Neural Transformation Learning for Deep Anomaly Detection Beyond Images

Chen Qiu [1 2] **Timo Pfrommer** [1] **Marius Kloft** [2] **Stephan Mandt** [3] **Maja Rudolph** [1]

Abstract

Data transformations (e.g. rotations, reflections, and cropping) play an important role in self-supervised learning. Typically, images are transformed into different views, and neural networks trained on tasks involving these views produce useful feature representations for downstream tasks, including anomaly detection. However, for anomaly detection beyond image data, it is often unclear which transformations to use. Here we present a simple end-to-end procedure for anomaly detection with *learnable* transformations. The key idea is to embed the transformed data into a semantic space such that the transformed data still resemble their untransformed form, while different transformations are easily distinguishable. Extensive experiments on time series show that our proposed method outperforms existing approaches in the one-vs.-rest setting and is competitive in the more challenging n-vs.-rest anomaly-detection task. On medical and cyber-security tabular data, our method learns domain-specific transformations and detects anomalies more accurately than previous work.

1. Introduction

Many recent advances in anomaly detection rely on the paradigm of data augmentation. In the self-supervised setting, especially for image data, predefined transformations such as rotations, reflections, and cropping are used to generate varying *views* of the data. This idea has led to strong anomaly detectors based on either transformation prediction (Golan & El-Yaniv, 2018; Wang et al., 2019b; Hendrycks et al., 2019) or using representations learned using these views (Chen et al., 2020) for downstream anomaly detection tasks (Sohn et al., 2021; Tack et al., 2020).

Unfortunately, for data other than images, such as time series or tabular data, it is much less well known which transformations are useful, and it is hard to design these transformations manually. This paper studies self-supervised anomaly detection for data types beyond images. We develop neural transformation learning for anomaly detection (NeuTraL AD): a simple end-to-end procedure for anomaly detection with *learnable* transformations. Instead of manually designing data transformations to construct auxiliary prediction tasks that can be used for anomaly detection, we derive a single objective function for jointly learning useful data transformations and anomaly thresholding. As detailed below, the idea is to learn a variety of transformations such that the transformed samples share semantic information with their untransformed form, while different views are easily distinguishable.

NeuTraL AD has only two components: a fixed set of learnable transformations and an encoder model. Both elements are jointly trained on a noise-free, deterministic contrastive loss (DCL) designed to learn faithful transformations. Our DCL is different from other contrastive losses in representation learning (Gutmann & Hyvärinen, 2010; 2012; Mnih & Kavukcuoglu, 2013; Oord et al., 2018; Bamler & Mandt, 2020; Chen et al., 2020) and image anomaly detection (Tack et al., 2020; Sohn et al., 2021), all of which use negative samples from a noise distribution. In contrast, our approach leads to a non-stochastic objective that neither needs any additional regularization or adversarial training (Tamkin et al., 2021) and can be directly used as the anomaly score.

Our approach leads to a new state of the art in anomaly detection beyond images. For time series and tabular data, NeuTraL AD significantly improves the anomaly detection accuracy. For example, in an epilepsy time series dataset, we raised the state-of-the-art from an AUC of 82.6% to 92.6% (+10%). On an Arrhythmia tabular dataset, we raised the F1 score by +2.9 percentage points to an accuracy of 60.3.

Our paper is structured as follows. We first discuss related work (Section 2) and present NeuTraL AD in Section 3. In Section 3.2, we discuss its advantages for neural transformation learning in comparison to other self-supervised learning objectives. Experimental results are presented in Section 4. Section 5 concludes this paper.

[1]Bosch Center for AI [2]TU Kaiserslautern [3]UC Irvine. Correspondence to: Maja Rudolph <majarita.rudolph@de.bosch.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

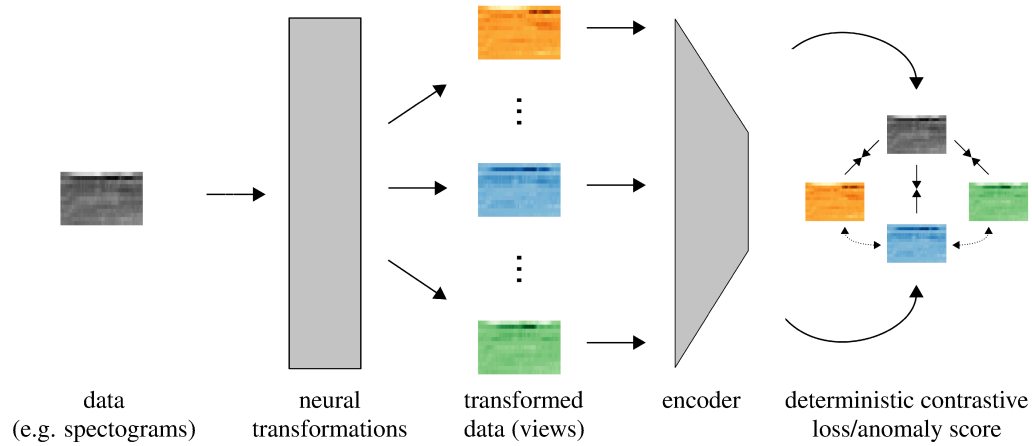


Figure 1. NeuTraL AD is a end-to-end procedure for self-supervised anomaly detection with *learnable* neural transformations. Each sample is transformed by a set of learned transformations and then embedded into a semantic space. The transformations and the encoder are trained jointly on a contrastive objective (Equation (2)), which is also used to score anomalies.

2. Related Work

Deep Anomaly Detection. Recently, there has been a rapidly growing interest in developing deep-learning approaches for anomaly detection (Ruff et al., 2021). While deep learning—by removing the burden of manual feature engineering for complex problems—has brought about tremendous technological advances, its application to anomaly detection is rather recent. Related work on deep anomaly detection includes deep autoencoder variants (Principi et al., 2017; Zhou & Paffenroth, 2017; Chen & Konukoglu, 2018), deep one-class classification (Erfani et al., 2016; Ruff et al., 2019a;b), deep generative models (Schlegl et al., 2017; Deecke et al., 2018), and outlier exposure (Hendrycks et al., 2018; Goyal et al., 2020).

