RCALAD: Regularized Complete Adversarially Learned Anomaly Detection

Zahra Dehghanian[1], **Saeed Saravani**[1], **Maryam Amirmazlaghani**[1], **Mohammad Rahmati**[1]

[1]Department of Computer Engineering Amirkabir University of Technology Tehran, Iran
{z.dehghanian, s.saravani, mazlaghani, rahmati}@aut.ac.ir

Abstract

The present study provides a comprehensive adversarial framework for anomaly detection in real-world problems based on generative adversarial neural networks' capability to model complex high-dimensional data. The proposed regularized complete adversarially learned anomaly detection (RCALAD) model, attempts to use the information of a complete cycle in order to improve the training process by adding a new joint discriminator to the structure. Moreover, to increase the distance between the anomalous sample and its reconstruction, a further supplementary distribution in the input space is employed to bias all the reconstructions toward the normal data distribution. In addition to the proposed model, two new anomaly scores are presented specialized for variation of datasets in this study. Experimental results demonstrate the superiority of the RCALAD model to other state-of-the-art models in the field of anomaly detection.

Keywords: Anomaly detection, Machine learning, Generative adversarial networks, Reconstruction error, Anomaly scores.

Introduction

Discovering dissimilar instances and rare patterns is one of the most essential tasks in real-world data. Such samples are referred to as anomalies, and identifying those instances is called anomaly detection (Yao et al. 2017). Anomalies play a crucial role in a variety of applications and are a vital part of any dataset. i.e. an unusual traffic pattern on a computer network can mean hacking the computer and transferring data to unauthorized destinations. Abnormal behaviors in credit card transactions may indicate fraudulent economic activities (Jiang et al. 2018), or abnormality in an MRI image may indicate the presence of a malignant tumor (Dai and Bikdash 2016). Despite the existence of statistical and machine learning-based methods, designing effective models for detecting anomalies in complex high-dimensional data space is still a major challenge (Zenati et al. 2018b).

Generative adversarial networks (GAN) are able to overcome this challenge and model the distribution of real-world data, which have high complexity and dimension; this results in a promising performance in the field of anomaly detection. In GAN, a generator network is contrasted with a discriminator network; the discriminator attempts to differentiate between the real data and data produced by the generator network. Generator and discriminator are trained simultaneously; the generator network G records the distribution of the data, and the discriminator D estimates the likelihood whether samples come from real data distribution or are generator made. The objective function of the generator G is maximizing the error probability of the network D. This structure leads to a two-player g mini-max games (Goodfellow et al. 2014).

The capability of adversarial neural networks to model natural images has been proven before (Radford, Metz, and Chintala 2015; Creswell et al. 2018). Recent years their usage increase significantly in processing speech and text (Zenati et al. 2018b), as well as medical images (Schlegl et al. 2017). In the present paper, an efficient and effective method based on adversarial generator networks is proposed, which is particularly designed for anomaly detection task. This method like many learning algorithms has two main steps; training and testing. In the training phase, the generator, encoder and discriminators are trained altogether and updated in turns to cover normal manifold . In test phase, normal and abnormal data pass into the model to be differentiated.

Related Work

Anomaly detection, also known as novelty detection and outlier detection, has been widely studied as reviewed in (Kaur and Singh 2016; Zimek, Schubert, and Kriegel 2012; Pimentel et al. 2014). The previous methods used in this field are generally divided into two main categories: methods based on representation learning and methods based on generative model.

The representation learning method learns a mapping for the main characteristics of normal data. One-class support vector machine finds the marginal boundary around the normal data (Schölkopf et al. 1999). The isolation forest method is one of the classic machine learning methods. In this method, the tree is built with randomly chosen features and the anomaly score is the average distance to the root (Ruff et al. 2018). Deep support vector data description (DSVDD) finds a hypersphere to enclose the representation of normal samples (Liang, Li, and Srikant 2017). Liu

and Gryllias constructed frequency domain features using cyclic spectral analysis and applied them in SVDD framework. This method has been proven robust against outliers and can achieve a high detection rate for bearing anomaly detection (Liu and Gryllias 2020). In (Golan and El-Yaniv 2018), researchers presented a new approach to identify imagery anomalies by training the model on normal images altered by geometric transformation. In this model, the classifier calculates the anomaly score using softmax statistics.

Generative models attempt to learn the reconstruction of the data and use this reconstruction to identify anomalous samples (Yang, Bozchalooi, and Darve 2020). For instance, auto-encoders model the normal data distribution, and the reconstruction error is used as the anomaly score (Nguyen et al. 2019; Pidhorskyi, Almohsen, and Doretto 2018). Deep structured energy-based models (DSEBM) learn an energy-based model and map each sample to an energy score (Zhai et al. 2016). Deep autoencoding Gaussian Mixture Model (DAGMM) estimates a mixed Gaussian Distribution by using an encoder for normal samples (Schlegl et al. 2019).

A recent line of anomaly detection work has focused on adversarial neural networks. This structure has been used to identify anomalies in optical coherence tomography images of the retina (Schlegl et al. 2017). In this work, the inverse mapping into the latent space was performed using the recursive backpropagation mechanism. In (Zenati et al. 2018a), which is a continuation of the previous work, the mapping to the latent space was performed by the encoder to reduce the computational complexity. In (Dumoulin et al. 2016), the pro-posed model was based on bidirectional GAN (BiGAN). In the article (Zenati et al. 2018b), anomaly detection was performed by adding the new discriminator in latent space to the adversarial structure to stabilize the training process. The deep convolutional autoencoder model is a classic autoencoder model, in which the encoder and decoder have a convolutional structure. The anomaly score in this model is L2-norm of reconstruction errors (Li et al. 2017).

In this section we reviewed and categorized the various methods which used to clarify abnormal data. In section 3 we will deep into fundamental basis required to deal with proposed model.

Preliminaries

In this section, we briefly elaborate the overall idea of generative adversarial network and then concentrate on the evolution of GAN-based anomaly detection algorithms. The GAN was first proposed in 2014 by Goodfellow et al (Goodfellow et al. 2014). In this network, the generator G is placed against the discriminator D. These networks are trained on M set of $\left\{\chi^{(i)}\right\}_{i=1}^{M}$ unlabeled samples. G maps selected samples from the latent space z to the input data space and D attempts to distinguish between the real data $x^{(i)}$ and the produced data. These two networks compete with each other; The generator G tries to imitate the distribution of the input data, while the discriminator tries to distinguish between the real samples and the data produced by the generator. In the training phase, the generator G and the discriminator D are alternatively optimized using stochastic gradient descent approach.

