Deep Learning based Anomaly Detection in Images: Insights, Challenges and Recommendations

Ahad Alloqmani[1], Yoosef B. Abushark[2], Asif Irshad Khan[3], Fawaz Alsolami[4]
Computer Science Department, Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—Deep learning-based anomaly detection in images has recently been considered a popular research area with numerous applications worldwide. The main aim of anomaly detection (i.e., Outlier detection), is to identify data instances that deviate considerably from the majority of data instances. This paper offers a comprehensive analysis of previous works that have been proposed in the area of anomaly detection in images through deep learning generally and in the medical field specifically. Twenty studies were reviewed, and the literature selection methodology was defined based on four phases: keyword filter, publish filter, year filter, and abstract filter. In this review, we highlight the differences among the studies included by considering the following factors: methodology, dataset, preprocessing, results and limitations. Besides, we illustrate the various challenges and potential future directions relevant to anomaly detection in images.**

Keywords*—*Anomaly detection; outlier detection; deep learning

I. Introduction

Identifying examples that deviate from what is typical or expected is the primary goal of anomaly detection and known as outlier detection [1]. Anomaly detection in images has recently been considered a popular research area with numerous applications in different fields ranging from the video surveillance field to medical fields [2] [3]. Anomalies arise due to various reasons such as data errors or data noises but sometimes indicate a new process that was previously unseen. Thus, anomaly detection is a crucial task, especially in medical image processing.

Many researchers tended to employ deep learning to detect abnormalities in images, due to the proliferation of deep neural networks, with unprecedented results across various applications. It can also deal with complicated features such as regions of interest points by examining every pixel in an image [4] [5].

In fact, deep learning-based anomaly detection have gained prominence and have been applied to various tasks, with the help of the technologies increasingly popular in the medical sector [3] [6–9]. This is because deep learning overcomes the issue of data being imbalanced, which may result in a bias towards the majority group (i.e., the negative case). Since the medical images for the negative cases are more than the positive ones, we believe that anomaly detection can be considered a better technique to be adopted than the binary classification [9].

There are several papers from different fields in the area of deep learning-based anomaly detection. We believe there is a gap in the literature about having reviews that state the gaps and limitations of the topic of interest of this article. Therefore, we opt to have a review article that collects and comprehensively analyzes recent works on deep learning-based anomaly detection in images. Hence, the community would be able to effortlessly understand the contributions and limitations of each study and to overcome these limitations in their future work.

This study aims to illustrate the state-of-the-art techniques for anomaly detection in images by reviewing recent studies that leverage deep learning techniques for anomaly detection. In our survey, we classify anomaly detection into two categories: general and medical fields in the context of medical anomalies. This study also discusses several factors that make the anomaly detection approach challenging. Such factors include the availability of labeled data, how to deal with noise that tends to be similar to the actual anomalies, and therefore, difficult to distinguish.

The significant contributions of this paper are as follows: (a) A comprehensive analysis of previous works that have been proposed in the area of anomaly detection in images through deep learning generally and in the medical field specifically by considering methodology, dataset, pre-processing, findings and limitations, outlining the difference between these studies. (b) Illustrate the various challenges and potential future directions relevant to anomaly detection in images.

The remainder of this article is organized as follows: the background of this study is given in Section II. In Section III we provide the necessary information for the reader to understand the rest of the article. Section IV discusses the literature selection methodology. Recent works of deep learning-based anomaly detection are reviewed in Section V. Observations and challenges are discussed in Section VI, while we conclude and provide the future work in Section VII and VIII.

II. Background

This section explains the necessary background to understand the various elements of this article. We briefly explain the elements of the context of this review (i.e., anomaly detection, deep learning, and automated medical image diagnosis).

A. Anomaly Detection

Anomaly detection, known as outlier detection, is defined as the process of identifying data instances that deviate

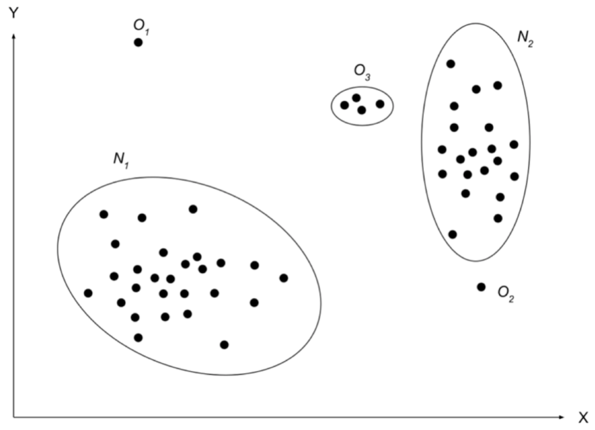

Fig. 1. Illustration of Anomalies in Two-Dimensional Dataset [5].

tremendously from other data instances [4]. As shown in Fig. 1, “N1” and “N2” are regions containing the majority of observations and are therefore considered to be normal data instance regions, while the “O3” area and the “O1” and “O2” data points are the few data points located far from the bulk of the data points. Given that “O3”, “O1”, and “O2” are therefore considered to be anomalies. They occur due to data errors but sometimes indicate a new basic process that was not previously known [5]. Anomaly detection plays an increasingly important role and is highlighted in different communities, including machine learning, computer vision, and data mining [4].

B. Deep Learning

In recent years there has been exponential development of deep learning and has been shown through several various application areas. Deep learning is considered a sub-domain of the machine learning field that aims to achieve good performance and flexibility [4]. As R. Chalapathy et al. stated in [5], deep learning achieves outstanding performance and flexibility than machine learning through learning to represent data as a nested hierarchy of concepts within the layers of a neural network. As Fig. 2 shows, deep learning outperforms the conventional approaches of machine learning considering the increased data scale [10].

