IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 January 2023, accepted 5 February 2023, date of publication 13 February 2023, date of current version 24 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244741

RESEARCH ARTICLE

Generative Adversarial Networks for Anomaly Detection in Biomedical Imaging: A Study on Seven Medical Image Datasets

MARZIEH ESMAEILI[1,2], AMIRHOSEIN TOOSI[3], (Graduate Student Member, IEEE), ARASH ROSHANPOOR[4], VAHID CHANGIZI[5], MARJAN GHAZISAEEDI[1], ARMAN RAHMIM[3,6], (Senior Member, IEEE), AND MOHAMMAD SABOKROU[2]

[1]Department of Health Information Management, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran 14177-44361, Iran
[2]School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran 19538-33511, Iran
[3]Department of Integrative Oncology, BC Cancer Research Institute, Vancouver, BC V5Z 1L3, Canada
[4]Department of Computer, Yadegar-e-Imam Khomeini, Janat-Abad Branch, Islamic Azad University, Tehran 14779-99651, Iran
[5]Department of Radiology and Radiotherapy Technology, School of Allied Health Sciences, Tehran University of Medical Sciences, Tehran 14177-44361, Iran
[6]Departments of Radiology and Physics, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada

Corresponding author: Marjan Ghazisaeedi (ghazimar@tums.ac.ir)

This work was supported by the Tehran University of Medical Sciences (TUMS), as part of a Ph.D. dissertation from TUMS, under Grant IR.TUMS.SPH.REC.1399.060.

ABSTRACT Anomaly detection (AD) is a challenging problem in computer vision. Particularly in the field of medical imaging, AD poses even more challenges due to a number of reasons, including insufficient availability of ground truth (annotated) data. In recent years, AD models based on generative adversarial networks (GANs) have made significant progress. However, their effectiveness in biomedical imaging remains underexplored. In this paper, we present an overview of using GANs for AD, as well as an investigation of state-of-the-art GAN-based AD methods for biomedical imaging and the challenges encountered in detail. We have also specifically investigated the advantages and limitations of AD methods on medical image datasets, conducting experiments using 3 AD methods on 7 medical imaging datasets from different modalities and organs/tissues. Given the highly different findings achieved across these experiments, we further analyzed the results from both data-centric and model-centric points of view. The results showed that none of the methods had a reliable performance for detecting abnormalities in medical images. Factors such as the number of training samples, the subtlety of the anomaly, and the dispersion of the anomaly in the images are among the phenomena that highly impact the performance of the AD models. The obtained results were highly variable (AUC: 0.475-0.991; Sensitivity: 0.17-0.98; Specificity: 0.14-0.97). In addition, we provide recommendations for the deployment of AD models in medical imaging and foresee important research directions.

INDEX TERMS Anomaly detection, artificial intelligence, machine learning, deep learning, unsupervised anomaly detection, generative adversarial networks, medical imaging, biomedical image processing.

I. INTRODUCTION

The primary aim of anomaly detection (AD) is to identify data samples that do not fit the overall data distribution (out-of-distribution samples) [1]. Anomalies can occur for a variety of causes, including noise in the data capture method, or new or previously unidentified aspects in the captured context [2]. In medical imaging, AD where the aim, in general, is to find structural and functional abnormalities in exposed organs and tissues, challenges rise due to the fact that gathering annotations (labels) is frequently time-consuming, expensive, and frequently impossible without a confident ground truth [3]. As a result, unsupervised and semi-supervised methods have received a great deal of attention in recent years in medical

The associate editor coordinating the review of this manuscript and approving it for publication was Sudipta Roy.

imaging. Generative adversarial networks (GANs) [4] as a recent branch of unsupervised learning methods can learn to model the distribution of highly complex medical imaging data. Hence, ideally, these methods do not suffer severely from imbalanced datasets, which is a well-known barrier in the field of medical imaging [5]. These characteristics made GANs to be an established solution for developing AD methods in medical imaging applications.

GAN-based AD has become one of the most popular unsupervised AD methods [6]. The main capability of GANs is the ability to learn the distribution of a dataset in order to generate new samples based on the learnt distribution of the original dataset to be hard to tell from the real data by the discriminator. However, the role of the discriminator as well as structural changes in GAN architectures for detecting anomalies is more prominent than the generator. Many works have developed GAN-based AD [7] on medical images and natural datasets.

Although recent unsupervised AD methods, especially GAN-based AD, have partly succeeded and reported promising results, deciding whether these methods are reliable on medical images with different abnormalities, pathologies, modalities, and resolutions is very difficult and challenging. On the other hand, it is difficult to compare the performance of these methods in detecting various abnormalities in medical images and choosing the best method. This is because most of these methods have been tested on natural datasets or some kind of medical imaging with specific pathology, modality, and resolution. This study is an attempt to show and compare the performance of recent unsupervised AD, especially GAN-based AD for medical image datasets.

Therefore, we present a comparison of three unsupervised ADs for 7 different medical image datasets. The datasets differ in terms of the number of samples, the type of abnormality and pathology, and the imaging modality. Also, the models differ in architectural complexity and loss functions. Therefore, it is possible to provide insights into the impact of different factors in terms of model-centric and data-centric issues and to highlight the main challenges in using them in medical imaging.

The remaining sections of the paper are organized as follows: in section II, we briefly explain the whole concept behind the AD task and its approaches. Then, in section III we go into details of GANs and more specifically the state-of-the-art GAN-based AD methods, and address the strengths and challenges they face. In section IV, we explain the methodology of our experiments containing datasets, models, and settings. Afterward, in section V, we show the findings from our experiments where we examined the performance of three different DL-based AD methods on 7 medical image datasets and discuss the results. Finally, in section VI, based on the merits and shortcomings of these methods, we then provide some recommendations for enhancing GAN-based AD on medical imaging data.

II. ANOMALY DETECTION

For decades, discovering data outliers (i.e., anomalies, novelties, out-of-distribution data) has been an active research area attracting scientists in statistics and data mining [1], [8]. This mainly owes both to the challenges in the AD process as well as the valuable insights and information achievable from anomalies in datasets [9].

Technically speaking, anomalies are data points that do not fit in the distribution of the majority of the dataset that are known to be normal [9]. Also, anomalous data samples are usually rare and unknown, making it difficult to fit into the definition of normal data samples. That is to say, it is assumed that normal samples all follow roughly the same distribution, while anomalous samples come from different distributions. This assumption may not necessarily be accurate for many datasets such as medical images. In medical imaging, given the large variety of normal instances, discovering abnormal data is indeed a challenging task. More specifically, an AD model suffers from both cases of a high number of false positive samples which results in low sensitivity of the model, and also a high number of false positive samples, which leads to lower specificity of the model. But despite all the challenges, having a model to be able to detect anomalies/novelties in medical images can play a significant role in providing a decision support system for the diagnosis of unknown or rare diseases [10].

Traditional AD such as Support Vector Data Description (SVDD) [11] and One-Class Support Vector Machines (OC-SVM) [12] are conventional unsupervised methods. These methods try to find a hyperplane including normal training samples to find rich normal features [13]. However, their performance is reported to be degraded on complex high dimensional datasets [14]. In order to reduce data dimensionality, one approach is selecting features based on expert opinion, while it is not necessarily the most optimal approach [15], another approach is through applying various feature selection and extraction techniques such as deep auto-encoder networks, principal component analysis (PCA) [16] and multidimensional scaling (MDS) [17].

Deep learning (DL)-based AD methods can be used in an end-to-end fashion, to learn more discriminative feature representations during training, from the normal input images without any prior knowledge imposed by human experts. Because of these fundamental properties, DL-based AD approaches are more generalizable and robust [15] particularly, in datasets with high dimensionality like medical images, where the manually designed feature engineering pipelines is highly time-consuming, labor-intensive, and inadequate.

In medical imaging, Image Biomarkers (IBs) are identified as statistical subclinical or clinical characteristics that can be derived from one or more modalities such as magnetic resonance imaging (MRI), computed tomography (CT), X-ray, ultrasound, positron emission tomography (PET), and single photon emission computed tomography (SPECT) [18].

The process of identifying, validating, and defining a well-defined subset of IBs requires rigorous statistical and clinical studies [19]. Despite the high potential of using IBs as abnormality indicators, only some of them have been adopted in clinical decision making. To mitigate this drawback in using hand-crafted image biomarkers, deep AD strategies can alleviate the limitations by automating this process.

The most popular deep AD approach relies on learning the distribution of normal images and extracting latent representations from them to be able to reconstruct normal images well [20]. Methods such as autoencoders (AEs) [21], [22], variational autoencoders (VAEs) [23], [24], [25], and GANs [26], [27], [28] are widely used to reconstruct normal images only. Therefore, anomalous images are expected not to be reconstructed well. Hence, the reconstructed error can be considered as the anomaly score [29], also information from latent space [30], [31] and discriminator [32] can help to better detect anomalous samples.

Although reconstruction-based AD methods are widely used and intuitive as well, they can be plagued with challenges such as computational cost for image reconstruction, mode collapse, non-convergence, and instability [33].

A recent approach, instead of learning representations of normal images from scratch, applies the representational power of pre-trained deep networks to learn the distribution of normal images [34]. Methods such as SPADE [35], PaDiM [36], Multi-KD [37], and PatchCore [38] use ImageNet [39] pre-trained features to extract meaningful representations describing an image or a patch image. The learned representations of normal images carry the different abstraction levels of information that capture from various intermediate layers [36], [40]. Therefore, the extracted feature representations of anomalous images are expected to result in a significant discrepancy with representations of normal images.

III. OVERVIEW OF GAN-BASED ANOMALY DETECTION

Generative modeling is an unsupervised ML-based process that automatically discovers and learns the inherent patterns in the source dataset to generate new samples that acceptably seem to come from the same (source/input) dataset [41]. GANs are essentially a branch of generative modeling approaches implemented based on deep neural networks or convolutional neural networks. A GAN model [4] reformulates the training process of a generative model as a supervised learning task using two sub-modules: a generator network "$\mathcal{G}$" and a discriminator network "$\mathcal{D}$". The generator module generates new samples while the discriminator module attempts to classify the generated samples as Real (coming from the source data domain) or Fake (generated samples) (Fig. 1).