The distribution of the input data is represented as $q(x)$, and $p(z)$ is considered as the generator distribution in the latent space. GAN training is done by finding a discriminator and a generator that can solve the saddle point problem as $\min_G \max_D V_{GAN}(D,G)$ and the $V_{GAN}(D,G)$ function is defined as:

$$V_{GAN} = E_{x\sim q(x)}[\log(D(x))] + E_{z\sim p(z)}[\log(1-D(G(z)))] \tag{1}$$

Solving this problem concludes that the generator distribution is equal to the true data distribution. It has been proved in (Goodfellow et al. 2014) that the global optimal discriminator will be obtained if and only if $P_G(x) = q(x)$. By P_G, we mean the distribution learned by the generator.

Adversarially Learned Inference (ALI) (Liu, Ting, and Zhou 2008) attempts to obtain the inverse mapping between input data space and latent space by modelling the joint distribution of encoder as $q(x,z) = q(x)e(z|x)$ and the distribution of generator as $p(x,z) = p(z)p(x|z)$ using the encoder E. Here, $e(z|x)$ is learned by the encoder. The objective function of the ALI model is as follows:

$$\min_{G,E}\max_{D} V_{ALI} = \\ E_{q(x,z)}[\log D(x,E(x))] + E_{p(x,z)}[\log(1-D(G(z),z)] \tag{2}$$

where D represents the discriminator, taking x and z as input, and its output value specifies probability of origination of the current inputs from the $q(x,z)$ distribution. Encoder, generator, and discriminator are in their optimal state only if $q(x,z) = p(x,z)$. This has been proved in (Liu, Ting, and Zhou 2008).

Although p and q distributions are apparent, in practice and during model training, they are not necessarily converging to the optimal point. This issue was attributed to the problem of cycle consistency which was defined as $\mathrm{G(E(x))} \approx \hat{\mathrm{x}}$ in(Zong et al. 2018). A new framework called ALICE was proposed to solve the above problem by adding the discriminator $\mathrm{D_{xx}}$ to the ALI network structure (Zong et al. 2018). The objective function of this model is as follows:

$$\min_{E,G}\max_{D_{xz},D_{xx}} V_{ALICE} = V_{ALI} \\ +E_{x\sim q(x)}\left[\log D_{xx}(x,x) + \log(1-D_{xx}(x,G(E(x))))\right] \tag{3}$$

This work demonstrates that using a discriminator $\mathrm{D_{xx}}$ can achieve the best reconstruction for the input data (Zong et al. 2018). In Adversarially Learned Anomaly Detection (ALAD) a conditional distribution was applied to the baseline ALICE model with the addition of another discriminator to stabilize the training process (Zenati et al. 2018b). To deep into the detail, a discriminator $\mathrm{D_{zz}}$ is added to the model to ensure the cycle consistency in the latent space, which tries to make the latent space variable and its reconstruction as analogous as possible. By assembling the block proposed in

(Zenati et al. 2018b) in ALICE framework, the cost function of the ALAD model will finally be as follows:

$$\begin{aligned} &\min_{G,E} \max_{D_{xz},D_{xx},D_{zz},} V_{ALAD} \\ &= V_{ALICE} + E_{z\sim p(z)} \left[\log\left(D_{zz}(z,z)\right) + \right. \\ &\left. \log\left(1 - D_{zz}(z, E(G(z)))\right)\right] \end{aligned} \tag{4}$$

In (Zenati et al. 2018b), it is declared that the training model will be stabilized by adding Lipschitz constraints to the discriminators of the GAN model. Moreover, it is shown in practice that with spectral normalization of the weight parameters, the network's performance will be improved (Zenati et al. 2018b).

Although the idea of ALAD helps stabilizing the cycle, still the latent and input space variables are scrutinized in two independent spaces, and the inherent dependence between the variables is ignored. More precisely, the x and its reconstruction are investigated in a separate process from the z and its corresponding reconstruction, while the reconstruction process of these two pairs of data is along with each other and affects each other directly. To model this dependence, we define a complete cycle and a new discriminator that utilizes this complete cycle information.

Moreover, another problem of ALAD is the careless assumption of the necessity of weak reconstruction for anomalous samples. In fact, in all the previous models, this assumption has been implied that if the model is trained with normal data, it will necessarily have a poor reconstruction for the anomalous data, while there is no constraint to bias the model towards generating poor reconstructions of anomalous samples. In the proposed RCALAD model, an attempt has been made to address this weakness. Also, this assumption should not be applied negligently. Here, by using the supplementary distribution $\sigma(\mathrm{x})$, the model bias all the reconstructions toward the normal data distribution and, with this procedure, tries to reduce the anomaly score for normal input while increase it for anomalous data.

Proposed Model

In this section, we propose a novel adversarial framework for anomaly detection that can overcome the aforementioned drawbacks of the existing methods. First, the problem of lack of complete cycle consistency and our solution to overcome this problem will be described. then, we will examine the issue of the necessity of weak reconstruction of anomalous sample, and the final proposed model designed to solve both problems will be introduced. At the end of this section, two new anomaly scores are presented based on the proposed model.

Complete Cycle Consistency

As mentioned before, in the ALAD model, the cycle consistency of the input data and latent space variable is examined in two independent procedures. This implies that the reconstruction of the latent space variable $\hat{z}$ and input data $\hat{x}$ are processed separately. It should be noted that here the variable z is a sample of the Gaussian distribution given as an input to the generator and it is not related to the input mapping in the latent space.

The Complete Cycle Consistency issue (CCC) declares that for each variable x in the input space if the encoder first estimates the inverse mapping to the latent space, which equals $E(x) = Z_x$ and the obtained representation is entered into the generator to generate the network reconstruction from the input variable $G\left(Z_x\right) = G(E(x)) = \hat{x}$ and the this reconstruction is given to the encoder network in order to calculate the reconstruction in the latent space, that is, $E(\hat{x}) = E\left(G\left(z_x\right)\right) = \hat{z}_{\hat{x}}$, it is logically expected from any reconstruction-based network that the two variables x and $\hat{x}$ as well as the two variables $\hat{z}_{\hat{x}}$ and z_x have the least possible difference. That is, the CCC issue is defined in such a way that, in any reconstruction-based model, for each input data and its mapping in the latent space, the network reconstruction for both variables should have a minimum error and maximum similarity with them.

In the ALAD model, the similarity between the input data and its reconstruction, as well as the similarity between z and its reconstruction, were examined independently and in two separate cycles. It was assumed that they are independent, while we know these two cycles are entirely dependent on each other, and the assumption of independence is not valid in these two issues. To solve this problem, we proposed to model the dependency by examining the CCC variables in the new discriminator D_{xxzz} and using the information flow in this chain to improve network training for anomaly detection in the best possible way. The difference between the input of D_{xxzz} and the input of D_{zz} used in the ALAD model is represented in Figure 1.