C. Automated Medical Image Diagnosis

In the field of medical image processing, automated diagnosis is the primary and most important task. Automated diagnosis is based on the detection of abnormal behavior in the images [11]. Still detect abnormalities such as malignant tumors from medical images, including mammograms or CT scan, are ongoing research problems that attract a lot of attention with applications in medical diagnosis [9].

III. Terminology

There are basic terminologies in the anomaly detection field, and they are as follows.

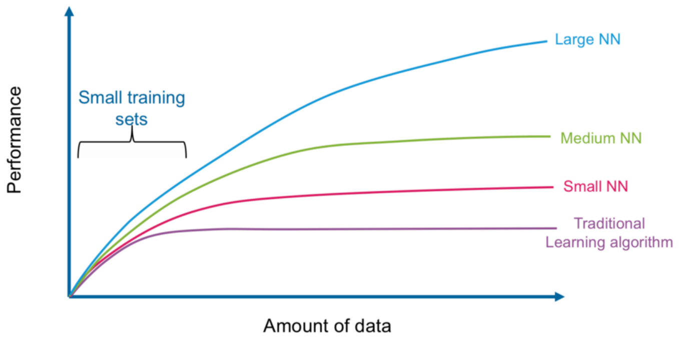

Fig. 2. Comparing the Performance of Deep Learning-based Algorithms Versus Traditional Algorithms [10].

A. Deep Learning

Deep learning is “learning feature hierarchies with features from higher levels of the hierarchy formed by the composition of lower-level features” [12]. Means deep learning learns layers of features.

B. Anomaly Detection

Anomaly detection is the process of identifying data instances that deviate from what is normal or expected data [1].

C. Semi-Supervised or (one-class classification) Deep Anomaly Detection

Defined as “a technique assumes that all training instances have only one class label” [5].

D. Unsupervised Deep Anomaly Detection

Unsupervised is “a technique that used automatic labeling of unlabeled data samples” [5].

E. Normal Data

Normal data are the majority of data instances (usually be the negative data in the medical field) [5] [9].

F. Anomalous/Abnormal data

Abnormal data are the deviants in data instances (usually be the positive/diseases data in the medical field)[5] [9].

G. Anomaly Score

is “describes the level of outlierness for each data point” [5].

IV. Literature Selection Methodology

In order to review the most important anomaly detection literature for this review, an existing selection methodology was having been adapted from [13]. This section provides a description of the process for selecting literature (see Fig. 3).

A. Keywords Filtering Stage

We started by selecting the related articles from the Google Scholar search engine, arXiv and bioRxiv using at least one of the following keywords in the title of the article: (1) anomaly detection, (2) anomaly detection in images, (3) anomaly detection in medical images, or (4) deep learning-based anomaly detection. Results from this stage 55 articles.

B. Publishers Filtering Stage

The methodology of the literature collection included article published by these publishers: (1) Springer, (2) IEEE, (3) Elsevier, (4) ACM, (5) ICLR, (6) SPIE, and (7) arXiv and bioRxiv preprints. Fig. 4 presents the percentage of articles for each publisher. Results from this stage reduced from 55 to 40.

C. Year Filtering Stage

The methodology of literature selection also focused on recent research articles in recent years by considering the following years only: (1) 2020, (2) 2019, and (3) 2018. Fig. 5 presents the percentage of articles for each year. Results from this stage reduced from 40 to 28.

D. Abstract Filtering Stage

An abstract reading was carried out in view of the 28 articles from the previous stage in order to identify only the most important articles that specifically study the deep learning-based anomaly detection in images and focus in particular on the medical field. Therefore, from the anomaly detection literature, 20 articles were chosen.

V. Recent Works of Deep Learning-Based Anomaly Detection

In this paper, twenty papers on detecting anomalies in images through deep learning generally and in the medical field specifically were reviewed. Fig. 6 presents the percentage of articles for each field.

A. General Field

This section will present some previous works of anomaly detection in terms of the general field.

The authors of this research [14], proposed Deep Semi-Supervised Anomaly Detection (Deep SAD). Furthermore, they presented an information-theoretic framework for deep anomaly detection, which as minimizing the entropy of the

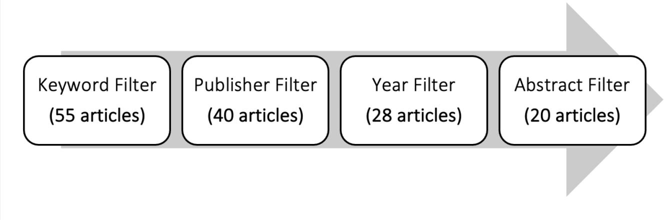

Fig. 3. Literature Selection Methodology.

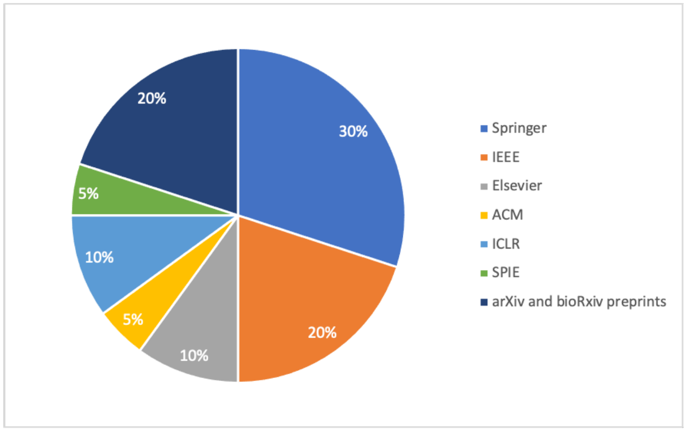

Fig. 4. Showed the Percentage Ratio of Articles for Each Publisher in 20 Articles.