Self-supervised anomaly detection has led to drastic improvements in detection accuracy (Golan & El-Yaniv, 2018; Hendrycks et al., 2019; Wang et al., 2019b; Sohn et al., 2021; Tack et al., 2020). For instance, Golan & El-Yaniv (2018) and Wang et al. (2019b) augment the data and learn to predict which transformation was applied. After training, the resulting classifier is used for anomaly detection.

An alternative approach to self-supervised anomaly detection is to train a classifier on a contrastive loss to tell if two views are of the same original image. This leads to strong representations (Chen et al., 2020), which can be used for anomaly detection (Sohn et al., 2021; Tack et al., 2020).

Bergman & Hoshen (2020) study how to extend self-supervised anomaly detection to other domains beyond images. Similar to Golan & El-Yaniv (2018) and Wang et al. (2019b), their approach is based on transformation prediction, but they consider the open-set setting. For tabular data, they use random affine transformations. We study the same datasets, but our method *learns* the transformations and achieves consistently higher performance (see Section 4).

Self-Supervised Learning. Self-supervised learning typically relies on data augmentation for auxiliary tasks (Doersch et al., 2015; Noroozi & Favaro, 2016; Misra et al., 2016; Zhang et al., 2017; Gidaris et al., 2018). The networks trained on these auxiliary tasks (e.g. patch prediction (Doersch et al., 2015), solving jigsaw-puzzles (Noroozi & Favaro, 2016), cross-channel prediction (Zhang et al., 2017), or rotation prediction (Gidaris et al., 2018)) are used as feature extractors for downstream tasks. While many of these methods are developed for images, Misra et al. (2016) propose temporal order verification as an auxiliary task for self-supervised learning of time series representations.

Contrastive Representation Learning. Many recent self-supervised methods have relied on the InfoMax principle (Linsker, 1988; Hjelm et al., 2018). These methods are trained on the task to maximize the mutual information (MI) between the data and their context (Oord et al., 2018) or between different "views" of the data (Bachman et al., 2019). Computing the mutual information in these settings is often intractable and various approximation schemes and bounds have been introduced (Tschannen et al., 2019). By using noise contrastive estimation (Gutmann & Hyvärinen, 2010; 2012) to bound MI, Oord et al. (2018) bridge the gap between contrastive losses for MI-based representation learning and the use of contrastive losses in discriminative methods for representation learning (Hadsell et al., 2006; Mnih & Kavukcuoglu, 2013; Dosovitskiy et al., 2015; Bachman et al., 2019; Chen et al., 2020). We also use a contrastive loss. But while the contrastive loss of Chen et al. (2020) (which is used for anomaly detection of images in Sohn et al. (2021); Tack et al. (2020),) contrast two views of the same sample with views of other samples in the mini-

batch, NeuTraL AD is tasked with determining the original version of a sample from different views of the same sample. The dependence on only a single sample is advantageous for scoring anomalies at test time and enables us to learn the data transformations (discussed further in Section 3.2).

Learning Data Augmentation Schemes. The idea of learning data augmentation schemes is not new. "AutoAugmentation" has usually relied on composing hand-crafted data augmentations (Ratner et al., 2017; Cubuk et al., 2019; Zhang et al., 2019; Ho et al., 2019; Lim et al., 2019). Tran et al. (2017) learn Bayesian augmentation schemes for neural networks, and Wong & Kolter (2020) learn perturbation sets for adversarial robustness. Though their setting and approach are different, our work is most closely related to Tamkin et al. (2021), who study how to generate views for representation learning in the framework of Chen et al. (2020). They parametrize their "viewmakers" as residual perturbations, which are trained adversarially to avoid trivial solutions where the views share no semantic information with the original sample (discussed in Section 3.2).

NeuTraL AD falls into the area of deep, self-supervised anomaly detection, with the core novelty of learning the transformations so that we can effectively use them for anomaly detection beyond images. Our method receives whole time series or tabular data as input. For time series, this is a remarkable difference to prevalent work on anomaly detection *within* time series (e.g. Shen et al., 2020), which output anomaly scores per time point, but not for the sequence as a whole. Additional approaches to time series anomaly detection include shallow (Hyndman et al., 2015; Baragona & Battaglia, 2007) and deep-learning approaches based on modeling (Munir et al., 2018), on autoencoders (Kieu et al., 2018), or one-class classification (Shen et al., 2020).

3. Neural Transformation Learning for Deep Anomaly Detection

We develop neural transformation learning for anomaly detection (NeuTraL AD), a deep anomaly detection method based on contrastive learning for general data types. We first describe the approach in Section 3.1. In Section 3.2, we provide theoretical arguments why alternative contrastive loss functions are less suited for transformation learning.

3.1. Proposed Method: NeuTraL AD

Our method, NeuTraL AD, is a simple pipeline with two components: a set of learnable transformations, and an encoder. Both are trained jointly on a deterministic contrastive loss (DCL). The objective has two purposes. During training, it is optimized to find the parameters of the encoder and the transformations. During testing, the DCL is also used to score each sample as either an inlier or an anomaly.

Learnable Data Transformations. We consider a data space $\mathcal{X}$ with samples $\mathcal{D} = \{x^{(i)} \sim \mathcal{X}\}_{i=1}^{N}$. We also consider K transformations $\mathcal{T} := \{T_1, ..., T_K | T_k : \mathcal{X} \to \mathcal{X}\}$. We assume here that the transformations are learnable, i.e., they can be modeled by any parameterized function whose parameters are accessible to gradient-based optimization and we denote the parameters of transformation T_k by θ_k. In our experiments, we use feed-forward neural networks for T_k. Note that in Section 3.2, we use the same notation also for fixed transformations (such as rotations and cropping).