During training a GAN model, these two modules are trained together in form of an adversarial zero-sum game [42]. That is, in a competitive scenario, the generator network competes against its adversary, the discriminator network. The generator network constantly produces fake samples, while its rival, the discriminator network, tries to differentiate between these generated samples and the samples that come from the training dataset [43].

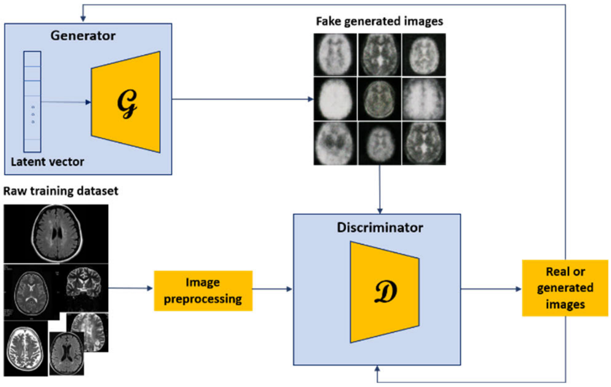

FIGURE 1. Generative adversarial network (GAN) architecture; It consists of two sub-modules - a generator network G and a discriminator network D – that are simultaneously trained. Using a latent vector as input the generator tries to produce realistic images. These, along with real images are then fed to the discriminator. The discriminator tries to differentiate between real and fake and outputs a probability for the image being real. Then both of them adapt their parameters to improve the generation and discrimination processes.

The ultimate goal of a GAN is to generate new (Fake) samples from the observed data that are indistinguishable by the discriminator network from the real ones that are coming directly from the input dataset. During the training of a GAN, the generator network $\mathcal{G}$ learns in form of a latent vector, a projection of the real data distribution into a latent space. Then after the training is successfully done, it uses this learnt latent vector in order to generate new images with a similar distribution of the input dataset. In addition, during the training, generator network $\mathcal{G}$ constantly receives signals from discriminator $\mathcal{D}$ based on whether the generated samples are close enough to the samples from the source dataset, and updates its parameters. As such, the learnt latent vector is known as a compression of the observed input data during training in a way that these projected points in the latent space can be used by the generator module in order to produce new samples close enough to the distribution of the input samples [44].

Therefore, GANs can capture the high-level concept of the majority of the given input dataset, particularly in complex datasets. Assuming that the majority of the input data correspond to the Normal samples, it is expected that the normal instances can be generated better than abnormal ones by a well-trained generator. This property of the generator and discriminator modules in GAN can be used to detect abnormal instances in a dataset. This is normally done by measuring an anomaly score [28]. By now, many GAN-based AD approaches are widely used in different areas such as industry, infrastructure, medicine, and other areas [45].

GANs in general are designed to generate realistic-looking synthetic data, regardless of being normal or anomalous.

As such, the challenge here is to find a way to employ GANs directly to AD. As mentioned earlier, latent space in the generator module of a trained GAN contains a high-level conceptual representation of the training data it is trained on. However, the main challenge in applying GAN on an AD task is to find the optimal latent space for the sample test image during the inferencing. This is mainly because conventional GAN architectures do not have mechanisms for inverse mapping from the image space to the latent space. Radford et al. [46] showed that there is a strong correlation between the latent space and the image space. In other words, the differences between normal and abnormal images are among the high-level properties of the data transferred to the latent space [28]. As such, there is a surge of efforts in addressing the inverse mapping from the image space to the latent space [47].

One of the first attempts at implementing GAN-based AD on medical images was AnoGAN [48]. In this work, they applied a Deep Convolutional Generative Adversarial Network (DCGAN) [46] along with a feature matching mechanism [49] to address the GANs' instability in AD. They applied AnoGAN on optical coherence tomography (OCT) image patches of healthy retinas. After the training, by solving an optimization problem, AnoGAN finds the optimal latent space and reconstructs the target image accordingly. They defined an anomaly score, consisting of the discriminator loss (Adversarial loss) and pixel-wise reconstruction loss (Contextual loss), in order to measure both the low dimensional and high dimensional dissimilarities between the generated image and the given input image. ADGAN [50] utilizes a Wasserstein GAN (WGAN) [51] to stabilize the training of the AD model. After the training phase, the anomaly score is calculated using the mean reconstruction loss of k latent points that are mapped to the image space while simultaneously adapting the generator parameters. As discussed earlier, here again the authors mentioned the tedious inverse mapping mechanism as the main challenge.

Many studies have tried to optimize the inverse mapping process by modifying the GAN's architecture and adding an auxiliary sub-module to simultaneously learn reverse mapping. Efficient-GAN [52] has been proposed to alleviate the computational complexity of the inference. It utilizes a vanilla GAN upon a bi-directional architecture proposed in [53], [54], which incorporated an encoder branch in order to do inferencing. The encoder alongside a generator and a discriminator simultaneously learns to infer underlying latent space. In the proposed approach, the discriminator separates two joint distributions; the given sample and the corresponding latent space (the output of the encoder) versus the original latent space and its generated synthetic sample (the output of the generator). In a follow-up study, ALAD [55] tried to improve the previous works, by incorporating two more discriminators by employing an architecture called ALICE [56]. They achieved more stable training and more accurate image reconstructions.

GANomaly [31] is developed to learn both image and latent representations jointly. The proposed model consists of an adversarial auto-encoder (Encoder-Decoder) as a generator to learn real image representations, followed by an encoder to learn the latent space representations, and a discriminator sub-module to classify fake and real images. GANomaly is based on a DCGAN architecture and feature matching to solve the problem of learning instability. The model is trained based on three different losses; an adversarial loss, a contextual loss, and an encoder loss to generate realistic looking images and optimize the encoding process. Here the anomaly score is defined based on the encoder loss. f-AnoGAN [32] is another proposed GAN-based AD model that uses a WGAN to capture more smooth representations, by employing an encoder for inverse mapping which is trained using the generator and the discriminator after the training phase was done.

Although employing additional networks like encoders or decoders for GANs may lead to a more efficient reconstruction of images, this may pose challenges including the need of training more networks and consequently having more parameters to be learned, hence, the well-known issues such as overfitting, or data memorizing problems [47].

GAN has proven to be effective as an unsupervised anomaly detection technique. It has overcome various challenges, such as a lack of adequately labeled datasets, unbalanced datasets, and a lack of anomalous data. These limitations are especially difficult to overcome in the field of medical imaging. However, the success rate of GAN-based AD approaches on medical images still needs to be investigated. To this end, in the next section, we conducted a number of experiments on using GAN-based AD approaches for anomaly detection on various publicly available medical imaging datasets in order to further investigate their power and address their challenges in medical imaging anomaly detection. More specifically, we applied 3 different GAN-based AD approaches on 7 public imaging datasets, namely a head hemorrhage CT images dataset, two different brain tumor MRI images datasets, a mammographic images dataset, a retinal OCT Images dataset, and a blood cancer images dataset.

IV. METHODS

In this section, we present the detail of our experiments on comparing three DL-based AD methods on seven publicly available medical image datasets. We investigated the performance of these three models from two main perspectives: model-centric and data-centric points of view. These seven datasets are picked each with different characteristics in order to have a fair comparison between the methods. The target datasets differ in terms of the number of samples, the image sizes, the interested organ or tissue, disease type, modality type, and the characteristics of the abnormality to be detected. On the other hand, three state-of-the-art AD models are picked for the purpose of comparison, each with different

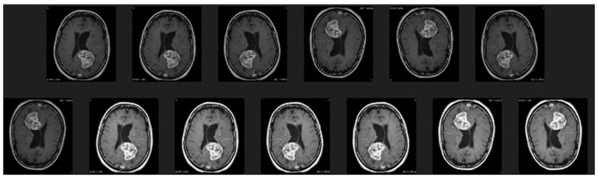

FIGURE 2. Examples of augmented images in the Br35H-MRI dataset [59].

structural and conceptual characteristics. In providing these comparisons, our goal was to highlight the challenges that AD approaches could face in the medical imaging area, from different points of view.

A. DATASETS

For this work, we gathered 7 medical imaging datasets from different modalities and focused on different organs/tissues. These datasets and their specifications are briefly described as follows:

1) HEAD HEMORRHAGE CT IMAGES (HEAD-CT)

This dataset [57] consists of 100 normal head CT single 2d slices together with 100 2d CT image slices each with a visible hemorrhage in the brain area. Each 2d slice belongs to a different patient. The hemorrhage abnormality is visually observable in each of the 100 slices. These images come in a range of dimensions and pixel sizes.

2) BRAIN TUMOR MRI IMAGES

This dataset [58] consists of 98 normal brain MRI images along with 155 images with brain tumors. These tumors are from a variety of sizes and locations. However, regardless of their variable size and distributed locations, they are visually observable. These images have a diverse range of spatial dimensions.

3) BRAIN TUMOR MRI IMAGES (BR35H-MRI)

Similar to the aforementioned dataset, these data [59] are a collection of 1500 normal brain MRI images and 1500 images that contain tumors. Again, the dataset contains variable-sized images with abnormal images that are visually detectable. The advantage of this data set compared to the previous one is the higher number of samples while being balanced in terms of normal or anomalous samples. However, it should be noted that the higher number of samples in this dataset is partly due to the use of data augmentation methods such as flipping vertically and horizontally and changing contrast (Fig. 2).

4) MAMMOGRAPHIC IMAGE ANALYSIS SOCIETY (MIAS-MAMMO) & MIAS-PATCHES-MAMMO

This collection [60] is composed of mammographic images divided into three classes normal, benign, and malignant. So, in order to formulate the dataset for AD, we considered both benign and malignant classes as abnormal, while the rest are remained as normal classes, resulting in 207 normal images and 115 abnormal images with dimensions 1024 × 1024. The abnormalities in the images are relatively subtle, making this dataset different from the above-mentioned in terms of visual observability. To form the MIAS-PATCHES-MAMMO dataset, a part of the original image with 120 × 120 dimensions as an image patch was extracted per each sample from the MIAS-Mammo dataset based on the given coordinates of the region of interest. That is 207 patches of normal images and 117 patches of abnormal samples.

