Advancing medical imaging with GAN-based anomaly detection

Nabila Ounasser[1], Maryem Rhanoui[2,3], Mounia Mikram[3], Bouchra El Asri[1]
[1]IMS Team, ADMIR Laboratory, Rabat IT Center, ENSIAS, Mohammed V University, Rabat, Morocco
[2]Laboratory "Health, Systemic, Process", Research Unit 4129, University Claude Bernard Lyon 1, Villeurbanne, France
[3]Meridian Team, LYRICA Laboratory, School of Information Sciences, Rabat, Morocco

Article Info

Article history:

Received Jan 30, 2024
Revised Mar 16, 2024
Accepted Mar 25, 2024

Keywords:

Anomaly detection
Artificial intelligence
Deep learning
Fracture detection
Generative adversarial networks
Machine learning
Medical imaging

ABSTRACT

Anomaly detection in medical imaging is a complex challenge, exacerbated by limited annotated data. Recent advancements in generative adversarial networks (GANs) offer potential solutions, yet their effectiveness in medical imaging remains largely uncharted. We conducted a targeted exploration of the benefits and constraints associated with GAN-based anomaly detection techniques. Our investigations encompassed experiments employing eight anomaly detection methods on three medical imaging datasets representing diverse modalities and organ/tissue types. These experiments yielded notably diverse results. The results exhibited significant variability, with metrics spanning a wide range (area under the curve (AUC): 0.475-0.991; sensitivity: 0.17-0.98; specificity: 0.14-0.97). Furthermore, we offer guidance for implementing anomaly detection models in medical imaging and anticipate pivotal avenues for future research. Results unveil varying performances, influenced by factors like dataset size, anomaly subtlety, and dispersion. Our findings provide insights into the complex landscape of anomaly detection in medical imaging, offering recommendations for future research and deployment.

Corresponding Author:

Nabila Ounasser
IMS Team, ADMIR Laboratory, Rabat IT Center, ENSIAS, Mohammed V University
Rabat, Morocco
Email: nabilaounasser81@gmail.com

1. INTRODUCTION

The intersection of artificial intelligence (AI) and medical imaging has led to groundbreaking advancements in disease diagnosis and ealthcare [1], [2]. Among the various AI methodologies, generative adversarial networks (GANs) have emerged as a powerful tool for the detection of anomalies [3]–[5]. GANs [6], which were originally designed for image generation, have demonstrated remarkable potential in capturing subtle irregularities in medical data, often imperceptible to the human eye [7]. These innovations have the potential to revolutionize diagnostic practices and enhance patient care [4].

Medical imaging [7], encompassing modalities such as X-rays, magnetic resonance imagines (MRIs), computed tomography (CT) scans, and more, is pivotal for clinicians in the assessment of various health conditions. However, the accurate identification of anomalies within these images, especially those of rare or subtle nature, poses a considerable challenge [8]. Diagnosing medical images demands significant time and expertise from doctors and highly qualified experts, diverting their attention from other critical medical tasks that require their specialized skills and attention. Also, traditional methods for anomaly detection have limitations, often failing to address the complex and nuanced patterns encountered in medical data.

Fractures pose a dual challenge in clinical practice, being both the most prevalent and time-consuming pathology to diagnose in medical imaging. These injuries are frequently encountered in emergency scenarios, demanding swift and precise diagnosis to ensure effective patient care. However, the interpretation of medical images, such as X-rays or CT scans, to identify fractures can be time-consuming and prone to errors. Doctors often face a high volume of radiographic images to analyze, leading to potential delays in diagnosis and treatment initiation. Moreover, the variability in fracture presentations, including subtle fractures or overlapping structures, further complicates the diagnostic process. Inaccurate or delayed diagnosis of fractures can have serious consequences for patient care, including prolonged pain and discomfort, impaired mobility, and increased risk of complications. Therefore, developing efficient and reliable methods for fracture detection in medical images is essential to improve patient outcomes and optimize healthcare delivery.

GAN is considered as a powerful deep learning family, it has garnered significant interest in several anomaly detection studies [9], [10] due to its innovative architecture, comprising a generator and a discriminator. The generator has the ability to produce synthetic data similar to existing data, effectively addressing the challenge of scarce data. The discriminator aims to distinguish and classify real data from generated data. Consequently, GANs have emerged as a disruptive force, offering the prospect of improved accuracy, precision, and early detection of abnormalities. The principle of GANs, involving the interplay between a generator and discriminator, has opened up new avenues for capturing the intricate relationships within medical images and distinguishing normal from abnormal data [9]. The generative capabilities of GANs are not limited to image synthesis but extend to the identification of outliers, making them a valuable asset in the diagnostic toolbox.

While the potential for GAN-based anomaly detection is evident in some fields [3], this paper delves into the evolving landscape of GAN-based anomaly detection in medical imaging specifically fracture detection. We embark on a comprehensive examination of the effectiveness of various GAN-based models in identifying fractures within diverse medical datasets. Through a systematic exploration of these models, we aim to provide valuable insights into their strengths and limitations. Furthermore, we explore the pivotal role of data augmentation and the potential of active learning strategies to enhance anomaly detection in the medical domain. Our study underscores the potential of GAN-based anomaly detection in the field of medical imaging. As AI continues to redefine healthcare practices, we envision that the insights presented here will contribute to the ongoing evolution of anomaly detection, ultimately enhancing diagnostic precision and patient outcomes.

The remainder of the paper is structured as follows: in section 2 provides an overview of related work in the same domain. Then we present overview of GANs in the third section. Then GAN-based anomaly detection methods are expounded in section 4. Section 5 comprises the experimental details, materials, datasets and the evaluation metrics essential for a comprehensive comparative analysis study. Models' results and the ensuing discussions are presented. Finally, the paper culminates with the conclusion and findings.

2. RELATED WORKS

In this section, the literature review encompasses two key components. Firstly, it delves into anomaly detection using GANs. Secondly, it explores various methods for augmenting data in the context of anomaly detection.