Deterministic Contrastive Loss (DCL). A key ingredient of NeuTraL AD is a new objective. The DCL encourages each transformed sample $x_k = T_k(x)$ to be similar to its original sample x, while encouraging it to be dissimilar from other transformed versions of the same sample, $x_l = T_l(x)$ with $l \neq k$. We define the score function of two (transformed) samples as

$$h(x_k, x_l) = \exp(\text{sim}(f_\phi(T_k(x)), f_\phi(T_l(x)))/\tau), \quad (1)$$

where τ denotes a temperature parameter, and the similarity is defined as the cosine similarity $\text{sim}(z, z') := z^T z'/\|z\|\|z'\|$ in an embedding space $\mathcal{Z}$. The encoder $f_\phi(\cdot) : \mathcal{X} \to \mathcal{Z}$ serves as a features extractor. The DCL is

$$\mathcal{L} := \mathbb{E}_{x\sim\mathcal{D}}\left[-\sum_{k=1}^{K} \log \frac{h(x_k, x)}{h(x_k, x) + \sum_{l\neq k} h(x_k, x_l)}\right]. \quad (2)$$

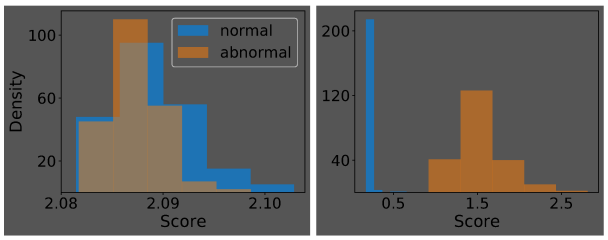
The term in the nominator of Equation (2) pulls the embedding of each transformed sample close to the embedding of the original sample. This encourages the transformations to preserve relevant semantic information. The denominator pushes all the embeddings of the transformed samples away from each other, thereby encouraging diverse transformations. The parameters of NeuTraL AD $\theta = [\phi, \theta_{1:K}]$ consist of the parameters ϕ of the encoder, and the parameters $\theta_{1:K}$ of the learnable transformations. All parameters θ are optimized jointly by minimizing Equation (2).

A schematic of NeuTraL AD is in Figure 1. Each sample is transformed by a set of learnable transformations and then embedded into a semantic space. The transformations and the encoder are trained jointly on the DCL (Equation (2)), which is also used to score anomalies.

Anomaly Score. One advantage of our approach over other methods is that our training loss is also our anomaly score. Based on Equation (2), we define an anomaly score $S(x)$ as

$$S(x) = -\sum_{k=1}^{K} \log \frac{h(x_k, x)}{h(x_k, x) + \sum_{l\neq k} h(x_k, x_l)}. \quad (3)$$

Since the score is deterministic, it can be straightforwardly evaluated for new data points x without negative samples.



(a) histogram before training (b) histogram after training

Figure 2. While the histogram of anomaly scores (computed using Equation (3)) is similar for inliers (blue) and anomalies (orange) before training, this changes drastically after training, and held-out inliers and anomalies become easily distinguishable. The data come from the SAD experiments described in Section 4

By minimizing the DCL (Equation (2)), we minimize the score (Equation (3)) for training examples (inliers). The higher the anomaly score, the more likely that a sample is an anomaly.

This concludes the proposed loss function. We stress that it is extremely simple and does not need any additional regularization. However, it is not trivial to see why other proposed self-supervised approaches are less well suited for anomaly detection without imposing constraints on the types of transformations. To this end, we establish theoretical requirements and desiderata for neural transformation learning.

3.2. A Theory of Neural Transformation Learning

To learn transformations for self-supervised anomaly detection we pose two requirements.

Requirement 1 (Semantics)**.** *The transformations should produce views that share relevant semantic information with the original data.*

Requirement 2 (Diversity)**.** *The transformations should produce diverse views of each sample.*

A valid loss function for neural transformation learning should avoid solutions that violate either of these requirements. There are plenty of transformations that would violate Requirement 1 or Requirement 2. For example, a constant transformation $T_k(x) = c_k$, where c_k is a constant that does not depend on x, would violate the semantic requirement, whereas the identity $T_1(x) = \cdots = T_K(x) = x$ violates the diversity requirement.

We argue thus that for self-supervised anomaly detection, the learned transformations need to negotiate the trade-off between semantics and diversity with the two examples as edge-cases on a spectrum of possibilities. Without semantics, i.e. without dependence on the input data, an anomaly detection method can not decide whether a new sample is normal or an anomaly. And without variability in learning transformations, the self-supervised learning goal is not met.

We now put the benefits of NeuTraL AD into perspective by comparing the approach with two other approaches that use data transformations for anomaly detection:

1. The first approach is the transformation prediction approach by Wang et al. (2019b). Here, $f_\phi(\cdot) : \mathcal{X} \to \mathbb{R}^K$ is a deep classifier[1] that outputs K values $f_\phi(x)_1 \cdots f_\phi(x)_K$ proportional to the log-probabilities of the transformations. The transformation prediction loss is a softmax classification loss,

$$\mathcal{L}_P := \mathbb{E}_{x\sim\mathcal{D}} \left[-\sum_{k=1}^{K} \log \frac{\exp f_\phi(x_k)_k}{\sum_{l=1:K} \exp f_\phi(x_k)_l} \right]. \quad (4)$$

2. We also consider (Chen et al., 2020), who define a contrastive loss on each minibatch of data $\mathcal{M} \subset \mathcal{D}$ of size $N = |\mathcal{M}|$. For each gradient step, they sample a minibatch and two transformations $T_1, T_2 \sim \mathcal{T}$ from the family of transformations, which are applied to all the samples to produce $x_k^{(i)} = T_k(x^{(i)})$. The loss function is given by $\mathcal{L}_C(\mathcal{M}) := \sum_{i=1}^{N} \mathcal{L}_C(x_1^{(i)}, x_2^{(i)}) + \mathcal{L}_C(x_2^{(i)}, x_1^{(i)})$, where

$$\mathcal{L}_C(x_1^{(i)}, x_2^{(i)}) := -\log h(x_1^{(i)}, x_2^{(i)}) \quad (5)$$
$$+ \log \left[\sum_{j=1}^{N} h(x_1^{(i)}, x_2^{(j)}) + \sum_{j=1}^{N} \mathbb{1}_{[j\neq i]} h(x_1^{(i)}, x_1^{(j)}) \right].$$

With hand-crafted, fixed transformations, these losses produce excellent anomaly detectors for images (Golan & El-Yaniv, 2018; Wang et al., 2019b; Sohn et al., 2021; Tack et al., 2020). Since it is not always easy to design transformations for new application domains, we study their suitability for *learning* data transformations.