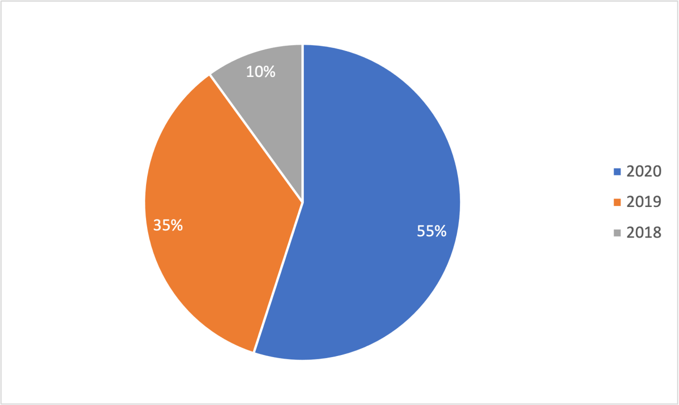

Fig. 5. Showed the Percentage Ratio of Articles for Each Year in the 20 Articles.

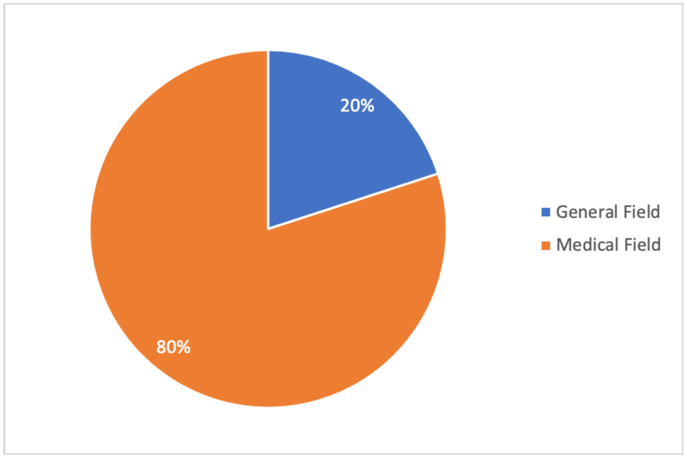

Fig. 6. Showed the Percentage Ratio of Articles for Each Field in 20 Articles.

latent distribution for normal data and maximizing the entropy of the latent distribution for anomalous data. The experiments were on several different public datasets and comparing their method with other previous methods. The results show that the method of this paper was on par or outperform other methods that compared it. The authors did not consider the problem of the difficulty availability of label anomalies.

This study [15] presented Iterative Training Set Refinement (ITSR), which is a novel method. An adversarial autoencoder architecture is geared to overcome the shortcomings of conventional autoencoders in the existence of anomalies in the training set. They used two public datasets, MNIST, and Fashion-MNIST datasets. The results show that their method has better accuracy than traditional autoencoders and adversarial autoencoders. However, they did not experiment with their method when there are noises in images which means do not consider preprocessing data. Also, they did not compare their result with other works or state-of-the-art methods.

This research [16] proposed a new framework and its instantiation Deviation Networks (DevNet) to take advantage of a few labeled anomalies with a prior probability to fulfill end-to-end differentiable learning of anomaly scores. Nine publicly available real data sets were used, and are from various critical fields, for example, fraud detection, disease detection, malicious URL detection, and intrusion detection. The experimental findings indicate that their current approach was more effective score than state-of-the-art competing methods. But the authors did not examine the lack of label anomalies data in the real world, particularly in medicine field.

On the contrary, using an unsupervised model is the proposed method of this paper [17], where the authors present a Deep Autoencoding Gaussian Mixture Model (DAGMM) for unsupervised anomaly detection. The experiment was applied to four public benchmark datasets and compared the results with state-of-the-art anomaly detection techniques. The results indicate that DAGMM exceeds state-of-the-art anomaly detection methods with a 14% improvement based on the standard F1 score. However, they did not test their method on images with noises to show the extent of its impact on the results.

B. Medical Field

This section will present some previous works of anomaly detection in terms of the medical field by considering the application area.

1) Breast: According to new research by [18], the authors introduced a new method that is a new measure for determining the effect of a particular sample on a task, allowing to detect samples outside of distribution. Their method integrated into a simple autoencoder CAE model for the abnormality recognition task. Examination of their method on Breast Magnetic Resonance Imaging (MRI) and Breast Full-Field Digital Mammography (FFDM) datasets. Experimental results demonstrate that the new method exhibits remarkable performance and outperforms the compared methods with accuracy 90.1% and 95.6% in MRI and FFDM datasets respectively. The experiments of the method are done on small datasets relatively.

The authors of this research [19] an architecture with two deep convolutional networks (R and M) proposed for irregular tissues in mammography images. They used three public datasets, the Mammographic Image Analysis Society (MIAS) and INbreast dataset for training their method. Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) dataset to test their method. The accuracy they achieve is 76% and 86% in MIAS and INbreast datasets respectively. However, the datasets used are of small size. Moreover, they did not consider processing the whole image in one step in the model.

This study [9] designed an autoencoder based on a deep neural network to detect an anomaly in medical images based on one-class classification. The INbreast dataset is used, and the performance was 84%. Also, this paper used a small dataset. Furthermore, they did not compare their result with other works or state-of-the-art methods.

2) Chest: In terms of the chest area, the confidence-aware anomaly detection (CAAD) model for viral pneumonia screening from non-viral pneumonia and healthy controls have been implemented in recent research [20] into a one-class classification-based anomaly detection challenge. Their model consists of a function extractor, a module for detecting anomalies, and a module for predicting confidence. Four datasets were used, which are X-VIRAL, X- COVID, public COVID-19, and lastly combine the X- COVID and Open-COVID datasets. The results show the accuracy is 87.57%, 83.61%, 94.93%, and 84.43% for datasets respectively. The only limitation of this research is it did not try to consider comparing without data preprocessing to see if there is much difference in results or not.