2.1. GANs for anomaly detection

In recent years, GANs have gained significant attention in the field of anomaly detection. Researchers have explored various approaches and applications of GAN-based anomaly detection, as summarized below. Early research in this domain primarily used GANs for data generation. Goodfellow [6] introduced GANs as a framework for generating realistic data samples. This concept was extended to create synthetic data representing the normal data distribution, with anomalies identified as deviations from these generated samples.

Subsequent works focused on utilizing GANs directly for anomaly detection. Schlegl *et al.* [11] introduced AnoGAN, combining GANs with gradient-based optimization to identify anomalies by generating similar data points. Variations, such as AlphaGAN [12], BiGAN [13] and GAAL [14].

In image anomaly detection, specialized GAN-based techniques have been developed. Akcay introduced GANomaly [10], and SkipGANomaly [15] for a GAN-based anomaly detection tasks. GANs to learn image representations for anomaly detection. This approach has been extended to various image data types, including medical and satellite imagery. Both semi-supervised and unsupervised approaches have been explored. Semi-supervised methods use labeled anomaly data during training for improved performance [9] while unsupervised methods aim to detect anomalies solely based on learned data distributions.

2.2. Data augmentation methods for anomaly detection

Data augmentation emerges as a strategic approach to address the challenge of data imbalance, particularly evident when there is a notable disparity in class distribution within a dataset, a common scenario in anomaly detection contexts. To tackle this issue in anomaly detection, Dixit and Verma [16] employed a conditional variational AE along with a centroide loss function, aiming to overcome limitations posed by a rare training samples. Furthermore, there is a growing interest in utilizing GANa for data augmentation [17], [18]. Alzantot *et al.* [17] introduced a generative model incorporating multiple long short-term memory (LSTM) networks and a mixture density network to generate synthetic sensory data. However, their primary objective was to substitute real datasets with syntetic samples to ensure patient privacy instead of focusing on improving diagnosis performance. In contrast, [18] employed an auxilliary classifier GAN (ACGAN) with staked convolution leyers to generate one-dimensionall raws signals from mechanical sonsor data.

In summary, data augmantation methods prove effective in mitigating the challenges posed by data imbalance and enhancing model classification accuracy. In conclusion, the suggested data augmentation methods, especially those rooted in GANs [18], enhance detection accuracy. However, it is crucial to recognize that the training of GANs is recognized for being time-intensive, primarily due to challenges related to stabillity and convergence, particularly in diverse robotic sonsor scenarios. As a result, our model incorporates the Waserstein distance with a gradient penalty to improve training stability and adopts an adaptive update approach to accelerate training convergence [19], [20].

3. GENERATIVE ADVERSARIAL NETWORKS

GAN introduced by Goodfellow [6] is considered as one of the most powerful member of the neural network family, due to realistic data-generation capacities Figure 1. The principal advantage of GANs is their ability to generate data. This property has enabled GANs to perform well in anomaly detection, image generation and image super-resolution, as well as other computer vision tasks.

GAN stands out as a potent member within the neural network category, specifically employed for unsupervised deep learning. It comprises two adversarial algorithms : generator and a discriminator. The generator is responsible for generating synthetic samples from the noise, residing in the latent z-space, while the discriminator is tasked with distinguishing between real and syntetic samples. we can schematize the GAN architecture as showen in Figure 1.

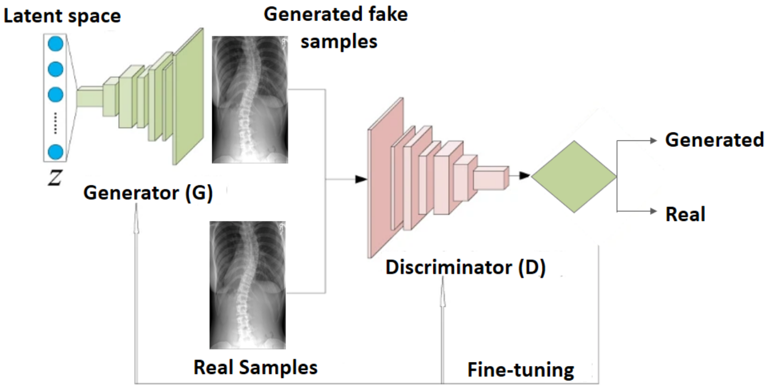

Figure 1. Architecture of a generative adversarial network

Training the two components introduces two main objectives:

- Discriminator maximizes the probability of asigning the correct label to the real data and the data produced by the generator.
- Generator minimizes the probability that the discriminator will predict that what it generates is false.

Learning a GAN takes place in several stages:

- From a random distribution, a noise is transformed to the G generator, which will then generate false pairs (x, y) with label y = 0.
- Discriminator D receives the false pair produced by generator G and the real pair with label y = 1.
- Each component has a different objective function, so the D discriminator calculates the loss for the false x and the real x and combines them as the final D loss. On the other hand, the G generator also calculates the loss of its noise as a G loss.
- Both losses return to their respective networks to learn about the loss and adjust their parameters.
- Optimization algorithm (Grad descent, ADAM, prop RMS, etc.) loops until a satisfactory level of performance is reached.

This mode of competition between the two components allows the generator to generate samples closer to the real data, and the discriminator becomes increasingly strong in distinguishing between false and true data. The objective function of GAN : as we have defined, the discriminator is a binary clasifier, so it aims to produce a high probability for real data and a low probability for false data (generator output). The variables will be defined as follows:

- z Noise vector ;
- $G(z)$ Generator output x_{fake}
- x Learning sample x_{real}
- $D(x)$ is the probability that x comes from the original data → The discriminator output for
- $x_{real} \rightarrow \mathrm{P}(x|x_{real}) \rightarrow \{0,1\}$
- $D(G(z)) \rightarrow$ Discriminator output for $x_{fake} \rightarrow \mathrm{P}(y|x_{fake}) \rightarrow \{0,1\}$