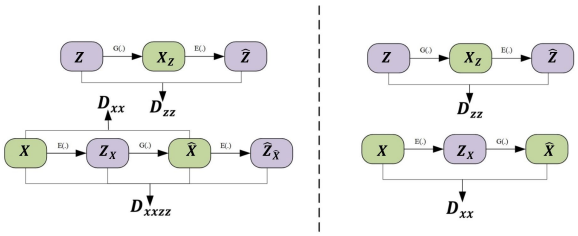

Figure 1: Using the variables of input data space and latent space in the cycle consistency of the ALAD network (right side) and the information of a complete cycle consistency in the proposed model (left side).

As can be seen in Figure 1, the ALAD model does not use the information of a complete cycle. In order to use the available information in a complete cycle, a new variable called $\hat{z}_{\hat{x}}$ is introduced. To calculate this variable, the inverse mapping of the input data x applied to the generator and the resulting inverse mapping is calculated again using the decoder. Hence, the complete cycle consistency will be provided in this model.

In order to ensure the condition of complete cycle consistency, the new D_{xxzz} discriminator is used with the joint input. It is noteworthy that the effectiveness of the joint discriminators has already been proven once in ALIGAN (Dumoulin et al. 2016). Actually, when adding the encoder to the GAN framework, two procedures can be scrutinized. The first one is adding an independent discriminator to train the

encoder, and the second one is changing the discriminator input from a single input mode to a joint input mode. It is proved that using joint discriminator obtains better results. According to the same idea, the input of the joint discriminator D_{xxzz} extracts the most information for model training.

This discriminator uses the quadruple of (x, x, z_x, z_x) as the real data and the quadruple of $(x, G(E(x)), z_x, E(G(z_x)))$ as fake data. This discriminator attempts to make the input x and network reconstruction as well as the inverse mapping of the input image in the latent space and its reconstruction by the encoder as close as possible to each other, so that a complete stable loop is provided and the model is trained and stabilized better.

Necessity of Weak Reconstruction

In reconstruction-based models, it has always been assumed that if the training and reconstruction of the normal data are properly done, the reconstruction of abnormal data will necessarily be weak and different from the input data. However, experiments indicate that it is not always the case, and sometimes the reconstructed anomalous sample is slightly similar to the input sample. Hence, it won't be easy to recognize it as an abnormal sample. In fact, in none of the previous models, there is no requirement or control condition to bias the model towards producing poor reconstruction for anomalous samples.

Essentially, this phenomenon occurs due to a weak mapping between the input and the latent space. Cause in the training phase, the encoder only learns the mapping of the normal samples to the latent space and, as a result, the corresponding space of z for the normal samples is well modeled, but in the test phase, given the fact that the model has not yet seen the rest of the space, including abnormal examples, it may map it onto an unknown point of the latent space. That is, in this case, there is no information regarding the mapping of anomalous data. One solution for this issue is mapping all the input space to the latent normal subspace. It means that to cover the data space as much as possible, the supplementary distribution called $\sigma(\mathrm{x})$ is used. samples form this distribution will cover the input data space it means that we produce some extra noisy samples and force the model to generate reconstruction in normal data manifold.As a result, network learns to reconstruct the normal data class for a relatively more expansive range of inputs. For anomalous sample the inverse mapping is close to the normal data latent space so its reconstruction will be similar to normal data. So a suitable distance is created between the anomalous sample and its reconstruction. This distance is regarded as an appropriate criterion for detecting abnormal samples. Figure 2 illustrates the training routine.

RCALAD Model

In this section, by combining both ideas presented in the previous sections, i.e., using the new variable of $\hat{z}_{\hat{x}}$ in the D_{xxzz} discriminator, as well as using the $\sigma(\mathrm{x})$ distribution and adding them to the basic model (Zenati et al. 2018b), our proposed model, RCALAD, will be introduced. In this

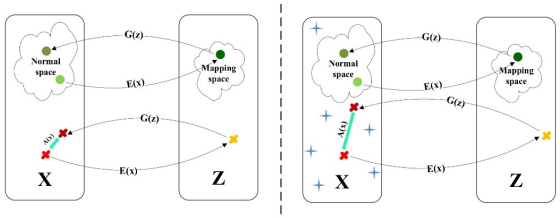

Figure 2: $\sigma(\mathrm{x})$ distribution effect in the model training process. In this figure, x shows the input data space and z represents the latent space. Also Green circles, red crosses and blue stars are assumed as normal samples, anomalous samples and $\sigma(\mathrm{x})$ distribution made samples. The cyan line indicates the value of the anomaly score. As can be seen in Figure 2, if $\sigma(\mathrm{x})$ is not present in the training process (on the left side of the figure), the anomaly score is lower than when this distribution is used (on the right side). It means that the distribution $\sigma(\mathrm{x})$ has biased the model towards the normal manifold.

network, issues of complete consistency cycle and the necessity of weak reconstruction are addressed simultaneously and it has been attempted to provide a comprehensive, practical and compatible framework for all anomaly detection problems. The outline of the proposed model can be seen in Figure 3.

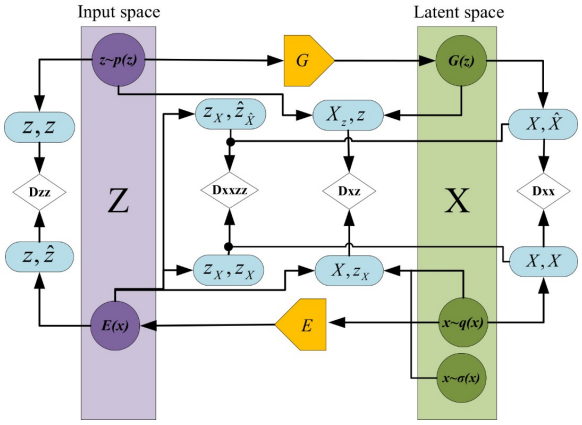

Figure 3: Overall structure of the RCALAD model.