This study [21] presented an abnormality detection method based on an autoencoder with uncertainty prediction. This method is able to reconstruct the image with pixel-wise uncertainty prediction. Two public chest X-ray datasets were used: RSNA Pneumonia Detection Challenge dataset and pediatric chest X-ray dataset. The area under the curve (AUC) was 89% and 78% for datasets respectively. There is no preprocessing data step.

In [22] an end-to-end architecture to determine a chest X-ray abnormal using generative adversarial one-class learning was proposed. It is similar to generative adversarial networks (GANs). Their architecture consists of a U-Net autoencoder, a CNN discriminator, and an encoder. The experiments were done on the NIH Clinical Center Chest X-ray dataset, and they achieve 80% accuracy to detect lung opacities. But their architecture results did not compare with other algorithms.

3) Brain: Since recently, researchers have shown an increased interest in Generative Adversarial Network (GAN) on deep learning. Accordingly, this paper [23] introduced unsupervised anomaly detection Generative Adversarial Network (MADGAN) method using multiple adjacent brain MRI slice reconstruction. This approach is capable of detecting various diseases at different stages on multi-sequence structural MRI. Two different datasets were used. The MRI dataset was extracted from the Open Access Series of Imaging Studies-3 (OASIS-3) and the second dataset was collected by the authors (National Center for Global Health and Medicine, Tokyo, Japan) which is brain metastasis and various disease MRI dataset. The results demonstrate that this method can detect anomaly detection at a very early stage with 72.7% and at a late stage with 89.4% in terms of area under the curve (AUC). But their method results did not compare with other algorithms.

A method of using GANs trained from multi-modal magnetic resonance images (MRI) as a 3-channel input is defined and demonstrated by the authors in [24]. Their model was used to detect tumour as an anomaly. The dataset was from The Cancer Imaging Archive. The resulting accuracies that differ substantially in the size of the anomaly have been observed. The area under the receiver operator characteristic curve (AUROC) was observed to be greater than 75% for anomaly sizes greater than 4 cm^2. The dataset consists of 20 patients, which is very small.

In [25] proposed a semi-supervised anomaly detection model to detect brain tumor abnormalities. The model consisted of four components which are the encoder-decoder part, the discriminator, latent regularizer, and auxiliary encoder. The model first has been tested on two benchmark datasets which are MNIST and CIFAR-10 for comparison with state-of-the-art methods. Then applied the model on the HCP database and BraTS dataset. Where using normal images from the HCP database as training data and the whole BraTS 2019 dataset as the test data. The results were 93%, 79.7% for MNIST and CIFAR-10 respectively. 99.4% for the BraTS dataset. There is no preprocessing data step.

4) Eye: In research [26], it proposed a novel P-Net for retina image anomaly detection. Their network architecture consisted of three modules which are structure extraction from the original image module, image reconstruction module, and structure extraction from the reconstructed image module. Two datasets have been used, which are Retinal Edema Segmentation Challenge Dataset (RESC) and Fundus Multi-disease Diagnosis Dataset (iSee). The result was 92.88% and 72.45% for both datasets, respectively. There is no preprocessing data step.

This study [27] proposed a transfer-learning-based approach for unsupervised anomaly detection. The methodology used a convolutional neural network as a feature extractor and Isolation Forest anomaly detection method as a classification. Two benchmark datasets (CIFAR-10 and SVHN) were used, and two retinal fundus image datasets, which are Retinopathy of Prematurity (ROP) and Diabetic Retinopathy (DR) were used. The results were 88.2%, 55.4% for CIFAR-10 and SVHN respectively. 77% and 74.5% for the ROP and DR respectively. The authors did not try to consider comparing without data preprocessing to see if there is much difference in results or not. Furthermore, the medical imaging performance results need improvement.

5) Abdomen: In another research that used an unsupervised model, the authors in [28] have considered the problem of other organs than the stomach in a gastric X-ray examination, which can be noisy and cause decadence of classification performance. Therefore, they proposed a deep learning-based anomaly detection model inspired by DAGMM as an organ classification task. The experiment was on one dataset, which is gastric X-ray images, and comparing with other approaches. The results show that their model outperforms the comparison models with 95.6% in terms of sensitivity. The limitation of this paper was having a small number of stomach images with barium leaks in the gastric X-ray examinations, which are not useful in gastritis detection.

6) Cardiac: Another application area of the medical field in [29] where the authors proposed the decision boundary-based anomaly detection model using improved AnoGan from ECG data. The proposed model achieves 94.75% in MIT-BIH Arrhythmia ECG dataset, which is the best performance compared with many different models. The authors did not consider testing the model without their data preprocessing to illustrate the difference ratio.

7) Musculoskeletal: This study [30] presented a preprocessing pipeline and survey unsupervised deep learning methods for an anomaly detection task. They were comparing these methods with each other with and without their preprocessing pipeline to demonstrate which algorithm is better for this task and also to show the effect of the presence of preprocessing pipeline on the performance. They work on a subset of the MURA dataset, which is X-Ray images of hands. The results illustrated that the best model is α-GAN based (GANs) approach with 60.7%, and the best model-based autoencoder is convolutional auto-encoder (CAE) with 57%. However, the experiments were on a small dataset because they did not use a full MURA dataset.

In [31] A new CNN model consisting of some previous CNN layers with the technique of weight standardization and a learning rate scheduler was proposed. The model name is GnCNNr, an acronym for Group Normalized Convolutional Neural Networks with Regularization. MURA dataset was used for experiments. This model was compared with the conventional deep learning methods: DenseNet, Inception, Inception v2 model. The Overall performance result was 89.9% in terms of area under the receiver operating characteristic curve (AUROC). This model was compared with only conventional deep learning methods and did not compare with other works.