At discriminator level D: $D(x) \rightarrow$ must be maximized and $D(G(z))$ must be minimized. At generator level G: $D(G(z))$ must be maximized. On the mathematical side:
At the discriminator level D:

$$D_{lossReal} = \log(D(x))$$

$$D_{lossFake} = \log(1 - D(G(x)))$$

$$D_{loss} = D_{lossReal} + D_{lossFake} = log(D(x)) + log(1 - D(G(x))) \tag{1}$$

the total cost of the loss is:

$$\frac{1}{m}\sum_{i=0}^{m} log(D(x^i)) + log(1 - D(G(z^i)))$$

at generator level G:

$$G_{loss} = log(1 - D(G(z))) = -log(D(G(z))) \tag{2}$$

the total cost of the loss.

$$\frac{1}{m}\sum_{i=0}^{m} log(1 - D(G(z^i))) = \frac{1}{m}\sum_{i=0}^{m} -log(D(G(z^i)))$$

The discriminator must classify false and true data, before calculating the final loss, so (1) must be maximized. For the generator, it must fool the discriminator by producing simulated data similar to the original data, so: (2) must be minimized $D(G(z)) = 1$. We move forward to compute the gradients along with their parameters and propagate them through their respective networks independently. As per Ian's publication, the ultimate equation in terms of expectation is presented below. D and G engage in the following adversarial game with the value function V(G, D):

$$\min_G \max_D V(D,G) = \mathbb{E}(log(D(x))) + \mathbb{E}(log(1 - D(G(z)))) \tag{3}$$

where $\mathbb{E}(log(D(x)))$ recognize real images and $\mathbb{E}(log(1 - D(G(z))))$ recognize generated images better. And $\max_G V(G)$ optimize G that can fool the discriminator the most.

4. METHOD: GANS FOR ANOMALY DETECTION

In this section, we present an overview of the GAN models utilized in our study. Our selection of these models was informed by a thorough review of scientific literature, where we identified them as the most appropriate algorithms for our research. Factors considered included the diverse characteristics of the datasets in terms of their types, sizes, and dimensions, as well as the specific objectives of our research task.

4.1. BiGAN

Bidirectional-GAN (BiGAN) [13] enhances generative capabilities by integrating an encoder network, enabling data generation from a latent space while simultaneously mapping real data to this latent space. This dual functionality is well-suited for anomaly detection, allowing both synthetic data generation and similarity measurement between real data and their latent representations. This approach improves accuracy and robustness in identifying anomalies by detecting data points that deviate from expected latent representations. Figure 2 illustrate AnoGAN, BiGAN, GANomaly and Skip-GANomaly's architectures.

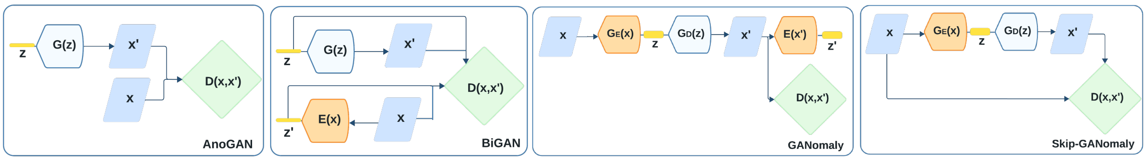

Figure 2. Architectures comparison: AnoGAN, BiGAN, GANomaly, Skip-GANomaly

4.2. AnoGAN

AnoGAN, short for "Anomalous GAN," [11] Figure 2 represents a pioneering effort in the adaptation of GANs for anomaly detection. It introduces an innovative approach by combining a GAN framework with gradient-based optimization. The model's core idea is to generate synthetic data samples that closely resemble the normal training data and then assess the dissimilarity between these generated samples and the real data to detect anomalies. AnoGAN marked a significant milestone in the evolution of GAN-based anomaly detection, serving as a foundation for subsequent research in the field.

4.3. GANomaly

GANomaly [10] Figure 2 builds upon the foundation laid by AnoGAN by incorporating an autoencoder network, enhancing the quality of the generated data samples. This hybrid model leverages the strengths of both GANs and autoencoders, with the GAN generator producing synthetic samples and the autoencoder quantifying the reconstruction errors. By doing so, GANomaly not only improves the quality of the generated data but also enhances the accuracy of anomaly detection. This approach has proven effective in addressing data imbalance and achieving better anomaly classification results.

4.4. Skip-GANomaly

Skip-GANomaly [15] Figure 2 is a groundbreaking anomaly detection model that enhances GANs with skip connections, improving feature extraction and gradient flow. This design empowers Skip-GANomaly to handle complex data distributions effectively, making it a robust choice for anomaly detection. While it delivers high accuracy in anomaly detection, its increased complexity can result in longer training times, and fine-tuning skip connections can be intricate. Nonetheless, Skip-GANomaly is a valuable tool for detecting subtle or rare anomalies in various applications, including medical imaging.

4.5. GAAL: generative adversarial active learning for unsupervised outlier detection

Liu *et al.* [14] introduce an alternative GAN representation, which assesses the posterior probability of test samples generated by the same generative model to identify an anomaly marker in medical images. In an unsupervised context, the generator produces anomalies that will be the discriminator input with real data. This will enable the discriminator to differentiate between normal and abnormal data. The models discussed above all consider the GAN as a variable extractor or reconstructor.

4.5.1. SOGAAL

In SO-GAAL [14], as illustrated in Figure 3, the G generator takes noise variables z as input to generate potential outliers, and the discriminator describes the dividing boundary between the two classes. At the start of training, the G generator is unable to produce a significant number of potential outliers, allowing the discriminator to create an approximate boundary between the generated data and the real data. After several iterations, the generator improves the mechanism for generating potential outliers that occur within or close to the real data. As a result, the dividing boundary created by the discriminator becomes more precise. In simpler terms, the generator enhances the accuracy of the discriminator by producing potentially informative outliers, essentially engaging in an active learning procedure.

