Chest Radiography Few-Shot Image Synthesis for Automated Pathology Screening Applications

Martim Quintas e Sousa*†, João Pedrosa*†, Joana Rocha*†, Sofia Cardoso Pereira*†,
Ana Maria Mendonça*† and Aurélio Campilho*†
*Institute for Systems and Computer Engineering, Technology and Science (INESC TEC) Porto, Portugal
†Faculty of Engineering of the University of Porto (FEUP), Porto, Portugal

2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669859

***Abstract*—Chest radiography is one of the most ubiquitous imaging modalities, playing an essential role in screening, diagnosis and disease management. However, chest radiography interpretation is a time-consuming and complex task, requiring the availability of experienced radiologists. As such, automated diagnosis systems for pathology detection have been proposed aiming to reduce the burden on radiologists and reduce variability in image interpretation. While promising results have been obtained, particularly since the advent of deep learning, there are significant limitations in the developed solutions, namely the lack of representative data for less frequent pathologies and the learning of biases from the training data, such as patient position, medical devices and other markers as proxies for certain pathologies. The lack of explainability is also a challenge for the adoption of these solutions in clinical practice.**

Generative adversarial networks could play a significant role as a solution for these challenges as they allow to artificially create new realistic images. This way, new synthetic chest radiography images could be used to increase the prevalence of less represented pathology classes and decrease model biases as well as improving the explainability of automatic decisions by generating samples that serve as examples or counter-examples to the image being analysed, ensuring patient privacy.

In this study, a few-shot generative adversarial network is used to generate synthetic chest radiography images. A minimum Fréchet Inception Distance score of 17.83 was obtained, allowing to generate convincing synthetic images. Perceptual validation was then performed by asking multiple readers to classify a mixed set of synthetic and real images. An average accuracy of 83.5% was obtained but a strong dependency on reader experience level was observed. While synthetic images showed structural irregularities, the overall image sharpness was a major factor in the decision of readers. The synthetic images were then validated using a MobileNet abnormality classifier and it was shown that over 99% of images were classified correctly, indicating that the generated images were correctly interpreted by the classifier. Finally, the use of the synthetic images during training of a YOLOv5 pathology detector showed that the addition of the synthetic images led to an improvement of mean average precision of 0.05 across 14 pathologies.

In conclusion, the usage of few-shot generative adversarial networks for chest radiography image generation was shown and tested in multiple scenarios, establishing a baseline for future experiments to increase the applicability of generative models in clinical scenarios of automatic CXR screening and diagnosis tools.

Email: joao.m.pedrosa@inesctec.pt

This work was funded by the ERDF - European Regional Development Fund, through the Programa Operacional Regional do Norte (NORTE 2020) and by National Funds through the FCT - Portuguese Foundation for Science and Technology, I.P. within the scope of the CMU Portugal Program (NORTE-01-0247-FEDER-045905) and UIDB/50014/2020.

I. Introduction

Chest radiography (CXR), also known as x-ray, is one of the most common medical imaging modalities globally, playing an essential role in screening, diagnosis and disease management. In comparison to other imaging modalities, it has significant advantages, namely its wide availability, low cost, portability and low radiation dosage. However, CXR interpretation is a time-consuming and complex task, requiring the availability of experienced radiologists. As such, computer-aided diagnosis (CAD) systems for CXR pathology detection have long been proposed, providing a valuable second opinion for radiologists. The advent of deep learning, as well as the release of large CXR datasets such as ChestXRay-8 [1] and CheXpert [2], have fostered the development of multi-disease detection approaches, while simultaneously improving performance in the detection of single pathologies [2].

While promising results have been obtained [2], there are significant limitations in current solutions. Firstly, the lack of representative data/annotations can hinder the robust training of deep learning approaches. In spite of the large available datasets such as ChestX-ray14 (224,316 images) and MIMIC-CXR (473,064 images), these datasets tend to be highly imbalanced with normal cases and/or more common pathologies being much more represented than other pathologies, which may lead to degraded performance in less-represented pathologies [2]. Furthermore, significant bias sources can be present in the data. For example, the presence of medical devices or even the position of the patient can be interpreted by the algorithm as a proxy for certain pathologies [3] which is highly undesirable as it could lead to unexpected misclassification of CXRs. Secondly, the lack of explainability of the decisions made by deep learning methods hinders the adoption of these techniques in clinical practice. Typical deep learning methods, which rely on convolutional networks and are the current state of the art in terms of performance, have a black-box behaviour and it is challenging to explain their decisions in a human-understandable way.

Generative adversarial networks (GAN) [4] could play a role in both these challenges as they allow to artificially create new examples of images from a learned distribution. In this way, new CXR images can be created which can be used for training, effectively increasing the prevalence of minority classes and decreasing biases of trained models. GANs can

also be used to improve the explainability of deep learning solutions by generating examples that have strong similarities to the image being analysed or, alternatively, examples representing the opposite (counterexamples), which are also of great importance to the understanding of the decision. In contrast to the use of real images, the use of artificial CXR images is particularly important to avoid privacy concerns associated to the use of real images in a clinical setting.

