# Self-Supervised Anomaly Detection via Neural Autoregressive Flows with Active Learning

**Jiaxin Zhang***
Oak Ridge National Laboratory
Oak Ridge, TN 37831

**Kyle Saleeby**
Oak Ridge National Laboratory
Oak Ridge, TN 37831

**Thomas Feldhausen**
Oak Ridge National Laboratory
Oak Ridge, TN 37831

**Sirui Bi**
Oak Ridge National Laboratory
Oak Ridge, TN 37831

**Alex Plotkowski**
Oak Ridge National Laboratory
Oak Ridge, TN 37831

**David Womble**
Oak Ridge National Laboratory
Oak Ridge, TN 37831

## Abstract

Many self-supervised methods have been proposed with the target of image anomaly detection. These methods often rely on the paradigm of data augmentation with predefined transformations such as flipping, cropping, and rotations. However, it is not straightforward to apply these techniques for non-image data, such as time series or tabular data, while the performance of the existing deep approaches has been under our expectation on tasks beyond images. In this work, we propose a novel active learning (AL) scheme that relied on neural autoregressive flows (NAF) for self-supervised anomaly detection, specifically on small-scale data. Unlike other generative models such as GANs or VAEs, flow-based models allow to explicitly learn the probability density and thus can assign accurate likelihoods to normal data which makes it usable to detect anomalies. The proposed NAF-AL method is achieved by efficiently generating random samples from latent space and transforming them into feature space along with likelihoods via invertible mapping. The samples with lower likelihoods are selected and further checked by outlier detection using Mahalanobis distance. The augmented samples incorporating with normal samples are used for training a better detector so as to approach decision boundaries. Compared with random transformations, NAF-AL can be interpreted as a likelihood-oriented data augmentation that is more efficient and robust. Extensive experiments show that our approach outperforms existing baselines on multiple time series and tabular datasets, and a real-world application in advanced manufacturing, with significant improvement on anomaly detection accuracy and robustness over the state-of-the-art.

## 1 Introduction

Anomaly detection, finding rare data that substantially differs from the majority of the data, is one of the essential problems in artificial intelligence. One typical anomaly detection setting is a one class classification, where the target is to detect samples as normal or anomalous. Many deep anomaly detection methods are recently proposed to solve one class classification tasks, specifically on image

*Corresponding author: zhangj@ornl.gov

NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications

benchmarks, with different scenarios, including supervised anomaly detection, unsupervised anomaly detection, and self-supervised anomaly detection [Ruff et al., 2021]. Here, we focus on the self-supervised setting where we have a training set of normal samples without anomalies and detect anomalies in the testing set which contains both normal and anomalous samples.

However, for data beyond images, such as tabular or time series data, we have several challenges in pursuing accurate and robust detection of anomalies. First, many recent advances in anomaly detection rely on data augmentation. Typical transformations, such as translation, rotation and reflection, are designed for images so that a strong detector is obtained based on the transformation predictions. Unfortunately, it is less well known which transformations are useful and hand-crafted transformation is not a straightforward task for non-image data [Bergman and Hoshen, 2020, Qiu et al., 2021]. Second, many tabular and time series data are from medical and healthcare. Small dataset size with sparse labels gives rise to unique difficulties which result that the anomaly detection performance is always under our expectation [Zong et al., 2018]. Third, although many deep anomaly detection methods show exceptional performance on large-scale image benchmarks, it is still a non-trivial task to handle small-scale tabular and time series data with high reliability and robustness [Pang et al., 2021]. This work aims at addressing these challenges in the scenario of self-supervised anomaly detection for data types beyond images. We develop a novel active learning scheme for effective data augmentation, which is a simple end-to-end procedure built upon a likelihood-based anomaly detection. The key idea is to leverage the advantages of neural autoregressive flows to assign likelihoods to normal data which enables to detect anomalies. Augmented samples with explicit likelihoods, drawn from the learned flow models can incorporate with original small data to improve the detector accuracy with high robustness.

Specifically, our proposed method consists of two core components: NAF anomaly detection framework (NAF-AD) and NAF-based active learning scheme (NAF-AL). Figure 1 visualizes the core idea behind our method. NAF-AD first performs data augmentation via random affine transformations and learns a feature space extracted by a neural network. The feature distribution of normal samples is captured by utilizing the latent space of a NAF model [Huang et al., 2018]. Unlike GANs or VAEs, flow-based models enable a bijective mapping between feature space and latent space in which each sample is assigned to a likelihood, which is used to derive a score function to decide if a sample is normal or anomalous. We propose NAF-AL by efficiently generating random samples from latent space and transforming them into feature space via bijective mapping. The samples with lower likelihoods are selected and further checked by outlier detection using Mahalanobis distance. The left effective samples are merged into normal data to approach decision boundaries for better detection. Compared with random transformations, NAF-AL can be interpreted as a likelihood-oriented data augmentation that is more active and efficient. As a result, we achieve superior performance in deep anomaly detection beyond images, specifically on small-scale tabular and time series data, with significant improvement on anomaly detection accuracy and robustness over the state-of-the-art.

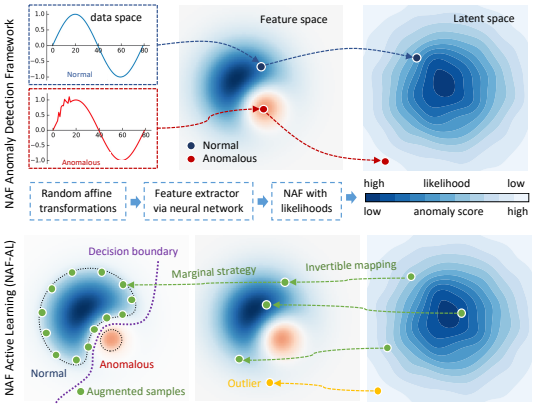


Figure 1: NAF-AL first transforms the data into multiple subspaces and learns a feature space by a neural network. Then we build an accurate density estimation via NAF and assign a higher likelihood to normal (a lower likelihood to anomaly) in latent space. NAF-AL is free to draw samples and transform them to feature space with explicit likelihoods via invertible mapping. Using a marginal strategy, these samples are partially selected to approach decision boundary by incorporating with normal data for improving the detector during training. This allows for more effective data augmentation.