The authors of this research [32] introduced a new Computer-Aided Diagnosis (CADx) model based on Deep Convolutional Neural Network (Deep CNN). This model identifies musculoskeletal abnormality detection from radiographs. Ensemble techniques were used to improve the model performance. For experiments, the MURA dataset was used with four types of study (Elbow, Finger, Humerus, and Wrist). The performance results were 86.45%, 82.13%, 87.15%, and 87.86% respectively. However, their model results did not compare with other works.

VI. Discussion

A. Overview

Many studies have worked on anomaly detection algorithms. Summary of the related studies on deep learning-based anomaly detection in images is presented in Table I and II for the general and medical fields respectively. After reviewing the studies, the following was observed. First, most researchers use deep learning other than machine learning. Because deep learning has better performance and can handle the complexity of images and large datasets efficiently. Second, most researchers either in general or medical fields have used unsupervised [9] [17–20] [23, 24] [26–30], or semi-supervised [14–16] [21, 22] [25] learning methods in an anomaly detection task. Third,

TABLE I. SUMMARY OF RECENT RELATED WORKS IN THE GENERAL FIELD

[Ref.] (Year)	Methodology	Dataset (#: Sample size)	Pre-processing	Results (%:Performance of the model used)	Limitation
[14] (2019)	Deep Semi-Supervised Anomaly Detection (Deep SAD). **Feature Extraction & Classification:** - (MNIST, Fashion-MNIS, CIFAR-10): convolutional neural networks (CNNs). - (benchmark datasets): Multi-Layer Perceptron (MLP) architectures.	1. MNIST: (70, 000). 2. Fashion-MNIS: (70, 000). 3. CIFAR-10: (60,000). - Other anomaly detection benchmark datasets: 4. arrhythmia: (452). 5. cardio: (1831). 6. satellite: (6435). 7. satimage-2: (5803). 8. shuttle: (49,097) 9. thyroid: (3772).	Standardize features to have zero mean and unit variance.	1. MNIST: 96.9 % 2. Fashion-MNIS: 91%. 3. CIFAR-10: 81.9%. 4. arrrhythmia: 75.9%. 5. cardio: 95 %. 6. satellite: 91.5%. 7. satimage-2: 99.9%. 8. shuttle: 98.4 %. 9. thyroid: 98.6%.	The difficulty availability of label anomalies.
[15] (2019)	A novel method called Iterative Training Set Refinement (ITSR)for anomaly detection in images. **Feature Extraction & Classification:** Adversarial autoencoders (AAE).	1. MNIST: (70,000). 2. Fashion-MNIST: (70,000).	NA	1.1 MNIST-(observed anomaly type): 91% 1.2 MNIST-(unobserved anomaly type): 90% 2.1.1 Fashion -MNIST: (T-shirt vs. Boot-observed anomaly type): 90% 2.1.2 Fashion -MNIST: (T-shirt vs. Boot-unobserved anomaly type): 80% 2.2.1 Fashion -MNIST: (T-shirt vs. Pullover-observed anomaly type): 80% 2.2.2 Fashion -MNIST: (T-shirt vs. Pullover-unobserved anomaly type): 80%	1. No Pre-processing data. 2. There is no comparison with different algorithms.
[16] (2019)	A novel anomaly detection framework and its instantiation Deviation Networks (DevNet). **Feature Extraction & Classification:** Multi-Layer Perceptron (MLP) network architectures.	1. donors: (619,326). 2. census: (299,285) 3. fraud: (284,807). 4. celeba: (202,599). 5. backdoor: (95,329). 6. URL: (89,063). 7. campaign: (41, 188). 8. news20: (10,523). 9. thyroid: (7,200).	For all data sets, missing values are replaced with the mean value in the corresponding feature, and categorical features are encoded by one-hot encoding.	1. donors: 100% 2. census: 68.6% 3. fraud: 92.6% 4. celeba: 87% 5. backdoor: 96.8% 6. URL: 94.1% 7. campaign: 67.9 % 8. news20: 81.7 % 9. thyroid: 78.7 %	The difficulty availability of label anomalies.
[17] (2018)	Deep Autoencoding Gaussian Mixture Model (DAGMM). **Feature Extraction & Classification:** Autoencoder and Gaussian Mixture Model (GMM).	1. KDDCUP: (494,021). 2. thyroid: (3772). 3. arrhythmia: (452). 4. KDDCUP-Rev: (121,597).	One-Hot Representation to encode categorical features in (KDDCUP) dataset.	1. KDDCUP: 93.69 % 2. thyroid: 47.82% 3. arrhythmia: 49.83% 4. KDDCUP-Rev: 93.80 %	Did not comparing without data preprocessing to show the difference.

most of the researches does not leverage a limited number of labeled anomalies as prior knowledge. Therefore, using this technique in future work is a good idea to avoid identifying anomalies as data noises or uninteresting data due to the lack of prior knowledge of the anomalies of interest and to increase the model's performance, as shown in [16]. Fourth, some studies used a small dataset [9] [18, 19] [24] [30]. So, there is a lack of used large datasets in an anomaly detection task. Fifth, data preprocessing is an essential technique to obtain good performance, as shown in [30]. Some researchers considered it [18] [27–30], and others are not. Sixth and finally, most studies consider the comparison with many different algorithms to illustrate the evaluation metrics of each of them, and that is an important aspect of evaluating the effectiveness of the model.

B. Challenges

There are numerous factors that make anomaly detection very challenging. First, handling the class imbalance of normal and abnormal data. Second, availability of labeled data. Third, there is often noise in the data that appears to be close to the actual anomalies and thus difficult to differentiate them [33]. Fourth, the exact concept of the anomaly varies with different areas of application. For example, fluctuations in body temperature are a small deviation from normal and might be an anomaly in the medical field. On the other hand, fluctuations in the value of a stock with a similar deviation might be normal in the stock market domain [33]. So, it is not straightforward to adapt a method developed in one field to another.

