One-Class Feature Learning Using Intra-Class Splitting

Patrick Schlachter, Yiwen Liao and Bin Yang
Institute of Signal Processing and System Theory
University of Stuttgart, Germany

***Abstract*—This paper proposes a novel generic one-class feature learning method based on intra-class splitting. In one-class classification, feature learning is challenging, because only samples of one class are available during training. Hence, state-of-the-art methods require reference multi-class datasets to pretrain feature extractors. In contrast, the proposed method realizes feature learning by splitting the given normal class into typical and atypical normal samples. By introducing closeness loss and dispersion loss, an intra-class joint training procedure between the two subsets after splitting enables the extraction of valuable features for one-class classification. Various experiments on three well-known image classification datasets demonstrate the effectiveness of our method which outperformed other baseline models in average.**

I. Introduction

One-class classification is a subfield in machine learning, aiming at identifying normal data from abnormal data using a training dataset consisting merely of samples from the normal class. It is more challenging than binary or multi-class classification in which samples from all classes are available during training. Various one-class classifiers were proposed and successfully applied to a wide range of applications including fault detection, novelty detection or anomaly detection [1], [2]. Schölkopf et al. [3] or Tax et al. [4] proposed state-of-the-art one-class classifiers which are characterized by a tight decision boundary around the training samples of the normal class in order to reject abnormal samples of different kinds during inference.

To achieve a tight decision boundary, one-class classifiers require an input feature space that fulfills the following two conditions. First, features of normal data must be compactly distributed. We call this requirement *closeness*. Second, normal and abnormal data must have large distances between each other in the feature space. This requirement is called *dispersion*. Both requirements are comparable to the linear discriminant analysis in clustering in which the sample distances within a cluster are minimized and the sample distances between clusters are maximized [5].

As the stated conditions are typically not fulfilled for high-dimensional data such as natural images, the performance of state-of-the-art one-class classifiers is quite limited. Hence, a feature extraction method is necessary for such classifiers to transform raw data into a suitable latent feature space satisfying the closeness and dispersion requirements. However, this is challenging, because abnormal data samples are not available during training. Therefore, few work was done on feature learning for one-class classification.

State-of-the-art feature learning methods for one-class classification are supported by samples from other classes. In particular, a reference dataset is used to pretrain a model on many other classes [6]. The underlying assumption is that real abnormal samples might be contained in these classes. However, this assumption is too strong, since the number of other classes is unlimited. If the reference dataset is not representative for real abnormal classes, then the decision boundary of a one-class classifier will be too loose. Therefore, the choice of a reference dataset is crucial. Indeed, a meaningful multi-class reference dataset is typically not available due to the nature of the one-class classification problem.

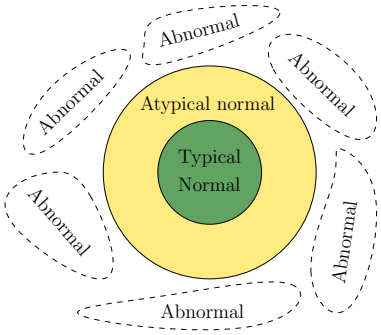

Fig. 1: In a certain latent space, typical normal data (*green*) is clustered and surrounded by atypical normal data (*yellow*). Abnormal data (*dashed lines*) is located outside this area and can be arbitrarily distributed.

In this paper, our goal is to learn a latent space that fulfills both closeness and dispersion requirements using only normal data. As a result, latent representations are optimized to the training class and not biased to any outliers. A solution for this goal does not only enable the use of conventional one-class classifiers to the learned feature space, but has also high potential for unsupervised learning and open set recognition [7]. Intuitively, we assume that in a certain latent space, normal data is clustered and abnormal data distributes around it. Fig. 1 illustrates this scenario.

In order to find such a latent space, our first idea is to split a normal training set into two subsets in an unsupervised manner and by using a similarity metric. One subset includes typical normal data, meaning the majority of a given training dataset, and the other subset consists of atypical normal data, meaning the minority of the training set. Hence, no reference dataset is necessary in contrast to [6]. By using an joint training procedure to maximize the dispersion between the typical and atypical normal data, a tight decision boundary around the normal data in the latent feature space can be achieved. This is our second key idea.