## 2 Related Work

**Deep Anomaly Detection.** Many recent advances have been proposed to use deep learning for anomaly detection. Ruff et al. [2021], Pang et al. [2021] provided a thorough survey and review on the

recent development of deep anomaly detection approaches. Related work on deep anomaly detection include one class classification [Ruff et al., 2018, Liznerski et al., 2020, Ruff et al., 2019], outlier exposure [Hendrycks et al., 2019a, Goyal et al., 2020], and out-of-distribution (OOD) detection [Ren et al., Hendrycks and Gimpel, 2017, Kirichenko et al., 2020].

There has been an increasingly growing interest in self-supervised scenarios since this supervision is easy to obtain in practical settings and also shows promising accuracy in detecting anomalies [Pang et al., 2019, Hendrycks et al., 2019b, Sohn et al., 2020, Tack et al., 2020, Li et al., 2021, Sehwag et al., 2020]. Self-supervised methods solve one or more classification-based auxiliary tasks (e.g., data transformations [Golan and El-Yaniv, 2018, Wang et al., 2019]), using normal data for training and the learned classifier is useful to detect anomalies. Bergman and Hoshen [2020] extended the work from Golan and El-Yaniv [2018], Wang et al. [2019] to investigate self-supervised anomaly detection for general data. This approach is established based on the open-set setting with affine transformations for tabular datasets. Qiu et al. [2021] followed a similar scope for anomaly detection but with learnable transformations, and demonstrated a higher performance.

**Likelihood (Density)-based Anomaly Detection.** Differing from the classification-based methods [Ruff et al., 2018, 2019, Bergman and Hoshen, 2020], likelihood-based methods offer a probabilistic view for anomaly detection. In this scenario, a flow-based model, learning a bijective mapping between data distributions and latent distributions via invertible neural networks, is an ideal candidate because it has significant advantages in explicit likelihood calculation and efficient sample generation [Kobyzev et al., 2020, Papamakarios et al., 2021]. Much recent effort has been made to improve model expressivity and computational efficiency that allow more accurate likelihood calculation and enable faster sampling [Rezende and Mohamed, 2015, Dinh et al., 2017, Kingma et al., 2016, Papamakarios et al., 2017, Kingma and Dhariwal, 2018, Grathwohl et al., 2018, Ho et al., 2019]. Huang et al. [2018] proposed a neural autoregressive flow (NAF) which is a universal approximator for density functions, and addresses the challenges in inverse AFs [Kingma et al., 2016].

Although the properties of normalizing flows are promising, flow-based models for anomaly detection have not raised much attention yet, although some works presented promising results using RealNVP [Rudolph et al., 2021], residual flows [Zisselman and Tamar, 2020] and conditional normalizing flows [Gudovskiy et al., 2021]. However, all of these works deal with large-scale image datasets. It remains an open, and sometimes contentious, debate as to why and why not flow-based models can guarantee anomaly detection [Kirichenko et al., 2020, Schirrmeister et al., 2020, Lan and Dinh, 2020].

**Active Learning for Anomaly Detection.** Deep anomaly detection tends to require immense amounts of computational and human resources for training and labeling. The design of effective training methods that require small labeled training sets is a fundamental research challenge [Tran et al., 2019]. To address this issue, two are particularly interesting: *data augmentation*, which artificially generates new samples for training, while *active learning* selects the most informative subset of unlabeled samples to be labeled. Although successful in image data, data augmentation does not utilize computational resources since the generated samples are not guaranteed to be informative [Shorten and Khoshgoftaar, 2019]. Active learning deals with this limitation through an iterative selection of small subsets while assessing how informative those subsets are for the training process. Recent advances on active learning rely on the incorporation of Bayesian approach [Gal et al., 2017, Tran et al., 2019] and deep generative models [Sinha et al., 2019, Liu et al., 2019].

Active learning strategies for anomaly detection [Stokes et al., 2008, Görnitz et al., 2009, Pelleg and Moore, 2004] which identify informative instances for labeling, have primarily only been explored for shallow detectors and could be extended to deep learning approaches [Pimentel et al., 2020, Trittenbach and Böhm, 2019]. Our goal is to integrate a likelihood-based detector with active learning, which leads to a more effective data augmentation scheme for designing anomaly detection that continuously improves via likelihood feedback loops, see Figure 1. This idea has not yet been explored for deep self-supervised anomaly detection.

## 3 NAF-AL Method

We develop a novel approach that relies on neural autoregressive flows with active learning (NAF-AL), which is designed for self-supervised anomaly detection beyond images.

### 3.1 Data Transformations in Self-Supervised Setting

Our method is built upon the self-supervised scenario which can be typically defined by

**Definition 1** *Assume all data $\mathcal{X}$ lies in space $\mathcal{S}_d$, where $d$ is the data dimension. Normal data $X$ lie in subspace $X \subset \mathcal{S}_d$ but all anomalies $X^*$ lie outside $X$. The task of self-supervised anomaly detection is to build a classifier $\mathcal{C}$ based on completely normal data, such that $\mathcal{C}(x) = 1$ if $x \in X$ and $\mathcal{C}(x) = 0$ if $x \in \mathcal{S}_d \setminus X$.*

In the self-supervised setting, data transformations $\mathcal{T} = \{T_1, ..., T_k | T_k : \mathcal{X} \to \mathcal{X}\}$ (e.g., translation, rotation and reflection), are often used to generate $K$ different views, which leads to a strong anomaly detector based on the transformation prediction or representations learned using these views. However, these transformations are not applicable to non-image data. To this end, we generalize the set of transformations to random affine transformations:

$$T_k(x) = \mathcal{A}_k(x) + b_k, \quad \mathcal{A}_k \sim \mathcal{N}(0, \boldsymbol{I}_d) \tag{1}$$

where $\mathcal{A}_k$ and $b_k$ are affine matrix and coefficient respectively, defined by random Gaussian distributions. The random affine transformation is a more general class that works for general data type with an unlimited number of transformations.

Since only normal data are used for training, we first transform the normal data $X$ into $K$ subspaces $X_1, ..., X_k$, and then learn a feature extractor $f_\theta(x)$ using a neural network parametrized by $\theta$, which maps the original normal data space $\mathcal{X}$ into a feature representation space $\tilde{\mathcal{X}}$. The probability of data point $x$ after transformation $k$ is denoted by $p(T_k(x) \in X_k)$. By assuming independence between different transformations $T_k$, the probability that $x$ is normal $p(x \in X)$ is the product of the probabilities that all transformed samples are in their respective subspace.

$$\mathcal{P}(x) = \log p(x \in X) = \sum_{k=1}^{K} \log p(T_k(x) \in X_k) \tag{2}$$

where $\mathcal{P}(x)$ computes the degree of anomaly of each data. Lower probabilities (likelihoods) indicate a more anomalous data. We will introduce how to explicitly calculate these probabilities (likelihoods) using flow-based models below.