We argue that $\mathcal{L}_P$ and $\mathcal{L}_C$ are less well suited for transformation learning than $\mathcal{L}$:

Proposition 1. *The 'constant' edge-case $f_\phi(T_k(x)) = Cc_k$, where c_k is a one-hot vector encoding the k^{th} position (i.e. $c_{kk} = 1$), tends towards the minimum of $\mathcal{L}_P$ (Equation (4)) as the constant C goes to infinity.*

Proposition 2. *The 'identity' edge-case $T_k(x) = x$ with adequate encoder f_ϕ is a minimizer of $\mathcal{L}_C$ (Equation (5)).*

The proof of these propositions is in Appendix A. The intuition is simple. Transformations that only encode which transformation was used make transformation prediction easy (Proposition 1), whereas the identity makes any two views of a sample identical, which can then be easily recognized as a positive pair by $\mathcal{L}_C$ (Proposition 2).

[1] even though here f_ϕ is a classifier, we refer to it as the encoder in the discussion below.

The propositions highlight a serious issue with using $\mathcal{L}_P$ or $\mathcal{L}_C$ for transformation learning and anomaly detection. Should the optimization reach the edge-cases of Propositions 1 and 2, $\mathcal{L}_P$ and $\mathcal{L}_C$ incur the same loss irrespective of whether the inputs are normal or abnormal data. Here are three remedies that can help avoid the trivial edge-cases: Through careful *parametrization*, one can define $T_k(\cdot;\theta_k)$ in a way that explicitly excludes the edge cases. Beware that the transformation family might contain other members that violate Requirement 1 or Requirement 2. The second potential remedy is *regularization* that explicitly encourages Requirements 1 and 2, e.g. based on the InfoMax principle (Linsker, 1988) or norm constraints. Finally, one can resort to *adversarial training*.

Tamkin et al. (2021) use all three of these remedies to learn "viewmakers" using the contrastive loss $\mathcal{L}_C$; They parametrize the transformations as residual perturbations, which are regularized to the ℓ_p ball and trained adversarially. In contrast, under NeuTraL AD there are no restrictions on the architecture of the transformations, as long as Equation (2) can be optimized (i.e. the gradient is well defined). The DCL is an adequate objective for training the encoder and transformations jointly as it manages the trade-off between Requirements 1 and 2.

Proposition 3. *The edge-cases of Propositions 1 and 2 do not minimize $\mathcal{L}$ (DCL, Equation (2)).*

The proof is in Appendix A. The numerator of Equation (2) encourages transformed samples to resemble their original version (i.e. the semantic requirement) and the denominator encourages the diversity of transformations. The result of our well-balanced objective is a heterogeneous set of transformations that model various relevant aspects of the data. The transformations and the encoder need to highlight salient features of the data such that a low loss can be achieved. After training, samples from the normal class have a low anomaly score while anomalies are handled less well by the model and as a result, have a high score.

Figure 2 shows empirical evidence for this. We observe that, while the histogram of anomaly scores (computed using Equation (3)) is similar for inliers and anomalies before training, this changes drastically after training, and held-out inliers and anomalies become easily distinguishable.

There's another advantage of using the DCL for self-supervised anomaly detection. Unlike most other contrastive losses, the "negative samples" are not drawn from a noise distribution (e.g. other samples in the minibatch) but constructed deterministically from x. Dependence on the minibatch for negative samples would need to be accounted for at test time. In contrast, the deterministic nature of Equation (3) makes it a simple choice for anomaly detection.

4. Empirical Study: Deep Anomaly Detection of Time Series and Tabular Data

We developed NeuTraL AD for deep anomaly detection beyond images, so we consider various application domains involving multiple data types. For image data, strong self-supervised baselines exist that benefit from hand-crafted transformations. We do not expect any benefit from using NeuTraL AD there. Our focus here is on time series and tabular data, which are important in many application domains of anomaly detection. Our study finds that NeuTraL AD improves detection accuracy over the state of the art.

4.1. Evaluation Protocol

We compare NeuTraL AD to prevalent shallow and deep anomaly-detection baselines using two evaluation protocols: the standard 'one-vs.-rest' and the more challenging 'n-vs.-rest' evaluation protocol. Both settings turn a classification dataset into a quantifiable anomaly-detection benchmark.

one-vs-rest. This evaluation setup has been used in virtually all papers on deep anomaly detection published at top-tier venues (e.g. Ruff et al., 2019a; Hendrycks et al., 2019; Ruff et al., 2018; Golan & El-Yaniv, 2018; Deecke et al., 2018; Akcay et al., 2018; Abati et al., 2019; Perera et al., 2019; Wang et al., 2019a; Bergman & Hoshen, 2020; Kim et al., 2019). For 'one-vs.-rest', the dataset is split by the N class labels, creating N one class classification tasks; the models are trained on data from one class and tested on a test set with examples from all classes. The samples from other classes should be detected as anomalies.

n-vs-rest. We also evaluate on the more challenging n-vs.-rest protocol, where n classes (for $1 < n < N$) are treated as normal and the remaining classes provide the anomalies in the test and validation set. By increasing the variability of what is considered normal data, one-class classification becomes more challenging.

4.2. Shallow and Deep Anomaly Detection Baselines

We study NeuTraL AD in comparison to a number of unsupervised and self-supervised anomaly detection methods.

Traditional Anomaly Detection Baselines. We chose three popular anomaly detection baselines: The one-class SVM (OC-SVM), a kernel-based method, isolation forest (IF), a tree-based model which aims to isolate anomalies (Liu et al., 2008), and local outlier factor (LOF), which uses density estimation with k-nearest neighbors.

Deep Anomaly Detection Baselines. Next, we include three deep anomaly detection methods, Deep SVDD (Ruff et al., 2018), which fits a one-class SVM in the feature space of a neural net, DROCC (Goyal et al., 2020), which fits a one-class classifier with artificial outlier exposure, and

DAGMM (Zong et al., 2018), which estimates the density in the latent space of an autoencoder.

Self-Supervised Anomaly Detection Baselines. We also choose two self-supervised baselines, which are technically also deep anomaly detection methods. GOAD (Bergman & Hoshen, 2020) is a distance-based classification method based on random affine transformations. Wang et al. (2019b) is a softmax-based classification method based on hand-crafted transformations, which show impressive performance on images. We adopt their pipeline to time series here by crafting specific time series transformations (fixed Ts, described in Appendix B).