TABLE II. Summary of Recent Related Works in the Medical Field

[Ref.] (Year)	**Application Area**	**Methodology**	**Dataset (#: Sample size)**	**Pre-processing**	**Results (%:Performance of the model used)**	**Limitation**
[18] (2020)	Breast	A new measure for determining the effect of a particular sample on a task, allowing to detect of samples outside of distribution. **Feature Extraction & Classification:** Convolutional Autoencoder CAE model.	1. Breast Magnetic Resonance Imaging (MRI): (2872). 2. Breast Full-Field Digital Mammography (FFDM): (304).	Image resizing.	1. MRI: 90.1% 2. FFDM: 95.6%	Used small datasets.
[19] (2019)		An architecture with two deep convolutional networks (R and M) based adversarial training. **Feature Extraction:** Reconstruction Network (R): Encoder-decoder networks. **Classification:** Matching Network (M): involves convolution and fully connected layers.	1. Mammographic Image Analysis Society (MIAS) dataset: (322). 2. INbreast dataset: (410).	NA	1. MIAS: 76% 2. INbreast: 86%	1. Used small datasets. 2. No Preprocessing data. 3. No process the whole image in one step.
[9] (2018)		An autoencoder based on a deep neural network. **Feature Extraction & Classification:** Autoencoder model.	INbreast dataset: (410).	NA	84%	1. Used small datasets. 2. No Preprocessing data. 3. There is no comparison with different algorithms.
[20] (2020)	Chest	Confidence-aware anomaly detection (CAAD) **Feature Extraction:** EfficientNet. **Classification:** Multi-Layer Perceptron (MLP) network architecture for anomaly detection network and four layers for Confidence prediction network.	1. X-VIRAL: (43,370). 2. X- COVID: (213). 3. Public COVID-19: (519). 4. Combine the X- COVID and Open-COVID: (2,706).	1. Image resizing. 2. Augmentation.	1. X-VIRAL: 87.57% 2. X- COVID: 83.61% 3. Public COVID-19: 94.93% 4. X- COVID and Open-COVID: 84.43%	Did not comparing without data preprocessing to show the difference.
[21] (2020)		Autoencoder with pixel-wise uncertainty prediction. **Feature Extraction & Classification:** Autoencoder.	1. RSNA Pneumonia Detection Challenge dataset: (26,684). 2. Pediatric chest X-ray dataset: (5,856).	NA	1. RSNA: 1.1 (normal vs. lung opacity): 89% 1.2 (normal vs. not normal): 78% 1.3 (normal vs. all) - (lung opacity and not normal): 83% 2. Pediatric: 78%	No Preprocessing data
[22] (2019)		An end-to-end architecture to determine a chest X-ray abnormal using generative adversarial one-class learning. **Feature Extraction & Classification:** U-Net autoencoder, CNN discriminator and second encoder.	The NIH Clinical Center Chest X-ray dataset: (112,120).	Image resizing.	1. Normal vs. Abnormal: 84.1% 2. Normal vs. Ling opacities: 80.2%	There is no comparison with different algorithms.

VII. Conclusion

This article presents a systematic study of recent research in general and medical fields on anomaly detection in images by considering methodology, dataset, pre-processing, findings and limitations, outlining the difference between these studies. The majority of anomaly detection studies focus on the medical field since it is the best technique than binary classification to cope with imbalanced data that is an issue in medical applications. The study concludes that most researchers used unsupervised or semi-supervised for anomaly detection. Further, most researchers used deep learning other than machine learning; Deep learning has better performance and can efficiently handle the complexity of images and large datasets. The limitation of this research is the limit of the number of literatures researched. While the authors used several databases, the ones used in the extensive index might not be exhaustive ones.

TABLE II. CONTINUED

[Ref.] (Year)	**Application Area**	**Methodology**	**Dataset (#: Sample size)**	**Pre-processing**	**Results (%:Performance of the model used)**	**Limitation**
[23] (2020)	Brain	Unsupervised Medical Anomaly Detection GAN using multiple adjacent brain MRI slice reconstruction (MADGAN). **Feature Extraction & Classification:** GAN with include U-Net.	1. MRI dataset extracted from the Open Access Series of Imaging Studies-3 (OASIS-3): (1,606 scans). 2. Brain metastasis and various disease MRI dataset collected by the authors (National Center for Global Health and Medicine, Tokyo, Japan): (193 scans).	NA	At a very early stage: 72.7% At a late stage: 89.4%	There is no comparison with different algorithms.
[24] (2020)		A method of using GANs trained from multi-modal magnetic resonance images (MRI) as a 3-channel input. **Feature Extraction & Classification:** GAN	Multi-modal magnetic resonance brain images MRI dataset from The Cancer Imaging Archive: (308).	NA	(AUC) was observed to be greater than 75% for anomaly sizes greater than 4 cm^2. Sensitivity (Sen): All tumours: 99% Area $>$4 cm^2: 99% Area $>$7 cm^2: 97%	1. Used small datasets. 2. There is no comparison with different algorithms.
[25] (2020)		A semi-supervised anomaly detection model to detect brain tumor abnormalities. **Feature Extraction & Classification:** The GAN-style architecture: the encoder-decoder part, the discriminator, auxiliary encoder, and latent regularizer.	1. MNIST dataset: (70,000). 2. CIFAR-10 dataset: (60,000). 3. HCP database -Training only-: (65 healthy patients). 4. BraTS dataset: (335 patients).	NA	1. MNIST: 93% 2. CIFAR-10: 79.7% 3. BraTS: 99.4%	No Preprocessing data.