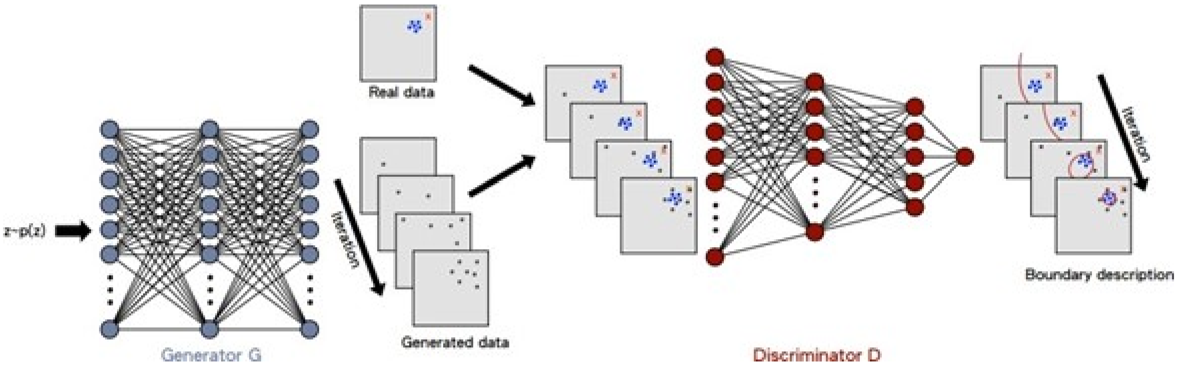

Figure 3. SOGAAL's architecture [14]

4.5.2. MOGAAL

The core concept behind MO-GAAL [14] involves each sub-generator, Gi, proactively learning the data generation mechanism for a specific subset of real data. MO-GAAL's performance and optimization process align with that of SO-GAAL during the initial iterations. However, MO-GAAL manages to sustain its high accuracy even after surpassing 400 iterations.

4.6. MadGAN

MadGAN [21] is an innovative anomaly detection model that extends the capabilities of GANs for anomaly identification. What sets MadGAN apart is its unique approach of incorporating multiple generators within the GAN framework. These generators collaborate to comprehensively capture the underlying data distribution. By training on both normal and anomalous data, MadGAN becomes adept at effectively distinguishing between the two. This multi-generator strategy significantly enhances anomaly detection accuracy, rendering MadGAN a promising solution for challenging anomaly detection tasks, especially when anomalies are subtle or rare in the dataset. MadGAN's ingenuity lies in its capacity to harness the diversity of multiple generators, ultimately improving the performance of anomaly detection and providing a robust and effective tool for this purpose.

4.7. RandGAN

RandGAN [22] is tailored for anomaly detection, with a specific emphasis on COVID-19 detection. It consists of two key components: a generator and a discriminator as illustrated in Figure 4. Notably, RandGAN's architecture incorporates Inception and residual blocks. To enhance its generalizability, RandGAN adopts a unique approach where images are randomly selected from the training class cohort and encoded into a lower-dimensional representation space using inception layers. This approach introduces variability in each iteration of the generator's training, encompassing both random noise vectors and real random image representations, making it effective for anomaly detection, particularly in COVID-19 identification.

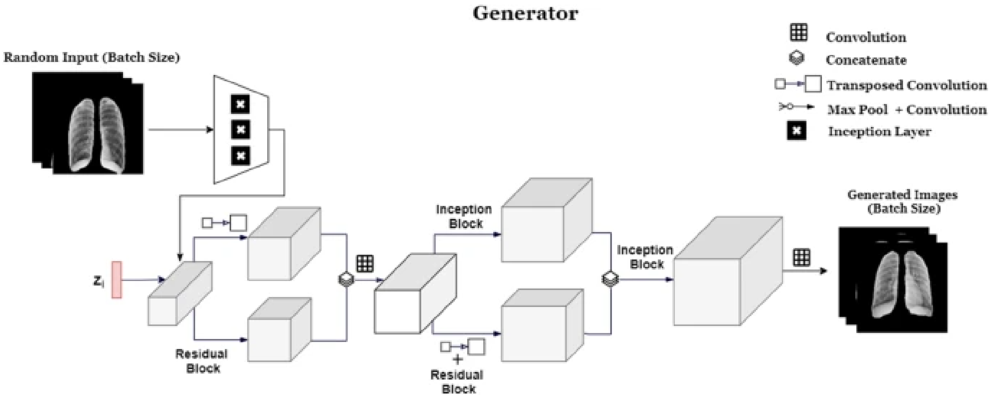

Figure 4. RandGAN's generator architecture [22]

5. EXPERIMENT

his section provides an evaluation of the proposed methodology using three diverse medical datasets, each characterized by different modalities. Additionally, detailed project code and resources utilized in this study are comprehensively outlined. They are openly accessible on GitHub at the following link: https://github.com/nabinabila/GAN- based-Anomaly-Detection.

5.1. Dataset description

In this study, we have gathered three unique medical imaging datasets, each representing a specific imaging modality and focusing on distinct anatomical structures. The subsequent section offers a succinct overview of these datasets, accompanied by their summarized characteristics, as presented in Table 1. These datasets are essential for our research, providing diverse insights into anomaly detection across various medical imaging modalities.

Table 1. Summary of medical image datasets characteristics

Dataset	Speciality	Total images	Number of normal images	Number of abnormal images	Number of organs	Number of classes	Modality
RibFrac	Rib fractures detection	660	20	640	1	6	CT Scan
MURA	Bone fractures detection	40,656	8,941	5,715	7	2	Radiograph
xVertSeg	Spinal fractures diagnosis	25	-	-	1	2	CT Scan

5.1.1. MURA

The MURA dataset [23] is a highly regarded resource in musculoskeletal imaging, encompassing a wide range of radiographic images across various anatomical regions, including the upper extremities, lower extremities, and torso Figure 5. It focuses on diverse musculoskeletal conditions, such as fractures, ligamentous injuries, and joint abnormalities, with each image meticulously annotated by radiologists. Serving as a valuable source of ground truth data for training and evaluation, the MURA dataset plays a vital role in advancing the field by facilitating algorithm development and evaluation tailored to musculoskeletal radiography. Researchers widely utilize this dataset to enhance diagnostic diagnosis systems, and strengthen computer-assisted fracture detection methods. Provided by the Stanford Program for Artificial Intelligence in Medicine, the dataset is accessible at [23].