### 3.2 Learning Likelihood by Autoregressive Flows

Normalizing flows (NF) are a flexible class of generative models that map a target distribution $p_X(x)$ into a base distribution in the latent space $p_Z(z)$ via an invertible transformation $f_\psi : \mathcal{Z} \to \mathcal{X}$ where $f_\psi$ is an invertible neural network parametrized by $\psi$. Based on the change of variable theorem, the likelihood for an input $x$ is

$$p_X(x) = p_Z(f_\psi^{-1}(x)) \left| \det \frac{\partial f_\psi^{-1}}{\partial x} \right|. \tag{3}$$

Flow-based models are typically trained by minimizing the negative log-likelihood of the training data $\mathcal{D}$ with respect to the parameters $\psi$ of the invertible transformation $f_\psi$.

$$\psi^* = \arg\min_\psi \{-\log p(\mathcal{D})\} = \arg\min_\psi \{-\log \prod_{x \in \mathcal{D}} p_X(x)\}.$$

Much effort in NFs focuses on designing expressive transformations while retaining efficient computing the determinant of the Jacobian $|\det \boldsymbol{J}|$. In particular, autoregressive flows (AFs) decompose a joint distribution $p_X(x)$ into a product of $m$ univartiate conditional densities:

$$p_X(x) = p_{X_1}(x_1) \prod_{i=2}^{m} p_{X_i | X_{<i}}(x_i | x_{<i}) \tag{4}$$

where each univariate density is parametrized by a NF. In particular, the transformation $f_\psi^{-1,(i)}$ can be decomposed via invertible transformers $t_\psi^{(i)}$ and conditioners $c_\psi^{(i)}$:

$$z_i = f_\psi^{-1,(i)}(x_{\leq i}) = t_\psi^{(i)}(x_i, c_\psi^{(i)}(x_{<i})). \tag{5}$$

The resulting flows have a lower triangular Jacobian and the invertibility of the flows as a whole depends on each $t_\psi^{(i)}$ being an invertible function of $x_i$ and each $c^{(i)}$ is an unrestricted function. RealNVP [Dinh et al., 2017] model each $t_\psi^{(i)}$ by using an affine transformation whose parameters are predicted by $c^{(i)}$. However, these models require complex conditioners and a composition of multiple flows due to their simplicity which leads to a limitation on expressiveness of $f_\psi$. Neural autoregressive flow (NAF) [Huang et al., 2018] was proposed by learning a complex bijection using a neural network monotonic in $x_i$. NAF is a universal approximator for explicitly learning likelihood with greater expressivity that allows it to better capture multimodal target distributions. The NAF architecture is illustrated by Figure 2.

We propose to utilize NAF to learn the distribution of feature space and train NAF using a maximum likelihood objective, which is equivalent to minimizing loss defined by

$$\mathcal{L}_{(\theta,\psi)}(\mathcal{D}_n) = -\frac{1}{|\mathcal{D}_n|}\sum_{x\in\mathcal{D}_n}\log p_X(x) \approx \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\|z\|_2^2}{2} - \log|\det \boldsymbol{J}_i|\right] + \text{const.} \tag{6}$$

where $n$ is the size of training data $\mathcal{D}$. During training, $\mathcal{L}(\mathcal{D}_n)$ is optimized for feature space $x$ of different transformations $\mathcal{T}$ of an input data. After training, the learned NAF model can be used to evaluate the log-likelihood of the testing dataset $\mathcal{D}_t$ that contains normal and anomalous samples.

We use the calculated likelihoods as a criterion to classify a sample as normal or anomalous. To pursue a robust anomaly score function $\mathcal{S}(x)$, we concatenate all the variable $z$ from multiple transformations $T_k(x) \in \mathcal{T}$ and average the negative log-likelihood as

$$\mathcal{S}(x) = -\mathbb{E}_{T_k\in\mathcal{T}}[\log p_Z(f_\psi^{-1}(\Upsilon))], \tag{7}$$

where $\Upsilon = f_\theta(T_k(x))$ represents the feature space extracted by a neural network $f_\theta$. If the anomaly score $\mathcal{S}(x)$ is above the threshold value $\xi$, the sample is identified as anomalous, otherwise as normal, which is given by

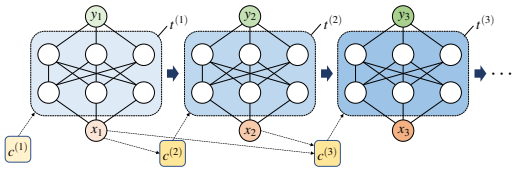


Figure 2: NAF architecture: each $c^{(i)}$ (a neural network) predicts pseudo-parameters for $t^{(i)}$, which in turns processes $x_i$.

$$\mathcal{Q}(x) = \begin{cases} 1 & \text{if } \mathcal{S}(x) \geq \xi \\ 0 & \text{if } \mathcal{S}(x) < \xi \end{cases}, \tag{8}$$

where $\mathcal{Q}(x) = 1$ indicates an anomaly. The threshold value $\xi$ is varied to calculate the AUC and F1 score in experiments. Figure 3 shows empirical evidence for NAF-AD method. While the histogram of anomaly scores (computed using Eq. (7)) is similar for inliers and anomalies before training, this changes drastically after training, and held-out inliers and anomalies become easily distinguishable.