978-9-0827-9703-9/19/$31.00

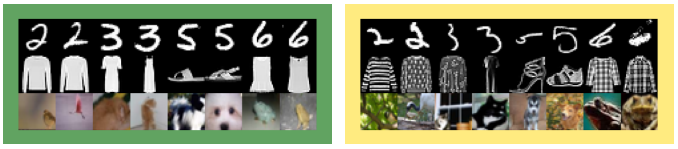

Fig. 2: *Green*: Exemplary typical normal samples with a higher SSIM score compared to their reconstructions. *Yellow*: Exemplary atypical normal samples having a lower SSIM score compared to their reconstructions. The atypical normal samples are more difficult to recognize or have more redundant details compared to typical normal ones.

One meaningful similarity metric for image data is the structural similarity (SSIM) [8]. For example, Fig. 2 illustrates the division of exemplary classes into typical and atypical normal samples using SSIM. In more detail, an autoencoder is first trained on the whole training set. Subsequently, the similarity metric between reconstructed and original data is calculated. Finally, the data with higher similarity is considered to be typical, whereas the samples with lower similarity are considered to be atypical normal data.

In general, this work is based on the assumption that unknown abnormal data has more common features with the atypical normal data rather than typical normal data in a certain latent space. In particular, we use the atypical normal samples from the given training dataset to model the unknown abnormal classes. This assumption is weaker than those in prior work such as [6], since we do not restrict the unseen abnormal data to a limited number of reference datasets, which may lead to a too loose decision boundary and a subsequent poor sensitivity. Instead, only the intra-class information from the normal class is utilized and a tight decision boundary is thus expected.

In summary, our main contributions in this paper are the following:

- *Intra-class splitting* is firstly introduced which is the key to solve feature learning for one-class classification. However, it is a generic method and not limited to one-class feature learning.
- A novel intra-class joint training strategy: Both typical and atypical normal subsets have their individual objectives as well as common objectives. These objectives are finally reformulated as a joint optimization problem.
- We empirically prove that *closeness* plays a key role in feature learning for one-class problems. Furthermore, it is shown that the combination of *closeness* and *dispersion* can achieve an even better feature extraction performance.

II. Proposed Method

A. Overview

As mentioned in the introduction, the goal is to learn a model which extracts features based on one class only, fulfilling both closeness and dispersion in the latent space. Therefore, the basic idea is to split the given normal data into typical and atypical normal samples and to use a three-stage joint training procedure. First, the foundations and strategy of intra-class splitting are explained. Subsequently, three desired characteristics of the latent space are briefly described. Then, the corresponding loss functions are introduced. Finally, an autoencoder-based network considering intra-class dispersion and closeness is presented.

B. Intra-Class Splitting

In a given normal class, not all samples are representative for this class as illustrated in Fig. 2. Accordingly, it is assumed that a given normal dataset is composed of two parts, typical normal samples and atypical normal samples. Typical normal samples are the most representative for the normal class and correspond to the majority of the given dataset. In contrast, the remaining samples are considered as atypical normal samples which may mislead the learning of a one-class classifier and thus are used to model the abnormal classes.

Intuitively, an approach to realize the splitting is utilizing neural networks with a bottleneck structure. By using a compression-decompression process such as in an autoencoder, input data is first transformed into a low-dimensional representation with information loss and then mapped back to the original data space. Hence, only the most important information of the given dataset is well maintained during this process. Accordingly, the samples contain more representative features if they are better reconstructed.

Formally, a given normal dataset χ is split by using a predefined similarity metric $f(\boldsymbol{x}, \hat{\boldsymbol{x}})$ and a ratio ρ where $\hat{\boldsymbol{x}}$ is the reconstruction of a sample $\boldsymbol{x} \in \chi$. In particular, the first $\rho\%$ samples with the lowest similarity scores are considered as atypical normal samples χ_{atypical}, while the others are considered as typical normal samples χ_{typical}.

C. Desired Characteristics of the Latent Space

1) Closeness: Input data typically distributes along a manifold in the original high-dimensional space. Therefore, it should naturally distribute in a small region in a low-dimensional latent space. In this low-dimensional space, all latent representations of normal data should be as close to each other as possible. In other words, the intra-class latent representations of normal data must have a high closeness among themselves.