5.1.2. xVertSeg

The xVertSeg challenge, Figure 6, aims to unite researchers interested in spinal imaging and analysis, focusing on vertebra segmentation and fracture classification. It provides a standardized clinical image database comprising 25 CT lumbar spine images with both non-fractured and fractured vertebrae.

For research and training, 15 images include segmentation masks and fracture classification scores, while 10 images are reserved for testing and evaluation by challenge organizers. It is obtained from the public SpineWeb repository 6.

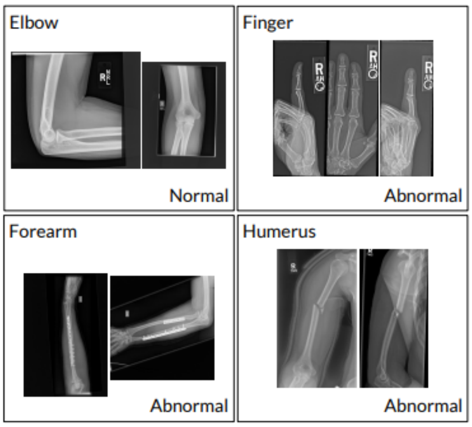

Figure 5. MURA dataset

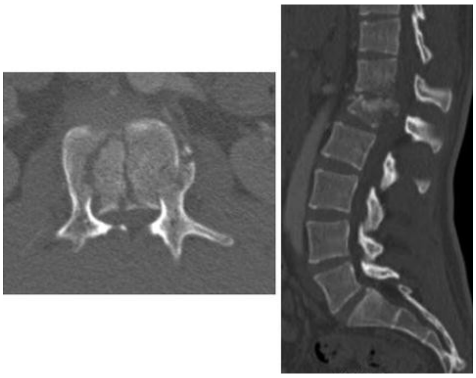

Figure 6. xVertSeg dataset

5.1.3. RibFrac

Diagnosing rib fractures [24] is crucial in clinical, forensic, and business settings like insurance claims. However, automated machine learning approaches for this task are limited. To bridge this gap, a benchmark dataset Figure 7 has been established. It includes about 5,000 rib fractures detected from 660 CT scans, with 420 scans for training (all with fractures), 80 for validation (20 without fractures), and 160 for evaluation. Each annotation provides a pixel-level mask for fracture regions and a four-type classification. The intention behind this challenge is to facilitate and advance the research and practical application of automated rib fracture detection and diagnosis.

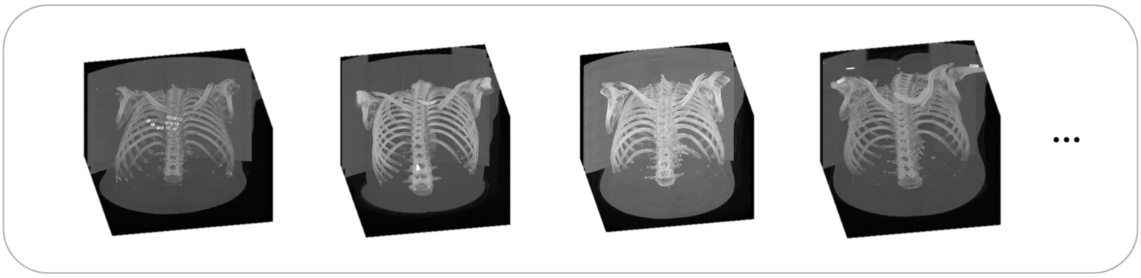

Figure 7. Rib dataset

5.2. Evaluation metrics

Accuracy: accuracy [25] is a straightforward metric for anomaly detection, calculating the ratio of correctly detected anomalies and normal instances to the total number of instances.

$$Accuracy = \frac{TruePositives + TrueNegatives}{TotalSamples}$$

Precision: precision assesses the model's ability to avoid false positives by measuring the proportion of correctly identified anomalies among predicted anomalies [25].

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

F1-score: harmonizes precision and recall [25], offering a concise summary of a model's anomaly detection performance, especially in imbalanced datasets or rare anomaly scenarios.

$$F1_{Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Detection speed: assesses the speed of anomaly detection in a model [25], crucial in real-time applications like network intrusion detection or industrial equipment monitoring.
These metrics provide insight into the model's ability to accurately detect fractures, minimize false positives, and efficiently process radiographic images.

5.3. Results and discussion

In this section, we conduct a thorough evaluation of GAN models for bone fracture detection in radiographs and CT scans, with a specific focus on critical evaluation metrics such as accuracy, precision, F1-score, and detection speed in Table 2. These metrics serve as essential indicators of the models' effectiveness and efficiency in identifying fractures, thereby supporting informed clinical decision-making. The models' performance is influenced by various factors, including architectural choices, layer design, padding, shape, normalization, activation functions, loss functions, optimizers, batch sizes, learning rates, pooling methods, and output layers. Achieving optimal results often required multiple rounds of fine-tuning. Some of our models incorporated computationally intensive layers and modules, resulting in extended training times that surpassed the capabilities of standard hardware or laptop configurations.

Table 2. Models evaluation

Dataset	Model	AUC	Precision	F1 score	Detection speed
RibFract	GANomaly	96.2	97.6	95.10	8.450
	SkipGANomaly	97.8	98.3	96.70	10.730
	BiGAN	88.5	88.9	83.60	13.230
	AnoGAN	89.0	89.8	88.80	11.030
	SOGAAL	72.1	76.9	73.50	17.450
	MOGAAL	79.8	79.9	80.20	8.880
	RandGAN	96.2	91.6	97.40	5.760
	MadGAN	97.6	96.9	98.00	2.370
MURA	GANomaly	86.1	87.3	88.10	2.987
	SkipGANomaly	90.1	90.3	90.70	5.837
	BiGAN	83.0	82.5	83.70	4.320
	AnoGAN	83.8	84.8	84.20	3.738
	SOGAAL	64.2	65.3	65.80	11.980
	MOGAAL	70.3	72.4	70.50	8.760
	RandGAN	84.2	87.7	84.18	3.291
	MadGAN	95.4	95.8	95.30	1.418
xVertSeg	GANomaly	90.7	91.4	90.80	4.570
	SkipGANomaly	92.3	93.5	92.90	9.340
	BiGAN	85.4	85.7	84.70	11.270
	AnoGAN	78.3	79.8	80.70	10.120
	SOGAAL	51.0	51.4	53.40	15.110
	MOGAAL	63.9	65.1	67.50	7.450
	RandGAN	91.3	89.2	90.00	5.870
	MadGAN	93.6	94.8	96.30	2.890