Given that insufficient data quantity/quality is a common issue in medical imaging, GANs have been widely applied in multiple image modalities but their use in CXR applications has been limited. Deep convolutional GANs (DCGAN) [5] have been successfully used for the generation of both normal and pathological CXR images [6], [7], [8]. An Auxiliary Classifier GAN (ACGAN) has also been applied for CXR image synthesis by Waheed et al. in the context of COVID-19 detection [9]. In all four studies, authors demonstrated that the addition of synthetic data during training of deep learning pathology detection classifiers improved classification performance. Other applications of GANs in CXR include rib suppression [10] and the conversion of pathological to normal CXRs and viceversa as an explainability tool [3].

In spite of the promising results obtained, traditional GANs rely on large quantities of data during training to guarantee the generation of realistic images. While this is not an issue for the synthesis of normal CXRs or the most represented pathologies, it can be problematic in minority classes or CXRs presenting rare pathologies with limited data. Given that these classes and pathologies would be the ones that would profit the most from additional data representation, it is of the utmost importance to validate strategies for CXR synthesis in limited data scenarios using few-shot learning GANs.

The goal of this work is thus to test and validate a high-fidelity few-shot learning approach for CXR image synthesis. For this purpose, a lightweight GAN was used to synthethize CXRs, which are then validated by radiologists and non-experts through perceptual analysis and their influence in training/inference scenarios in pathology screening is tested.

II. Methods

A. Dataset

The CXR images used in this study were obtained from the VinDr-CXR dataset [11], a public dataset collected from two major hospitals in Vietnam, Hospital 108 and Hanoi Medical University Hospital. The dataset consists of 18,000 postero-anterior (PA) CXR scans, of which 15,000 were manually annotated by three radiologists. Both the localization of critical findings and the classification of common thoracic diseases was performed by each radiologist independently.

B. Chest Radiography Image Synthesis

CXR image synthesis was performed using a model based on the lightweight GAN (LWGAN) proposed in Liu et al. [12]. The LWGAN is a model inspired on the DCGAN with a minimalistic design with a single convolution layer on each resolution of the generator. It also features skip-layer channel-wise excitation layers, inspired on residual structures [13], to improve gradient transfer across layers without increasing computational cost. The discriminator features an autoencoding block to reconstruct the original image at a downsampled resolution, as well as a high resolution crop, allowing to generate more representative features for the discriminator. For additional details on the LWGAN architecture the reader is referred to the original publication [12].

For CXR image synthesis, the model was modified to generate/discriminate 1-channel images, compatible with CXR grayscale images. Furthermore, to improve the modeling of long-range dependencies in the network, global self-attention (GSA) modules as proposed by Shen et al. [14] were added to the generator and discriminator modules.

Training was performed using a combination of the hinge adversarial loss [15], a mean squared error loss $\mathcal{L}_{AE}$ for the discriminator's autoenconder branches [12] and a gradient penalty loss $\mathcal{L}_{GP}$ [16]:

$$\begin{aligned}\mathcal{L}_D = &- \mathbb{E}_{x\sim I_{real}}\left[min(0, -1 + D(x))\right] \\ &- \mathbb{E}_{x\sim G(z)}\left[min(0, -1 - D(x))\right] \\ &+ \mathcal{L}_{AE} + \lambda\mathcal{L}_{GP}\end{aligned} \tag{1}$$

$$\mathcal{L}_G = -\mathbb{E}_{z\sim\mathcal{N}}\left[D(G(z)\right] \tag{2}$$

where D and G are the discriminator and generator modules, I_{real} is the set of real images, z is a noise vector sampled from a normal distribution $\mathcal{N}$ and λ is a weight given to the gradient penalty loss. $\mathcal{L}_{AE}$ and $\mathcal{L}_{GP}$ are defined as:

$$\mathcal{L}_{AE} = -\mathbb{E}_{x\sim I_{real}}\left[\sum_b \|D_b(x) - P_b(x)\|\right] \tag{3}$$

$$\mathcal{L}_{GP} = -\mathbb{E}_{x\sim I_{real}}[(\|\nabla_x D(x)\| - 1)^2] \tag{4}$$

where $D_b(x)$ is the output of the autoencoding branch b of the discriminator including the processing function applied to the intermediate feature map and P_b is the processing function applied to sample x for autoenconding branch b and $\nabla_x D(x)$ is the gradient of $D(x)$ with regard to input x.

III. Experiments

A. Chest Radiography Image Synthesis

A total of 2,000 CXRs, randomly selected from the dataset, were used to train the LWGAN. Only CXRs labelled as "No Finding" were chosen given that they are representative of generic CXR structures. CXRs were resampled to 512×512 resolution as state-of-the-art pathology detection approaches typically use CXRs at resolutions of 512×512 or lower [2]. Data augmentation, namely translation, horizontal flip, cutouts and brightness changes, was applied to the images during training with a probability of 0.25. The weight factor λ was empirically set to 10 and training was performed with a batch size of 6 and a learning rate of 0.0002 for a maximum of 150 epochs. Four different models were tested: one without GSA

modules and the remaining with GSA modules on the 32×32 layers, on layers 32×32 up to 128×128 and on layers 32×32 up to 512×512.

To evaluate the progression of the generative module during training, the Fréchet Inception Distance (FID) score [17] was computed from a set of 5,000 real CXRs and a set of 5,000 synthetic CXRs. The final generative model was then selected based on the epoch with the minimum FID score. The Inception Score (IS) [18] and the Kernel Inception Distance (KID) [19] were also calculated for the final generative model using the same set of 5,000 synthetic CXRs.