Anomaly Detection Baselines for Time Series. Finally, we also include two baselines that are specifically designed for time series data: The RNN-based model (RNN) directly models the data distribution $p(x_{1:T}) = \prod p(x_t|x_{<t})$ and uses the log-likelihood as the anomaly score. Details on the architecture are in Appendix B. LSTM-ED (Malhotra et al., 2016) is an encoder-decoder time series model where anomaly score is based on the reconstruction error.

4.3. Anomaly Detection of Time Series

Our goal is to detect abnormal time series on a *whole-sequence level*. This is a different set-up than novelty detection within time series, but also very important in practice.

For example, one might want to detect abnormal sound or find production quality issues by detecting abnormal sensor measurements recorded over the duration of producing a batch. Other applications are sports and health monitoring; an abnormal movement pattern during sports can be indicative of fatigue or injury; whereas anomalies in health data can point to more serious issues.

We study NeuTraL AD on a selection of datasets that are representative of these varying domains. The datasets come from the UEA multivariate time series classification archive[2] (Bagnall et al., 2018).

Time Series Datasets

- SAD: Sound of ten Arabic digits, spoken by 88 speakers. The dataset has 8800 samples, which are stored as 13 Mel Frequency Cepstral Coefficients (MFCCs). The data is zero-padded to have the same time length of 50.
- Naval air training and operating procedures standardization (NATOPS): The data is originally from a motion detection competition of various movement patterns used to control planes in naval air training. The data has six classes of distinct actions. Each sample is a sequence of x, y, z coordinates for eight body parts of length 51.
- Character trajectories (CT): The data consists of 2858 character samples from 20 classes, captured using a WACOM tablet. Each instance is a 3-dimensional pen tip velocity trajectory. The data is truncated to the length of the shortest, which is 182.
- Epilepsy (EPSY): The data was generated with healthy participants simulating four different activities: walking, running, sawing with a saw, and seizure mimicking whilst seated. The data has 275 cases in total, each being a 3-dimensional sequence of length 203.
- Racket sports (RS): The data is a record of university students playing badminton or squash whilst wearing a smart watch, which measures the x, y, z coordinates for both the gyroscope and accelerometer. Sport and stroke types separate the data into four classes. Each sample is a 6-d sequence with a length of 30.

We compare NeuTraL AD to all baselines from Section 4.2 on these datasets under the one-vs-rest setting. Additionally, we study how the methods adapt to increased variability of inliers by exploring SAD and NATOPS under the n-vs-rest setting for a varying number of classes n considered normal.

Implementation Details We consider the following parametrizations of the learnable transformations: *feed forward* $T_k(x) := M_k(x)$, *residual* $T_k(x) := M_k(x) + x$, and *multiplicative* $T_k(x) := M_k(x) \quad x$, which differ in how they combine the learnable masks $M_k(\cdot)$ with the data[3]. The masks M_k are each a stack of three residual blocks of 1d convolutional layers with instance normalization layers and ReLU activations, as well as one 1d convolutional layer on the top. For the multiplicative parameterization, a sigmoid activation is applied to the masks. All bias terms are fixed as zero, and the learnable affine parameters of the instance normalization layers are frozen. For a fair comparison, we use the same number of 12 transformations in NeuTraL AD, GOAD, and the classification-based method (fixed Ts) for which we manually designed appropriate transformations. In Section 4.5 we make a more detailed comparison of various design choices for NeuTraL AD and one of our findings is that its anomaly detection results are robust to the number of learnable transformations.

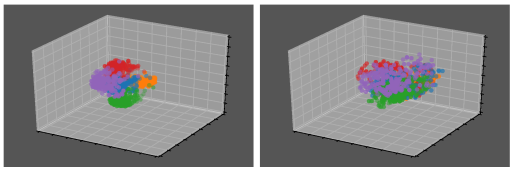
The same encoder architecture is used for NeuTraL AD, Deep SVDD, DROCC, and with slight modification to achieve the appropriate number of outputs for DAGMM and transformation prediction with fixed Ts. The encoder is a stack of residual blocks of 1d convolutional layers. The number of blocks depends on the dimensionality of the data and is detailed in Appendix B. The encoder has output dimension 64 for all experiments.

[2] from which we selected datasets on which supervised multi-class classification methods achieve strong results (Ruiz et al., 2020). Only datasets with separable classes can be repurposed for one-class classification

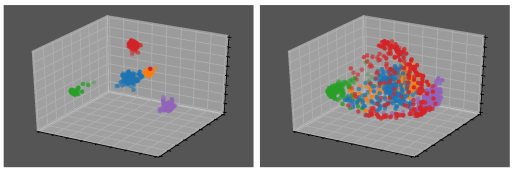
[3] We use 10% of the test set as the validation set to allow parameterization selection.

Table 1. Average AUC with standard deviation for one-vs-rest anomaly detection on time series datasets.

	OC-SVM	IF	LOF	RNN	LSTM-ED	Deep SVDD	DAGMM	GOAD	DROCC	fixed Ts	NeuTraL AD
SAD	95.3	88.2	98.3	81.5±0.4	93.1±0.5	86.0±0.1	80.9±1.2	94.7±0.1	85.8±0.8	96.7±0.1	**98.9**±0.1
NATOPS	86.0	85.4	89.2	89.5±0.4	91.5±0.3	88.6±0.8	78.9±3.2	87.1±1.1	87.2±1.4	78.4±0.4	**94.5**±0.8
CT	97.4	94.3	97.8	96.3±0.2	79.0±1.1	95.7±0.5	89.8±0.7	97.7±0.1	95.3±0.3	97.9±0.1	**99.3**±0.1
EPSY	61.1	67.7	56.1	80.4±1.8	82.6±1.7	57.6±0.7	72.2±1.6	76.7±0.4	85.8±2.1	80.4±2.2	**92.6**±1.7
RS	70.0	69.3	57.4	84.7±0.7	65.4 ±2.1	77.4±0.7	51.0±4.2	79.9±0.6	80.0±1.0	**87.7**±0.8	86.5±0.6



(a) normal data on $\mathcal{X}$ (b) anomalies on $\mathcal{X}$



(c) normal data on $\mathcal{Z}$ (d) anomalies on $\mathcal{Z}$

Figure 3. 3D visualization (projected using PCA) of how the original samples (blue) from the SAD dataset and the different views created by the neural transformations of NeuTraL AD (one color per transformation type) cluster in data space (Figures 3a and 3b) and in the embedding space of the encoder (Figures 3c and 3d). The crisp separation of the different transformations of held-out inliers (Figure 3c) in contrast to the overlap between transformed anomalies (Figure 3d) visualizes how NeuTraL AD is able to detect anomalies.