Preprocessing emerged as a pivotal element in achieving successful outcomes in our DL tasks. Following the careful selection of GAN models based on benchmarking, meticulous data preparation became imperative. The image size emerged as a crucial parameter influencing fracture detection accuracy. Several treatments were applied to our dataset. Furthermore, in addressing data limitations, we employed data augmentation techniques to expand the dataset during training. It's worth noting that we exercised caution when considering rotation methods, extensive compression, and shear, as these techniques had the potential to impact bone fracture diagnosis performance. By carefully examining these factors and conducting a series of rigorous experiments, our aim was to fine-tune models and improve the precision of our findings. By carefully examining these factors and conducting a series of rigorous experiments, our aim was to fine-tune models and improve the precision of our findings.

Table 2 presents the results obtained from our experiments involving eight AD methods across the three medical image datasets. Evaluation metrics such as precision, recall, F1-score, ROC AUC, and PR AUC are used to assess performance. One notable observation from the summarized findings in Table 2 is the considerable variability in the performance of GAN models across different datasets. The effectiveness of a GAN-based anomaly detection model in capturing the concept of ”normality” from the training data significantly impacts its performance. This relationship is intricately linked to both the inherent characteristics of the provided dataset and the specific mechanisms employed by the model for this purpose.

The models exhibited a noteworthy level of accuracy, with their performance ranging from 0.7% to 0.954%. These outcomes underscore the substantial potential of the GAN-based approach in enhancing fracture diagnosis within the field of orthopedics. The most favorable results were achieved when working with the MURA dataset, notable for its extensive training sample size. It is worth noting that GAN-based methods exhibited nearly consistent performance across all seven different anatomical organs. However, when examining the RibFrac dataset, we encountered a specific challenge: if the training dataset’s size is not adequately substantial relative to the GAN-based anomaly detection model’s capacity, the discriminator module may tend to excessively memorize the training data during the training process. This situation can lead to overfitting and, in turn, a collapse of the model. Consequently, the quality of the generated images deteriorates, which is a significant concern, particularly in the field of medical imaging, where data collection is a resource-intensive and expensive endeavor. One potential solution to tackle this challenge and enhance the performance and robustness of GANs is data augmentation. Research has shown that data augmentation, when applied to both real and generated images, can substantially enhance the performance of GANs. The impact may not be as pronounced if data augmentation is exclusively applied to real images.

In the context of GANs, assessing the quality of generated images from both normal and abnormal samples is crucial for evaluating the performance of anomaly detection models and explaining their decisions. MRI typically offers superior anatomical detail and sharper images for soft tissues, while CT scans provide comprehensive views of cortical bones with enhanced contrast. With smaller training datasets, GANs may struggle to capture fine details, leading to more generalized representations. For instance, in the xVertSeg dataset, GANs reconstruct bones effectively due to the region’s high contrast, despite potential limitations in capturing finer details.

Detailly, in Table 2, it becomes apparent that MadGAN and SkipGANomaly consistently deliver the highest accuracy scores. The architectural design of these models plays a pivotal role in their performance. MadGAN’s incorporation of multiple discriminators enhances its ability to detect fractures by recognizing anomalous patterns in radiographs. On the other hand, Skip-GANomaly’s innovative inclusion of skip connections within the GAN structure enhances feature extraction and gradient flow, making it robust in handling complex data distributions, particularly in the field of medical imaging. Subsequently, GANomaly and RandGAN achieve respectable accuracy scores of 0.861% and 0.842%, respectively. GANomaly leverages GANs to reconstruct normal data and quantify anomalies through anomaly scores, offering a flexible and effective approach to anomaly detection in medical imaging. RandGAN introduces randomness using inception layers, enhancing generalizability and excelling in detecting subtle or rare anomalies, although this randomness occasionally leads to unpredictability in the synthetic data generation process, contributing to the margin of error. AnoGAN and BiGAN exhibit relatively lower performance in fracture detection. This disparity in performance is attributed to their architectural design, which is not inherently optimized for fracture detection. Furthermore, the high number of convolutional layers in these models leads to extended prediction times. AnoGAN’s performance is also influenced by the quality of the AE, which sometimes struggles to capture complex multimodal distributions in medical images. Additionally, training a BiGAN is computationally intensive and slower in convergence compared to other models.

In the case of the GAAL method, MOGAAL outperforms SOGAAL. Initially, SOGAAL reaches maximum accuracy (AUC=1) within the first 100 iterations, but its accuracy substantially declines afterward, leading to the ”Mode collapse problem.” MO-GAAL, consisting of k sub-generators and a discriminator, proves more efficient than SOGAAL. MO-GAAL mirrors SO-GAAL’s performance and optimization process in the early iterations but maintains higher accuracy even after exceeding 400 iterations.

It is noteworthy that the runtime of AnoGAN, BiGAN, SOGAAL, MOGAAL, and SkipGANomaly may not be suitable for real-time anomaly detection due to their iterative algorithms. While SkipGANomaly's performance is commendable, the increased complexity results in longer training times, and fine-tuning the skip connections can be intricate. In contrast, MadGAN exhibits a shorter processing time, making it a favorable choice for applications that require speed detection.

Ultimately, the selection of the model should align with the specific requirements of the medical imaging anomaly detection task. BiGAN and Skip-GANomaly are ideal for complex data distributions, while RandGAN and MadGAN excel in detecting subtle anomalies. AnoGAN and GANomaly offer flexibility, MOGAAL and SOGAAL present strong options for active learning in scenarios with limited labeled data. Factors such as dataset characteristics, and available computational resources should guide the model choice. If the objective is to develop a platform for detecting fractures across various organs, MadGAN and SkipGANomaly emerge as robust models to integrate into the framework.