B. Perceptual Validation

In order to complement the quantitative evaluation metrics, qualitative evaluation methods were used, based on the perceptual validation of the images. For this purpose, a mixed set of real images and artificially generated images was created and evaluated by six readers independently: two radiologists (Rad1 and Rad2), two PhD students in the field of medical image analysis acquainted with CXR (PhD1 and PhD2) and two readers inexperienced with CXR and medical imaging (Ixp1 and Ixp2). Images were presented in a random order using a graphical user interface designed for this purpose. Each reader was required to classify each CXR in terms of authenticity, i.e. real or fake. The two radiologists were also required to classify each CXR in terms of abnormality, i.e. normal or pathological.

A total of 100 randomly selected CXRs were used for this validation, of which 50 were artificially generated, 25 were real images labeled as "No Finding" and 25 were real images not labeled as "No Finding", i.e. pathological images. Real images were downsampled to 512×512 to avoid readers from using resolution as a criteria for distinguishing real and synthetic images.

C. Abnormality Classification Inference

In order to establish if the generated images contain features representative of normal CXR images, an abnormality classification model trained on the VinDr-CXR dataset was used. This model is based on a MobileNet architecture [20] with one output node corresponding to the probability that the CXR presented any of the pathologies annotated on VinDr-CXR. For training, the 15,000 CXRs of VinDR-CXR were randomly divided into train (60%), validation (20%) and test (20%) sets, preserving the approximate prevalence of each pathology as much as possible between the three divisions.

Two CXR sets, both composed uniquely of normal CXRs, were given to the classification model for inference: 1) 6,000 synthetic CXRs; 2) 2,120 real CXRs of the test set labeled as "No Finding".

D. Training of a Pathology Detection

In order to establish if the generated images are of sufficient quality for the training of deep learning models, a YOLOv5 object detection architecture [21] was used with the same division of the dataset into training, validation and test as in Section III-C. Additionally, the 6,000 synthetic CXRs generated in Section III-C were used in one of three different training strategies: 1) using only real pathological images; 2) using all real images, both normal and pathological; 3) using both real pathological images and artificially generated normal images. The performance of each of these three scenarios was evaluated in terms of average precision (AP) [22] computed from the precision-recall curve for each of the 14 pathologies annotated on VinDR-CXR at an intersection over union > 0.4.

GSA	FID	IS	KID
None	24.13	2.047 ± 0.0267	0.017 ± 0.001
32×32	**17.83**	2.109 ± 0.034	$\mathbf{0.012 \pm 0.001}$
32×32-128×128	77.22	$\mathbf{2.318 \pm 0.051}$	0.082 ± 0.002
32×32-512×512	52.39	2.065 ± 0.047	0.058 ± 0.002

TABLE I: Quantitative evaluation metrics for each of the CXR image generation models. Bold indicates the best score obtained for each metric.

Reader	Authenticity			Abnormality				
	Acc	Spc	Sns	Acc	Spc	Sns	SpcR	SpcA
Rad1	98	98	98	87	83	100	64	92
Rad2	80	96	64	84	80	96	72	84
PhD1	100	100	100	-	-	-	-	-
PhD2	65	60	70	-	-	-	-	-
Ixp1	91	88	76	-	-	-	-	-
Ixp2	67	76	58	-	-	-	-	-

TABLE II: Perceptual validation results. Acc, Spc and Sns indicate accuracy, specificity and sensitivity, wheras SpcR and SpcA indicate specificity for the real and artificial images respectively. All values in percentage.

IV. Results

A. Chest Radiography Image Synthesis

Table I shows the quantitative evaluation metrics obtained for each of the final CXR image generative modules. It can be seen that the best results in terms of FID and KID are obtained with the model containing GSA modules on the 32×32 layers, whereas the best IS is obtained with GSA modules on the 32×32-128×128 layers. Because the FID and KID are more complete metrics for GAN evaluation, the 32×32 GSA model was used for all experiments in the remainder of this study.

B. Perceptual Validation

Figure 2 shows the confusion matrices of the six readers in terms of authenticity and abnormality classification whereas Table II shows the accuracy, sensitivity and specificity of each reader. Additionally, for abnormality classification, the specificity in abnormality classification was computed separately for real and artificial images.

As seen in Table II, the radiologists obtained the best overall authenticity classification performance, followed by the PhD students and lastly by the inexperienced readers. Nevertheless, a high discrepancy between readers 1 and 2 in each group can be observed, with the best authenticity classification performance obtained by PhD1. Regarding the additional task of abnormality classification, both radiologists correctly labeled

