real images given for testing are preprocessed and fed to the ResNet101 model for feature extraction as well. The trained classifier is then used to predict the labels (used or not used) for the test feature vectors. The prediction is driven by a majority count of 1's or 0's. EigenValue Analysis is then performed to store the maximum eigenvalue for each feature vector of the used training images.

6 Results

Table 1 depicts the official results with the F1-Score being the competition's official metric.

Table 1. Submission runs and their respective scores

Submission#	Accuracy	Precision	Specificity	Recall	F1-Score	TP	TN	FP	FN
1	50.5	50.3	22	79	61.4	79	22	78	21
2	49.5	49.5	44	55	52.1	55	44	56	45
3	52.5	53.7	69	36	43.1	36	69	31	64

The relational model has performed the best among the three similarity based approaches achieving a F1-Score of 61.4 denoted by Submission 1. The relational model used ResNet-101 as the backbone model for feature extraction. We notice a higher number of false positives 78 when compared to the other approaches. This might stem from the fact that the decision boundary was very low or the model was not very complex. Future improvements that can be undertaken are to use a more complex model, ones preferably pre-trained on medical image data and tuning the hyperparameters.

Submission 2 talks about the scores achieved by the Hierarchical clustering approach. It gives us an F1-Score of 52.1 which is the second best amongst the three. The threshold for the number of clusters was decided after plotting a dendogram and taking an approximate threshold.

Submission 3 portrays the Support Vector Machine's performance with an F1-Score of 43.1. The SVM has a higher number of false negatives-64 when compared to other approaches. This might be because the SVM has a very high sensitivity and might also be because the feature vectors that represent the data need to be more comprehensive. A more complex model can be deployed and the hyperparameters can be tunes for further improvement in scores.

7 Conclusion

This paper proposes three different approaches based on ResNet backbones as image feature extractors, and explores different techniques to compare the so obtained feature vectors toward identifying fingerprinting patterns between images used to train a generative model and the images generated by that model. The relation model and SVM approaches show promise, and the authors believe that further work in this direction can yield good results. In addition, the relational model approach sought to fine-tune the feature extraction process to fingerprints across real and generated images. All the more, the relation model unkike the SVM approach, merges features while conserving spatial context. Accordingly, the relation model approach attained the best performance. Hence, a fine-tuned feature extractor along with a custom distance metric adapts most congruously to the task, and with larger amounts of data, hyperparameter tuning, and perhaps a deeper network architecture, it could produce better performance. The proposed approaches as a whole suggest that training deep learning models to extract fingerprinting features by conditioning on the similarity between the representation vectors is a direction worth pursuing, and that higher dimensional representations along with a sufficient amount of data are more effective in identifying these fingerprints.

Acknowledgments

The authors would like to express their gratitude to the Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineer-

ing, Chennai, India (https://www.ssn.edu.in/) for providing the GPU resources for model training and testing.

References

1. Alexandra-Georgiana Andrei, Ahmedkhan Radzhabov, I.C.V.K.B.I., Müller, H.: 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 (September 18-21 2023)
2. CaliŃski, T.: Dendrogram. Wiley StatsRef: Statistics Reference Online (2014)
3. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**, 273–297 (1995)
5. Farris, J.S.: On the cophenetic correlation coefficient. Systematic Zoology **18**(3), 279–285 (1969)
6. Han, S., Qubo, C., Meng, H.: Parameter selection in svm with rbf kernel function. In: World Automation Congress 2012. pp. 1–4. IEEE (2012)
7. Ionescu, B., Müller, H., Drăgulinescu, A., Yim, W., Ben Abacha, A., Snider, N., Adams, G., Yetisgen, M., Rückert, J., Garcıa Seco de Herrera, A., Friedrich, C.M., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Hicks, S.A., Riegler, M.A., Thambawita, V., Storås, A., Pål Halvorsen, N.P., Schöler, J., Jha, D., Andrei, A., Radzhabov, A., Coman, I., Kovalev, V., Stan, A., Ioannidis, G., Manguinhas, H., Ştefan, L., Constantin, M.G., Mihai Dogariu, J.D., Popescu, A.: 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 (September 18-21 2023)
8. Kench, S., Cooper, S.J.: Generating 3d structures from a 2d slice with gan-based dimensionality expansion (2021)
9. Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., Sun, J.: Detnet: Design backbone for object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 334–350 (2018)
10. Nataraj, L., Mohammed, T.M., Chandrasekaran, S., Flenner, A., Bappy, J.H., Roy-Chowdhury, A.K., Manjunath, B.S.: Detecting gan generated fake images using co-occurrence matrices (2019)
11. Nielsen, F., Nielsen, F.: Hierarchical clustering. Introduction to HPC with MPI for Data Science pp. 195–211 (2016)
12. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1199–1208 (2018)
13. Targ, S., Almeida, D., Lyman, K.: Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029 (2016)
14. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)