### 3.3 Active Learning with Marginal Strategy

Formally, active learning is used to automatically select the most informative subset of unlabeled training samples and label them by an oracle. One of the key enabling techniques is uncertainty sampling, which uses one classifier to identify unlabeled samples with least confidence [Zhu et al., 2009]. However, active learning for self-supervised anomaly detection is different from the original scope. To this end, we devise a novel active learning scheme to query low-confidence decisions, hence guiding the detector with augmented normal samples in the training process. Such a marginal strategy can be expressed by adding the sample $x^*$ that is close to the likelihood-based decision boundary:

$$x^* = \underset{x\in\{x_1,\ldots,x_n\}}{\arg\min}\frac{|\mathcal{L}(x)|}{\Omega}, \quad \Omega = \max_i|\mathcal{L}(x_i)|. \tag{9}$$

This is achieved by generating samples $x_j^{aug}$ from the learned NAF model, calculating the log-likelihood $\mathcal{L}(x_j^{aug})$ and retaining the samples with lower likelihoods if $\mathcal{L}(x_j^{aug}) < \mathcal{Q}_\alpha(\mathcal{L}(x_i))$ where $\mathcal{Q}_\alpha$ is referred to as the $\alpha$-quantile ($\alpha = 0.9 \sim 0.95$ is a hyperparameter) and $\mathcal{L}(x_i)$ is the log-likelihood of current training samples. It is critical to design an appropriate likelihood level set trade-off between aggressive boundary and conservative boundary, see Figure 6 for more discussion.

The sample with lower likelihoods are desired but it is probably an outlier (anomalous sample) if the likelihood is too small. Thus we have an additional step to check the sample by outlier detection, which is done by Mahalanobis distance $M(x)$ [Lee et al., 2018, Ren et al., 2021]

$$M(x)^2 = (x-\mu)^T \Sigma^{-1} (x-\mu). \tag{10}$$

We define a threshold $\delta_M$ that is Mahalanobias distance at $\chi^2_{0.05}$. The samples with $M(x) > \delta_M$ are rejected. We eventually determine the augmented samples $x^*$ by solving the optimization problem in Eq. (9) subject to two constrains:

$$\mathcal{L}(x^*) < \mathcal{Q}_\alpha(\mathcal{L}(x_i)), \quad M(x^*) < \delta_M. \tag{11}$$

These samples $x^*$ are selected as normal samples and added to update the training process. Appropriate stopping criterion for active learning is a trade-off issue between training cost and effectiveness of the detector. We set up a criterion to stop if the number of training iteration increase to five due to time cost limitation or the augmented samples size reaches 50% of the original training size given the concern of sample efficiency. In summary, we propose an active data augmentation by leveraging these synthesized normal samples based on meaningful likelihoods, without changing the original training, testing, and proportion of anomalies. Note that the augmented samples generated from the NAF are located on the feature space rather than the original data space. The details of the NAF-AL are provided in Algorithm 1.

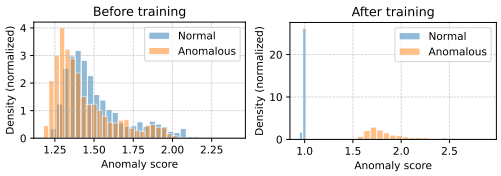


Figure 3: Histogram of anomaly score before training and after training. The data come from the Cardio. experiments in Table 5.

**Algorithm 1** The NAF-AL algorithm

**Require**: training and testing datasets $\mathcal{D}_n, \mathcal{D}_t$, number of transformations $K$, feature extractor $f_\theta$, NAF model $f_\psi$, hyperparameters $\alpha, \delta_m$ and stopping criterion in active learning
**while** Active learning stopping criterion **do**
  // `Training process`
  Transform each training sample according to Eq. (1): $T_1(x_i), T_2(x_i), ..., T_k(x_i) \leftarrow x_i,\ i = 1, ..., n$
  Extract feature representation $\tilde{T}_k(x_i) \leftarrow T_k(x_i),\ k = 1, ..., K$ via a neural network model $f_\theta$
  Concatenate all different affine transformations $\tilde{T}_k(x_i)$
  Evaluate the NAF model $f_\psi$ for $z$ and $|\det \boldsymbol{J}|$
  Minimize the loss $\mathcal{L}$ in Eq. (6) to update $\theta, \phi$
  // `Testing process`
  Transform testing sample by all transformations 1 to $K$: $T_1(x_t), ..., T_k(x_t) \leftarrow x_t,\ t = 1, ..., n_t$
  Calculate the log likelihood $p_Z \leftarrow f_\psi^{-1}(f_\theta(T_k(x_t)))$
  Concatenate transformations and average the log-likelihoods to compute anomaly score $\mathcal{S}(x)$ in Eq. (7)
  // `Active learning process`
  Draw samples $x_j^{aug}$ from the learned NAF model $f_\psi$
  Evaluate the log-likelihood of new augmented samples $\mathcal{L}(x_j^{aug})$ and original training samples $\mathcal{L}(x_i)$
  Retain the samples $x_j^{aug}$ with lower likelihoods if $\mathcal{L}(x_j^{aug}) < \mathcal{Q}_\alpha(\mathcal{L}(x_i))$ according to Eq. (9)
  Check the outlier samples of $x_j^{aug}$ if the Mahalanobis distance $M(x_j^{aug}) < \delta_M$ based on Eq. (11)
  Add the new samples $x_j^{aug}$ with normal labels to the current training dataset, then iteratively update the training process
**end while**
**return** Anomaly score $\mathcal{S}(x)$

## 4 Experiments

We provide several experiments to demonstrate the effectiveness of our NAF-AL approach on deep anomaly detection beyond images. Most image datasets are large-scale and have existing strong baselines so we do not expect significant improvement via our NAF-AL method. Instead, our focus is on *small-scale* tabular data and time series data.

### 4.1 Anomaly Detection Baseline Methods

We compare our NAF-AL with a couple of anomaly detection methods, including:

- *Shallow AD baselines*: Isolation Forest (IForest) [Liu et al., 2008] uses a tree-based model to isolate anomalies. Local Outlier Factor (LOF) [Breunig et al., 2000] utilizes density estimation with $k$-nearest neighbors. One-Class SVM (OC-SVM) [Schölkopf et al., 1999] is a kernel-based approach for one-class classification.
- *Deep AD baselines*: Deep Autoencoding Gaussian Mixture Model (DAGMM) [Zong et al., 2018] uses latent space to estimate density. Deep Support Vector Data Description (DSVDD) [Ruff et al., 2018] is a distance-based method with one-class SVM in the feature space. Feature Bagging Autoencoder (FB-AE) [Chen et al., 2017] is an ensemble method with autoencoders as the base classifier. GOAD [Bergman and Hoshen, 2020] is a self-supervised classification-based method. Neural Transformation Learning for Anomaly Detection (NeuTral) [Qiu et al., 2021], is a self-supervised method with learned transformations. For time series data, we also include LSTM-ED [Malhotra et al., 2016] which is an encoder-decoder model to detect anomalies based on reconstruction error.