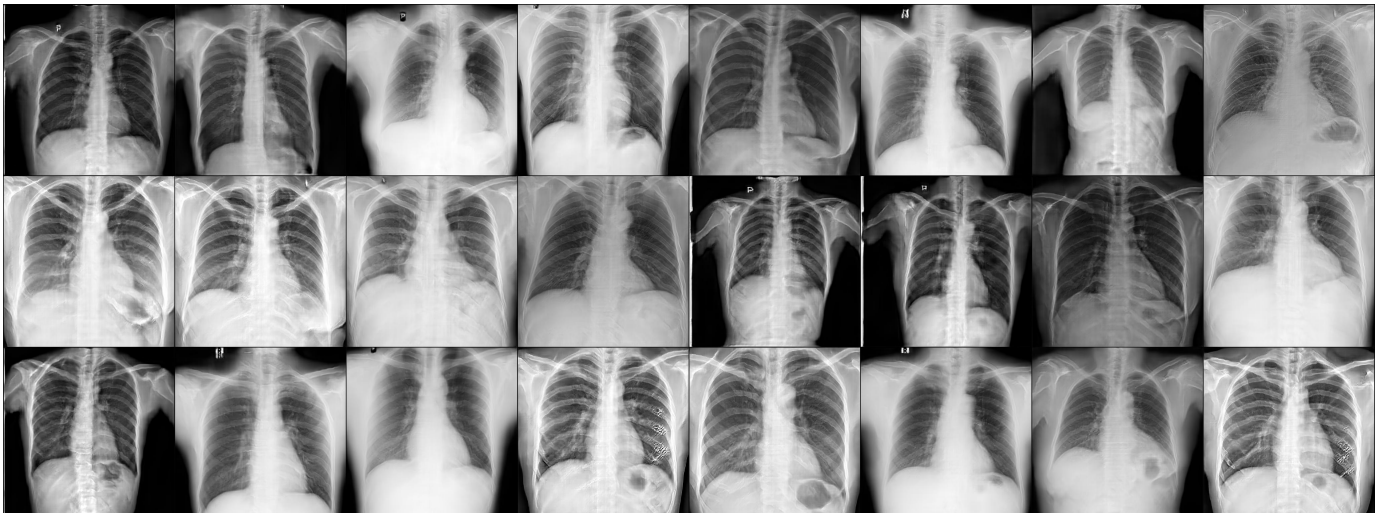


Fig. 1: Randomly sampled synthetic CXRs generated by the model with GSA modules on the 32×32 layer.

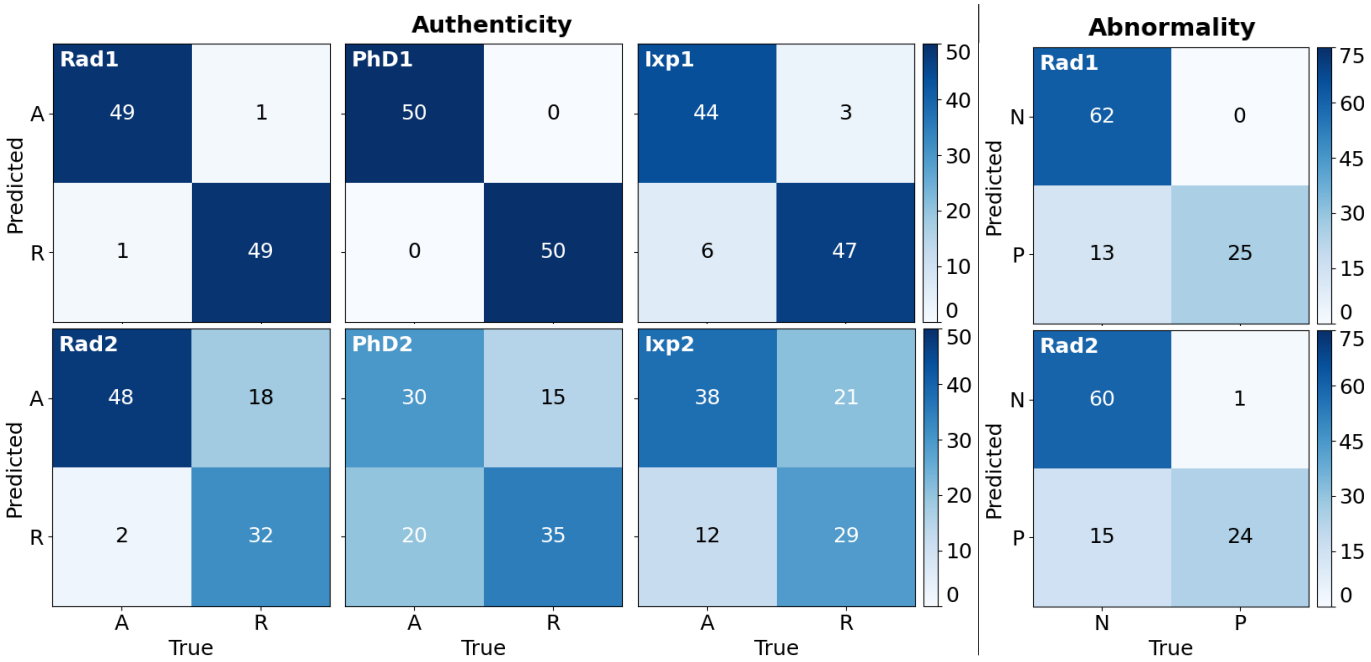


Fig. 2: Authenticity and abnormality classification confusion matrices of all six readers. A and R in the authenticity classification indicate artificial and real CXR images respectively whereas N and P in the abnormality classification indicate normal and pathological CXR images respectively.

most of the pathological images, leading to a high sensitivity. However, both radiologists misclassified a similar number of normal images as pathological, leading to a lower specificity. Comparing the specificity for real and synthetic images in separate, it can be seen that a significantly higher specificity was obtained for the synthetic images, indicating that real images were more often misclassified as pathological by the radiologists than synthetic images.