As in Figure 3, an encoder and the generator are trained in the general structure of the adversarial neural network. The inverse mapping from the input data space to the latent space is obtained simply by using the encoder E in the proposed structure. Here, a joint discriminator called D_{xxzz} is used to train both generator and encoder networks simultaneously. This discriminator determines whether an input variable pair is derived from the input data x distribution and its corresponding point in the latent space or if it is generated by the generator G and sampled from the latent space of z. In order to satisfy the condition of cycle consistency in the input data space, D_{xx} and D_{zz} discriminators are used so cycle consistency in both latent space and input data space will be modeled independently. The D_{xxzz} is introduced to use all

the information in a complete cycle. That is, in addition to examining both variables x and z and their reconstruction in the corresponding space, their joint distribution is used in D_{xxzz} during the process of detecting anomalous samples so more information is available to determine whether the input data is anomalous or not. This network is responsible for determining between quadruple samples of (x, x, Z_x, Z_x) and $(x, G(E(x)), z_x, E(G(z_x)))$ and tries to extract x and the reconstruction provided by the network as well as to make the mapping of the input image in the latent space of z_x and the reconstruction of the output of the generating network by the encoder of $E(G(z_x))$ as close as possible. The $\sigma(\mathrm{x})$ block is added to this model to cover the maximum latent space. By using this distribution, new samples in the input data space are generated and then mapped into the latent space of normal data. Finally, the objective function of the proposed model is as follows:

$$\begin{gathered}\min_{G,E}\max_{D_{xxzz},D_{xz},D_{xx},D_{zz}} V_{RCALAD} \\ (D_{xxzz}, D_{xz}, D_{xx}, D_{zz}, E, G) = \\ V_{ALAD} + E_{x\sim\sigma(x)}\left[\log\left(1 - D_{xz}(x, E(x))\right)\right] \\ + E_{x\sim\mathrm{q(x)}}\left[\log D_{xxzz}(x, x, E(x), E(x))\right] + \\ E_{\boldsymbol{x}\sim q(x)}\left[1 - \log D_{xxzz}(x, G(E(x)), E(x), E(G(E(x))))\right]\end{gathered} \tag{5}$$

Anomaly Scores

In this study the major aim of presenting the proposed model is to detect anomalies based on input data reconstruction. Our study presents a model that is used to detect anomalies in input data by reconstructing normal samples accurately, while anomalous samples are reconstructed in a weak manner. One of the key elements in anomaly detection is the definition of the anomaly score for calculating the distance between the input sample and the reconstruction provided by the network (Zenati et al. 2018b). Some of anomaly scores that were used in previous models is as follows:

$$\begin{gathered}A_{L_1}(x) = \|x - \hat{x}\|_1 \\ A_{L_2}(x) = \|x - \hat{x}\|_2 \\ A_{\text{Logits}}(x) = log\left(D_{xx}(x, \hat{x})\right) \\ A_{\text{Features}}(x) = \|f_{xx}(x, x) - f_{xx}(x, \hat{x})\|_1\end{gathered} \tag{6}$$

Here, logit means the raw output of the discriminators, while feature means the output of the layer preceding logit.In our proposed model since the D_{xxzz} adds the ability of extracting new information to the model which has not been introduced in any of the previous scores, there is a strong need to define the new anomaly scores to use this ability. In this paper, two new anomaly scores are introduced based on the information mentioned in the D_{xxzz}.

The first anomaly score presented in this paper is called $A_{fm}(x)$. This score uses the D_{xxzz} discriminator feature space to calculate distance between samples and their reconstruction. For this purpose, the output of second to last layer is used as features. Our anomaly score is defined as follows:

$$A_{fm}(x) = \|f_{xxzz}(x, x, z_x, z_x) - f_{xxzz}(x, \hat{x}, z_x, \hat{z}_{\hat{x}})\|_1 \tag{7}$$

In this equation, $f(.)$ represents the activation function of the second to the last layer in the D_{xxzz} discriminator structure. The concept used in the definition of this score is using the confidence level of discriminator on the quality of the reconstructions provided by the network. In other words, if reconstruction is performed well, the sample belongs to the trained normal data of the network. Thus, the higher value of this criterion means the greater difference in reconstructions and so higher possibility of input data's abnormality. The second point in this article is presented with the aim of maximizing the use of information in the model for anomaly detection. In this section, the A_{all} criterion is defined. The score is the sum of the outputs of all discriminators, including D_{xx}, D_{zz} and D_{xxzz}.

In fact, since all the discriminators in the proposed model are trained only based on the normal samples and the reconstruction for all the input data space is biased towards the normal data space, it is expected that input image data and its reconstruction looks different, and the discriminators can easily identify these anomalous inputs. The mathematical expression of this criterion is given in the following equation:

$$A_{all}(x) = D_{xxzz}(x, \hat{x}, z_x, \hat{z}_{\hat{x}}) + D_{xx}(x, \hat{x}) + D_{zz}(z_x, \hat{z}_{\hat{x}}) \tag{8}$$

The criterion A_{all} tries to utilize all discriminators' information. During the training phase the discriminators learn to pay attention to the difference between the pairs of (x, x) and $(x, \hat{x})$ as well as the pairs of (z_x, z_x) and $(z_x, \hat{z}_{\hat{x}})$. It means, the farther $\hat{x}$ from x or $\hat{z}_{\hat{x}}$ from z_x, it will be easier for the discriminators to recognize the data origin. In the proposed model, by adding the distribution of $\sigma(\mathrm{x})$ and biasing all the reconstruction towards the normal data distribution, the reconstruction error for abnormal data is increased and the discriminators' output can be considered as a reliable criterion for abnormality detection. Finally, the recommended anomaly scores can be viewed according to the Algorithm 1.

Experiments

This section compares the proposed anomaly detection model with prominent anomaly detection models. To test the models on a fair basis, the reported outcomes for all the implemented models are based on tabular data obtained from the average of ten runs, and, for each class of imagery data, it is based on the average of three runs. Also, all hyperparameters are set as described in (Zenati et al. 2018b). The anomaly score used in tabular data is A_{all} score and, for imagery data, it is A_{fm} score. The reason for choosing these scores based on data type will be discussed later. Moreover, the ALAD model is implemented and the results of the best anomaly score A_{fm} are reported. For other models, the available results are adopted from (Makhzani and Frey 2015).

Datasets

In order to evaluate the performance of the proposed model and scrutinize its efficiency from different viewpoints, various data sets with diverse characteristics are used. The pro-

Algorithm 1: Process of calculating anomaly scores in Regularized Complete Adverarially Learned Anomaly Detection

Input: $x \sim p_{x_{\text{Test}}}(x)$, E, G, D_{xx}, D_{zz}, D_{xxzz}, f_{xxzz} where f_{xxzz} is the feature layer of D_{xxzz}