2) Dispersion: While closeness is necessary for intra-class latent representations of normal data, abnormal data must be as far away from normal data as possible in the latent space. Hence, abnormal data should have a high dispersion comparing with normal data in this space. As mentioned in the introduction, a training set is assumed to consist of typical and atypical normal data. As illustrated in Fig. 1, atypical normal samples are forced to distribute far away from typical normal data. Hence, we assume that the unknown abnormal data that behave more like these atypical normal data will also lie far from the normal data.

3) Reconstruction information: While transforming raw data to the latent space, the information contained in high-dimensional data should be retained in order to ensure that the latent space is a compressed representation of the raw data. Thus, we utilize the autoencoder structure as constraints on maintaining high-dimensional information. Equivalent to the ordinary autoencoder, the reconstruction ability of latent representations is quantified by a reconstruction loss.

D. Training

As closeness and dispersion are opposite to each other, we designed a joint training procedure for feature learning. It is performed by three stages using the loss functions defined in Section II-E.

1) First Stage: All data from the training set is used to train an autoencoder only with a regular reconstruction loss $\mathcal{L}_{\text{rec}}$. During this stage, the network is considered to be a deep convolutional autoencoder. Note that no constraints are performed on the latent representations at this stage.

2) Second Stage: Once the network is trained at the first stage, the similarity between the reconstructed data and the original data is calculated. Afterwards, according to a predefined ratio ρ, the first $\rho\%$ of the data with the lowest similarity scores are chosen as the atypical normal samples, whereas the remaining data is typical normal.

3) Third Stage: To learn proper latent representations, the model is trained with different objectives regarding typical and atypical normal samples, respectively.

More precisely, the closeness loss $\mathcal{L}_{\text{cls}}$ acts as the objective for only typical normal samples to force the corresponding latent representations to be close to each other. In contrast, one dispersion loss $\mathcal{L}_{\text{disp},1}$ is designed for only atypical normal samples to keep them far away from each other in the latent space. Another dispersion loss $\mathcal{L}_{\text{disp},2}$ is taken as an objective by both typical and atypical normal samples to maximize the distances between typical and atypical latent representations. Moreover, a regular reconstruction loss is applied as the objective for all the training samples to retain the high-dimensional information.

The two different training subsets, typical and atypical normal samples, have their own objectives. Nevertheless, all loss terms are reformulated in one unified loss function defined as

$$\mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{cls}} + \beta_1\mathcal{L}_{\text{disp},1} + \beta_2\mathcal{L}_{\text{disp},2} \ , \tag{1}$$

where α, β_1 and β_2 balance the ratio between all four loss terms.

E. Loss Functions

First, the notations used in the loss functions are introduced.

- The tensor $\boldsymbol{X}$ denotes a batch of images from the training set. B is the number of images in one minibatch. $\boldsymbol{X}_j$ represents the j-th image in the given minibatch. $\boldsymbol{x}_j$ is the vectorized form of $\boldsymbol{X}_j$.
- The matrix $\boldsymbol{Z}$ denotes a batch of latent representations with L being the dimension of the latent vectors. The vector $\boldsymbol{z}_j$ is the j-th latent representation of $\boldsymbol{Z}$.
- The mapping from the raw data space to the latent space performed by the encoder is denoted as $f_{\text{enc}}(\cdot)$. Accordingly, the decoder's mapping from the latent space to the original data space is denoted as $f_{\text{dec}}(\cdot)$. The function $f(\cdot)$ describes a cascade of these two mappings.

Based on these notations, we define three loss functions for each of the three characteristics of the latent space described in Section II-C.

1) Reconstruction Loss: An ordinary mean squared error (MSE) loss is used as reconstruction loss

$$\mathcal{L}_{\text{rec}} = \frac{1}{B}\sum_{j=1}^{B} ||\boldsymbol{x}_j - \hat{\boldsymbol{x}}_j||^2 \ , \tag{2}$$

where $\hat{\boldsymbol{x}}_j = f_{\text{dec}}(f_{\text{enc}}(\boldsymbol{x}_j))$ is the reconstruction of $\boldsymbol{x}_j$. The minimization of $\mathcal{L}_{\text{rec}}$ forces the original data to be well reconstructed from the latent space. Correspondingly, the high-dimensional information is retained during the training. Moreover, this term helps to avoid trivial solutions, e.g. all-zero latent representations.