5) RETINAL OCT IMAGES

This dataset [61] consists of four classes of images being normal, Choroidal neovascularization (CNV), Diabetic macular edema (DME), and Drusen. There are 26315 normal images for training and 242 images for testing per class. The dataset contains images with a variety of dimensions. The abnormalities range from those with subtle appearances to ones that are visually recognizable.

6) BLOOD CANCER DATASET (C-NMC-LEUKEMIA)

This dataset [62] consists of 3389 normal cell images for training and 648 normal cell images and 1219 images with Acute lymphoblastic leukemia (ALL) for testing. The images' dimensions are 450 × 450. Anomalous cells in the images are hard to distinguish visually. Characteristics of all these seven publicly available datasets are summarized in TABLE 1.

B. MODELS

We trained three unsupervised AD methods on the seven medical image datasets mentioned in the previous section. These AD methods were trained solely on normal samples and then validated on a portion of the data containing both normal and anomalous samples. TABLE 2 outlines the main characteristics of the AD methods used in this study. Also, Fig. 3, Fig.4, and Fig.5 show the architecture of the models. These AD models are detailed as follows:

1) F-ANOGAN

f-AnoGAN [32] is a GAN-based AD model. At first, f-AnoGAN trains a generator G and a discriminator D same as the ones employed in WGAN, to capture a more smooth representation of the input samples. Then, in the second phase, an encoder E is trained by the trained G and D to map the images to the latent space, then reconstruct them (Fig.3). For this purpose, in the GAN training phase, the objective function is the adversarial loss based on feature matching (L_{adv-fm}) which are defined as:

$$L_{adv-fm} = \|f\,(x) - f(G\,(E(x)))\|^2\,,$$

where $f\,(x)$ shows the output of an intermediate layer of D for input x that leads to fooling D with generated images, and in the encoder training phase, the objective function is the

TABLE 1. The characteristics of the medical image datasets.

Dataset	Image size	Number of classes	Number of normal images	Number of abnormal images	Modality
Head-CT [57]	Different image sizes	2	Normal: 100 Train: 80, Test: 20	Hemorrhage: 100	CT
Brain Tumor MRI [58]	Different image sizes	2	Normal: 98 Train: 80, Test: 18	Tumor: 155	MRI
Br35H-MRI [59]	Different image sizes	2	Normal: 1500 Train: 1200, Test: 300	Tumor: 1500	MRI
MIAS-Mammo [60]	1024×1024	3	Normal: 207 Train: 167, Test: 40	Abnormal: 115 Benign: 64, Malignant: 52	Mammography
MIAS-Patches-Mammo [60]	120×120	3	Normal: 207 Train: 167, Test: 40	Abnormal: 117	Mammography
Retinal OCT Images [61]	Different image sizes	4	Train: 26315 Test: 242	CNV: 242 DME: 242 Drusen:242	OCT
C-NMC-Leukemia [62]	450×450	2	Train: 3389 Test: 648	ALL: 1219	Microscopic images

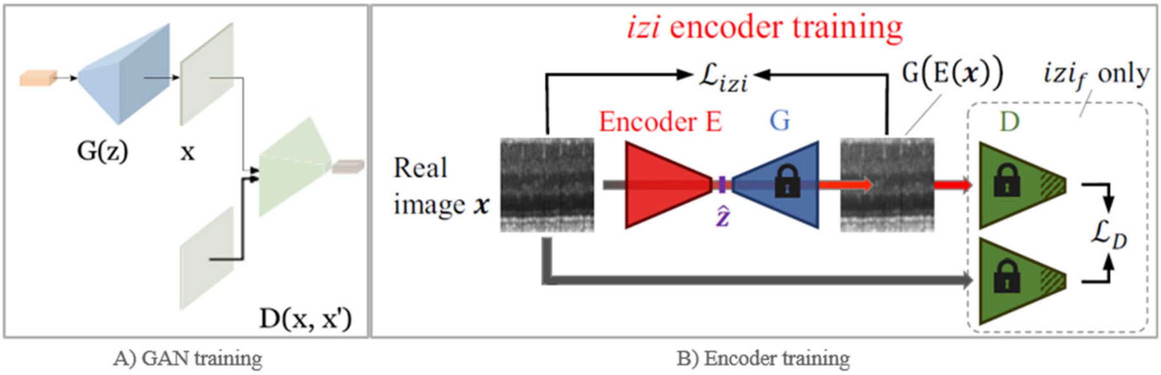

FIGURE 3. The architecture of f-AnoGAN [32].

combination of L_{adv-fm} and image reconstruction loss (L_{izi})

$$L_{izi} = \frac{1}{n} \|x - G(E(x))\|^2,$$
$$L_{enc} = L_{izi} + kL_{adv-fm}.$$

The anomaly score is then measured based on the combination of L_{adv-fm} and L_{izi} losses.

2) GANOMALY

Similar to f-AnoGAN, GANomaly [31] model is also a GAN-based AD. During the training phase, the model learns the distribution of both normal images and their corresponding latent spaces jointly through an encoder-decoder-encoder architecture (Fig. 4). To achieve this, an adversarial auto-encoder as a generator G_E and G_D, followed by an encoder E, and a discriminator D are trained simultaneously. Through the encoder and decoder networks, the generator learns the representation of the input data, $z = G_E(x)$, and reconstructs the input image, $x' = G_D(z)$. The second sub-network is the encoder network that compresses the reconstructed image, $z' = E(x')$ from the previous step. Contrary to other auto-encoder-based methods, where the bottleneck features are used to reduce the latent vectors, this sub-network expressly learns to minimize the distance which later is used to do AD during the test time. The last sub-module – the discriminator network – aims to categorize the input sample and the corresponding reconstructed one as real or fake $D(x, x')$, respectively. The objective function in this process is the combination of three loss functions; adversarial loss (L_{adv-fm}),

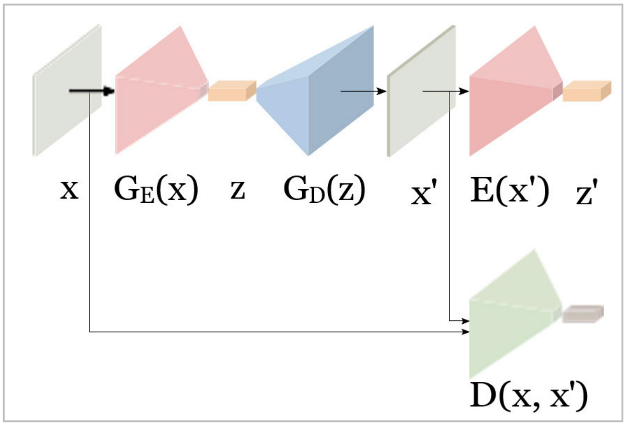

FIGURE 4. The architecture of GANomali [31].

contextual loss (L_{con}), and encoder loss (L_{enc}) which are defined as:

$$L_{adv-fm} = \|f(x) - f(G(x))\|_2,$$

where $f(x)$ shows the output of an intermediate layer of D for input x that leads to fooling D with generated images.

$$L_{con} = \|x - G(x)\|_1,$$

that shows the distance between the input x and its reconstructed image x' in the image space, and

$$L_{enc} = \|G_E(x) - E(G(x))\|_2,$$

that shows the distance between the input x and its reconstructed image x' in the feature space.

As assumed that the generator is trained to encode features of the generated image for normal samples. Therefore, the generator minimizes L_{enc}, so L_{enc} is considered as the anomaly score.

3) MULTI-KD

Multi-KD [37] utilizes a knowledge distillation technique to detect anomalous data by transferring the intermediate knowledge of a pre-trained VGG-16 as a source network S to a smaller network as a cloner C (Fig. 5). S is pre-trained on the ImageNet dataset, and C is similar to S's structure but smaller. C is trained to learn the knowledge of the S model on the normal samples at pixel and semantic levels from different critical intermediate layers CP. The notion of knowledge is defined as the value and direction of all activation values that transfer from selected layers of S to C. Therefore, based on this definition, two loss functions L_{val} and L_{dir} are defined as:

$$L_{val} = \sum_{i=1}^{N_{\mathrm{CP}}} \frac{1}{N_i} \sum_{j=1}^{N_i} \left(a_s^{\mathrm{CP}_i}(j) - a_c^{\mathrm{CP}_i}(j)\right)^2,$$

where N_i represents the number of neurons in layer CP_i, $a^{\mathrm{CP}_i}(j)$ is the value of the j-th activation of layer CP_i, and N_{CP} shows the number of critical layers.

$$L_{dir} = 1 - \sum_i \frac{\mathrm{vec}\left(a_s^{\mathrm{CP}_i}\right)^T . \mathrm{vec}\left(a_c^{\mathrm{CP}_i}\right)}{\left\|\mathrm{vec}\left(a_s^{\mathrm{CP}_i}\right)\right\| \left\|\mathrm{vec}\left(a_c^{\mathrm{CP}_i}\right)\right\|},$$

where the result of vec(x) function is a 1-D vector of x.

Hence, the discrepancy of the intermediate behavior of S and C is formulated by a combination of two loss functions L_{val} and L_{dir} as L_{total}, then used to detect anomalies at the testing time.

C. EXPERIMENTAL SETTINGS

In all experiments, the training set only contains normal images and the test set contains normal and abnormal images. For all seven datasets, the number of training and test samples is shown in TABLE 1. Additionally, all images of datasets are resized to 128 × 128 pixels and normalized to the range [0,1].

A publicly available unofficial implementation of f-AnoGAN[1] and official implementations of GANomaly[2] and Multi-KD[3] via PyTorch [63] were employed. In all experiments, the batch size was set to 32 and the learning rate for f-AnoGAN and GANomaly were set to 0.0002 and for Multi-KD was set to 0.001. Other parameters were fixed as default values while training.