Results. The results of NeuTraL AD in comparison to the baselines from Section 4.2 on time series datasets from various fields are reported in Table 1. NeuTraL AD outperforms all shallow baselines in all experiments and outperforms the deep learning baselines in 4 out of 5 experiments. Only on the RS data, it is outperformed by transformation prediction with fixed transformations, which we designed to understand the value of learning transformations with NeuTraL AD vs using hand-crafted transformations. The results confirm that designing the transformations only succeeds sometimes, whereas with NeuTraL AD we can learn the appropriate transformations. The learned transformations also give NeuTraL AD a competitive advantage over the other self-supervised baseline GOAD which uses random affine transformations. The performance of the shallow anomaly detection baselines hints at the difficulty of each anomaly detection task; the shallow methods perform well on SAD and CT, but perform worse than the deep learning based methods on other data.

What does NeuTraL AD learn? For visualization purposes, we train NeuTraL AD with 4 learnable transformations on the SAD data. Figure 3 shows the structure in the data space $\mathcal{X}$ and the embedding space of the encoder $\mathcal{Z}$ after training. Held-out data samples (blue) are transformed by each of the learned transformations $T_k(x) = M_k(x) \quad x$ to produce $K = 4$ different views of each sample (the transformations are color-coded by the other colors). Projection to three principal components with PCA allows for visualization in 3D. In Figures 3a and 3b, we can see that the transformations already cluster together in the data space, but only with the help of the encoder, the different views of inliers are separated from each other Figure 3c. In comparison, the anomalies and their transformations are less structured in $\mathcal{Z}$ (Figure 3d), visually explaining why they incur a higher anomaly score and can be detected as anomalies.

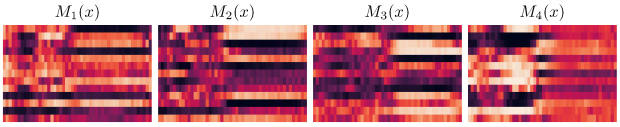


Figure 4. NeuTraL AD learns dissimilar masks for SAD spectrograms. Dark horizontal lines indicate where M_1 and M_2 mask out frequency bands almost entirely, while the bright spot in the middle left part of M_4 indicates that this mask brings the intermediate frequencies in the first half of the recording into focus.

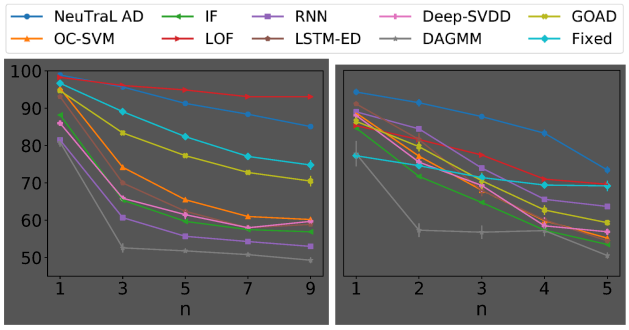
The learned masks $M_{1:4}(x)$ of one inlier x are visualized in Figure 4. We can see that the four masks are dissimilar from each other, and have learned to focus on different aspects of the spectrogram. The masks take values between 0 and 1, with dark areas corresponding to values close to 0 that are zeroed out by the masks, while light colors correspond to the areas of the spectrogram that are not masked out. Interestingly, in M_1, M_2, and M_3 we can see 'black lines' where they mask out entire frequency bands at least for part of the sequence. In contrast, M_4 has a bright spot in the middle left part of the spectrogram; it creates views that focus on the content of the intermediate frequencies at the first half of the recording.

How do the methods cope with an increased variability of inliers?

To study this question empirically, we increase the number of classes n considered to be inliers. We test all methods on SAD and NATOPS under the n-vs-rest setting with vary-

Table 2. Average AUC with standard deviation for n-vs-rest ($n = N - 1$) anomaly detection on time series datasets

	OC-SVM	IF	LOF	RNN	LSTM-ED	Deep SVDD	DAGMM	GOAD	DROCC	fixed Ts	NeuTraL AD
SAD	60.2	56.9	**93.1**	53.0±0.1	58.9±0.5	59.7±0.5	49.3±0.8	70.5±1.4	58.8±0.5	74.8±1.3	85.1±0.3
NATOPS	57.6	56.0	71.2	65.6±0.4	56.9±0.7	59.2±0.8	53.2±0.8	61.5±0.7	60.7±1.6	70.8±1.3	**74.8**±0.9
CT	57.8	57.9	**90.3**	55.7±0.8	50.9±1.2	54.4±0.7	47.5±2.5	81.1±0.1	57.6±1.5	63.0±0.6	87.4±0.2
EPSY	50.2	55.3	54.7	74.9±1.5	56.8±2.1	52.9±1.4	52.0±1.0	62.7±0.9	55.5±1.9	69.8±1.6	**80.5**±1.0
RS	55.9	58.4	59.4	75.8±0.9	63.1 ±0.6	62.2±2.1	47.8±3.5	68.2±0.9	60.9±0.2	**81.6**±1.2	80.0±0.4



(a) AUCs on SAD (b) AUCs on NATOPS

Figure 5. AUC result of n-vs-all experiments on SAD and NATOPS with error bars (barely visible due to significance). NeuTraL AD outperforms all baselines on NATOPS and all deep learning baselines on SAD. LOF, a method based on k-nearest neighbors, outperforms NeuTraL AD, when $n > 3$ on SAD.

ing n. Since there are too many combinations of normal classes when n approaches $N - 1$, we only consider combinations of n consecutive classes. From Figure 5 we can observe that the performance of all methods drops as the number of classes included in the normal data increases. This shows that the increased variance in the nominal data makes the task more challenging. NeuTraL AD outperforms all baselines on NATOPS and all deep-learning baselines on SAD. It is interesting that LOF, a method based on k-nearest neighbors, performs better than our method (and all other baselines) on SAD when n is larger than three. We also include quantitative results for $n = N - 1$ under the n-vs-rest setting for all time series datasets, where only one class is considered abnormal, and the remaining $N - 1$ classes are normal. The results are reported in Table 2. We can see, the performance of all deep learning based methods drops as the variability of normal data increases. Our method outperforms other deep learning methods on 4 out of 5 datasets. On RS, it is outperformed by transformation prediction with hand-crafted transformations. The results are consistent with the experiments under one-vs.-rest setting in Table 1. The traditional method LOF performs better than deep learning methods on CT and SAD.

