NEURAL AUTOREGRESSIVE REFINEMENT FOR SELF-SUPERVISED OUTLIER DETECTION BEYOND IMAGES

Anonymous authors
Paper under double-blind review

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. However, it is not straightforward to apply these techniques to non-image data, such as time series or tabular data. Here we propose a novel data refinement (DR) scheme that relies on neural autoregressive flows (NAF) for self-supervised anomaly detection. 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-DR 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 incorporated with normal samples are used to train a better detector to approach decision boundaries. Compared with random transformations, NAF-DR 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 tabular and time series datasets, and one real-world application, significantly improving accuracy and robustness over the state-of-the-art baselines.

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 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.

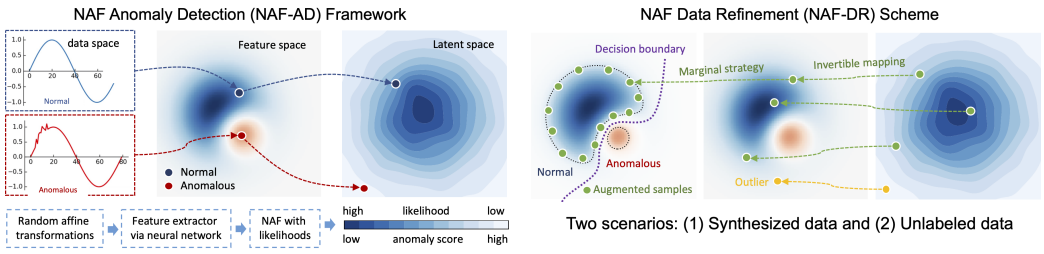

Figure 1: NAF-DR first transforms the data into multiple subspaces and learns a feature space by a neural network. Then we build an accurate density estimation via NAF and assign a higher likelihood to normal (a lower likelihood to anomaly) in latent space. NAF-DR is free to draw samples and transform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

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 & 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 data refinement scheme (NAF-DR). 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-DR 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-DR 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.

2 Related Work

Deep Anomaly Detection. Many recent advances have been proposed to use deep learning for anomaly detection. Ruff 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., 2019; Hendrycks & 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 & El-Yaniv, 2018; Wang et al., 2019)), using normal data for training and the learned classifier is useful to detect anomalies. Bergman & Hoshen (2020) extended the work from Golan & 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 & 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). Much recent effort has been made to improve model expressivity and computational efficiency that allow more accurate likelihood calculation and enable faster sampling (Rezende & Mohamed, 2015). 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). Recently, flow-based models have been explored for deep AD on large-scale image datasets and show promising results (Rudolph et al., 2021; Zisselman & Tamar, 2020; Gudovskiy et al., 2021). Inspired by these studies, our work leverages the benefits of NFs (exact density estimation, efficient sampling, and inference) but focuses on addressing the AD challenges in general data types beyond images, specifically given limited labeled data.

Data Augmentation and Refinement. 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 & 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 in active learning rely on the incorporation of the Bayesian approach (Gal et al., 2017; Tran et al., 2019) and deep generative models (Sinha et al., 2019). Active learning strategies for anomaly detection (Stokes et al., 2008; Görnitz et al., 2009; Pelleg & 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 & Böhm, 2019) but they are not feasible due to limited budget or high cost in practice. Unlike conventional data augmentation or active learning, our goal is to integrate a likelihood-based detector with novel data refinement, 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 self-supervised anomaly detection.

3 NAF-DR METHOD

3.1 DATA TRANSFORMATIONS IN SELF-SUPERVISED SETTING

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 $x \in \mathcal{S}_d \setminus X$. Our method is built upon the self-supervised scenario which can be typically defined above. 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:

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

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 types 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 θ, 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 the anomaly of each data. Lower probabilities (likelihoods) indicate 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_ψ is an invertible neural network parametrized by ψ. 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 ψ of the invertible transformation f_ψ.

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

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 univariate 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 an NF. In particular, the transformation $f_\psi^{-1,(i)}$ can be decomposed via invertible transformation neural network $t_\psi^{(i)}$ and condition neural network $c_\psi^{(i)}$:

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

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 $c_\psi^{(i)}$ and a composition of multiple flows due to their simplicity which leads to a limitation on expressiveness of f_ψ.

Neural autoregressive flow (NAF) (Huang et al., 2018) was proposed by learning a complex bijection using a neural network monotonic in x_i, which is a universal approximator for explicitly learning likelihood with greater expressivity that allows it to better capture multimodal target distributions.