### 4.2 Tabular Data Experiments

Tabular data plays a crucial role in anomaly detection applications since many medical, health, and cybersecurity data come in this format. However, many important areas, e.g., medical, only have small-scale data because the data collection is time-consuming, while labeling relied on expert opinion is expensive.

**Datasets.** We focus on six tabular datasets, including four small-scale medical datasets, Arrhythmia, Cardiotocograph, Lymphography and Thyroid from the Outlier Detection Datasets (ODDs) repository[2], and two cybersecurity datasets, KDD and KDDRev from the empirical studies of Zong et al. [2018], Bergman and Hoshen [2020], Qiu et al. [2021] which are used to show our potential to large-scale datasets. Table 4 shows the key statistics (data size, dimension, and anomaly ratio) of the tabular datasets, and all relevant details of the datasets can be found in the Appendix. Following the setting of Zong et al. [2018], we train all models on 50% of the normal data and evaluate the performance on testing data containing the rest of normal data as well as all the anomalies.

**Implementation Details.** We use a standard normal distribution to generate random affine transformation matrices for each case. Similar as the setting in Bergman and Hoshen [2020], we use 256 transformations for small-scale medical datasets and 64 for large-scale datasets (KDD and KDDRev). For the feature extractor, we used fully-connected hidden layers (1 layer with 8 hidden nodes for the small-scale datasets and 5 layers with 128 hidden nodes for large-scale datasets) with leaky-ReLU activations, as well as one 1d convolutional layer on the top. The NAF model consists of 4 flow blocks with 2 layers (128 hidden units) for small data and 8 flow blocks with 3 layers (1024 hidden units). We optimized the network parameters using Adam with a learning rate of 0.001.

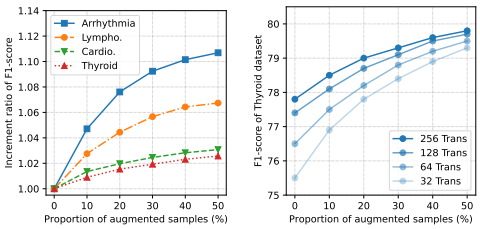


Figure 4: Left: F1 score increment ratio (NAF-AL/NAF-AD) as a function of the proportion of augmented samples from active learning. Right: the effect of varying number of transformations on the F1-score with a specific proportion of augmented samples.

Considering the training cost limit, we set up a stop criterion by using a maximum number of iteration ($N_{\max} = 5$) for active learning. For each iteration, we draw a branch of samples where the sample size equals the training sample size, then rank the samples based on their likelihoods and finally reject the larger 90% samples (retain 10% samples with lower likelihoods near the decision boundaries). These samples are further checked by Mahalanobis distance criterion if it is an outlier. After that, we combine these augmented samples with the existing normal samples to update training.

[2] http://odds.cs.stonybrook.edu/

Table 1: F1-score (%) for anomaly detection on tabular datasets.

| | Arrhythmia | Cardio. | Lympho. | Thyroid | KDD | KDDRev |
|---|---|---|---|---|---|---|
| IForest | 57.4 | 79.5 | 60.4 | 46.9 | 90.7 | 90.6 |
| LOF | 50.0 | 75.3 | 62.9 | 52.7 | 83.8 | 81.6 |
| OC-SVM | 45.8 | 72.6 | 58.7 | 38.9 | 79.5 | 83.2 |
| DAGMM | 49.8 | 74.4 | 61. | 47.8 | 93.7 | 93.8 |
| DSVDD | 53.9±3.1 | 80.1±1.9 | 64.1±1.9 | 70.8±1.8 | 99.0±0.1 | 98.6±0.2 |
| FB-AE | 51.5±1.6 | 78.9±1.1 | 66.2±1.2 | 75.0±0.8 | 92.7±0.3 | 95.9±0.4 |
| GOAD | 52.0±2.3 | 79.7±1.5 | 66.8±1.4 | 74.5±1.1 | 98.4±0.2 | 98.9±0.3 |
| NeuTraL | 60.3±1.1 | - | - | 76.8±1.9 | **99.3**±0.1 | **99.1**±0.3 |
| NAF-AD | 55.2±1.1 | 81.3±1.1 | 67.3±1.2 | 77.8±1.1 | 97.9±0.2 | 98.2±0.2 |
| NAF-AL | **61.1**±0.9 | **84.0**±1.0 | **71.3**±0.9 | **79.8**±0.8 | 98.5±0.1 | **99.0**±0.2 |

The implementation details of the baseline methods are replicated from the existing studies [Zong et al., 2018, Bergman and Hoshen, 2020], as we report their results with mean and standard deviation (if they provide). We also implement these baselines for two additional small-scale datasets (Cardio. and Lympho.) using their official code (if they have, otherwise keep the relevant cell blank).

**Results.** The results of NAF-AL in comparison to all baseline methods on tabular data are shown in Table 1. We follow the configuration of previous work [Zong et al., 2018, Bergman and Hoshen, 2020, Qiu et al., 2021] to report results in terms of F1-scores.

- *Small-scale datasets:* all medical datasets are small with a low anomaly to normal ratio. Our NAF-AD performs reasonably well as similar to most baselines. Our NAF-AL outperforms all baselines on these small-scale datasets thanks to the benefits from active learning. Compared with GOAD which is a classification-based method, our probabilistic flow-based model is competitive even without the help of active learning. NeuTral beats our NAF-AD in the Arrhythmia dataset but underperforms our NAF-AL method. Since our NAF-AL is flexible to incorporate any number of transformations such that our robustness (with a smaller variance) is better than NeuTral.
- *Large-scale datasets:* The deep baselines show superior performance than the shallow methods in this case. NAF-AD is slightly lower than NeuTral, GOAD, and DSVDD but NAF-AL is still competitive. One explanation is that the performance improvement in such a large dataset is not significant as the small-scale case discussed above. The large datasets, having different dynamics from very small datasets found by Bergman and Hoshen [2020], are probably not well-suited to the probabilistic methods.

### 4.3 Time Series Data Experiments

Differing from novelty detection within time series (point or group anomalies), we aim to detect abnormal time series on a *whole time sequence*. In other words, the whole time series data is labeled by normal or anomalies. This scenario is also important in practice. For example, we identify abnormal facility operations by detecting abnormal sensor measurements over the whole time-series signals in scientific applications. Anomalies in medical, health, and sport monitoring may indicate injury, disease, or more serious issues.