Figure 3 shows six examples of real and synthetic CXRs used in the perceptual validation. The two leftmost figures show synthetic images which were classified as real by at least one radiologist. The two center figures show artificial CXRs identified as artificial due to distorted ribs (top figure, red arrow) and a distorted clavicle (bottom figure, red arrow). Furthermore, the two center figures were classified as pathological by the radiologists, namely due to an apical assymmetry (top figure, yellow arrow) and cardiomegaly (bottom figure, yellow arrow). The two rightmost figures show real and normal CXRs which were labeled as real and pathological by the radiologists due to cardiomegaly (top figure, yellow arrow) and aortic enlargement (bottom figure, yellow arrow). Note that these CXRs are labeled as "No Finding" in VinDr-CXR, meaning

Image Set	Prediction	
	Normal	Pathological
Real	93.35	6.65
Artificial	99.12	0.88

TABLE III: Abnormality classification prediction of the real and artificial images. All values in percentage.

that these findings had not been identified.

C. Abnormality Classification Inference

Figure 4 shows the relative frequency histograms of the predicted abnormality probability by the MobileNet architecture for the real and artificial images. It can be seen that the set of real images has a higher incidence of very low probability images, corresponding to a prediction of normality. However, the real image set also has a higher incidence of higher probability predictions, corresponding to a prediction of pathology. This is corroborated by Table III which shows the percentage of images predicted as normal and pathological from the real and pathological CXRs. While the majority of images in both sets were correctly classified as normal, a significantly higher proportion of images were incorrectly classified as pathological in the set of real images when compared to the artificial images.

D. Training of a Pathology Detection

Figure 5 shows the AP obtained for each of the pathological classes in each of the three training strategies. It can be seen that the best overall performance was obtained with the model trained with both real normal and real pathological CXRs with a mean AP of 0.381, followed by the model trained with real pathological CXRs and artificial normal CXRs with a mean AP of 0.329, and lastly the model trained only with real pathological CXRs with a mean AP of 0.279.

V. Discussion

A. Chest Radiography Image Synthesis

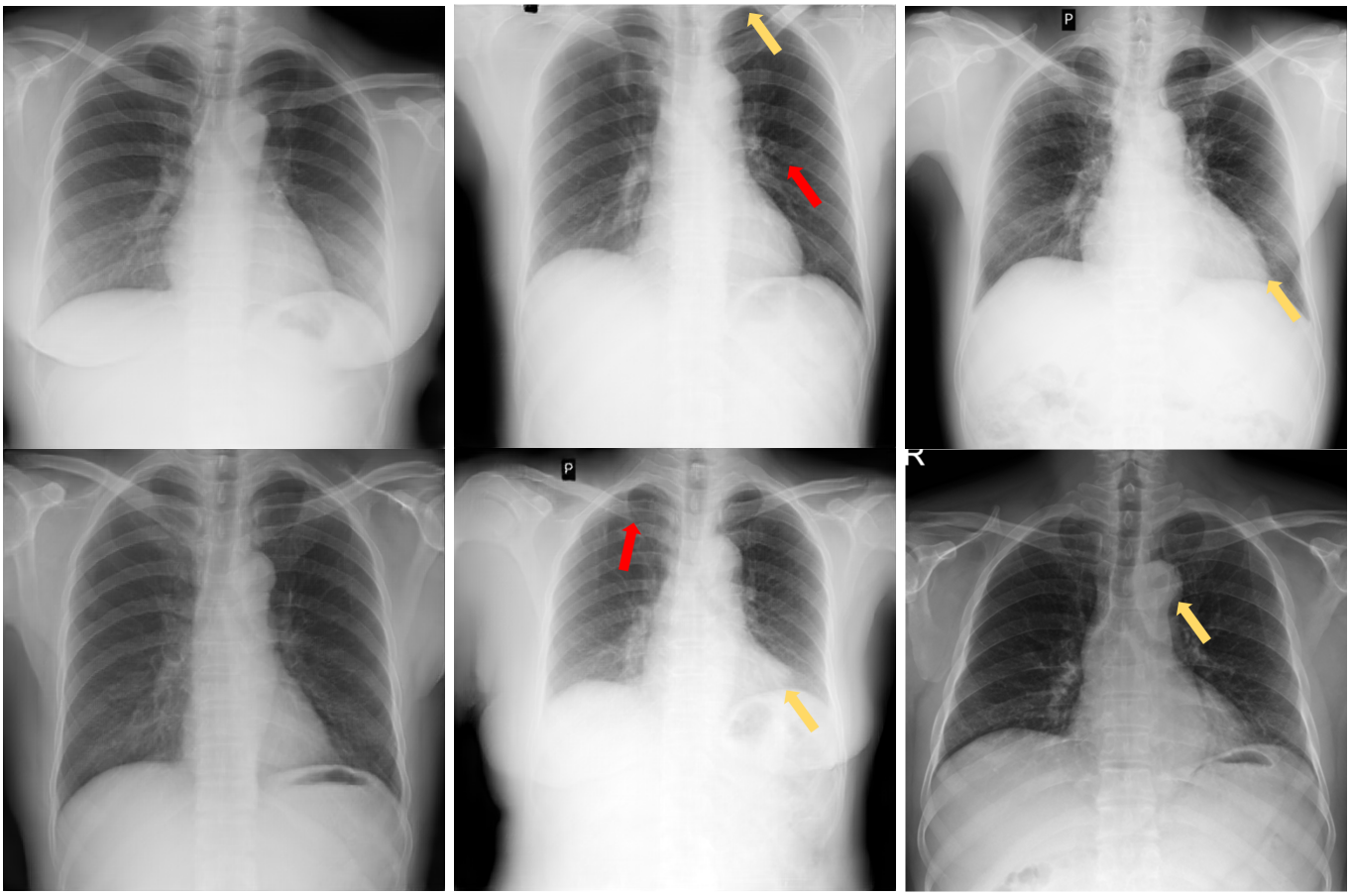
Table I shows that the addition of the GSA modules in the lower resolution layers was beneficial for CXR image generation. However, this was only true for the 32×32 GSA model and both FID and KID metrics were worse when the GSA modules were added to higher resolution layers. It would be expected that adding GSA layers to higher resolution would improve the representation of higher resolution details in the generated images. However, this was not the case and the convergence of the model was significantly degraded in these experiments. Compared to previous work in CXR synthetization, a lower FID was reported in [7] using a DCGAN trained on normal CXRs. However, the fact that a different dataset and Inception layer were used to calculate FID means that a direct comparison is not straightforward.