3.3 LIKELIHOOD-BASED ANOMALY SCORE

We propose to utilize the NAF model to learn the distribution of feature space and train the NAF by maximum likelihood estimate (MLE), 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, flow-based model parameters ψ and feature extractor parameters θ in $\mathcal{L}_{(\theta,\psi)}(\mathcal{D}_n)$ are simultaneously 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

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

where $f_\theta(T_k(x))$ represents the feature space parametrized by a neural network f_θ. The anomaly score $\mathcal{S}(x)$ measures the anomaly degree, which is used to compare with the ground truth labels and calculate the quantitative metrics, e.g., AUC and F1 score.

3.4 DATA REFINEMENT SCHEME

Formally, active learning is used to automatically select the most informative subset of unlabeled training samples and label them by an oracle (domain expert), where the cost is often high in practice. To this end, we devise a novel data refinement scheme by leveraging the benefits of the NAF model on efficient sampling and exact likelihood-based inference to query low-confidence decisions, hence guiding the detector with augmented normal data in the training process. Figure 2 shows the proposed data refinement scheme with a marginal strategy in two different scenarios:

Synthesized data: if labeled data is small and no unlabeled data is available, the NAF model is able to generate synthesized data that follow the learned distribution of (normal) training data. Some of these data are selected and then labeled by likelihood-based inference from the NAF model, which approximates the capability of an oracle, which is very expensive practically. The labeled normal data will merge with the normal training data together to retrain and update the NAF model.

Unlabeled data: In a practical scenario, labeled data is sparse but large unlabeled data are often available. Instead of human labeling by domain experts, a batch of data from an unlabeled data pool is selected and then labeled via likelihood-based inference by leveraging the advances of the NAF model. Then these data with normal labels are merged with the training data pool for retrain.

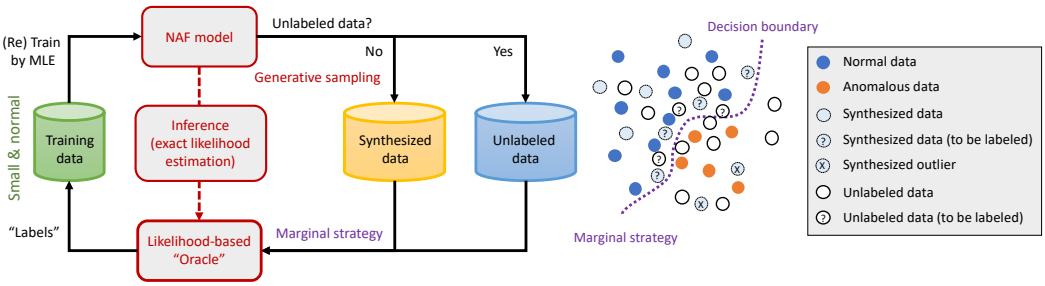

Figure 2: Data refinement using synthesized data or unlabeled data, and marginal strategy illustration.

Marginal strategy: One of the key enabling techniques is uncertainty sampling, which uses one classifier to identify unlabeled samples with the least confidence. However, data refinement for self-supervised anomaly detection is different from the original scope of active learning. We thus propose a marginal strategy by selecting the sample x^* that is close to the likelihood-based decision boundary:

$$x^* = \underset{x \in \{x_1, \ldots, x_m\}}{\arg\min} \frac{|\mathcal{L}(x)|}{\Omega}, \quad \Omega = \max_j |\mathcal{L}(x_j)|, \quad j = 1, ..., m \tag{8}$$

where m is the batch size of new samples from unlabeled data or synthesized data, shown in Fig. 2. Ω is the maximum of likelihood value, which is used for normalization. Benefiting from high sampling and inference efficiency, we can calculate the log-likelihood of new samples $\mathcal{L}(x_j^{aug})$ and retain samples with lower likelihoods if $\mathcal{L}(x_j^{aug}) < \mathcal{Q}_\alpha(\mathcal{L}(x_k^c))$ or $\mathcal{L}(x_j^{aug}) > \mathcal{Q}_{1-\alpha}(\mathcal{L}(x_k^c))$ where $\mathcal{Q}_\alpha$ is the α-quantile ($\alpha \in [0.05, 0.1]$ is a hyperparameter) and $\mathcal{L}(x_k^c)$ is the log-likelihood of current training data $x_k^c \in X, k = 1, ..., n$. It is critical to design an appropriate likelihood level set trade-off between aggressive boundary and conservative boundary, see Figure 3 for more discussion.

The samples with lower likelihoods are desired but it is probably an outlier (anomalous data, see Fig.2) if the likelihood is too small. Thus we have an additional step to check the samples by outlier detection, which is done by Mahalanobis distance $M(x)$ with a defined threshold δ_M that is Mahalanobis distance at $\chi^2_{0.05}$. As a result, the samples with $M(x) > \delta_M$ are rejected. We eventually determine the augmented samples x^* by solving the optimization problem in Eq. 8 subject to the low-confidence and outlier constraints. These samples x^* are selected as normal samples and added to update the training.