**Datasets.** We focus on five multivariate time series datasets from the UEA multivariate time series classification archive [3] which has been widely used for anomaly detection tasks [Zhang et al., 2020, Ruiz et al., 2021, Jiao et al., 2020, Zerveas et al., 2021, Qiu et al., 2021]. The datasets include two relatively large cases, Character Trajectories (CT) and Spoken Arabic Digits (SAD), and three small-scale cases, Epilepsy (EPSY), NATOPS, and Racket Sports (RS). Table 5 provides the necessary information, e.g., data size, dimension, data length, and the number of classes, and more details are offered in the Appendix.

**Evaluation Protocol.** We evaluate NAF on these benchmark datasets based on two protocols:

[3] http://www.timeseriesclassification.com/ and more details can be found in Bagnall et al. [2018].

- *one-vs-rest:* The goal is to create $N$ one class classification tasks by splitting the dataset by $N$ class labels. The anomaly detection models are trained on data from one class and tested on data from the rest of classes. The class used for training is labeled as normal data, while the other classes are labeled as anomalies.
- *m-vs-rest:* This protocol is more challenging because multiple classes $m$ ($1 < m < N$) are labeled as normal and the rest of classes are treated as anomalies. In this case, the normal data is no longer from one class such that the variability increases significantly.

Table 2: Mean and standard deviation of AUC for one-vs-rest tasks

| | CT | EPSY | NATOPS | RS | SAD |
|---|---|---|---|---|---|
| IForest | 94.3 | 67.7 | 85.4 | 69.3 | 88.2 |
| LOF | 97.8 | 56.1 | 89.2 | 57.4 | 98.3 |
| OC-SVM | 97.4 | 61.1 | 86.0 | 70.0 | 95.3 |
| DAGMM | 89.8±0.7 | 72.2±1.6 | 78.9±3.2 | 51.0±4.2 | 80.9±1.2 |
| DSVDD | 95.7±0.5 | 57.6±0.7 | 88.6±0.8 | 77.4±0.7 | 86.0±0.1 |
| FB-AE | 96.3±0.3 | 80.1±0.4 | 89.9±1.2 | 78.0±0.7 | 93.9±0.1 |
| GOAD | 97.7±0.1 | 76.7±0.4 | 87.1±1.1 | 79.9±0.6 | 94.7±0.1 |
| LSTM-ED | 79.0±1.1 | 82.6±1.7 | 91.5±0.3 | 65.4±2.1 | 93.1±0.5 |
| NeuTraL | 99.3±0.1 | 92.6±1.7 | 94.5±0.8 | 86.5±0.6 | **98.9**±0.1 |
| NAF-AD | 97.8±0.1 | 90.4±0.5 | 91.9±0.5 | 85.4±0.6 | 97.8±0.1 |
| NAF-AL | **99.5**±0.1 | **93.2**±0.3 | **95.7**±0.3 | **88.2**±0.5 | 98.4±0.1 |

**Implementation Details.** The implementation details of most baselines are replicated from Qiu et al. [2021] and we implemented the FB-AE method using their official code. For the time series datasets in Table 5, the first four are small-scale while the SAD dataset is slightly large. We therefore use a similar setting in small-scale tabular datasets for the time series AD tasks. For the $m$-vs-rest tasks, we use the same setting $m = N - 1$, which makes the task more challenging.

**Results.** Table 2 shows the results of NAF-AD and NAF-AL in comparison to the shallow and deep AD baselines on multiple time series experiments summarized in Table 5. NAF-AL outperforms all baselines in CT, EPSY, NATOPS, and RS experiments. In most cases, the performance from NAF-AD is already competitive and further improved by augmented samples from active learning. Only on the SAD dataset, our NAF-AL is outperformed by NeuTral with learned transformations which have an advantage over the random transformations, while our active learning improvement looks marginal because its dataset size is larger than the other experiments. Our NAF-AL shows a superior performance close to LOF but still better than the other deep baselines, like GOAD and FB-AE. The shallow baselines perform worse on the small-scale datasets, like EPSY, NATOPS, and RS, but show better on CT and SAD. Our NAF-AL can well handle both scenarios although it is designed for addressing the specific challenges from small data.

The results of the $m$-vs-rest tasks are shown in Table 3. In this case, NAF-AL outperforms all baselines on EPSY, NATOPS, and RS experiments. LOF performs best on CT and SAD and is also competitive in one-vs-rest tasks in Table 2. It is interesting to see this KNN-based method that outperforms all deep baselines. Compared with the deep baseline, our NAF-AL shows superior performance on 4 out of 5 experiments. On SAD, NAF-AL is only slightly lower than NeuTral but still very competitive. Although the $m$-vs-rest is more challenging, the results are consistent with the performance under one-vs-rest tasks in Table 2.

### 4.4 A real-world application to advanced manufacturing

Our NAF-AL method is naturally generalized to solve anomaly detection problems on general data. In this section, we demonstrate our capability to study self-supervised anomaly detection on images. This problem is a challenging real-world application in advanced manufacturing where only sparse labeled data is provided because the labeling process is time-consuming and needs expert's help.

**Datasets.** We collected data from an Okuma MU-8000V Laser EX hybrid manufacturing system using a coaxial meltpool camera, as shown in Figure 5. This camera is used to monitor the shape and size of the meltpool during the directed energy deposition process. A total number of 7913 meltpool images are collected but we only have 104 labeled normal data. There are three subsets: Cube T4,