4.4. Anomaly Detection of Tabular Data

Tabular data is another important application area of anomaly detection. For example, many types of health data come in tabular form. To unleash the power of self-supervised anomaly detection for these domains, Bergman & Hoshen (2020) suggest using random affine transformation. Here we study the benefit of *learning* the transformations with NeuTraL AD. We base the empirical study on tabular datasets used in previous work (Zong et al., 2018; Bergman & Hoshen, 2020) and follow their precedent of reporting results in terms of F1-scores.

Tabular Datasets. We study the four tabular datasets from the empirical studies of Zong et al. (2018); Bergman & Hoshen (2020). The datasets include the small-scale medical datasets Arrhythmia and Thyroid as well as the large-scale cyber intrusion detection datasets KDD and KDDRev (see Appendix C for all relevant details). We follow the configuration of (Zong et al., 2018) to train all models on half of the normal data, and test on the rest of the normal data as well as the anomalies.

Baseline Models. We compare NeuTraL AD to shallow and deep baselines outlined in Section 4.2, namely OC-SVM, IF, LOF, and the deep anomaly detection methods Deep SVDD, DAGMM, GOAD, and DROCC.

Implementation details. The implementation details of OC-SVM, LOF, DAGMM, and GOAD are replicated from Bergman & Hoshen (2020), as we report their results. The implementation of DROCC is from their official code.

The learnable transformations are again parameterized multiplicatively $T_k(x) = M_k(x) \quad x$, with the masks M_k consisting of 3 bias-free linear layers with intermediate ReLU activations and sigmoid activation for the output layer. The number of learnable transformations is 11 for Arrhythmia, 4 for Thyroid, and 7 for KDD and KDDRev.

We use a comparable encoder architecture for NeuTraL AD and Deep SVDD of 3 (4 for KDD and KDDRev) linear layers with ReLU activations. The output dimensions of the encoder are 12 for Thyroid and 32 for the other datasets.

Results. The results of OC-SVM, LOF, DAGMM, and GOAD are taken from (Bergman & Hoshen, 2020). The results of DROCC were provided by Goyal et al. (2020) [4]. NeuTraL AD outperforms all baselines on all datasets. Compared with the self-supervised anomaly detection baseline

[4]Their empirical study uses a different experimental setting while the results reported here are consistent with prior works.

Table 3. F1-score (%) with standard deviation for anomaly detection on tabular datasets (choice of F1-score consistent with prior work).

	Arrhythmia	Thyroid	KDD	KDDRev
OC-SVM	45.8	38.9	79.5	83.2
IF	57.4	46.9	90.7	90.6
LOF	50.0	52.7	83.8	81.6
Deep SVDD	53.9±3.1	70.8±1.8	99.0±0.1	98.6±0.2
DAGMM	49.8	47.8	93.7	93.8
GOAD	52.0±2.3	74.5±1.1	98.4±0.2	98.9±0.3
DROCC	46	27	-	-
NeuTraL AD	**60.3**±1.1	**76.8**±1.9	**99.3**±0.1	**99.1**±0.1

GOAD, we use much fewer transformations.

4.5. Design Choices for the Transformations

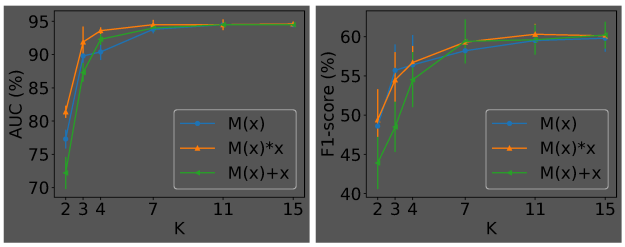
In Section 3.2, we have discussed the advantages of NeuTraL AD for neural transformation learning; no regularization or restrictions on the transformation family are necessary to ensure the transformations fulfill the semantics and diversity requirements defined in Section 3.2.

In this section, we study the performance of NeuTraL AD under various design choices for the learnable transformations, including their parametrization, and their number K. We consider the following parametrizations: *feed forward* $T_k(x) := M_k(x)$, *residual* $T_k(x) := M_k(x) + x$, and *multiplicative* $T_k(x) := M_k(x) \quad x$, which differ in how they combine the learnable masks $M_k(\cdot)$ with the data.

In Figure 6 we show the anomaly detection accuracy achieved with each parametrization, as K varies from 2 to 15 on the time series data NATOPS and the tabular data Arrhythmia. For large enough K, NeuTraL AD is robust to the different parametrizations, since DCL ensures the learned transformations satisfy the semantic requirement and the diversity requirement. The performance of NeuTraL AD improves as the number k increases, and becomes stable when K is large enough. When $K \leq 4$, the performance has a larger variance, since the learned transformations are not guaranteed to be useful for anomaly detection without the guidance of any labels. When K is large enough, the learned transformations contain with high likelihood some transformations that are useful for anomaly detection. The transformation based methods (including NeuTraL AD) have roughly K times the memory requirement as other deep learning methods (e.g. Deep SVDD). However, the results in Figure 6 show that even with small K NeuTraL AD achieves competitive results.

5. Conclusion

We propose a self-supervised anomaly detection method with learnable transformations. The key ingredient is a novel training objective based on a deterministic contrastive loss, which encourages the learned transformations to produce diverse views that each share semantic information with the original sample, while being dissimilar. This unleashes the power of self-supervised anomaly detection to various data types including time series and tabular data. Our extensive empirical study finds, that on these data types, learning transformations and detecting outliers with NeuTraL AD improves over the state of the art.



(a) AUC on NATOPS (b) F1-score on Arrhythmia

Figure 6. The outlier detection accuracy in terms of AUC of NeuTraL AD on NATOPS and in terms of F1-score of NeuTraL AD on Arrhythmia increases as the number of transformations K increases, but stabilizes when a certain threshold is reached ($K >\approx 10$). With enough transformations, NeuTraL AD is robust to the parametrization of the transformations.