Our research is centered on the development of a computer-aided diagnosis (CAD) system, specifically designed for the detection of fractures. We evaluate various methodologies, with a particular focus on GAN-based approaches, to identify the most effective technique for incorporation into the CAD system. This chosen methodology aims to improve the CAD system's efficiency and accuracy in detecting fractures within medical imaging, thereby aiding healthcare professionals in making faster and more precise diagnostic decisions.

By leveraging the capabilities of GAN models, such as their proficiency in generating synthetic data and identifying anomalies, our CAD system is positioned to significantly enhance the interpretation of medical images by doctors. The integration of these advanced machine learning techniques into clinical settings is anticipated to elevate patient care and outcomes, thereby advancing the effectiveness of healthcare services.

Our study stands out due to several key strengths. Technically, GANs are highlighted for their exceptional capability to create synthetic data that is indistinguishable from real data, which is critical for improving anomaly detection accuracy in medical imaging [26], [27]. Moreover, our research diverges from other studies by utilizing a diverse array of datasets rather than depending solely on the MURA dataset [3], [28]. This approach not only broadens the scope of our findings but also ensures a thorough evaluation of GAN-based anomaly detection across various medical imaging fields. Additionally, unlike studies that focus on a single modality [3], [24], [28], our research incorporates multiple imaging modalities, such as radiographs and CT scans. This multi-modal strategy increases the robustness of our results. Furthermore, by diagnosing conditions across different organs, our study transcends the limitations of single-task frameworks and adopts a multi-task approach. This broadens the understanding of GAN-based anomaly detection in medical imaging, setting the stage for the development of more effective and versatile diagnostic tools in the healthcare sector.

Presently, the primary impediment to integrating AI-based solutions into healthcare systems is their limited generalization capacity. Anomaly detection [3], [5] grapples with generalization challenges. In the broader context, ensuring ”trustworthiness” in AI algorithms necessitates considering a multitude of factors. Crucially, adherence to best practices for AI model development and validation is imperative. These guidelines encompass critical aspects of study design, data collection, model development, training, testing, and evaluation. Furthermore, it is essential to distinguish proof-of-concept assessments from more comprehensive technical and clinical evaluations.

6. CONCLUSION

The use of GANs for anomaly detection shows significant promise and growth. This paper provides a thorough investigation into the effectiveness of GAN-based approaches for detecting anomalies in medical images. Through extensive experiments on diverse medical image datasets, we gained valuable insights into the capabilities and limitations of various GAN-based models for anomaly detection. BiGAN, AnoGAN, GANomaly, Skip-GANomaly, RandGAN, MadGAN, MOGAAL, and SOGAAL showed varying effectiveness in detecting anomalies across different organs and imaging modalities. These findings promise to assist medical professionals in early and accurate diagnosis across various conditions. Harnessing the full potential of GAN-based anomaly detection in medical imaging faces challenges, notably interpretability and model opacity. These issues hinder widespread adoption, requiring the integration of XAI methods to enhance transparency and trust, especially in critical healthcare settings. Our research contributes to the growing body of knowledge on GAN-based anomaly detection in medical imaging. Our findings highlight both successes and challenges, indicating a promising path with significant potential to impact medical diagnostics.

REFERENCES

[1] A. S. Rawat, A. Rana, A. Kumar, and A. Bagwari, “Application of multi layer artificial neural network in the diagnosis system: a systematic review,” *IAES International Journal of Artificial Intelligence*, vol. 7, no. 3, pp. 138–142, 2018, doi: 10.11591/ijai.v7.i3.pp138-142.

[2] M. Mahmood, B. Al-Khateeb, and W. M. Alwash, “A review on neural networks approach on classifying cancers,” *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 317–326, 2020, doi: 10.11591/ijai.v9.i2.pp317-326.

[3] N. Ounasser, M. Rhanoui, M. Mikram, and B. El Asri, “Generative and autoencoder models for large-scale mutivariate unsupervised anomaly detection,” *Smart Innovation, Systems and Technologies*, vol. 237, pp. 45–58, 2022, doi: 10.1007/978-981-16-3637-0_4.

[4] N. Ounasser, M. Rhanoui, M. Mikram, and B. El Asri, “Enhancing computer-assisted bone fractures diagnosis in musculoskeletal radiographs based on generative adversarial networks,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, pp. 960–966, 2023, doi: 10.14569/IJACSA.2023.01407104.

[5] N. Ounasser, M. Rhanoui, M. Mikram, and B. El Asri, “Anomaly detection in orthopedic musculoskeletal radiographs using deep learning,” *Lecture Notes in Networks and Systems*, vol. 693 LNNS, pp. 93–102, 2023, doi: 10.1007/978-981-99-3243-6_8.

[6] I. J. Goodfellow *et al.*, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.

[7] S. Roy, T. Meena, and S. J. Lim, “Demystifying supervised learning in healthcare 4.0: a new reality of transforming diagnostic medicine,” *Diagnostics*, vol. 12, no. 10, 2022, doi: 10.3390/diagnostics12102549.

[8] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, “Deep learning for anomaly detection: a review,” *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–38, 2021, doi: 10.1145/3439950.

[9] F. Di Mattia, P. Galeone, M. De Simoni, and E. Ghelfi, “A survey on GANs for anomaly detection,” *arXiv preprint arXiv:1906.11632*, 2019, [Online]. Available: http://arxiv.org/abs/1906.11632.

[10] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, “GANomaly: semi-supervised anomaly detection via adversarial training,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11363 LNCS, pp. 622–637, 2019, doi: 10.1007/978-3-030-20893-6_39.

[11] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10265 LNCS, pp. 146–147, 2017, doi: 10.1007/978-3-319-59050-9_12.

[12] K. Raza and N. K. Singh, “A tour of unsupervised deep learning for medical image analysis,” *Current Medical Imaging Reviews*, vol. 17, no. 9, pp. 1059–1077, 2022, doi: 10.2174/18756603mtezonzmk0.