Output: $A_{all}(x)$, $A_{fm}(x)$, where A is the anomaly score

1: procedure INFERENCE
2: $z_x \leftarrow E(x)$ Encode samples, Construct latent Embedding
3: $\widehat{x} \leftarrow G(z_x)$ Reconstruct samples
4: $\hat{z}_{\hat{x}} \leftarrow E(\widehat{x})$ Reconstruct latent Embedding
5: $A_{fm}(x) \leftarrow \| f_{xxzz}(x, x, z_x, z_x) - f_{xx}(x, \widehat{x}, z_x, \hat{z}_{\hat{x}}) \|_1$
6: $A_{all}(x) \leftarrow D_{xxzz}(x, \widehat{x}, Z_x, \hat{z}_{\hat{x}}) + D_{xx}(x, \hat{x}) + D_{zz}(z_x, \hat{z}_{\hat{x}})$
7: **return** $A_{all}(x), A_{fm}(x)$
8: **end** procedure

posed method is tested on the available imagery and tabular datasets. For tabular datasets, four datasets, including kddcup99(Dua and Graff 2017), arrhythmia(Dua and Graff 2017), thyroid(Dua and Graff 2017) and musk(Dua and Graff 2017) are used. Kddcup99 is a dataset related to network penetration. Arhythmia is a medical collection related to cardiac arrhythmia with 16 classes. Also, thyroid is a three-class data related to thyroid disease. The musk dataset was created to classify six classes on molecular musk. These four datasets, 20, 15, 2.5 and 3.2 percent of the data are anomalous samples, respectively. Hence, in the test phase, after calculating the anomaly score, the aforementioned proportion of the data that has the highest anomaly score is classified as anomaly. In order to assess the proposed model on these datasets F1, Recall and Precision criterion is used.

Two datasets of CIFAR10(Krizhevsky, Hinton et al. 2009) and SVHN(Netzer et al. 2011) are considered for the imagery dataset. Both of these datasets have ten classes and, each time, one class is considered as the normal class and the other nine classes as the abnormal class like previous works. The criterion used to evaluate the model on imagery data is area under the receiver operating curve (AUROC). For all the data which have been used, 80% of the data are selected as training data and 20% as test data. 25% of training data is selected as validation data. It is noteworthy that, in the training phase, all the anomalous samples are eliminated from the training data.

Experiments on the Tabular Datasets

Evaluation results of the proposed RCALAD model and other state-of-the-art models on tabular data of kddcup99, arrhythmia, thyroid and musk are summarized in Table 1. The structures used in the generator, discriminator and encoder networks are all fully connected layers with nonlinear activation functions. It should be noted that, in this step, $N(0, I)$ distribution is used as $\sigma(\mathrm{x})$.

In order to make a clear comparison between different models, the error bar is used in the last row of Table 1. As can be seen in this table, the proposed model has a successful performance on the arrhythmia and musk datasets, compared to other models. Also on KDD dataset our model is the best according to F1 criteria, but on thyroid dataset secures the second place due to the extraordinary performance of the IF model. The reason for this phenomenon can be attributed to the nature of the data in this dataset since there are various features in this data set, only a few of them are in-formative; therefore, the results of classic models such as IF, which are based on feature selection, are better. An idea to improve the proposed model results on thyroid dataset is to use models such as IF in the preprocessing stage to select more effective characteristics for training the model.

Experiments on the imagery Datasets

In this section, the performance of the proposed model on CIFAR10 and SVHN imagery data is scrutinized in two separate tables.

As in Tables 2 and 3, the proposed model has significantly improved the results on CIFAR10 dataset. The results obtained on the majority of classes have the best performance and, on other classes, the outcomes are comparable with other models. in addition to being superior in seven classed compared to other models, the proposed model also performs the best in the average of all classed as well.

Ablation Studies

In this section we examine effectiveness of each component added to the basic model on both types of datasets. In these experiments, the average results of the model are repeated in the presence and absence of the discriminator D_{xxzz} and the supplementary distribution $\sigma(\mathrm{x})$.

According to Tables 4 and 5, the proposed RCALAD model achieves the highest efficiency in the presence of both parts. In scrutinizing the role of the D_{xxzz} discriminator, this discriminator has improved the accuracy on the CIFAR10 dataset to an optimal level but has not made significant improvement on the SVHN dataset. Concerning the role of $\sigma(\mathrm{x})$ distribution, this distribution has performed well on the CIFAR10 dataset and improved the AUROC criterion. However, the SVHN dataset has reduced the AUROC criterion by a small amount compared to the base model. Still, its presence in the final model has led to extract new information and a more comprehensive view.

Evaluating the Sufficiency of the $\mathbf{D_{xxzz}}$

By adding the D_{xxzz} discriminator, is there a need for D_{xx} and D_{zz} discriminators or not? To answer this question correctly, we perform some experiments that their results are summarized in Table 6. In fact, in this section, in addition to the above-mentioned question, the result of adding D_{xxzz} discriminator in basic models such as ALI and ALICE are investigated.

In Table 6 we examine effect of all discriminators on different models. According to this table and as expected from the theoretical results, adding the D_{xxzz} discriminator to the general frameworks had the highest efficiency. Subsequently, eliminating D_{xx} damages the model less since some of the information it extracts is covered with the D_{xxzz} discriminator. But considering that D_{zz} examines the similarity

Model	KDDCUP			Arhythmia			Thyroid			Musk		
	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1
IF	92.16	93.73	92.94	51.47	54.69	53.03	**70.13**	**71.43**	**70.27**	47.96	47.72	47.51
OC-SVM	74.57	85.23	79.54	53.97	40.82	45.18	36.39	42.39	38.87	–	–	–
DSEBM.	85.12	64.72	73.28	15.15	15.13	15.10	4.04	4.03	4.03	–	–	–
DSEBMe	86.19	64.46	73.99	46.67	45.65	46.01	13.19	13.19	13.19	–	–	–
AngGAD	87.86	82.97	88.65	41.18	43.75	42.42	44.12	46.87	45.45	3.06	3.10	3.10
DAGMM	92.97	94.22	93.69	49.09	50.78	49.83	47.66	48.34	47.82	–	–	–
ALAD	94.27	**95.77**	95.01	50.00	53.13	51.52	22.92	21.57	22.22	58.16	59.03	58.37
DSVDD	89.81	94.97	92.13	35.32	34.35	34.79	22.22	23.61	23.29	–	–	–
RCALAD	**95.36**	95.62	**95.49**	**58.82**	**62.50**	**60.60**	53.76	51.53	52.62	**62.96**	**63.33**	**63.14**
error bar	0.28	0.29	0.28	6.6	6.8	5.8	4.3	2.7	2.8	5.06	2.53	2.62

Table 1: Output results of the proposed model in comparison with the basic models on the tabular data set.