B. Perceptual Validation

In terms of the perceptual validation, it was observed that real and synthetic images could in the majority of cases be distinguished in spite of the realism of the synthetic images. Nevertheless, a comparison of the performances of radiologists, PhDs and inexperienced readers indicates a strong effect related to the experience of the readers as radiologists obtained the highest accuracy, followed by PhDs and only then the inexperienced individuals.

Regarding the disparities observed between the two readers in each of the groups, it was clear by debriefing with each of the readers after the classification that different strategies were used. While all readers identified structural irregularities in synthetic CXRs, Rad1, PhD1 and Ixp1 took into account the overall appearance and sharpness of the CXR as an indicator of authenticity. In spite of the downsampling applied to the real CXRs, synthetic images were more blurred than real CXRs and while that might not be apparent in Figure 3, it could be identified in a one-to-one comparison in a large monitor. The other three participants - Rad2, PhD2 and Ixp2 - however did not use this factor for authenticity classification. Both Rad2 and PhD2 noted during the debriefing that blurred CXR images or with lower quality can occur in a clinical environment and that this factor alone could not be used to determine if a CXR was real or synthetic. Instead, this second group focused solely on structural irregularities in the CXRs. The main irregularities found were ripples and undulations in the edges of the bones and particularly the ribs, distortion of the clavicles, asymmetries in the images and angulation of other anatomical structures, with some examples shown in Figure 3. This difference in the method used during the classification justifies why Rad2, PhD2 and Ixp2 obtained an average accuracy of 70.6%, which is significantly lower than the average 96.3% obtained by Rad1, PhD1 and Ixp1.

As for the results of the abnormality classification by the radiologists, a very high sensitivity was obtained, meaning that radiologists correctly identified the majority of pathological images. However, specificity was much lower for both radiologists with a significant proportion of normal images being labeled as pathological. Furthermore, most of the normal images labeled as pathological were common between the two radiologists. In the debriefing after classification, both Rad1 and Rad2 confirmed that these images presented findings indicative of pathology, as shown in the examples of Figure 3. Looking separately at the specificity in the sets of synthetic and real CXRs, it can be seen that the presence of findings indicative of pathology was most frequent in the real images, where an average specificity of 68% was obtained compared to 88% for the synthetic CXRs. Additionally, the findings identified by the radiologists on the synthetic CXRs, were mainly distortions in anatomical structures, whereas with real images, these findings were identifiable pathologies belonging to the VinDr-CXR dataset classes such as cardiomegaly, aortic enlargement and enlarged mediastinum or lung opacities and nodules. The fact that these findings are part of the VinDr-CXR dataset classes, and should thus have been annotated on the dataset, means that both Rad1 and Rad2 are in disagreement with the three radiologists that annotated each of these CXRs in the VinDr-CXR dataset. This is once more a testament to the inherent challenge of CXR analysis and the inherent variability



(a) Synthetic CXRs incorrectly classified as real.

(b) Synthetic CXRs correctly classified as artificial but classified as pathological due to structural irregularities.

(c) Real and normal CXRs correctly classified as real but classified as pathological.

Fig. 3: Examples of real and synthetic CXRs and their authenticity and abnormality classification by radiologists. Red and yellow arrows indicate respectively the structural irregularities and pathological findings identified by radiologists as contributing to their decision regarding authenticity and abnormality classification.

between radiologist and is a reminder of the importance of the development of automatic solutions for CXR screening and pathology detection which can improve the variability and performance of radiologists in this task. Furthermore, and particularly since VinDr-CXR is the only large volume dataset with manual annotations, it is important to be aware that this does not preclude the existence of mislabeled CXRs and future studies should have this into account.

C. Abnormality Classification Inference

From Figure 4, it can be concluded that even though human readers could correctly distinguish the majority of synthetic and real images, the synthetic image features can correctly be interpreted by the binary classification model, with the large majority of these images being classified as normal and having a distribution similar to that of real normal images.