Acknowledgements

MK acknowledges support by the Carl-Zeiss Foundation, by the German Research Foundation (DFG) award KL 2698/2-1, and by the Federal Ministry of Science and Education (BMBF) awards 01IS18051A and 031B0770E.

SM was supported by the National Science Foundation (NSF) under Grants 1928718, 2003237 and 2007719, by Qualcomm, and by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0021. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA.

We acknowledge helpful discussions with Dennis Wagner and Luis Augusto Weber Mercado's contribution to Figure 1 and thank Sachin Goyal and Prateek Jain for being impressively responsive by email and for updating their experimental results.

References

Abati, D., Porrello, A., Calderara, S., and Cucchiara, R. Latent space autoregression for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 481–490, 2019.

Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P. Ganomaly: Semi-supervised anomaly detection via adver-

sarial training. In *Asian conference on computer vision*, pp. 622–637. Springer, 2018.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

Bagnall, A., Dau, H. A., Lines, J., Flynn, M., Large, J., Bostrom, A., Southam, P., and Keogh, E. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Bamler, R. and Mandt, S. Extreme classification via adversarial softmax approximation. In *International Conference on Learning Representations*, 2020.

Baragona, R. and Battaglia, F. Outliers detection in multivariate time series by independent component analysis. *Neural computation*, 19(7):1962–1984, 2007.

Bergman, L. and Hoshen, Y. Classification-based anomaly detection for general data. In *International Conference on Learning Representations*, 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X. and Konukoglu, E. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. In *MIDL Conference book*. MIDL, 2018.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., and Kloft, M. Image anomaly detection with generative adversarial networks. In *Joint european conference on machine learning and knowledge discovery in databases*, pp. 3–17. Springer, 2018.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., and Brox, T. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., and Leckie, C. High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition*, 58:121–134, 2016.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Golan, I. and El-Yaniv, R. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pp. 9758–9769, 2018.

Goyal, S., Raghunathan, A., Jain, M., Simhadri, H. V., and Jain, P. Drocc: Deep robust one-class classification. In *International Conference on Machine Learning*, pp. 3711–3721. PMLR, 2020.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Gutmann, M. U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(2), 2012.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pp. 15663–15674, 2019.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

Ho, D., Liang, E., Chen, X., Stoica, I., and Abbeel, P. Population based augmentation: Efficient learning of augmentation policy schedules. In *International Conference on Machine Learning*, pp. 2731–2741. PMLR, 2019.

Hyndman, R. J., Wang, E., and Laptev, N. Large-scale unusual time series detection. In *2015 IEEE international conference on data mining workshop (ICDMW)*, pp. 1616–1619. IEEE, 2015.

Kieu, T., Yang, B., and Jensen, C. S. Outlier detection for multidimensional time series using deep neural networks. In *2018 19th IEEE International Conference on Mobile Data Management (MDM)*, pp. 125–134. IEEE, 2018.

Kim, K. H., Shim, S., Lim, Y., Jeon, J., Choi, J., Kim, B., and Yoon, A. S. Rapp: Novelty detection with reconstruction along projection pathway. In *International Conference on Learning Representations*, 2019.

Lim, S., Kim, I., Kim, T., Kim, C., and Kim, S. Fast autoaugment. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Linsker, R. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008.

Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., and Shroff, G. Lstm-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*, 2016.

Misra, I., Zitnick, C. L., and Hebert, M. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pp. 527–544. Springer, 2016.

Mnih, A. and Kavukcuoglu, K. Learning word embeddings efficiently with noise-contrastive estimation. *Advances in neural information processing systems*, 26:2265–2273, 2013.

Munir, M., Siddiqui, S. A., Dengel, A., and Ahmed, S. Deepant: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, 7:1991–2005, 2018.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Perera, P., Nallapati, R., and Xiang, B. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2898–2906, 2019.

Principi, E., Vesperini, F., Squartini, S., and Piazza, F. Acoustic novelty detection with adversarial autoencoders. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 3324–3330. IEEE, 2017.

Ratner, A. J., Ehrenberg, H. R., Hussain, Z., Dunnmon, J., and Ré, C. Learning to compose domain-specific transformations for data augmentation. *Advances in neural information processing systems*, 30:3239, 2017.

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

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K.-R., and Kloft, M. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2019a.

Ruff, L., Zemlyanskiy, Y., Vandermeulen, R., Schnake, T., and Kloft, M. Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4061–4071, 2019b.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., and Müller, K.-R. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., and Bagnall, A. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, pp. 1–49, 2020.

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

Shen, L., Li, Z., and Kwok, J. Timeseries anomaly detection using temporal hierarchical one-class network. In *Advances in Neural Information Processing Systems*, 2020.

Sohn, K., Li, C.-L., Yoon, J., Jin, M., and Pfister, T. Learning and evaluating representations for deep one-class classification. In *International Conference on Learning Representations*, 2021.

Tack, J., Mo, S., Jeong, J., and Shin, J. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *34th Conference on Neural Information Processing Systems (NeurIPS) 2020*. Neural Information Processing Systems, 2020.

Tamkin, A., Wu, M., and Goodman, N. Viewmaker networks: Learning views for unsupervised representation learning. In *International Conference on Learning Representations*, 2021.

Tran, T., Pham, T., Carneiro, G., Palmer, L., and Reid, I. A bayesian data augmentation approach for learning deep models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 2794–2803, 2017.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2019.

Wang, J., Sun, S., and Yu, Y. Multivariate triangular quantile maps for novelty detection. In *Advances in Neural Information Processing Systems*, pp. 5060–5071, 2019a.

Wang, S., Zeng, Y., Liu, X., Zhu, E., Yin, J., Xu, C., and Kloft, M. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In *Advances in Neural Information Processing Systems*, pp. 5962–5975, 2019b.

Wong, E. and Kolter, J. Z. Learning perturbation sets for robust machine learning. *arXiv preprint arXiv:2007.08450*, 2020.

Zhang, R., Isola, P., and Efros, A. A. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1058–1067, 2017.

Zhang, X., Wang, Q., Zhang, J., and Zhong, Z. Adversarial autoaugment. In *International Conference on Learning Representations*, 2019.

Zhou, C. and Paffenroth, R. C. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 665–674, 2017.

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