2) Closeness Loss: The closeness requirement states that each typical normal latent representation $\boldsymbol{z}_j$ should have a small distance to any another randomly chosen typical normal latent representation $\boldsymbol{z}_{i\neq j}$, where arbitrary distance metrics can be used for this purpose. Specifically, it is a metric based on the Euclidean distance

$$\mathcal{L}_{\text{cls}} = \frac{1}{B}\sum_{j=1}^{B} \sqrt{\frac{1}{L}||\boldsymbol{z}_j - \boldsymbol{z}_{i\neq j}||^2} \ , \tag{3}$$

with $\boldsymbol{z}_j = f_{\text{enc}}(\boldsymbol{X}_j)$ and $\boldsymbol{z}_{i\neq j} = f_{\text{enc}}(\boldsymbol{X}_{i\neq j})$. This minimization has the same expectation as if we calculated all distances from other latent representations, since the mini-batch stochastic gradient decent method is used in this work.

The above loss function alone has the tendency to map all raw data, both from the normal class and the abnormal classes, to a small region in the latent space. Fig. 3 shows this effect for the dataset Fashion-MNIST. As a result, latent representations of normal and abnormal data distribute in a mixture.

3) Dispersion Loss: The dispersion loss between these typical and atypical normal subsets is defined as

$$\begin{aligned}\mathcal{L}_{\text{disp}} =& \beta_1 \cdot \underbrace{\Big(-\frac{1}{B}\sum_{j=1}^{B} \sqrt{\frac{1}{L}||\boldsymbol{z}_{j,\text{atypical}} - \boldsymbol{z}_{i\neq j,\text{atypical}}||^2} \Big)}_{\mathcal{L}_{\text{disp},1}} \\ &+ \beta_2 \cdot \underbrace{\Big(-\frac{1}{B}\sum_{j=1}^{B} \sqrt{\frac{1}{L}||\boldsymbol{z}_{j,\text{atypical}} - \boldsymbol{z}_{j,\text{typical}}||^2} \Big)}_{\mathcal{L}_{\text{disp},2}} \ ,\end{aligned} \tag{4}$$

where $\boldsymbol{z}_{j,\text{atypical}} = f_{\text{enc}}(\boldsymbol{X}_{j,\text{atypical}})$ and $\boldsymbol{z}_{j,\text{typical}} = f_{\text{enc}}(\boldsymbol{X}_{j,\text{typical}})$. $\boldsymbol{X}_{j,\text{typical}}$ denotes randomly chosen typical normal samples. The minimization of $\mathcal{L}_{\text{disp},1}$ forces latent representations of atypical normal data to be far away from each other which results in a dispersion among atypical normal latent representations. Moreover, minimizing $\mathcal{L}_{\text{disp},2}$ leads to large distances between typical and atypical normal samples.

F. Network Architecture

The proposed feature learning method is modeled by a deep convolutional autoencoder based on AlexNet [9]. Both encoder and decoder are used during the training procedure. Once the model is trained, only the encoder part is utilized as a feature extractor.

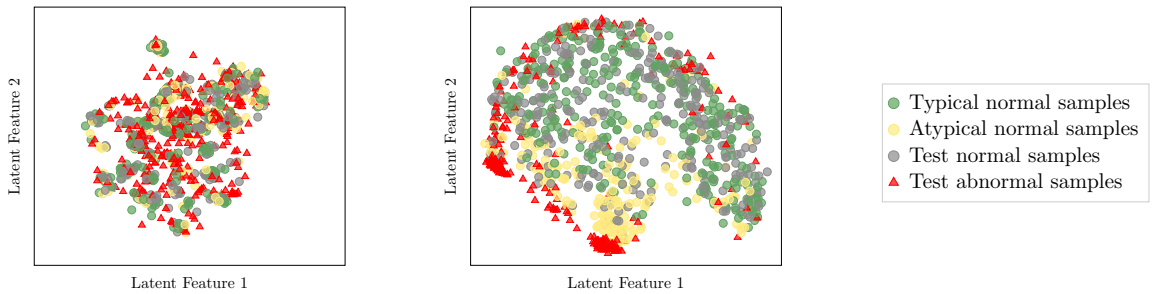

Fig. 3: If only trained with the closeness loss, the network tends to learn a too simple function to map all samples to a small region *(left)*. Accordingly, the normal data *(green, yellow and gray)* and abnormal data *(red)* distribute in an indistinguishable mixture. However, with the dispersion loss *(right)*, the typical normal samples *(green)* distribute compactly, while the atypical normal samples *(yellow)* distribute far away or around the typical normal ones. Finally, the abnormal *(red)* samples are thus also far away from the normal ones.