As discussed in the previous section, a significant amount of images in the VinDr-CXR dataset are labeled as "No Finding" while, according to Rad1 and Rad2, they show findings indicative of pathology. This mislabeling can once more be seen in the histograms shown in Figure 4 and Table III, where the classification model classifies 6.65% of the real images as abnormal, including some with a high probability of abnormality. The synthetic images, however, do not follow this distribution, with less than 1% of CXRs being classified as pathological. This is in agreement with the higher specificity in abnormality classification by the radiologists in synthetic images in comparison to real images (Table II). Given that the LWGAN was trained on VinDr-CXR data which also likely has a percentage of mislabeled CXRs, it could be expected that synthetic CXRs also showed signs of pathology. However, since the LWGAN learns a statistical representation of the data, the pathological cases are likely considered as outliers during training, leading the generator to mostly generate CXRs which are classified as normal by the binary abnormality classifier and by the radiologists.

D. Training of a Pathology Detection

Regarding the training of the YOLOv5 for pathology detection with synthetic images, when compared to the performance

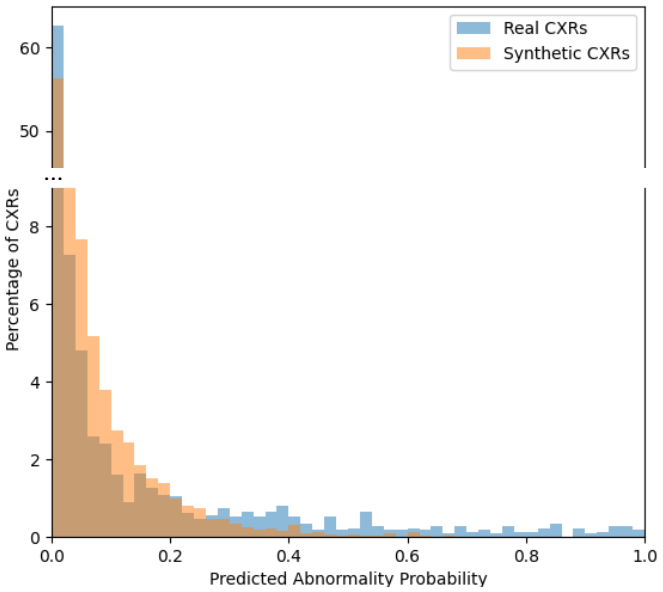


Fig. 4: Relative frequency histogram for the real and artificial according to the binary abnormality classification model.

of the model trained with only real pathological images, Figure 5 shows a clear improvement in performance in almost all classes. Nevertheless, when compared to the model trained with both real normal and real pathological CXRs, it can be seen that this model outperforms the one trained with synthetic images, even though approximately the same number of normal CXRs was used during training. This indicates that, although the synthetic images are not on par with real images, they still provide an improvement on the overall performance of the classification model for pathological classes. Although one could hope that synthetic and real images would provide the same performance improvement, the fact that 2,000 CXRs were used to train the LWGAN instead of the 6,000 real normal CXRs used to train the YOLOv5 is a clear disadvantage for the model trained with synthetic images. Furthermore, and as seen in Section V-B, the trained LWGAN does not seem able to generate images that completely represent and pose as real images, which could have an influence on the result of the classifier trained on synthetic CXRs and the lower performance when compared to the model trained with real normal CXRs.

Compared to previous similar work, the results presented in [6], [7], [8] and [9] show an overall improvement in CXR classification accuracy for both pathological and normal images in models trained with real images supplemented with artificially generated samples as well. However, the number of training samples in the work by Salehinejad et al. was significantly larger than the one used in the LWGAN models.

E. Limitations

Furthermore, while the aim of this manuscript was to study use cases of GANs in CXR, one of the main applications of GANs is the generation of synthetic data for training, particularly for pathological cases in minority classes. However, in this study only normal cases were used for CXR generation, which is a major limitation. As discussed in Section III-A, only normal CXRs were used as these are representative of the generic CXR structures. However, pathological CXRs naturally contain additional challenges, one of which is that, while normal CXRs necessarily belong to a single class - "No Finding" - pathological images typically show multiple findings pertaining to different pathologies and thus multiple classes which might be an issue for data representation during training. Feature disentanglement or conditional GANs could be a possible solution by allowing for multiple pathologies to be generated by the same generative model, which has the advantage that features from prevalent pathologies can be used in the training and generation of images from minority classes. Alternatively, the local annotations of VinDr-CXR could be used to train a patch-based GAN [23] which would be able to learn to generate any of the pathologies in a given region, independently of other pathologies present in that same image.

An additional limitation is the resolution used for image generation. Although the 512×512 resolution appears to be enough for training classification models in current state of the art, future architectures may rely on higher resolutions. Furthermore, if the goal is to fully represent the native data, even regarding resolution, a significant improvement must still be done. While it was not possible to generate higher resolution images with the current LWGAN architecture, it would be important to develop architectures that can work with increasing resolutions and progressively growing GANs are a promising approach [24].

Finally, while the model with GSA modules in the lower resolution layers proved to be the best, the structural irregularities observed by the radiologists during the perceptual validation point to the fact that the synthetic CXRs are not on par with real CXRs. The addition of GSA layers in higher resolution layers could have helped in fixing these issues, but the convergence of the model was not as efficient as in the 32×32 GSA model. It would thus be important for future work to continue investigating few-shot GAN architectures and training schematics that can better generate realistic synthetic data.