The f-AnoGAN and GANomaly models were implemented using PyTorch 1.8.1+cu102 and Python 3.8.1, and the Multi-KD model was implemented using PyTorch 1.6.0 and Python 3.6.12. All experiments were performed using an NVIDIA GeForce RTX 2060 GPU with CUDA 11.2 and CUDNN 8.0.3.

V. RESULTS AND DISCUSSION

Here we report the obtained results of the experiments carried out using three AD methods on the datasets. To evaluate the performance of the AD methods, we computed Precision, Recall or Sensitivity, Specificity, F1-Score, the area under the curve of the Receiver Operating Characteristic curve (ROC AUC), and the area under the curve of the Precision-Recall curve (PR AUC) using True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) for each experiment which are defined in the following equations:

$$\begin{aligned}
\mathrm{Precision} &= \mathrm{TP} \big/ (\mathrm{TP} + \mathrm{FP}), \\
\mathrm{RecallorSensitivity} &= \mathrm{TP} \big/ (\mathrm{TP} + \mathrm{FN}), \\
\mathrm{Specificity} &= \mathrm{TN} \big/ (\mathrm{TN} + \mathrm{FP}), \\
\mathrm{F1} - \mathrm{Score} &= 2 \times \frac{(\mathrm{Precision} \times \mathrm{Recall})}{(\mathrm{Precision} + \mathrm{Recall})}, \\
\mathrm{TruePositiveRate} &= \mathrm{Sensitivity}, \\
\mathrm{FalsePositiveRate} &= 1 - \mathrm{Specificity}.
\end{aligned}$$

TABLE 3 presents the results of three AD methods obtained on each of the 7 medical image datasets in terms of Precision, Recall, F1-Score, ROC AUC, and PR AUC. For a more accurate comparison of the results, the Sensitivity and Specificity were also calculated and shown in Fig. 6 and Fig. 7. The ROC curve and PR curve of all experiments are shown in Fig.8.

The first thing that can be observed from the results summarized in TABLE 3 is the significant difference in the performance of an AD model on different datasets as well as the different performances of different models on the same dataset.

The success of a DL-based AD model in grasping the concept of normality from the training set significantly affects its performance, which is in fact related to both the supplied dataset's properties and the model mechanism used for this purpose. The higher the variation of the image dataset, particularly the medical images, makes defining the concept of normality more difficult. Therefore, the challenges of AD in

[1] https://github.com/A03ki/f-AnoGAN
[2] https://github.com/samet-akcay/ganomaly
[3] https://github.com/Niousha12/Knowledge_Distillation_AD

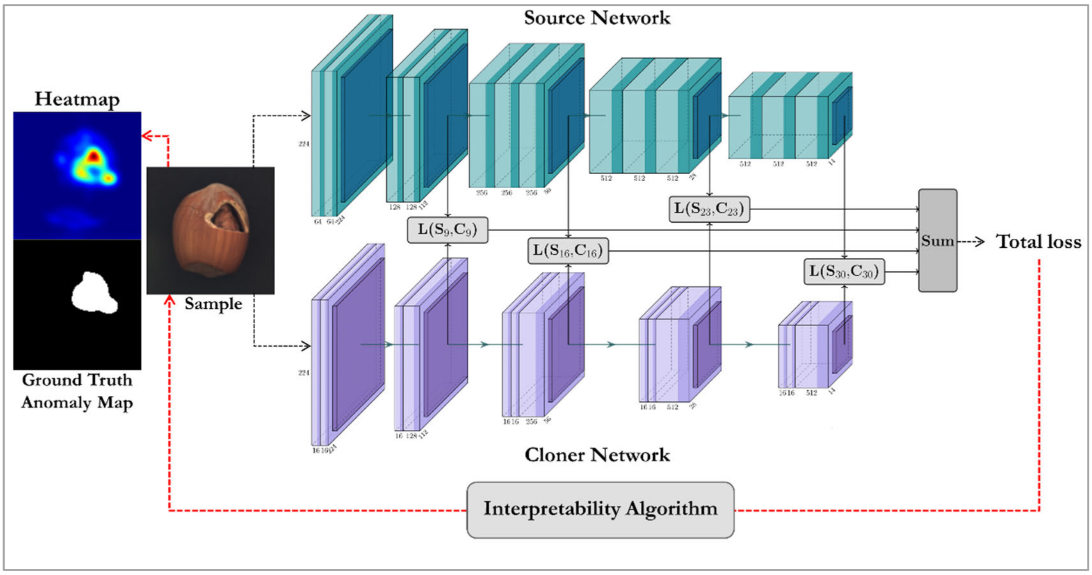

FIGURE 5. The architecture of Multi-KD [37].

TABLE 2. Summary of the unsupervised AD models.

	Training phase	Testing phase	Anomaly score
f-AnoGAN, Schlegl et al. [32]	• Two-stage training: GAN training & Encoder training • First, a WGAN was trained on normal images. • Second, an encoder was trained by the trained WGAN to map images to the latent space with the combination of adversarial loss L_{adv-fm} and reconstruction loss L_{izi}.	• The trained WGAN and encoder are utilized to map the given image x to the latent space z, then the generator generates x' as a reconstruction of x using z. • The adversarial loss L_{adv-fm} and deviation of x and x' as the reconstruction loss $L_{img-rec}$ are calculated.	As the GAN and encoder were trained on normal images, it is expected the generated sample was more similar to normal images. Therefore, the combination of L_{adv-fm} and L_{izi} is considered as anomaly score. Abnormal images result in a larger amount of deviation and normal images result in a smaller amount of deviation.
GANomaly, Akcay et al. [31]	• One-stage training on normal images. • An adversarial autoencoder (Encoder-Decoder) as a generator, followed by an encoder, and a discriminator are trained simultaneously to learn the distribution of both normal images and their corresponding latent spaces jointly. • The objective function in this process is the combination of three loss functions; adversarial loss based on feature matching (L_{adv-fm}), contextual loss (L_{con}), and encoder loss (L_{enc}). • The generator and discriminator are based on a DCGAN.	• The trained adversarial autoencoder is utilized to map the given image x to the latent space z, then generate x' as a reconstruction of x using z. Afterward, the trained encoder is utilized to map x' to the latent space z'. • The deviation of z and z' is calculated as the encoder loss L_{enc}.	Here, it is assumed that the generator is trained to encode features of the generated image for normal samples. Therefore, the generator minimizes the encoder loss (L_{enc}), so L_{enc} is considered as anomaly score.
Multi-KD, Salehi et al. [37]	• Based on the knowledge distillation technique. • A pre-trained VGG16 network on ImageNet is utilized as a source network S. Also, a cloner network C is trained on normal images using transferred knowledge of intermediate layers of S in various abstraction levels with calculating the discrepancy of their intermediate behavior based on L_{val} and L_{dir} that is formulated by a total loss function.	• The given image x is fed to the trained S and C. • The discrepancy of their important intermediate layers is calculated as L_{val} and L_{dir}.	As S only transfers knowledge of normal images to C, C's behavior with abnormal images is different from S's. Therefore, this distance is considered as anomaly score that is a a combination of two loss functions L_{val} and L_{dir} as L_{total}.

Abbreviations: $\boldsymbol{L_{adv-fm}}$: Adversarial loss based on feature matching; $\boldsymbol{L_{izi}}$: Image reconstruction loss; $\boldsymbol{L_{con}}$: Contextual loss; $\boldsymbol{L_{enc}}$: Encoder loss; $\boldsymbol{L_{val}}$: Activation values loss; $\boldsymbol{L_{dir}}$: Activation directions loss.

medical imaging datasets necessitate more thorough concerns regarding both the data-centric and model-centric points of view.

The loss functions for the GAN-based AD method during training as well as the training loss function of the Multi-KD method are shown in Fig. 9.

TABLE 3. Results of Precision, Recall, F1-Score, ROC AUC, and PR AUC for the AD methods on the medical image datasets. We highlight the best results in blue color and the lowest results in red color.

	Precision			Recall			F1-Score			ROC AUC			PR AUC		
	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
Head-CT	**0.861**	**0.840**	0.855	**0.990**	**1.000**	**1.000**	0.921	**0.913**	**0.922**	**0.761**	0.739	**0.698**	0.918	**0.921**	**0.885**
Brain Tumor MRI	0.917	**0.896**	**0.934**	1.000	1.000	1.000	0.957	**0.945**	**0.966**	**0.729**	**0.704**	0.727	**0.953**	0.950	**0.935**
Br35H-MRI	**0.913**	0.939	**0.960**	**0.974**	0.988	**0.989**	**0.942**	0.963	**0.974**	0.922	**0.905**	**0.974**	0.982	**0.966**	**0.994**
MIAS-Mammo	0.742	0.742	0.742	1.000	1.000	1.000	0.852	0.852	0.852	**0.526**	**0.596**	0.578	**0.781**	0.800	**0.814**
MIAS-Patches-Mammo	**0.745**	**0.770**	0.758	**1.000**	**0.974**	0.991	**0.854**	**0.860**	0.859	**0.490**	0.585	**0.628**	**0.714**	0.773	**0.843**
CNV-OCT	**0.746**	0.867	**0.967**	**0.909**	0.917	**0.963**	**0.819**	0.892	**0.965**	**0.886**	0.922	**0.991**	0.863	**0.851**	**0.992**
DME-OCT	**0.689**	0.785	**0.942**	0.905	**0.831**	**0.934**	**0.782**	0.807	**0.938**	**0.842**	**0.842**	**0.981**	0.808	**0.778**	**0.982**
Drusen-OCT	**0.544**	0.600	**0.871**	**0.921**	0.905	**0.897**	**0.684**	0.722	**0.884**	**0.663**	0.718	**0.934**	0.659	**0.649**	**0.938**
C-NMC-Leukemia	0.665	**0.805**	**0.654**	0.988	**0.916**	**1.000**	0.795	**0.857**	**0.791**	**0.475**	**0.824**	0.481	**0.609**	**0.851**	0.614

Abbreviations: **M1**: f-AnoGAN; **M2**: GANomaly; **M3**: Multi-KD; **ROC AUC**: The area under the curve of the Receiver Operating Characteristic curve; **PR AUC**: The area under the curve of the Precision-Recall curve.