III. Experiments

The proposed method was used to extract latent features of different datasets which were subsequently fed into a one-class classifier. OCSVMs achieved state-of-the-art performance on extracted features, so an OCSVM was used to perform one-class classification. In one experiment, images of one class were selected as normal data, while the images of all remaining classes were considered as abnormal data and were not available during training. Each experiment was repeated five times with different initializations.

A. Experimental Setup

1) Datasets: The proposed method was evaluated on the datasets MNIST [10], Fashion-MNIST (FMNIST) [11] and CIFAR-10 [12]. All three datasets are composed of 10 different classes. The number of training data for all experiments was set to 4000. The numbers of normal and abnormal samples in the test set were 1000 and 9000, respectively[1]. For example, in one individual experiment, if digit 2 from dataset MNIST was considered as normal data, then the training set consisted of 4000 different images of digit 2. Furthermore, the test set consisted of 9000 other images of digit 2 and 1000 images of the other nine digits from 0 to 9 except of digit 2.[2]

Before training, all image pixels were normalized to the range $[0, 1]$ by min-max scaling. Note that images from MNIST and Fashion-MNIST have the size of 28×28, while images from CIFAR-10 have the size of $32 \times 32 \times 3$.

2) Hyperparameters: The proposed model was implemented with TensorFlow and Keras [13]. L2-regularization was used for every convolutional layer with a regularization parameter of 10^{-6}. The training minibatch size was 64 for all experiments. The ratio ρ for choosing atypical normal samples was set to 10, unless otherwise stated. α, β_1 and β_2 defined in Section II-E were chosen to be 1, 10^{-5} and 10^{-5}, unless otherwise specified. The dimension of the latent space was set to 64. The one-class classifier used in this work was the one-class support vector machine (OCSVM) with $\nu = 0.1$ and $\gamma = \frac{1}{\#\text{features}}$.

[1]The MNIST dataset has about but not exactly 1000 normal and 9000 abnormal samples for each individual experiment.

[2]There are 20% samples of the test set randomly selected as validation set to choose a proper classification threshold for the metric balanced accuracy.

3) Baseline Models: To the best of our knowledge, few prior works were designed for image-level one-class feature extraction. Therefore, the following conventional feature extraction methods were used as shallow baseline models:

- *Original features (Original):* The original images were vectorized as the input for OCSVM. Correspondingly, samples from MNIST and FMNIST have 784 features and those from CIFAR-10 have 3072 features.
- *Principle Component Analysis (PCA):* The first 64 PCA components of each input image were used as the input for OCSVM.
- *Histogram of oriented gradients (HOG):* HOG features [14] for each sample were extracted as the input for OCSVM. The length of HOG features for the samples from MNIST and FMNIST was 144 and for CIFAR-10 was 324.

Furthermore, the following deep feature extraction methods were considered as baselines in this work:

- *Pretrained Features (ImageNet):* Features extracted by a VGG19 [15] pretrained on ImageNet [16] were used as the input for OCSVM.
- *Convolutional Autoencoder (CAE):* Features extracted by a regular autoencoder without any constraints on the latent space were used for subsequent one-class classification.
- *Autoencoder with closeness loss (CLS):* An autoencoder was trained with the proposed closeness loss as a regularization but without intra-class splitting to extract features for one-class classification.

Note that both the baselines CAE and CLS shared the same architecture with the proposed method.

4) Metrics: In one-class classification, imbalanced data is common. Hence, we used balanced accuracy as metric, because it allows a fair comparison between our method and the baseline models also for imbalanced datasets. The balanced accuracy is defined as

$$\text{BACC} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right). \tag{5}$$

with the terms true positive (TP), true negative (TN), false negative (FN) and false positive (FP). In this work, the normal

TABLE I: Balanced Accuracy in %.