Normal	DCAE	DSEBM	DAGMM	IF	AnoGAN	ALAD	RCALAD
Airplane	59.1 ± 5.1	41.4 ± 2.3	56.0 ± 6.9	60.1 ± 0.7	67.1 ± 2.5	64.7 ± 2.6	$\mathbf{72.8 \pm 0.8}$
auto.	$\mathbf{57.4} \pm 2.9$	57.1 ± 2.0	56.0 ± 6.9	50.8 ± 0.6	54.7 ± 3.4	45.7 ± 0.8	50.2 ± 0.3
Bird	48.9 ± 2.4	61.9 ± 0.1	53.8 ± 4.0	49.2 ± 0.4	52.9 ± 3.0	67.0 ± 0.7	$\mathbf{72.6} \pm 0.2$
Cat	58.4 ± 1.2	50.1 ± 0.4	51.2 ± 0.8	55.1 ± 0.4	54.5 ± 1.9	59.2 ± 1.1	$\mathbf{64.2} \pm 0.9$
Deer	54.0 ± 1.3	73.2 ± 0.2	52.2 ± 7.3	49.8 ± 0.4	65.1 ± 3.2	72.7 ± 0.6	$\mathbf{74.9} \pm 0.5$
Dog	$\mathbf{62.2 \pm 1.8}$	60.5 ± 0.3	49.3 ± 3.6	58.5 ± 0.4	60.3 ± 2.6	52.8 ± 1.2	60.1 ± 1.1
Frog	51.2 ± 5.2	68.4 ± 0.3	64.9 ± 1.7	42.9 ± 0.6	58.5 ± 1.4	69.5 ± 1.1	$\mathbf{75.3} \pm 0.4$
Horse	58.6 ± 2.9	53.3 ± 0.7	55.3 ± 0.8	55.1 ± 0.7	$\mathbf{62.5 \pm 0.8}$	44.8 ± 0.4	56.6 ± 0.2
Ship	$\mathbf{76.8 \pm 1.4}$	73.9 ± 0.3	51.9 ± 2.4	74.2 ± 0.6	75.8 ± 4.1	73.4 ± 0.4	$\mathbf{77.5} \pm 0.3$
Truck	$\mathbf{67.3 \pm 3.0}$	63.6 ± 3.1	54.2 ± 5.8	58.9 ± 0.7	66.5 ± 2.8	43.2 ± 1.3	52.6 ± 0.6
Mean	59.4	60.3	54.4	55.5	61.8	59.3	**65.7**

Table 2: Output results of the proposed model compared to the basic models on the CIFAR10 dataset.

Normal	OCSVM	DSEBMr	DSEBMe	IF	ANOGAN	ALAD	RCALAD
0	52.0 ± 1.6	56.1 ± 0.2	53.4 ± 1.8	53.0 ± 0.6	57.3 ± 0.4	58.7 ± 0.9	$\mathbf{60.4} \pm 0.1$
1	48.6 ± 5.3	52.3 ± 0.9	52.1 ± 0.3	51.2 ± 0.9	57.0 ± 0.8	$\mathbf{62.8 \pm 1.7}$	59.2 ± 0.3
2	49.7 ± 7.7	51.9 ± 0.8	51.8 ± 0.4	52.3 ± 0.1	53.1 ± 0.4	$\mathbf{55.2 \pm 2.3}$	54.9 ± 0.1
3	50.9 ± 1.4	51.8 ± 0.4	51.7 ± 0.5	52.2 ± 0.3	52.6 ± 0.4	53.8 ± 3.3	$\mathbf{55.8 \pm 1.9}$
4	48.4 ± 5.2	52.5 ± 0.1	52.4 ± 0.2	49.1 ± 0.6	53.9 ± 0.5	58.0 ± 0.1	$\mathbf{58.5 \pm 0.2}$
5	51.1 ± 2.6	52.4 ± 2.3	52.3 ± 2.6	52.4 ± 0.8	52.8 ± 0.1	56.1 ± 0.9	$\mathbf{56.2 \pm 0.4}$
6	50.1 ± 3.9	52.1 ± 1.8	52.2 ± 1.8	51.8 ± 0.2	53.2 ± 0.0	57.4 ± 0.6	$\mathbf{59.4 \pm 0.5}$
7	49.6 ± 1.3	53.4 ± 0.9	55.3 ± 1.1	52.0 ± 0.4	55.0 ± 0.0	$\mathbf{58.8 \pm 0.3}$	58.0 ± 0.4
8	45.0 ± 4.2	51.9 ± 0.3	52.5 ± 0.6	52.3 ± 0.8	52.2 ± 0.7	$\mathbf{55.2} \pm 0.4$	$\mathbf{56.1 \pm 0.5}$
9	52.5 ± 3.9	55.8 ± 1.7	52.7 ± 1.4	53.7 ± 0.6	53.1 ± 0.1	57.3 ± 0.6	$\mathbf{58.3} \pm 0.2$
Mean	50.2	52.9	52.4	51.6	54.0	57.3	57.7

Table 3: Output results of the proposed model compared to the basic models on the SVHN dataset.

of z and its reconstruction in an independent cycle, it is evident that eliminating it reduces the accuracy. As can be seen, it can be concluded that using these three discriminators are regarded as the most effective, and the D_{xxzz} discriminator alone does not cover all aspects. It is noteworthy that this statement is true on both imagery and tabular datasets.

Scores Evaluation

In this section, the proposed anomaly scores are evaluated and compared with the anomaly scores presented in previous works (Zenati et al. 2018b). As in Table 7, on the tabular data, the raw output of the D_{xxzz} discriminator A_{all} has the best results compared to other anomaly scores. As is clear in Table 8, the performance of the feature-based score A_{fm} on imagery data is significant. This difference in the performance of the scores can be owing to the difference in the number of features of these two types of datasets. Given the fact that the number of features on tabular data is less than those of imagery data, the D_{xxzz} discriminator is able to extract and recognize abnormal samples properly. However, in the imagery data set, the output of the latent layer before the logit layer contains richer information to distinguish between normal and abnormal data. This improves the performance of the A_{fm} score.

Model	AUROC
CIFAR-10	
Baseline(ALAD)	0.593 ± 0.017
Baseline + D_{xxzz}(CALAD)	0.634 ± 0.018
Baseline + σ(x)(RALAD)	0.642 ± 0.012
Baseline + D_{xxzz} + σ(x)(RCALAD)	$\mathbf{0.657 \pm 0.016}$
SVHN	
Baseline(ALAD)	0.573 ± 0.016
Baseline + D_{xxzz}(CALAD)	0.576 ± 0.014
Baseline + σ(x)(RALAD)	0.568 ± 0.018
Baseline + D_{xxzz} + σ(x)(RCALAD)	$\mathbf{0.577 \pm 0.019}$

Table 4: Effects of the various proposed sections in improving the results of the imagery datasets.