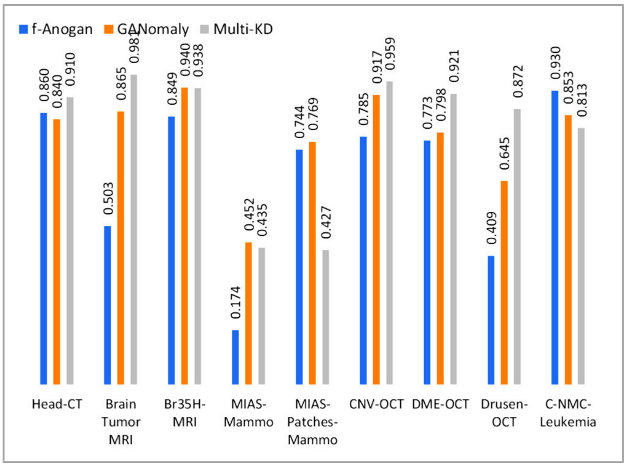

FIGURE 6. The sensitivity of the AD methods on the medical image datasets.

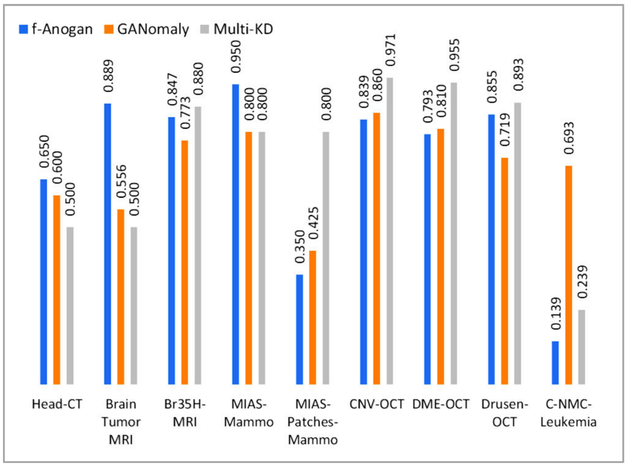

FIGURE 7. The specificity of the AD methods on the medical image datasets.

A. HEAD-CT, BRAIN TUMOR MRI, AND BR35H-MRI DATASETS

The achieved results show that the performance of the models on two datasets of Head-CT and Brain Tumor MRI are more or less similar, and relatively low, mostly owing to the similarly small number of samples in both datasets. When the training dataset is not large enough with respect to the capacity of the AD model, during training, the discriminator module tends to simply memorize the training data, resulting in overfitting. Hence the model will collapse. Consequently, the quality of the generated images deteriorates [64]. This could be an important setback, especially in the medical imaging area where data collection is an expensive process.

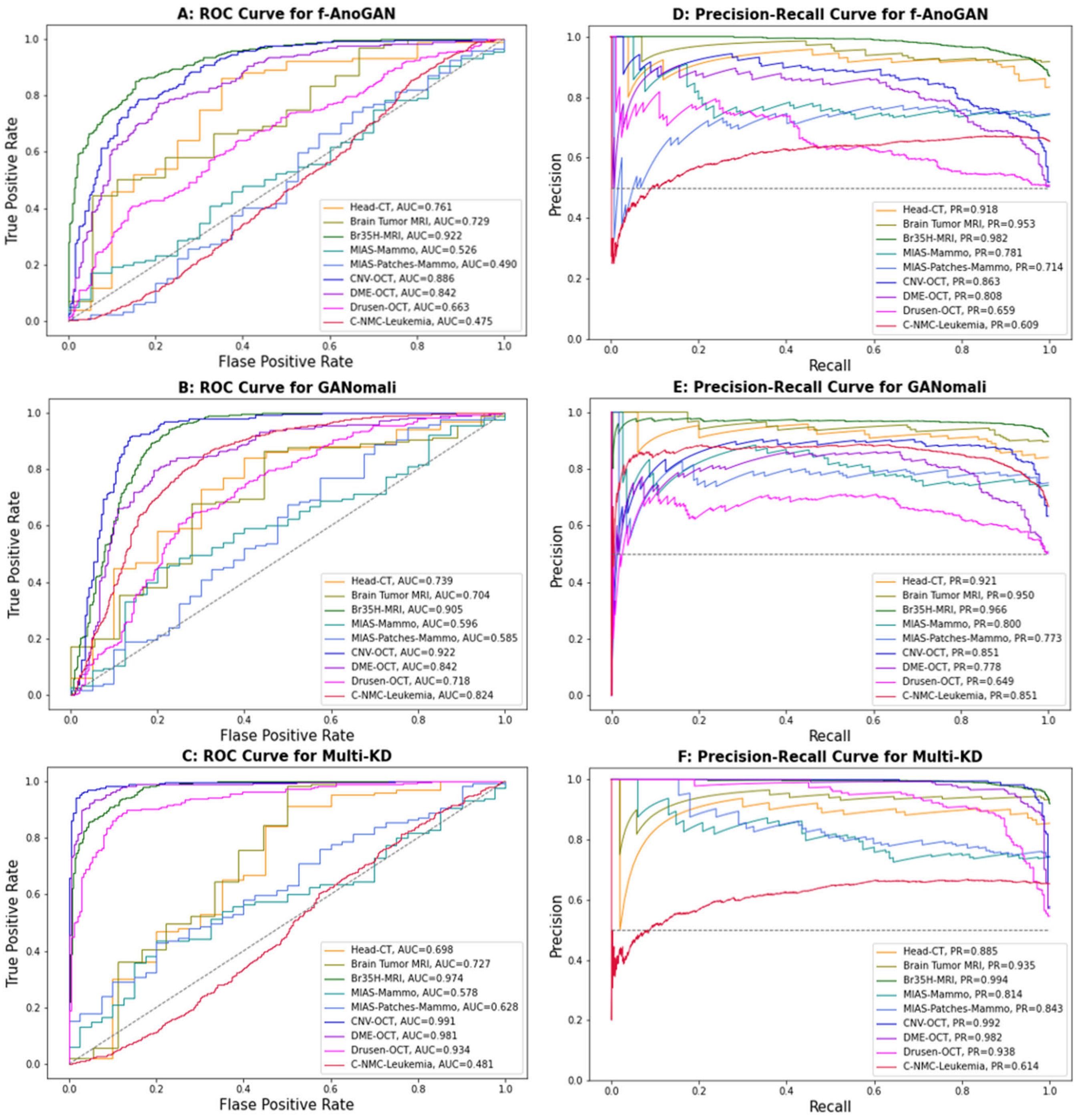

FIGURE 8. ROC curves for A) f-AnoGAN, B) GANomali, and C) Multi-KD, and Precision-Recall curves for D) f-AnoGAN, E) GANomali, and F) Multi-KD.

One possible solution to address this challenge to improve GANs' performance and robustness is data augmentation. Studies demonstrated that data augmentation; if done for both real and generated images, can boost GANs' performance. Whereas it could not be the same effect if it has been done only for real images [65], [66]. Our results however on the Br35H-MRI dataset which contains augmented brain MRI images (Fig. 2) show much higher performance compared to the Brain Tumor MRI dataset with a lower number of samples, regardless of the fact that data augmentation has only been done on the input data, indicating that data augmentation can help to improve the AD models performance, even if it is done solely on the input data and not on the constructed data.

In GANs, the quality of images generated from both normal and abnormal samples could be taken as a clue for