Table 3: Mean and standard deviation of AUC for $m$-vs-rest tasks

| | CT | EPSY | NATOPS | RS | SAD |
|---|---|---|---|---|---|
| IForest | 57.9 | 55.3 | 56.0 | 58.4 | 56.9 |
| LOF | **90.3** | 54.7 | 71.2 | 59.4 | **93.1** |
| OC-SVM | 57.8 | 50.2 | 57.6 | 55.9 | 60.2 |
| DAGMM | 47.5±2.5 | 52.0±1.0 | 53.2±0.8 | 47.8±3.5 | 49.3±0.8 |
| DSVDD | 54.4±0.7 | 52.9±1.4 | 59.2±0.8 | 62.2±2.1 | 59.7±0.5 |
| FB-AE | 77.2±0.3 | 63.0±1.2 | 60.8±0.9 | 65.3±1.1 | 70.8±1.3 |
| GOAD | 81.1±0.1 | 62.7±0.9 | 61.5±0.7 | 68.2±0.9 | 70.5±1.4 |
| LSTM-ED | 50.9±1.2 | 56.8±2.1 | 56.9±0.7 | 63.1±0.6 | 58.9±0.5 |
| NeuTraL | 87.0±0.2 | 80.5±1.0 | 74.8±0.9 | 80.0±0.4 | 85.1±0.3 |
| NAF-AD | 86.7±0.2 | 77.3±0.8 | 71.3±0.6 | 78.9±0.4 | 83.0±0.4 |
| NAF-AL | 89.3±0.2 | **81.7**±0.7 | **75.8**±0.5 | **82.7**±0.3 | 83.9±0.3 |

T5, and T6, and we have 2450 total with 31 normal data for T4, 2323 total with 40 normal data for T5, and 3140 total with 33 normal data for T6. The image resolution is 576×704. The normal data is labeled by domain experts from the Manufacturing Demonstration Facility at Oak Ridge National Laboratory. This manufacturing data can be used as a benchmark dataset for future testing and comparison in AI/ML community.

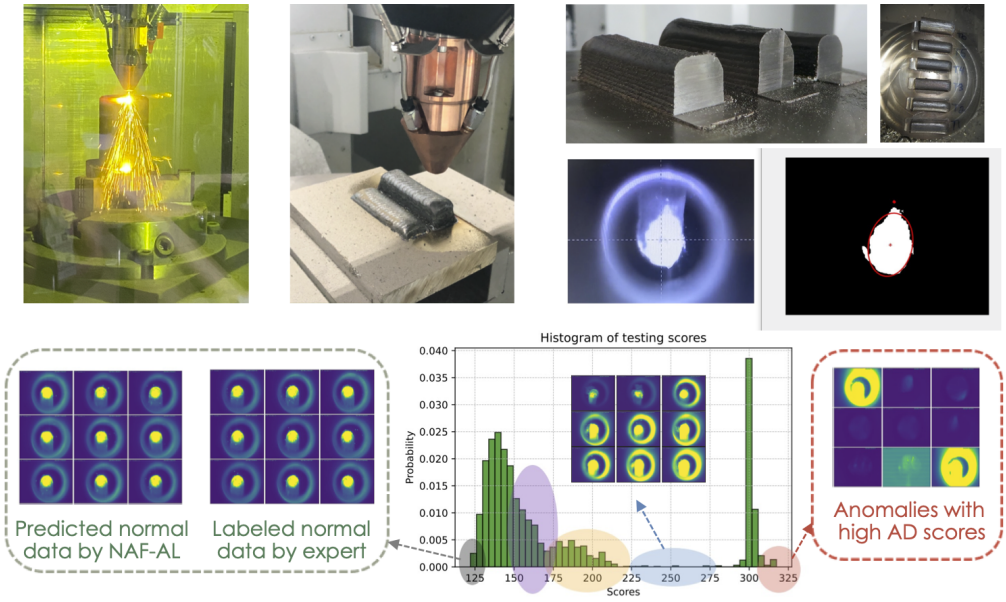


Figure 5: A real-world demonstration in advanced manufacturing. Our objective is to determine correct processing parameters for deposition of stainless steel. We utilized the coaxial laser camera to capture meltpool images to detect normal and anomalies via our proposed NAF-AL method.

**Results.** Our proposed NAF-AL method is well-suited to such problems with sparse labels. We use the limited normal data for training and sequentially identify more normal data with higher likelihoods via an active learning scheme. Using this way, our prediction will be gradually improved as we update our training models with more normal data. After five active learning iterations, the histogram of anomaly scores is shown in Figure 5. From the distribution, we can easily identify normal and anomalous data since our training model has clearly separated the normal and anomalies. The predicted normal data with lower scores show a very consistent pattern with the data identified by experts and the anomalies with higher scores show a substantially different pattern from the normal data. Our method also enables automatically labeling based on limited initial sparse labels, which will significantly improve the working efficiency in anomaly detection for advanced manufacturing.

## 5 Conclusion

We propose a likelihood-based active learning method for self-supervised anomaly detection on small data beyond images. The key contribution is to develop a new active learning strategy that benefits from efficient sampling and explicit likelihoods from neural autoregressive flows. We demonstrate the novel method on several tabular, time series benchmarks and real-world application in advanced manufacturing with superior performance over the state of the art. We plan to generalize NAF-AL to further improve the anomaly detection performance on general large-scale datasets.


## Acknowledgements

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR), Applied Mathematics program; and by the Artificial Intelligence Initiative at the Oak Ridge National Laboratory (ORNL). This work used resources of the Oak Ridge Leadership Computing Facility, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.


## A Discussion

***How does the active learning work with NAF?*** Figure 6 shows the data augmentation scheme via active learning in NAF. Ideally, we expect the augmented samples are near the decision boundary as close as possible. However, the decision boundary is typically unknown and difficult to determine. Instead, we pursue an ideal level set of the data likelihood (black dash line), which enables us to detect normal and anomalies. For each iteration in active learning, an *aggressive* strategy is to only retain the samples with very low likelihoods but this way will expand the likelihood boundary (red dash line) across the decision boundary. Under this way, the likelihood of anomalous data (red dots) in testing may lie at the same level set as the normal data (green dots), which confuses the detector (anomalous $\rightarrow$ normal) and hurt the detection performance. On the contrary, one can choose a *conservative* strategy by rejecting the samples with relatively low likelihoods but this way will shrink the likelihood boundary which is far away from the decision boundary. The potential issue is that the likelihood of normal samples in testing tends to be smaller and these samples are probably labeled as anomalous data (normal $\rightarrow$ anomalous). To deal with this trade-off issue, We propose a marginal strategy that sequentially augments samples with a small proportion in each iteration and adaptively pushes them to the boundary, while controlling the likelihood level set by detecting outliers with Mahalanobis distance. This scheme effectively avoids active learning from being too aggressive or too conservative.