Model	Precision	Recall	F1
KDD99			
Baseline(ALAD)	0.94 ± 0.008	$\mathbf{0.957 \pm 0.006}$	0.950 ± 0.007
Baseline + D_{xxzz}(CALAD)	$\mathbf{0.959 \pm 0.004}$	0.957 ± 0.007	$\mathbf{0.958 \pm 0.005}$
Baseline + σ(x)(RALAD)	0.943 ± 0.005	0.955 ± 0.004	0.949 ± 0.004
Baseline + D_{xxzz} + σ(x)(RCALAD)	0.953 ± 0.007	0.956 ± 0.005	0.954 ± 0.006
Arrhythmia			
Baseline(ALAD)	0.500 ± 0.049	0.531 ± 0.047	0.515 ± 0.048
Baseline + D_{xxzz}(CALAD)	0.574 ± 0.021	0.605 ± 0.0224	0.575 ± 0.021
Baseline + σ(x)(RALAD)	0.546 ± 0.035	0.565 ± 0.039	0.555 ± 0.037
Baseline + D_{xxzz} + σ(x)(RCALAD)	$\mathbf{0.588 \pm 0.42}$	$\mathbf{0.625 \pm 0.41}$	$\mathbf{0.606 \pm 0.41}$
Thyroid			
Baseline(ALAD)	0.229 ± 0.067	0.215 ± 0.067	0.222 ± 0.067
Baseline + D_{xxzz}(CALAD)	0.529 ± 0.071	$\mathbf{0.518 \pm 0.075}$	0.523 ± 0.073
Baseline + σ(x)(RALAD)	0.431 ± 0.039	0.457 ± 0.043	0.443 ± 0.041
Baseline + D_{xxzz} + σ(x)(RCALAD)	$\mathbf{0.537 \pm 0.054}$	0.515 ± 0.057	$\mathbf{0.526 \pm 0.055}$
Musk			
Baseline(ALAD)	0.500 ± 0.068	0.531 ± 0.070	0.515 ± 0.069
Baseline + D_{xxzz}(CALAD)	0.574 ± 0.026	0.605 ± 0.027	0.575 ± 0.026
Baseline + σ(x)(RALAD)	0.546 ± 0.051	0.565 ± 0.051	0.555 ± 0.051
Baseline + D_{xxzz} + σ(x)(RCALAD)	$\mathbf{0.629 \pm 0.011}$	$\mathbf{0.633 \pm 0.016}$	$\mathbf{0.631 \pm 0.013}$

Table 5: Effects of the various proposed sections in improving the results of the tabular datasets.

model	D_{zz}	D_{xx}	D_{xxzz}	Prec.	Recall	F1
KDD99						
ALAD	yes	yes	no	0.942 ± 0.008	$\mathbf{0.957 \pm 0.006}$	0.950 ± 0.007
ALI + D_{xxzz}	no	no	yes	0.938 ± 0.007	0.951 ± 0.010	0.944 ± 0.009
ALI + D_{zz}+ D_{xxzz}	yes	no	yes	0.946 ± 0.005	0.955 ± 0.004	0.950 ± 0.004
ALICE + D_{xxzz}	no	yes	yes	0.941 ± 0.005	0.945 ± 0.008	0.947 ± 0.006
CALAD	yes	yes	yes	$\mathbf{0.959 \pm 0.004}$	0.957 ± 0.007	$\mathbf{0.958 \pm 0.005}$
RCALAD	yes	yes	yes	0.953 ± 0.007	0.956 ± 0.005	0.954 ± 0.006
Arrhythmia						
ALAD	yes	yes	no	0.500 ± 0.049	0.531 ± 0.047	0.515 ± 0.048
ALI + D_{xxzz}	no	no	yes	0.522 ± 0.054	0.529 ± 0.049	0.525 ± 0.052
ALI+ D_{zz} + D_{xxzz}	yes	no	yes	0.571 ± 0.033	0.582 ± 0.028	0.576 ± 0.031
ALICE + D_{xxzz}	no	yes	yes	0.543 ± 0.052	0.561 ± 0.044	0.551 ± 0.048
CALAD	yes	yes	yes	0.574 ± 0.021	0.605 ± 0.022	0.575 ± 0.021
RCALAD	yes	yes	yes	$\mathbf{0.588 \pm 0.042}$	$\mathbf{0.625 \pm 0.041}$	$\mathbf{0.606 \pm 0.041}$

Table 6: Assessing the performance of the model in the presence or absence of each of the discriminators.

Score	Precision	Recall	F1 score
KDD99			
A_{L_1}	0.9081 ± 0.0638	0.9108 ± 0.0638	0.9094 ± 0.0638
A_{L_2}	0.9011 ± 0.0155	0.9004 ± 0.0157	0.9007 ± 0.0156
A_{Logits}	0.9169 ± 0.0162	0.9168 ± 0.0164	0.9168 ± 0.0163
$A_{Features}$	0.9127 ± 0.0029	0.9177 ± 0.0039	0.9151 ± 0.0034
A_{fm}	$\mathbf{0.9327 \pm 0.0017}$	$\mathbf{0.9377 \pm 0.0017}$	$\mathbf{0.9301 \pm 0.0017}$
A_{all}	0.9231 ± 0.0018	0.9207 ± 0.0018	0.9218 ± 0.0018
Arrhythmia			
A_{L_1}	0.3529 ± 0.0148	0.3750 ± 0.0164	0.3636 ± 0.0256
A_{L_2}	0.3529 ± 0.0107	0.3750 ± 0.0108	0.3636 ± 0.0107
A_{Logits}	0.5588 ± 0.0334	0.5937 ± 0.0386	0.5757 ± 0.0359
$A_{Features}$	0.2325 ± 0.0029	0.2500 ± 0.0029	0.2424 ± 0.0029
A_{fm}	0.4411 ± 0.0013	0.4687 ± 0.0013	0.4545 ± 0.0013
A_{all}	$\mathbf{0.6176 \pm 0.0208}$	$\mathbf{0.6562 \pm 0.0221}$	$\mathbf{0.6363 \pm 0.0214}$
Thyroid			
A_{L_1}	0.4981 ± 0.0028	0.4908 ± 0.0024	0.4994 ± 0.0024
A_{L_2}	0.5011 ± 0.0330	0.5004 ± 0.0318	0.5007 ± 0.0324
A_{Logits}	0.4969 ± 0.0142	0.4968 ± 0.0144	0.4968 ± 0.0143
$A_{features}$	0.5127 ± 0.0119	0.5177 ± 0.0119	0.5151 ± 0.0119
A_{fm}	0.5227 ± 0.0083	0.5123 ± 0.0083	0.5174 ± 0.0083
A_{all}	$\mathbf{0.5376 \pm 0.0029}$	$\mathbf{0.5153 \pm 0.0029}$	$\mathbf{0.5262 \pm 0.0029}$
Musk			
A_{L_1}	0.5979 ± 0.0103	0.5931 ± 0.0109	0.5954 ± 0.0106
A_{L_2}	0.6008 ± 0.0021	0.6018 ± 0.0028	0.6013 ± 0.0024
A_{Logits}	0.5868 ± 0.0124	0.5897 ± 0.0127	0.5882 ± 0.0125
$A_{Features}$	0.5824 ± 0.0011	0.5883 ± 0.0019	0.5883 ± 0.0015
A_{fm}	0.6111 ± 0.0481	0.6187 ± 0.0468	0.6148 ± 0.04744
A_{all}	$\mathbf{0.6296 \pm 0.0013}$	$\mathbf{0.6333 \pm 0.0013}$	$\mathbf{0.6314 \pm 0.0013}$