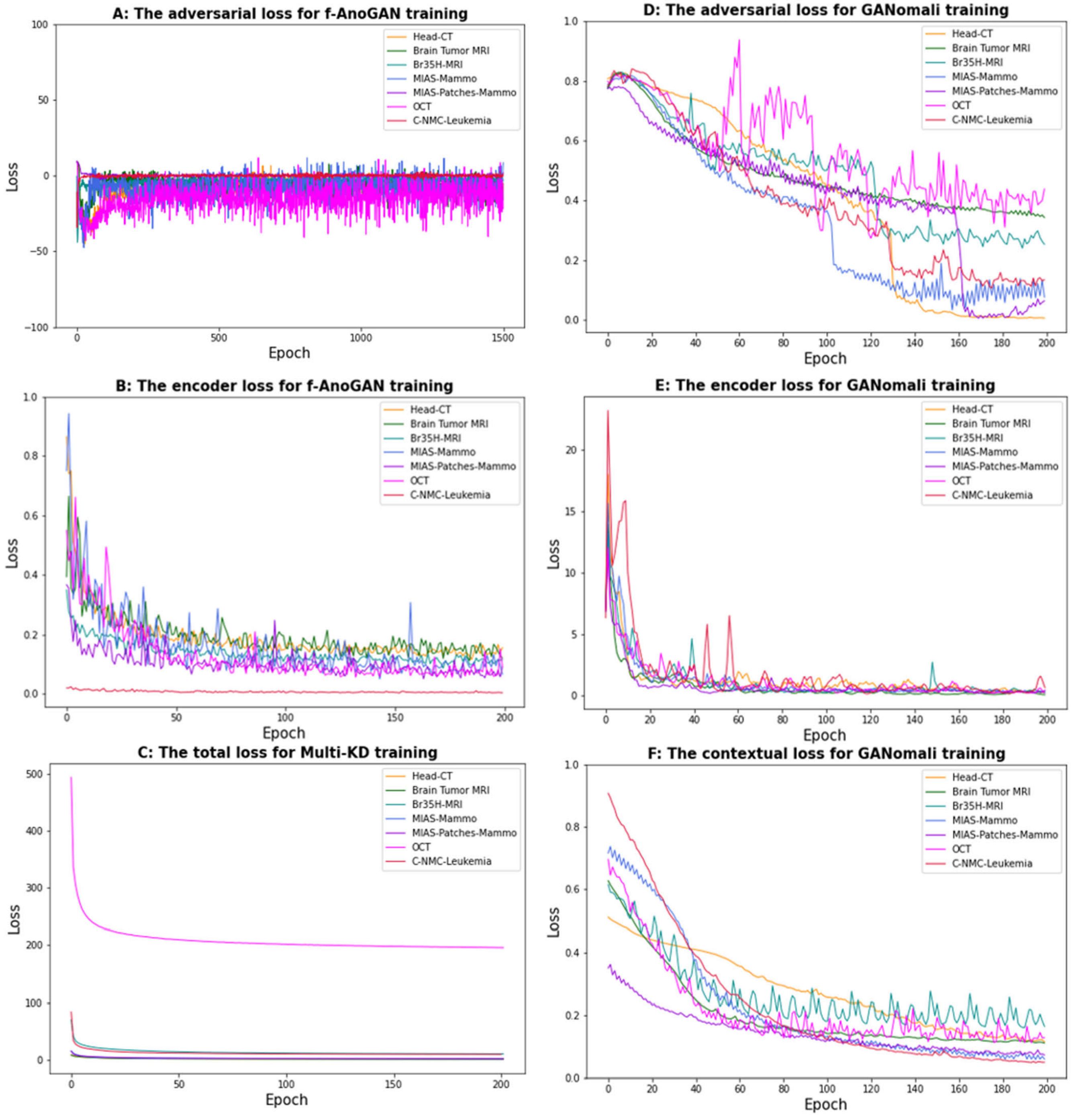

FIGURE 9. Graph of loss functions during AD models training. A) The adversarial loss and B) the encoder loss for f-AnoGAN training, C) the total loss for Multi-KD training, and D) the The adversarial loss, E) the encoder loss, and F) the contextual loss for GANomali training.

evaluating the performance of the AD model and consequently explaining the final decision made by the model. Fig. 10 and Fig. 11 present some normal and abnormal samples of Head-CT, Brain Tumor MRI, and Br35H-MRI datasets along with their corresponding reconstructions using two examined GAN-based AD models. MRI modality in general depicts anatomy in higher detail and provides better contrast and sharper image, especially for the soft tissues; in contrast, CT gives a better holistic picture of the cortical bones with higher contrast [67].

As discussed earlier, it can be observed that with a lower training size, GANs might fail in learning the details and as a result, it only provides the general shape of the brain. For instance, in the case of CT images (Head-CT dataset), the cortical bone is well reconstructed, especially due to the high contrast of the region, while in MRI images (Brain Tumor

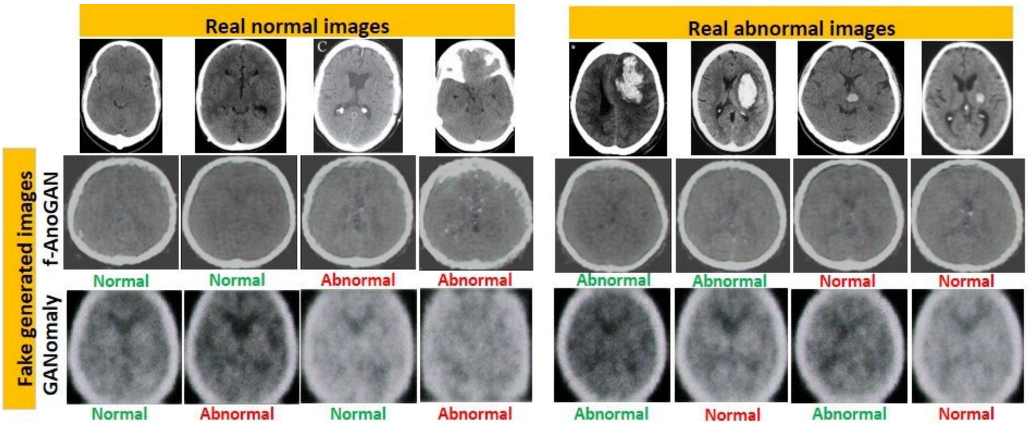

FIGURE 10. Examples of the Head-CT dataset (the first row) and their reconstructed images by f-AnoGAN (the second row) and GANomaly (the third row) along with their predicted labels. We marked labels with green if the predicted label matches the true label, and red otherwise.

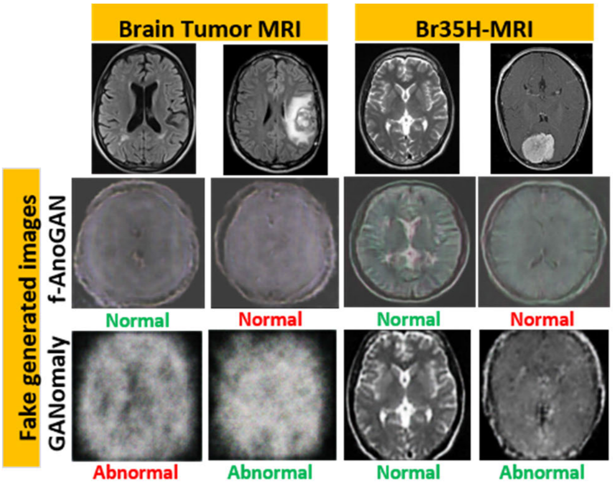

FIGURE 11. Normal and abnormal examples of the Brain Tumor MRI and Br35H-MRI datasets (the first row) and their reconstructed images by f-AnoGAN (the second row) and GANomaly (the third row) along with their predicted labels. Labels are marked green if the prediction matches the true label, and red if it does not.

MRI dataset) it is not the case. Notably, by applying data augmentation only on real samples (Br35H-MRI dataset), GAN was able to learn and reconstruct images with more details, especially normal samples.

We should mention that the results obtained by the Multi-KD method on all three brain datasets are almost similar to what we achieved using GAN-based AD methods.

B. MIAS-MAMMO AND MIAS-PATCHES-MAMMO DATASETS

Results show that all three AD methods performed relatively poor on the MIAS-Mammo dataset despite the uniform and high resolutions of the images. Moreover, given the high resolution of the images in this dataset, we extracted patches from the regions of interest in the images, to better cope with the dispersion of abnormalities in the breast tissue. We prepared the patches dataset aiming at increasing the accuracy of the AD models, yet, the obtained results remained still relatively low. One possible explanation could be the fact that feeding the AD networks with extracted patches as opposed to feeding the whole slice to the model during training limits the model's ability to comprehensively model tissue composition and global information might not be considered. The performance of the Multi-KD method has been slightly better than the GAN-based methods on the MIAS-Patches-Mammo dataset. Similar to what has been reported in [37] regarding the high performance of this method in texture AD on the MVTecAD dataset [68]. Some image samples of the MIAS-Mammo and MIAS-Patches-Mammo datasets and their reconstructions are illustrated in Fig. 12.

Unlike brain abnormalities that are mostly detectable, breast cancer anomalous tissue detection using mammography images is prone to errors [69]. The variability in lesion morphology makes breast cancer detection and its characterization challenging [70], [71]. There are important factors affecting cancer detection including size, shape, density, margins, subtlety, and also location of the lesions [70]. Overall, due to the complexity of the breast cancer lesions' abnormality and their similarity to the normal tissue, these methods have failed to detect the abnormality well.

In these cases, anomalies are semantically close to a normal distribution and are not far from the concept of normality. In fact, the different definition of abnormality here does not correspond to the general premise of most AD methods. Recently, some studies [72] have addressed to this challenge as ‘‘near novelty'' detection.

C. C-NMC-LEUKEMIA DATASETS

Similarly, the results obtained on the C-NMC-Leukemia dataset show the inability of the methods, except for the GANomaly method, in detecting abnormalities in histological images, where similar to the mammography images,

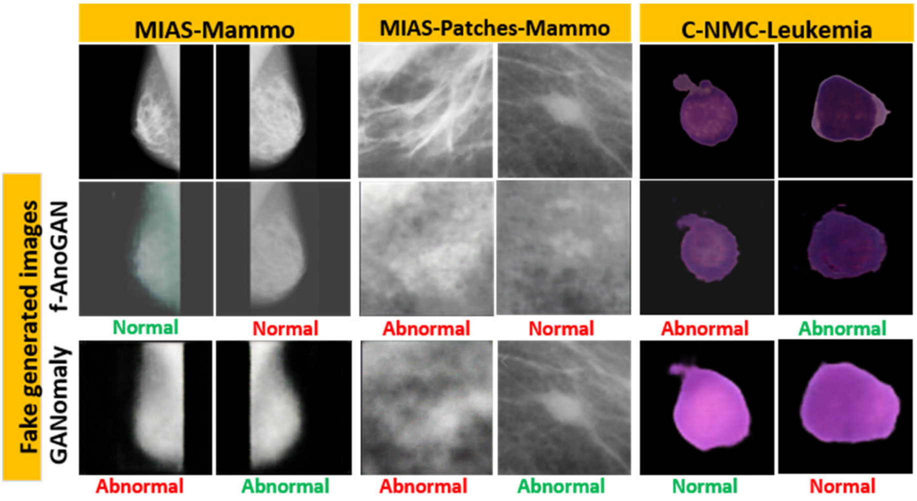

FIGURE 12. Normal and abnormal examples of the MIAS-Mammo, MIAS-Patches-Mammo, and C-NMC-Leukemia datasets (the first row) and their reconstructed images by f-AnoGAN (the second row) and GANomaly (the third row) along with their predicted labels. We marked labels with green color if the prediction label matches the true label, and red if it was not the case.

the abnormalities are very subtle. The diagnosis of Acute lymphoblastic leukemia (ALL) requires the evaluation of a variety of morphological and histological parameters, including the shape, size, and heterogeneity of cells, the volume of cytoplasm, and the number of nuclei [73]. The complexity of the abnormality and its similarity to normal cells could be one of the effective factors in the failure of distance-based AD methods. However, the acceptable performance of the GANomaly method on this dataset, compared to its performance on the MIAS-Mammo dataset as well as the lack of details in resulting reconstructed images, needs further investigation. Fig. 12 shows some normal and abnormal samples of the C-NMC-Leukemia dataset and their corresponding reconstructions.