TABLE II. CONTINUED

[Ref.] (Year)	**Application Area**	**Methodology**	**Dataset (#: Sample size)**	**Pre-processing**	**Results (%:Performance of the model used)**	**Limitation**
[26] (2020)	Eye	A novel P-Net methodology is proposed by the researcher for the detection of anomalies in retina images. **Feature Extraction & Classification:** U-Net autoencoder and Discriminator architecture.	1. Retinal Edema Segmentation Challenge Dataset (RESC): NA. 2. Fundus Multi-disease Diagnosis Dataset (iSee): (10,000).	NA	1. RESC: 92.88% 2. iSee: 72.45%	No Preprocessing data.
[27] (2019)		This research applied transfer-learning based method for unsupervised anomaly detection. **Feature Extraction:** CNN: Inception-ResNet-v2 network **Classification:** Isolation unsupervised anomaly detection method (Isolation Forest method).	1. CIFAR-10 dataset: (60,000). 2. SVHN dataset (99,289): 3. Retinopathy of Prematurity (ROP): (5511). 4. Diabetic Retinopathy (DR): (11,741).	CIFAR-10 & SVHN: Rescaling the images to [0, 1]. ROP & DR: Squared cropped to cut the neutral background and resized images to 256 pixels.	1. CIFAR-10: 88.2% 2. SVHN: 55.4% 3. ROP: 77% 4. DR: 74.5%	1. Did not comparing without data preprocessing to show the difference. 2. Medical imaging performance results need improvement.

TABLE II. CONTINUED

[Ref.] (Year)	Application Area	Methodology	Dataset (#: Sample size)	Pre-processing	Results (%:Performance of the model used)	Limitation
[28] (2020)	Abdomen	Deep Autoencoding Gaussian Mixture Model (DAGMM). **Feature Extraction & Classification:** Convolutional Autoencoder and Gaussian Mixture Model (GMM).	Gastric X-ray dataset: (48,012).	Image resizing.	- Sensitivity (Sen): 95.6 % - Specificity (Spe): 98% - Harmonic mean of sensitivity and specificity (HM): 96.8%	There are still a minimal amount of stomach images in the gastric X-ray examinations with barium leakage that are not successful in addressing gastritis detection.
[29] (2020)	Cardiac	This research suggested a decision boundary-based Anomaly Detection model using improved AnoGan that uses ECG dataset. **Feature Extraction & Classification:** AnoGan.	MIT-BIH Arrhythmia ECG dataset: (85,717).	1. Filtering using Hamilton algorithm. 2. R-peak Detection and ECG Data Segmentation for signal processing. 3. Gray Scale Conversion and resize for image processing. 4. Segmentation is performed in the range between 0.3 seconds and 0.4 seconds on the basis of R-peak.	94.75%	Did not comparing without data preprocessing to show the difference.

TABLE II. CONTINUED

[Ref.] (Year)	Application Area	Methodology	Dataset (#: Sample size)	Pre-processing	Results (%:Performance of the model used)	Limitation
[30] (2020)	Musculoskeletal	Unsupervised anomaly detection in X-ray images. **Feature Extraction & Classification:** CAE, VAE, DCGAN, BiGAN, α-GAN.	MURA dataset containing only X-ray images of hands: (5,543).	Introduced preprocessing pipeline 1. Cropping. 2. Localization (single shot multibox detector - SSD) 3. Hand Segmentation using Photoshop's. 4. Augmentation. 5. Padding & Centring. 6. Min-Max Normalization	1. CAE: 57 % 2. VAE: 48.3% 3. DCGAN: 53% 4. BiGAN: 54.9% 5. α-GAN: 60.7%	1. The segmentation manually. 2. Used small datasets.
[31] (2020)		A Group Normalized Convolutional Neural Networks with Regularization (GnCNNr) model. **Feature Extraction & Classification:** New CNN model- (GnCNNr).	MURA dataset containing images of hand, wrist, humerus, shoulder, elbow, finger and forearm: (40,561).	1. Images used are of fixed size. 2. Increased channels. 3. Normalization. 4. Data Augmentation.	1. Hand: 83.5% 2. Wrist: 93.2% 3. Humerus: 92.4% 4. Shoulder: 85.6% 5. Elbow: 90.6% 6. Finger: 88.8% 7. Forearm: 92.6%	There is no comparison with different works, just comparing with conventional deep learning methods.
[32] (2019)		This research proposed a novel Computer-Aided Diagnosis (CADx) model based on Deep Convolutional Neural Network (Deep CNN). **Feature Extraction & Classification:** VGG-19 and ResNet.	MURA dataset containing mages of elbow, finger, humerus, and wrist: (22,938).	1. Image normalization. 2. Gaussian blur. 3. Histogram equalization. 4. Adaptive thresholding.	1. Elbow: 86.45% 2. Finger: 82.13% 3. Humerus: 87.15% 4. Wrist: 87.86%	There is no comparison with different works.

VIII. Future work

As future work, we would establish an anomaly detecting mechanism utilizing deep learning techniques for detecting breast cancer.

References

[1] N. Sarafijanovic-Djukic and J. Davis, "Fast distance-based anomaly detection in images using an inception-like autoencoder," in *International Conference on Discovery Science*, pp. 493–508, Springer, 2019.

[2] X. Xie, C. Wang, S. Chen, G. Shi, and Z. Zhao, "Real-time illegal parking detection system based on deep learning," in *Proceedings of the 2017 International Conference on Deep Learning Technologies*, pp. 23–27, 2017.

[3] W. Shi, G. Yan, Y. Li, H. Li, T. Liu, C. Sun, G. Wang, Y. Zhang, Y. Zou, and D. Wu, "Fetal brain age estimation and anomaly detection using attention-based deep ensembles with uncertainty," *NeuroImage*, vol. 223, p. 117316, 2020.