Table 7: Comparing the performance of the proposed anomaly scores with other scores on tabular data

Anomaly Score	AUROC
SVHN	
A_{L_1}	0.5778 ± 0.0141
A_{L_2}	0.5636 ± 0.0251
A_{Logits}	0.5369 ± 0.0785
$A_{Features}$	0.5763 ± 0.0367
A_{fm}	$\mathbf{0.5778 \pm 0.0161}$
A_{all}	0.5768 ± 0.0251
CIFAR-10	
A_{L_1}	0.6341 ± 0.0321
A_{L_2}	0.6327 ± 0.0782
A_{Logits}	0.6297 ± 0.0643
$A_{Features}$	0.6312 ± 0.0368
A_{fm}	$\mathbf{0.6573 \pm 0.0194}$
A_{all}	0.6477 ± 0.0227

Table 8: Comparing the performance of the proposed anomaly scores with other scores on imagery datasets.

Conclusion

This paper introduces a new model for anomaly detection based on adversarial neural networks. In the proposed model, an encoder was used for the inverse mapping of the input data space, and the discriminator D_{xx} was employed to satisfy the condition of cycle consistency. In order to stabilize the training process of the adversarial generator network, discriminator D_{zz} was adopted in the adversarial structure of the model. To use the information of a complete cycle in the proposed model, the $\hat{z}_{\hat{x}}$ variable was introduced. Consequently, the discriminator D_{xxzz} was employed to benefit from maximum information flow in the proposed model.

Moreover, supplementary distribution $\sigma(\mathrm{x})$ was utilized to bias the network output toward the normal data distribution accompanying two novel anomaly scoring specified for different data types. The outcomes of the experiments demonstrated the effectiveness of the proposed model in the field of anomaly detection, as well as its superiority over other state-of-the-art models on tabular and imagery datasets. In spite of the impressive and brilliant results of the proposed RCALAD model, like other GAN-based models, it has a robustness problem when dealing with a variety of anomaly classes. This issue could be improved by applying robustness methods that exist in (Salimans et al. 2016; Chalapathy, Menon, and Chawla 2017).

References

Chalapathy, R.; Menon, A. K.; and Chawla, S. 2017. Robust, deep and inductive anomaly detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 36–51. Springer.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1): 53–65.

Dai, X.; and Bikdash, M. 2016. Distance-based outliers method for detecting disease outbreaks using social media. In *SoutheastCon 2016*, 1–8. IEEE.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository.

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

Golan, I.; and El-Yaniv, R. 2018. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, 31.

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

Jiang, C.; Song, J.; Liu, G.; Zheng, L.; and Luan, W. 2018. Credit card fraud detection: A novel approach using aggregation strategy and feedback mechanism. *IEEE Internet of Things Journal*, 5(5): 3637–3647.

Kaur, R.; and Singh, S. 2016. A survey of data mining and social network analysis based anomaly detection techniques. *Egyptian informatics journal*, 17(2): 199–216.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, C.; Liu, H.; Chen, C.; Pu, Y.; Chen, L.; Henao, R.; and Carin, L. 2017. Alice: Towards understanding adversarial learning for joint distribution matching. *Advances in neural information processing systems*, 30.

Liang, S.; Li, Y.; and Srikant, R. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

Liu, C.; and Gryllias, K. 2020. A semi-supervised Support Vector Data Description-based fault detection method for rolling element bearings based on cyclic spectral analysis. *Mechanical Systems and Signal Processing*, 140: 106682.

Liu, F. T.; Ting, K. M.; and Zhou, Z.-H. 2008. Isolation forest. In *2008 eighth ieee international conference on data mining*, 413–422. IEEE.

Makhzani, A.; and Frey, B. J. 2015. Winner-take-all autoencoders. *Advances in neural information processing systems*, 28.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Nguyen, D. T.; Lou, Z.; Klar, M.; and Brox, T. 2019. Anomaly detection with multiple-hypotheses predictions. In *International Conference on Machine Learning*, 4800–4809. PMLR.

Pidhorskyi, S.; Almohsen, R.; and Doretto, G. 2018. Generative probabilistic novelty detection with adversarial autoencoders. *Advances in neural information processing systems*, 31.

Pimentel, M. A.; Clifton, D. A.; Clifton, L.; and Tarassenko, L. 2014. A review of novelty detection. *Signal processing*, 99: 215–249.

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

Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *International conference on machine learning*, 4393–4402. PMLR.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Langs, G.; and Schmidt-Erfurth, U. 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54: 30–44.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Schmidt-Erfurth, U.; and Langs, G. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, 146–157. Springer.

Schölkopf, B.; Williamson, R. C.; Smola, A.; Shawe-Taylor, J.; and Platt, J. 1999. Support vector method for novelty detection. *Advances in neural information processing systems*, 12.

Yang, Z.; Bozchalooi, I. S.; and Darve, E. 2020. Regularized cycle consistent generative adversarial network for anomaly detection. *arXiv preprint arXiv:2001.06591*.

Yao, D.; Shu, X.; Cheng, L.; and Stolfo, S. J. 2017. Anomaly detection as a service: challenges, advances, and opportunities. *Synthesis Lectures on Information Security, Privacy, and Trust*, 9(3): 1–173.

Zenati, H.; Foo, C. S.; Lecouat, B.; Manek, G.; and Chandrasekhar, V. R. 2018a. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*.

Zenati, H.; Romain, M.; Foo, C.-S.; Lecouat, B.; and Chandrasekhar, V. 2018b. Adversarially learned anomaly detection. In *2018 IEEE International conference on data mining (ICDM)*, 727–736. IEEE.

Zhai, S.; Cheng, Y.; Lu, W.; and Zhang, Z. 2016. Deep structured energy based models for anomaly detection. In *International conference on machine learning*, 1100–1109. PMLR.

Zimek, A.; Schubert, E.; and Kriegel, H.-P. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5): 363–387.

Zong, B.; Song, Q.; Min, M. R.; Cheng, W.; Lumezanu, C.; Cho, D.; and Chen, H. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*.