D. RETINAL OCT DATASETS

The best results were obtained on the Retinal OCT dataset, which has a large number of training samples, especially compared to other datasets. GAN-based methods provide almost similar performance over all three different types of anomalies (CNV, DME, and Drusen), while the performance of the Multi-KD method is constantly better on all three aforementioned subsets. However, results on the Drusen anomaly sub-type are relatively less than the others, presumably due to the subtler nature of this type of anomaly.

E. FUTURE DIRECTIONS

After manually inspecting the images and comparing the correctly and incorrectly detected samples by the models, there are cases that do not seem to be hard for the models to detect but were not correctly detected by the models, even the ones with higher performances. In fact, there are images that contain very obvious abnormalities, or normal images that are visually normal but have not been identified properly. On the other hand, there are images that contain very subtle and hardly visible abnormalities, or normal images that are noisy and/or suspicious but have been correctly detected. These observations suggest that the black-box nature of DL-based models is an important barrier to their adoption, mitigating expert users' trust, especially in highly critical-safety settings such as in the healthcare ecosystem [74], [75]. Explaining the inference process and final decision of the models by explainable AI (XAI) approaches [76] could be an essential tool for both end users and model designers.

As an instance in point, diffusion models [77], [78] have recently received much attention as generative models in a variety of fields. Diffusion models typically consist of a forward process aiming at slowly corrupting the input image using an added noise, and a reverse process to reconstruct them in a step-by-step manner, in order to learn the distribution of the latent representation of input images. These non-adversarial generative models are proved to be more stable and able to model small datasets more effectively [79]. Investigating their ability as AD models is an area of future research.

Currently, the main obstacle to incorporating AI-based solutions into healthcare systems is their lack of generalization power [80]. Unsupervised AD also suffers from poor generalization [81]. Overall, numerous factors need to be considered to enable 'trustworthiness' in AI algorithms [82]. Importantly, best practice guidelines for AI model development and validation [83], [84] need to be followed. These include important considerations for study design, data, model development, model training, model testing, and evaluation. For instance, different biases need to be anticipated and very diverse datasets considered to mitigate such issues. Proof-of-concept evaluations should also be distinguished from more technical and clinical evaluations.

VI. CONCLUSION

This study demonstrated the unreliability of recent unsupervised DL-based AD methods on medical images. For this purpose, we applied 3 unsupervised DL-based AD methods with different structures and loss functions on 7 datasets of medical images with different abnormalities, pathologies, modalities, and the number of samples. Therefore, we established an almost thorough comparison between these methods and showed that their performance can be varied in different medical images, which led to some insights. The existing challenges were discussed in detail from both model-centric and data-centric points of view.

All in all, none of the methods performed well enough to be used in clinical applications. This we attribute to the diversity of anomalies in the field of biomedical imaging and challenges with the generalizability of AI methods. Our experiments showed that the effects of abnormality characteristics should be carefully considered along with the mechanisms involved in the selected AD methods. In the design and development of AD algorithms in medical images, it is suggested to consider factors such as the subtlety of the anomaly, the spread of the anomaly, tissue-related anomalies such as breast cancer in mammography images/blood cell cancers, as well as imaging modalities in which the contrast difference between the anomaly regions and normal regions is relatively similar. Hence, there is a significant need for further investigations and deployment of more robust, generalizable, and trustworthy AD models.

REFERENCES

[1] C. C. Aggarwal, "An introduction to outlier analysis," in *Outlier Analysis*. Cham, Switzerland: Springer, 2017, pp. 1–34.

[2] L. Ruff, "A unifying review of deep and shallow anomaly detection," *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, May 2021.

[3] T. Ching, "Opportunities and obstacles for deep learning in biology and medicine," *J. Roy. Soc. Interface*, vol. 15, no. 141, 2018, Art. no. 20170387.

[4] I. Goodfellow, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[5] 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, p. 2549, Oct. 2022.

[6] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, 2021.

[7] F. Di Mattia, P. Galeone, M. D. Simoni, and E. Ghelfi, "A survey on GANs for anomaly detection," 2019, *arXiv:1906.11632*.

[8] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[9] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*.

[10] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn, and R. H. Mak, "Artificial intelligence in cancer imaging: Clinical challenges and applications," *CA, Cancer J. Clinicians*, vol. 69, no. 2, pp. 127–157, Mar. 2019.

[11] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Mach. Learn.*, vol. 54, no. 1, pp. 45–66, Jan. 2004.

[12] B. Schölkopf, A. J. Smola, and F. Bach, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.

[13] M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M. Hossein Rohban, and M. Sabokrou, "A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges," 2021, *arXiv:2110.14051*.

[14] R. Chalapathy, A. K. Menon, and S. Chawla, "Anomaly detection using one-class neural networks," 2018, *arXiv:1802.06360*.

[15] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Rev. Cancer*, vol. 18, pp. 500–510, May 2018.

[16] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Phil. Trans. R. Soc. A*, vol. 374, Apr. 2016, Art. no. 20150202.

[17] M. A. Cox and T. F. Cox, "Multidimensional scaling," in *Handbook of Data Visualization*. Cham, Switzerland: Springer, 2008, pp. 315–347.

[18] D. C. Sullivan, "Metrology standards for quantitative imaging biomarkers," *Radiology*, vol. 277, no. 3, pp. 813–825, 2015, doi: 10.1148/radiol.2015142202.

[19] J. P. B. O'Connor, "Imaging biomarker roadmap for cancer studies," *Nature Rev. Clin. Oncol.*, vol. 14, no. 3, pp. 169–186, Mar. 2017, doi: 10.1038/nrclinonc.2016.162.

[20] J. Yang, R. Xu, Z. Qi, and Y. Shi, "Visual anomaly detection for images: A survey," 2021, *arXiv:2109.13157*.

[21] D. Davletshina, V. Melnychuk, V. Tran, H. Singla, M. Berrendorf, E. Faerman, M. Fromm, and M. Schubert, "Unsupervised anomaly detection for X-ray images," 2020, *arXiv:2001.10883*.

[22] D. T. Nguyen, Z. Lou, M. Klar, and T. Brox, "Anomaly detection with multiple-hypotheses predictions," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4800–4809.

[23] S. Venkataramanan, K.-C. Peng, R. V. Singh, and A. Mahalanobis, "Attention guided anomaly localization in images," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2020, pp. 485–503.

[24] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[25] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps, "Towards visually explaining variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8642–8651.

[26] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3379–3388.

[27] S. Pidhorskyi, R. Almohsen, D. A. Adjeroh, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 6823–6834.

[28] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-GANomaly: Skip connected and adversarially trained encoder–decoder anomaly detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[29] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[30] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 481–490.

[31] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2018, pp. 622–637.

[32] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.*, vol. 54, pp. 30–44, May 2019.

[33] D. Saxena and J. Cao, "Generative adversarial networks (GANs) challenges, solutions, and future directions," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–42, 2021.

[34] L. Bergman, N. Cohen, and Y. Hoshen, "Deep nearest neighbor anomaly detection," 2020, *arXiv:2002.10445*.

[35] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," 2020, *arXiv:2005.02357*.

[36] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: A patch distribution modeling framework for anomaly detection and localization," in *Proc. Int. Conf. Pattern Recognit.*, Cham, Switzerland: Springer, 2021, pp. 475–489.

[37] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14902–14912.

[38] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14318–14328.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[40] O. Rippel, P. Mertens, and D. Merhof, "Modeling the distribution of normal data in pre-trained deep features for anomaly detection," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 6726–6733.

[41] H. Gm, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "A comprehensive survey and analysis of generative models in machine learning," *Comput. Sci. Rev.*, vol. 38, Nov. 2020, Art. no. 100285.

[42] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, "Generative adversarial networks: Introduction and outlook," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 4, pp. 588–598, Sep. 2017.

[43] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[44] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, and M. Esmaeili, "Clinical data sharing using generative adversarial networks," *Connected Health*, vol. 1, no. 3, pp. 98–100, 2022.

[45] X. Xia, X. Pan, N. Li, X. He, L. Ma, X. Zhang, and N. Ding, "GAN-based anomaly detection: A review," *Neurocomputing*, vol. 493, pp. 497–535, Jul. 2022.

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

[47] A. Creswell and A. A. Bharath, "Inverting the generator of a generative adversarial network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 1967–1974, Jul. 2019.

[48] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, Cham, Switzerland: Springer, 2017, pp. 146–157.

[49] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[50] L. Deecke, R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft, "Image anomaly detection with generative adversarial networks," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Cham, Switzerland: Springer, 2018, pp. 3–17.

[51] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[52] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient GAN-based anomaly detection," 2018, *arXiv:1802.06222*.

[53] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," 2016, *arXiv:1605.09782*.

[54] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, "Adversarially learned inference," 2016, *arXiv:1606.00704*.

[55] H. Zenati, M. Romain, C.-S. Foo, B. Lecouat, and V. Chandrasekhar, "Adversarially learned anomaly detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 727–736.

[56] C. Li et al., "Alice: Towards understanding adversarial learning for joint distribution matching," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5501–5509.

[57] F. Kitamura. *Head CT Hemorrhage*. Accessed: Dec. 4, 2022. [Online]. Available: https://www.kaggle.com/datasets/felipekitamura/head-ct-hemorrhage

[58] N. Chakrabarty. *Brain MRI Images for Brain Tumor Detection*. Accessed: Dec. 4, 2022. [Online]. Available: https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection

[59] A. Hamada. (2020). *Br35H : Brain Tumor Detection*. [Online]. Available: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection

[60] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, and P. Taylor. *Mammographic Image Analysis Society (MIAS) Database v1.21 [Dataset]*. Accessed: Dec. 4, 2022. [Online]. Available: https://www.repository.cam.ac.uk/handle/1810/250394

[61] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled optical coherence tomography (OCT) and chest X-ray images for classification, Mendeley data, V2," Tech. Rep., 2018, doi: 10.17632/rscbjbr9sj.2.