[4] G. Pang, C. Shen, L. Cao, and A. v. d. Hengel, "Deep learning for anomaly detection: A review," *arXiv preprint arXiv:2007.02500*, 2020.

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

[6] P. Seeböck, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunović, S. Klimscha, G. Langs, and U. Schmidt-Erfurth, "Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 87–98, 2019.

[7] H. Choi, S. Ha, H. Kang, H. Lee, D. S. Lee, A. D. N. Initiative, *et al.*, "Deep learning only by normal brain pet identify unheralded brain anomalies," *EBioMedicine*, vol. 43, pp. 447–453, 2019.

[8] S. Xu, H. Wu, and R. Bie, "Cxnet-m1: Anomaly detection on chest x-rays with image-based deep learning," *IEEE Access*, vol. 7, pp. 4466–4477, 2018.

[9] Q. Wei, Y. Ren, R. Hou, B. Shi, J. Y. Lo, and L. Carin, "Anomaly detection for medical images based on a one-class classification," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575, p. 105751M, International Society for Optics and Photonics, 2018.

[10] A. Ng, "Nuts and bolts of building ai applications using deep learning," *NIPS Keynote Talk*, 2016.

[11] M. Abbass, K.-C. Kwon, N. Kim, S. A. Abdelwahab, N. Haggag, F. Ibrahim, Y. Mahrous, A. Seddik, A. Khalil, Z. Elsherbeeny, *et al.*, "Anomaly detection from medical signals and images using advanced convolutional neural network," *Research Square*, 2020.

[12] Y. Bengio, *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[13] D. A. Alboaneen, D. Alsaffar, A. Alateeq, A. Alqahtani, A. Alfahhad, B. Alqahtani, R. Alamri, and L. Alamri, "Internet of things based smart mirrors: A literature review," in *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*, pp. 1–6, IEEE, 2020.

[14] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft, "Deep semi-supervised anomaly detection," *arXiv preprint arXiv:1906.02694*, 2019.

[15] L. Beggel, M. Pfeiffer, and B. Bischl, "Robust anomaly detection in images using adversarial autoencoders," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 206–222, Springer, 2019.

[16] G. Pang, C. Shen, and A. van den Hengel, "Deep anomaly detection with deviation networks," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 353–362, 2019.

[17] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International Conference on Learning Representations*, 2018.

[18] L. Gao and S. Wu, "Response score of deep learning for out-of-distribution sample detection of medical images," *Journal of Biomedical Informatics*, vol. 107, p. 103442, 2020.

[19] M. Ahmadi, M. Sabokrou, M. Fathy, R. Berangi, and E. Adeli, "Generative adversarial irregularity detection in mammography images," in *International Workshop on PRedictive Intelligence In MEdicine*, pp. 94–104, Springer, 2019.

[20] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, *et al.*, "Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection," *IEEE transactions on medical imaging*, 2020.

[21] Y. Mao, F.-F. Xue, R. Wang, J. Zhang, W.-S. Zheng, and H. Liu, "Abnormality detection in chest x-ray images using uncertainty prediction autoencoders," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 529–538, Springer, 2020.

[22] Y.-X. Tang, Y.-B. Tang, M. Han, J. Xiao, and R. M. Summers, "Abnormal chest x-ray identification with generative adversarial one-class classifier," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1358–1361, IEEE, 2019.

[23] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. A. Milacski, S. Koshino, E. Sala, H. Nakayama, and S. Satoh, "Madgan: unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction," *arXiv preprint arXiv:2007.13559*, 2020.

[24] S. Benson and R. Beets-Tan, "Gan-based anomaly detection in multi-modal mri images," *bioRxiv*, 2020.

[25] N. Wang, C. Chen, Y. Xie, and L. Ma, "Brain tumor anomaly detection via latent regularized adversarial network," *arXiv preprint arXiv:2007.04734*, 2020.

[26] W. Luo, Z. Gu, J. Liu, and S. Gao, "Encoding structure-texture relation with p-net for anomaly detection in retinal images," *In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science*, vol. 12365, 2020.

[27] K. Ouardini, H. Yang, B. Unnikrishnan, M. Romain, C. Garcin, H. Zenati, J. P. Campbell, M. F. Chiang, J. Kalpathy-Cramer, V. Chandrasekhar, *et al.*, "Towards practical unsupervised anomaly detection on retinal images," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pp. 225–234, Springer, 2019.

[28] R. Togo, H. Watanabe, T. Ogawa, and M. Haseyama, “Deep convolutional neural network-based anomaly detection for organ classification in gastric x-ray examination,” *Computers in Biology and Medicine*, vol. 123, p. 103903, 2020.

[29] D.-H. Shin, R. C. Park, and K. Chung, “Decision boundary-based anomaly detection model using improved anogan from ecg data,” *IEEE Access*, vol. 8, pp. 108664–108674, 2020.

[30] D. Davletshina, V. Melnychuk, V. Tran, H. Singla, M. Berrendorf, E. Faerman, M. Fromm, and M. Schubert, “Unsupervised anomaly detection for x-ray images,” *arXiv preprint arXiv:2001.10883*, 2020.

[31] M. Goyal, R. Malik, D. Kumar, S. Rathore, and R. Arora, “Musculoskeletal abnormality detection in medical imaging using gncnnr (group normalized convolutional neural networks with regularization),” *SN Computer Science*, vol. 1, no. 6, pp. 1–12, 2020.

[32] T. C. Mondol, H. Iqbal, and M. Hashem, “Deep cnn-based ensemble cadx model for musculoskeletal abnormality detection from radiographs,” in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, pp. 392–397, IEEE, 2019.

[33] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.