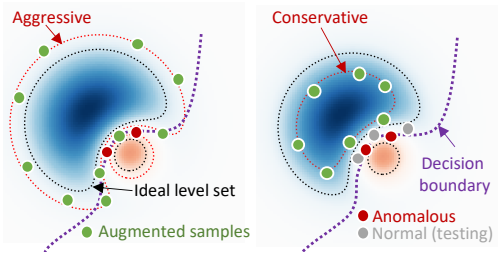


Figure 6: Sample (for training) augmentation via active learning in NAF: aggressive strategy (left) vs conservative strategy (right).

***How does NAF-AL improve the AD performance?*** Figure 4 shows the improvement of F1-score with respect to the proportion of augmented samples ($\lambda = N_{\text{aug}}/N_{\text{train}}$). We choose the NAF-AD results as the base for four small-scale experiments in tabular datasets and compare the increment of F1-score via five active learning iterations. All experiments show a consistent trend as augmented samples are gradually added to the training. Arrhythmia and Lympho. experiments show better improvement since their original training data is very small. The improvement tends to converge if more iterations are used but we choose 50% as a threshold given training cost limitation. Figure 4 (right) shows the effect of transformations on active learning. Although a smaller number of transformations increases the classification error (also reported by Bergman and Hoshen [2020]), our NAF-AL can decrease the error via sample augmentation and achieve an equivalent accuracy by using fewer transformations and thus reduce the computational cost.

## B Tabular and time series data information and statistics

The statistical information of the tabular datasets and time series dataset are provided in Table 4 and 5 respectively.

Table 4: Statistical information of the tabular benchmark datasets

| Dataset | Data size | Dim | Anomaly ratio | Domain |
|---|---|---|---|---|
| Arrhythmia | 274 | 452 | 0.15 | Medical |
| Cardio. | 1831 | 21 | 0.096 | Medical |
| Lympho. | 148 | 18 | 0.04 | Medical |
| Thyroid | 3772 | 6 | 0.025 | Medical |
| KDD | 494,021 | 120 | 0.2 | Cybersecurity |
| KDDRev | 121,597 | 120 | 0.2 | Cybersecurity |

Table 5: Statistical information of the time series benchmark datasets

| Dataset | Data size | Dim | Length | Classes |
|---|---|---|---|---|
| Character Trajectories | 2858 | 3 | 182 | 20 |
| Epilepsy | 275 | 3 | 206 | 4 |
| NATOPS | 360 | 24 | 51 | 6 |
| Racket Sports | 303 | 6 | 30 | 4 |
| Spoken Arabic Digits | 8800 | 13 | 93 | 10 |

## References


Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

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

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

Jinghui Chen, Saket Sathe, Charu Aggarwal, and Deepak Turaga. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM international conference on data mining*, pages 90–98. SIAM, 2017.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2017.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9781–9791, 2018.

Nico Görnitz, Marius Kloft, and Ulf Brefeld. Active and semi-supervised data domain description. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 407–422. Springer, 2009.

Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. Drocc: Deep robust one-class classification. In *International Conference on Machine Learning*, pages 3711–3721. PMLR, 2020.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2018.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. *arXiv:2107.12571*, 2021.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019a.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32:15663–15674, 2019b.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pages 2722–2730. PMLR, 2019.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087. PMLR, 2018.

Yang Jiao, Kai Yang, Shaoyu Dou, Pan Luo, Sijia Liu, and Dongjin Song. Timeautoml: Autonomous representation learning for multivariate irregularly sampled time series. *arXiv preprint arXiv:2010.01596*, 2020.

Diederik P Kingma and Prafulla Dhariwal. Glow: generative flow with invertible $1\times$ 1 convolutions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10236–10245, 2018.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29:4743–4751, 2016.

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in Neural Information Processing Systems*, 33, 2020.

Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Charline Le Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. *arXiv preprint arXiv:2012.03808*, 2020.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

Yezheng Liu, Zhe Li, Chong Zhou, Yuanchun Jiang, Jianshan Sun, Meng Wang, and Xiangnan He. Generative adversarial active learning for unsupervised outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1517–1528, 2019.

Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus Robert Muller. Explainable deep one-class classification. In *International Conference on Learning Representations*, 2020.

Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *Anomaly Detection Workshop at 33rd International Conference on Machine Learning*, 2016.

Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 353–362, 2019.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2335–2344, 2017.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Dan Pelleg and Andrew Moore. Active learning for anomaly and rare-category detection. *Advances in neural information processing systems*, 17:1073–1080, 2004.

Tiago Pimentel, Marianne Monteiro, Adriano Veloso, and Nivio Ziviani. Deep active learning for anomaly detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

Chen Qiu, Timo Pfrommer, Marius Kloft, Stephan Mandt, and Maja Rudolph. Neural transformation learning for deep anomaly detection beyond images. In *International conference on machine learning*. PMLR, 2021.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1907–1916, 2021.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2019.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

Alejandro Pasos Ruiz, Michael Flynn, James Large, Matthew Middlehurst, and Anthony Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35(2):401–449, 2021.

Robin Schirrmeister, Yuxuan Zhou, Tonio Ball, and Dan Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. *Advances in Neural Information Processing Systems*, 33, 2020.

Bernhard Schölkopf, Robert C Williamson, Alexander J Smola, John Shawe-Taylor, John C Platt, et al. Support vector method for novelty detection. In *NIPS*, volume 12, pages 582–588. Citeseer, 1999.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2020.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.

Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. In *International Conference on Learning Representations*, 2020.

Jack W Stokes, John Platt, Joseph Kravis, and Michael Shilman. Aladin: Active learning of anomalies to detect intrusions. 2008.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in Neural Information Processing Systems*, 33:11839–11852, 2020.

Toan Tran, Thanh-Toan Do, Ian Reid, and Gustavo Carneiro. Bayesian generative active deep learning. In *International Conference on Machine Learning*, pages 6295–6304. PMLR, 2019.

Holger Trittenbach and Klemens Böhm. One-class active learning for outlier detection with multiple subspaces. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 811–820, 2019.

Siqi Wang, Yijie Zeng, Xinwang Liu, En Zhu, Jianping Yin, Chuanfu Xu, and Marius Kloft. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 5962–5975, 2019.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2114–2124, 2021.

Xuchao Zhang, Yifeng Gao, Jessica Lin, and Chang-Tien Lu. Tapnet: Multivariate time series classification with attentional prototypical network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6845–6852, 2020.

Jingbo Zhu, Huizhen Wang, Benjamin K Tsou, and Matthew Ma. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on audio, speech, and language processing*, 18(6):1323–1331, 2009.

Ev Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13994–14003, 2020.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.