[62] A. Gupta and R. Gupta, "ALL challenge dataset of ISBI 2019 [data set]. The cancer imaging archive," 2019, doi: 10.7937/tcia.2019.dc64i46r.

[63] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8024–8035.

[64] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 12104–12114.

[65] S. Zhao, Z. Liu, J. Lin, J.-Y. Zhu, and S. Han, "Differentiable augmentation for data-efficient GAN training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7559–7570.

[66] Z. Zhao, Z. Zhang, T. Chen, S. Singh, and H. Zhang, "Image augmentations for GAN training," 2020, *arXiv:2006.02595*.

[67] J. Vymazal, A. M. Rulseh, J. Keller, and L. Janouskova, "Comparison of CT and MR imaging in ischemic stroke," *Insights Imag.*, vol. 3, no. 6, pp. 619–627, Dec. 2012, doi: 10.1007/s13244-012-0185-9.

[68] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9592–9600.

[69] M. Esmaeili, S. M. Ayyoubzadeh, Z. Javanmard, and S. R. Niakan Kalhori, "A systematic review of decision aids for mammography screening: Focus on outcomes and characteristics," *Int. J. Med. Informat.*, vol. 149, May 2021, Art. no. 104406, doi: 10.1016/j.ijmedinf.2021.104406.

[70] E. U. Ekpo, M. Alakhras, and P. Brennan, "Errors in mammography cannot be solved through technology alone," *Asian Pacific J. Cancer Prevention*, vol. 19, no. 2, pp. 291–301, Feb. 2018, doi: 10.22034/apjcp.2018.19.2.291.

[71] M. Esmaeili, S. M. Ayyoubzadeh, N. Ahmadinejad, M. Ghazisaeedi, A. Nahvijou, and K. Maghooli, "A decision support system for mammography reports interpretation," *Health Inf. Sci. Syst.*, vol. 8, no. 1, p. 17, Dec. 2020, doi: 10.1007/s13755-020-00109-5.

[72] H. Mirzaei, M. Salehi, S. Shahabi, E. Gavves, C. G. M. Snoek, M. Sabokrou, and M. H. Rohban, "Fake it till you make it: Towards accurate near-distribution novelty detection," 2022, *arXiv:2205.14297*.

[73] V. G. Nikitaev, A. N. Pronichev, E. V. Polyakov, A. V. Mozhenkova, N. N. Tupitsin, and M. A. Frenkel, "Textural characteristics of bone marrow blast nucleus images with different variants of acute lymphoblastic leukemia," *J. Phys., Conf. Ser.*, vol. 945, Jan. 2018, Art. no. 012008.

[74] K. Vodrahalli, R. Daneshjou, T. Gerstenberg, and J. Zou, "Do humans trust advice more if it comes from AI? An analysis of human-AI interactions," 2021, *arXiv:2107.07015*.

[75] A. Balagopalan, H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, and M. Ghassemi, "The road to explainability is paved with bias: Measuring the fairness of explanations," 2022, *arXiv:2205.03295*.

[76] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, Cham, Switzerland: Springer, 2019, pp. 563–574.

[77] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," 2022, *arXiv:2209.00796*.

[78] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," 2022, *arXiv:2209.04747*.

[79] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, "AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 650–656.

[80] T. Eche, L. H. Schwartz, F.-Z. Mokrane, and L. Dercle, "Toward generalizability in the deployment of artificial intelligence in radiology: Role of computation stress testing to overcome underspecification," *Radiol., Artif. Intell.*, vol. 3, no. 6, Nov. 2021, Art. no. e210097.

[81] Y. Zhang, J. Wang, Y. Chen, H. Yu, and T. Qin, "Adaptive memory networks with self-supervised learning for unsupervised anomaly detection," *IEEE Trans. Knowl. Data Eng.*, early access.

[82] N. Hasani, M. A. Morris, A. Rahmim, R. M. Summers, E. Jones, E. Siegel, and B. Saboury, "Trustworthy artificial intelligence in medical imaging," *PET Clinics*, vol. 17, no. 1, pp. 1–12, Jan. 2022.

[83] T. J. Bradshaw, R. Boellaard, J. Dutta, A. K. Jha, P. Jacobs, Q. Li, C. Liu, A. Sitek, B. Saboury, P. J. H. Scott, P. J. Slomka, J. J. Sunderland, R. L. Wahl, F. Yousefirizi, S. Zuehlsdorff, A. Rahmim, and I. Buvat, "Nuclear medicine and artificial intelligence: Best practices for algorithm development," *J. Nucl. Med.*, vol. 63, no. 4, pp. 500–510, Apr. 2022.

[84] A. K. Jha, T. J. Bradshaw, I. Buvat, M. Hatt, P. Kc, C. Liu, N. F. Obuchowski, B. Saboury, P. J. Slomka, J. J. Sunderland, R. L. Wahl, Z. Yu, S. Zuehlsdorff, A. Rahmim, and R. Boellaard, "Nuclear medicine and artificial intelligence: Best practices for evaluation (the RELAINCE guidelines)," *J. Nucl. Med.*, vol. 63, no. 9, pp. 1288–1299, Sep. 2022.

MARZIEH ESMAEILI received the M.Sc. degree in information security from the Amirkabir University of Technology (Tehran Polytechnic). She is currently pursuing the Ph.D. degree in medical informatics with the Tehran University of Medical Sciences (TUMS).

Her primary research interests include deep learning and computer vision. She focuses on detecting anomaly or novelty in biomedical imaging using deep learning solutions, especially unsupervised approaches.

AMIRHOSEIN TOOSI (Graduate Student Member, IEEE) received the Ph.D. degree in computer and control engineering from the Polytechnic University of Turin.

During his Ph.D. research project, he worked on feature fusion methods for pattern recognition problems in computer vision, mainly focused on providing deep learning feature fusion-based solutions for biometrics liveness detection. He is currently a Postdoctoral Research Fellow with Qurit Lab, BC Cancer Research Institute, Vancouver, BC, Canada. He is also a former Computer Vision/Deep Learning Specialist with Abinsula s.r.l, Torino, Italy. He is also a former Postdoctoral Researcher with the Computer Graphics and Vision Group, Department of Computer and Control Engineering, Polytechnic University of Turin, where he worked on deep learning-based 3D multi-person tracking and pose estimation from multi-view cameras.

ARASH ROSHANPOOR received the Ph.D. degree in medical informatics from the Tehran University of Medical Sciences (TUMS), in 2018.

He is currently an Assistant Professor, a Senior Medical Data Scientist, and a Machine/Deep Learning Specialist, mainly in the health and medicine industry. He has more than eight years of theoretical and practical experience in delivering insights via data-driven techniques and developing data-intensive solutions, collaborating with some companies, hospitals, health research centers, and high-ranked universities of medical sciences. He has more than 13 years of experience teaching computer science, machine learning, and data science courses, and has instructed in more than six data mining, machine learning, and deep learning workshops to make academic people ready for implementing business projects. Moreover, he has collaborated as a supervisor in more than eight research and business projects. He built more than four successful projects for solving real-world problems. He is also involved in developing a system for detecting types of brain tumors with Faraadid Company.

VAHID CHANGIZI received the Ph.D. degree in medical physics from the Tehran University of Medical Sciences (TUMS), in 2005.

He is currently a Professor in radiology and radiotherapy technology with TUMS, where he was the Vice President (2014–2017) and the President (2017–2020), with the School of Allied Medical Sciences. He is also the Chair of the Department of Radiology and Radiotherapy Technology. He has published many journal articles and translated books related to his field.

MARJAN GHAZISAEEDI received the Ph.D. degree in health information management from the Iran University of Medical Sciences (IUMS).

She is currently an Associate Professor in health information management with the Tehran University of Medical Sciences (TUMS). From 2017 to 2020, she was the Vice President of the School of Allied Medical Sciences, TUMS, and the Chair of the Department of Health Information Management. She has published over 160 journal articles and three books. Her research interests include electronic health records, health information management, and related areas.

ARMAN RAHMIM (Senior Member, IEEE) received the M.Sc. degree in condensed matter physics and the Ph.D. degree in medical imaging physics from The University of British Columbia (UBC). Following the doctoral studies, he was recruited by Johns Hopkins University (JHU), leading the high-resolution brain PET imaging physics program and was pursuing research with the Department of Radiology and Electrical Engineering He is currently a Professor in radiology and physics with UBC, and a Distinguished Scientist and Provincial Medical Imaging Physicist with BC Cancer Research Institute. In 2018, he was recruited back to Vancouver, where he pursues research in molecular imaging and therapy. He has published a book and over 210 journal articles and delivered more than 130 invited lectures worldwide. He was awarded the John S. Laughlin Young Scientist Award by the American Association of Physicists in Medicine (AAPM), in 2016, and the Presidential Distinguished Service Award by SNMMI, in 2022, for significant contributions to the field of nuclear medicine and molecular imaging. Since 2020, he has been the Chair of the SNMMI Artificial Intelligence (AI) Task Force. He has been the Chair of the SNMMI Dosimetry-AI working group, since 2022.

MOHAMMAD SABOKROU received the master's and Ph.D. degrees in artificial intelligence. He has held various academic positions, including a Faculty Member with the Institute for Research in Fundamental Sciences, a Senior Researcher with the University of Oulu, and a Senior Postdoctoral Researcher with the Institute for Research in Fundamental Science. He is currently a highly experienced and accomplished AI researcher and scientist with a strong background in academic and industry settings. Throughout his career, he has demonstrated expertise in various areas, including activity recognition, outlier detection, video anomaly detection, and self-supervised video representation. He has published numerous papers in highly regarded journals and conferences, including the IEEE Transactions on Neural Network and Learning Systems, *Computer Vision and Image Understanding*, ICCV, CVPR, and ICLR.