VI. Conclusion

In conclusion, a few-shot CXR image generator was proposed in this study and its application in CXR pathology/abnormality detection scenarios tested. It was shown that the LWGAN was able to generate convincing CXRs, but the perceptual validation revealed that most artificial CXRs were detected, with a dependency on reader experience level. While structural irregularities were found, the overall image appearance and sharpness was a major factor in the decision of half of the readers. In spite of this, the image features of the synthetic CXRs were successfully interpreted by a binary abnormality classifier, which correctly identified no findings indicative of pathology. Furthermore, it was shown that using synthetic CXRs in the training of a CXR pathology detector led to a significant improvement of performance. As such,

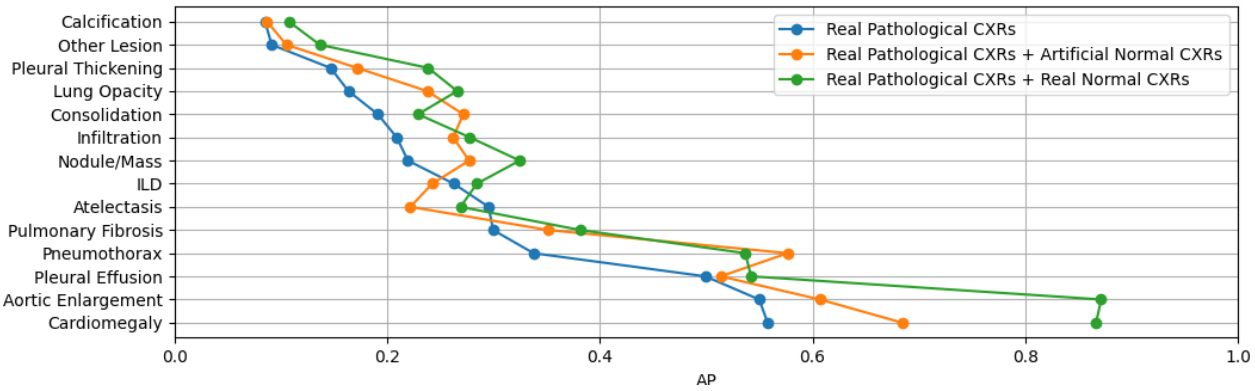


Fig. 5: Pathology detection performance in each of the training scenarios for the 14 pathologies. Pathologies were sorted in terms of performance when trained with real pathological CXRs.

and in spite of the challenges it poses, few-shot CXR image generation could come to play a role in pathology detection in CXR, increasing performance and robustness.

Acknowledgment

The authors would like to thank Joana Silva, MD of the Administração Regional de Saúde do Norte (ARSN), Pedro Sousa, MD of the Centro Hospitalar de Vila Nova de Gaia/Espinho (CHVNGE), Frederica Quintas e Sousa and Alexandre Quintas e Sousa, for their valuable contributions to the perceptual validation and this study.

References

[1] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2097–2106.

[2] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 590–597.

[3] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *medRxiv*, 2020.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[5] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[6] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 990–994.

[7] S. Kora Venu and S. Ravula, "Evaluation of deep convolutional generative adversarial networks for data augmentation of chest x-ray images," *Future Internet*, vol. 13, no. 1, p. 8, 2021.

[8] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Chest x-ray generation and data augmentation for cardiovascular abnormality classification," in *Medical Imaging 2018: Image Processing*, vol. 10574. International Society for Optics and Photonics, 2018, p. 105741M.

[9] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "CovidGAN: Data augmentation using auxiliary classifier GAN for improved COVID-19 detection," *IEEE Access*, vol. 8, pp. 91 916–91 923, 2020.

[10] J. Liang, Y.-X. Tang, Y.-B. Tang, J. Xiao, and R. M. Summers, "Bone suppression on chest radiographs with adversarial learning," in *Medical Imaging 2020: Computer-Aided Diagnosis*, vol. 11314. International Society for Optics and Photonics, 2020, p. 1131409.

[11] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. Tong, D. H. Dinh *et al.*, "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *arXiv preprint arXiv:2012.15029*, 2020.

[12] B. Liu, Y. Zhu, K. Song, and A. Elgammal, "Towards faster and stabilized GAN training for high-fidelity few-shot image synthesis," in *International Conference on Learning Representations*, 2020.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[14] Z. Shen, I. Bello, R. Vemulapalli, X. Jia, and C.-H. Chen, "Global self-attention networks for image recognition," *arXiv preprint arXiv:2010.03019*, 2020.

[15] D. Tran, R. Ranganath, and D. M. Blei, "Deep and hierarchical implicit models," *arXiv preprint arXiv:1702.08896*, vol. 7, no. 3, p. 13, 2017.

[16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," *arXiv preprint arXiv:1704.00028*, 2017.

[17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[18] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.

[19] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," *arXiv preprint arXiv:1801.01401*, 2018.

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[21] G. Jocher, A. Stoken, J. Borovec, A. Chaurasia, L. Changyu, V. Laughing, A. Hogan, J. Hajek, L. Diaconu, Y. Kwon *et al.*, "Ultralytics/YOLOv5," 2021. [Online]. Available: https://github.com/ultralytics/yolov5

[22] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[23] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, "PD-GAN: Probabilistic diverse GAN for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9371–9381.

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.