Dataset	HOG	PCA	ImageNet	CAE	Original	CLS	Ours
MNIST	64.3 ±0.0	75.2 ±0.0	68.7 ±0.0	85.2 ±0.7	84.2 ±0.0	84.7 ±5.6	**91.3 ±0.7**
FMNIST	81.1 ±0.0	81.7 ±0.0	65.2 ±0.0	81.6 ±1.0	82.3 ±0.0	84.8 ±2.5	**87.7 ±0.5**
CIFAR-10	53.8 ±0.0	57.1 ±0.0	53.7 ±0.0	56.9 ±1.4	56.5 ±0.0	54.6 ±3.0	**60.6 ±1.2**

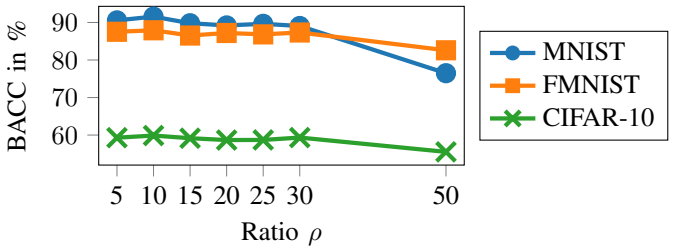

Fig. 4: Balanced accuracy vs. ratio ρ.

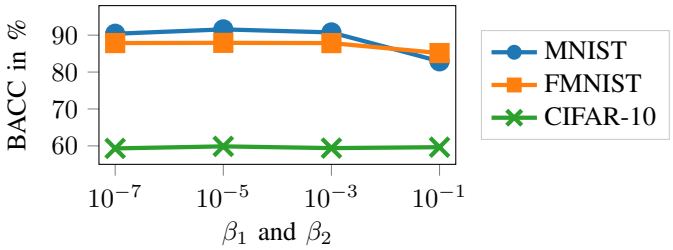

Fig. 5: Balanced accuracy vs. β_1 and β_2.

class (the known class during training) was considered as the negative class. In contrast, the abnormal classes (unknown classes during training) were considered as positive classes.

B. Results and Discussion

Table I shows the results averaged over all classes. Per class, the performance was determined by the mean and standard deviation of the balanced accuracies for different initializations. The results indicate that the proposed method performed best compared to all baseline models on all datasets. In particular, our method significantly outperformed the deep baseline models, i.e. the ImageNet and CAE features. Although CLS, a regular autoencoder only with the proposed closeness regularization, achieved a good performance, adding an additional dispersion constraint further improved the performance in all experiments.

Fig. 4 shows the balanced accuracies averaged over classes in relation to different ratios ρ. In general, the proposed method was not sensitive to the choice of the ratio ρ for splitting the training data. However, the balanced accuracies tended to be lower if the ratio was higher, e.g. $\rho = 50$, because a larger ratio indicates that more samples are considered to be atypical normal. This will lead to a low true negative rate.

Fig. 5 illustrates the balanced accuracies averaged over classes in relation to different values of β_1 and β_2. In the range of 10^{-7} to 10^{-3}, the performance of the proposed method was stable for the conducted experiments. In contrast, the balanced accuracies on MNIST and FMNIST decreased with $\beta_1 = \beta_2 = 10^{-1}$, because the dispersion loss term is dominant. This is not desirable for an OCSVM.

IV. Conclusion

In this work, we presented a novel generic feature learning method for one-class classification. To learn a proper latent space, normal training samples were split into typical and atypical normal data. These two subsets were used to train a network with different losses in an joint way under the constraints of the proposed closeness and dispersion requirements. As a result, the trained feature extractor enabled to extract highly discriminative features between normal and abnormal data. Various and intensive experiments on three image datasets used the extracted features as the input for an OCSVM to perform one-class classification. In all conducted experiments, the method outperformed all other baseline models. In particular, our method showed a large improvement over ImageNet features, CAE and the original features. An ongoing work is to realize the proposed method in an end-to-end way since it has already shown its efficient feature extraction ability for one-class classification.

References

[1] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," in *Artificial Intelligence and Cognitive Science*, L. Coyle and J. Freyne, Eds. Springer, 2010, pp. 188–197.
[2] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "Review: A review of novelty detection," *Signal Process.*, vol. 99, 2014.
[3] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
[4] D. M. Tax and R. P. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.
[5] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*. Springer, 2009.
[6] P. Perera and V. M. Patel, "Learning Deep Features for One-Class Classification," 2018.
[7] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013.
[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, pp. 600–612, 2004.
[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[11] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.
[12] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.
[13] M. Abadi, P. Barham *et al.*, "Tensorflow: a system for large-scale machine learning," in *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*. USENIX Association, 2016, pp. 265–283.
[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893.
[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.