[13] J. Donahue, T. Darrell, and P. Krähenbühl, “Adversarial feature learning,” *arXiv preprint arXiv:1605.09782*, 2016.

[14] Y. Liu *et al.*, “Generative adversarial active learning for unsupervised outlier detection,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1517–1528, 2020, doi: 10.1109/TKDE.2019.2905606.

[15] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, “Skip-GANomaly: skip connected and adversarially trained encoder-decoder anomaly detection,” *Proceedings of the International Joint Conference on Neural Networks*, vol. 2019-July, 2019, doi: 10.1109/IJCNN.2019.8851808.

[16] S. Dixit and N. K. Verma, “Intelligent condition-based monitoring of rotary machines with few samples,” *IEEE Sensors Journal*, vol. 20, no. 23, pp. 14337–14346, 2020, doi: 10.1109/JSEN.2020.3008177.

[17] M. Alzantot, S. Chakraborty, and M. Srivastava, “SenseGen: a deep learning architecture for synthetic sensor data generation,” in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2017*, 2017, pp. 188–193, doi: 10.1109/PERCOMW.2017.7917555.

[18] S. Shao, P. Wang, and R. Yan, “Generative adversarial networks for data augmentation in machine fault diagnosis,” *Computers in Industry*, vol. 106, pp. 85–93, 2019, doi: 10.1016/j.compind.2019.01.001.

[19] P. Malik, M. Pathania, and V. K. Rathaur, “Overview of artificial intelligence in medicine,” *Journal of family medicine and primary care*, vol. 8, no. 7, pp. 2328–2331, 2019, doi: 10.4103/jfmpc.jfmpc_440_19.

[20] S. Benjamens, P. Dhunnoo, and B. Meskó, “The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database,” *NPJ Digital Medicine*, vol. 3, no. 1, p. 118, 2020, doi: 10.1038/s41746-020-00324-0.

[21] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S. K. Ng, “MAD-GAN: multivariate anomaly detection for time series data with generative adversarial networks,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11730 LNCS, pp. 703–716, 2019, doi: 10.1007/978-3-030-30490-4_56.

[22] S. Motamed, P. Rogalla, and F. Khalvati, “RANDGAN: randomized generative adversarial network for detection of COVID-19 in chest X-ray,” *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021, doi: 10.1038/s41598-021-87994-2.

[23] P. Rajpurkar *et al.*, “MURA: large dataset for abnormality detection in musculoskeletal radiographs,” *arXiv preprint arXiv:1712.0695*, 2017, [Online]. Available: http://arxiv.org/abs/1712.06957.

[24] L. Jin *et al.*, “Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet,” *eBioMedicine*, vol. 62, p. 103106, Dec. 2020, doi: 10.1016/j.ebiom.2020.103106.

[25] D. Joshi and T. P. Singh, “A survey of fracture detection techniques in bone X-ray images,” *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4475–4517, 2020, doi: 10.1007/s10462-019-09799-0.

[26] N. Elmrabit, F. Zhou, F. Li, and H. Zhou, “Evaluation of machine learning algorithms for anomaly detection,” in *International Conference on Cyber Security and Protection of Digital Services, Cyber Security 2020*, 2020, pp. 1–8, doi: 10.1109/CyberSecurity49315.2020.9138871.

[27] S. S. Sinthura, Y. Prathyusha, K. Harini, Y. Pranusha, and B. Poojitha, “Bone fracture detection system using CNN algorithm,” in *2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019*, 2019, pp. 545–549, doi: 10.1109/ICCS45141.2019.9065305.

[28] D. Davletshina *et al.*, “Unsupervised anomaly detection for X-ray images,” *arXiv preprint arXiv:2001.10883*, 2020, [Online]. Available: http://arxiv.org/abs/2001.10883.

BIOGRAPHIES OF AUTHORS

Nabila Ounasser possesses a bachelor's degree in Mathematics and a State Engineering Diploma in Data and Knowledge from the School of Information Sciences (ESI). Presently, she is enrolled in a Ph.D. program at ENSIAS (École Nationale Supérieure d'Informatique et d'Analyse des Systèmes) within the Computer Science Department. Her doctoral research revolves around the utilization of Artificial Intelligence for anomaly detection, conducted within the IT Architecture and Model Driven Systems Development (IMS) team. Her research interests primarily encompass artificial intelligence and computer vision. She can be contacted at email: nabilaounasser81@gmail.com.

Maryem Rhanoui is an Associate Professor specializing in Computer Sciences and Data Engineering. She received an Engineering degree in Computer Science then a Ph.D. degree from ENSIAS, Mohammed V University, Rabat 2015. Her research interests include artificial intelligence, knowledge extraction and decision making, and medical data analysis. She can be contacted at email: mrhanoui@gmail.com.

Mounia Mikram is an Associate Professor of Computer Sciences and Mathematics at the School of Information Sciences, Rabat since 2010. She received her master degree from Mohammed V University Rabat (2003) and her Ph.D. degree from Mohammed V University, Rabat, and Bordeaux I University (2008). Her research interests include pattern recognition, computer vision, biometrics security systems, and artificial intelligence. She can be contacted at email: mmikram@esi.ac.ma.

Bouchra El Asri currently serves as the Teaching Research Director at ENSIAS (École Nationale Supérieure d'Informatique et d'Analyse des Systèmes). Previously, she held the position of Technical Director at Cyber Machine. She has effectively overseen two significant projects for prominent national institutions. Moreover, she occupies various pivotal roles, such as Department Head of Software Engineering and Coordinator of the Software Engineering program at ENSIAS, where she is also tasked with enhancing the program. Actively engaged within the institution, she participates in the governing council, pedagogical committee, and budget monitoring committee. Notably, she played a critical role in transitioning the Software Engineering program to online teaching during the COVID-19 pandemic. With a robust research background, she has supervised and continues to supervise numerous doctoral theses in software architecture and data management across healthcare, industry, and education sectors. Her expertise and contributions extend to scientific committees, doctoral study centers, and teaching modules within ENSIAS's Software Engineering program. She can be contacted at email: bouchra.elasri@ensias.um5.ac.ma or b.elasri@um5s.net.ma.