Stopping criterion: the new augmented samples are selected as normal samples and added to the current training dataset for retraining. An appropriate stopping criterion for data refinement is a trade-off between training cost and effectiveness of the detector. We set up two criteria: (1) if the retrain times increase to the maximum, $T_{\max}$ or (2) the augmented data size reaches a specific ratio $R_{\max}$ of the original training size n given the concern of sample efficiency. The data refinement

update will be stopped as long as either criterion is triggered. Here, we choose $T_{\max} = 5$ and $R_{\max} = 50\%$.

Algorithm 1 The NAF-DR algorithm

Require: Training and testing datasets $\mathcal{D}_n, \mathcal{D}_t$, number of transformations K, feature extractor f_θ, NAF model f_ψ, hyperparameters α, δ_m and initialize T = 0 and R = 0 in the stopping criterion
while $T \leq T_{\max}$ or $R \leq R_{\max}$ **do**
  // `Training process`
  Transform each training sample according to Eq. 1: $T_1(x_i), T_2(x_i), ..., T_k(x_i) \leftarrow x_i,\ x_i \in \mathcal{D}_n$
  Extract feature representation $\tilde{T}_k(x_i) \leftarrow T_k(x_i),\ k = 1, ..., K$ via a neural network model f_θ
  Concatenate all different affine transformations $\tilde{T}_k(x_i)$
  Evaluate the NAF model f_ψ for z and $|\det \boldsymbol{J}|$ and minimize the loss $\mathcal{L}$ in Eq. 6 to update θ, ϕ
  // `Data refinement process (for training data)`
  Draw samples x_j^{aug} from the learned NAF model f_ψ or from unlabeled data pool in Fig. 2.
  Evaluate the log-likelihood of new augmented samples $\mathcal{L}(x_j^{aug})$ and current training samples $\mathcal{L}(x_k^c)$
  Retain samples x_j^{aug} with lower likelihoods if $\mathcal{L}(x_j^{aug}) < \mathcal{Q}_\alpha(\mathcal{L}(x_k^c))$ or $\mathcal{L}(x_j^{aug}) > \mathcal{Q}_{1-\alpha}(\mathcal{L}(x_k^c))$
  Check the outliers of x_j^{aug} if the Mahalanobis distance $M(x_j^{aug}) > \delta_M$
  Merge new samples x_j^{aug} with normal labels to current training dataset, $\mathcal{D}_n = \mathcal{D}_n \cup x_j^{aug}$ for retrain
  // `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 loglikelihood $p_Z \leftarrow f_\psi^{-1}(f_\theta(T_k(x_t)))$
  Concatenate all transformations and average the negative log-likelihoods to compute $\mathcal{S}(x)$ in Eq. 7
end while

4 EXPERIMENTS

Most image datasets are large-scale and have existing strong baselines so we do not expect significant improvement via our NAF-DR method. Instead, our focus is on *small-scale* multiple benchmark tabular data and time series data, as well as one real-world application in the scientific facility.

4.1 ANOMALY DETECTION BASELINE METHODS

- *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 & Hoshen, 2020) is a self-supervised classification-based method. Neural 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 is important in anomaly detection applications since 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[1], and two cybersecurity datasets, KDD and KDDRev from the empirical studies of Zong et al. (2018); Bergman & Hoshen (2020); Qiu et al. (2021) which are used to show our potential to large-scale datasets. The key statistics (data size, dimension, and anomaly ratio) of the tabular

[1] `http://odds.cs.stonybrook.edu/`

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 the 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 & 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.

Considering the training cost limit, we set up a stop criterion by using a maximum number of iterations ($N_{\max} = 5$) for data refinement. 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 likelihood and finally reject the larger 90% samples (retain 10% samples with lower likelihoods near the decision boundaries). These samples are further checked by the Mahalanobis distance criterion if it is an outlier. After that, we combine these augmented samples with the existing normal samples to update training.

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-DR	**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 & 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-DR 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 & 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. NAF-DR outperforms all baselines on these small-scale datasets thanks to the benefits of data refinement. Compared with GOAD which is a classification-based method, our probabilistic flow-based model is competitive even without the help of data refinement. NeuTral beats our NAF-AD in the Arrhythmia dataset but underperforms our NAF-DR method. Since our NAF-DR is flexible to incorporate any number of transformations such that our robustness is better than NeuTral.

- *Large-scale datasets:* The deep baselines show superior performance compared with the shallow methods in this case. NAF-AD is slightly lower than NeuTral, GOAD, and DSVDD but NAF-DR is still competitive. One explanation is that the performance improvement in such a large dataset is not significant as in the small-scale case discussed above. The large datasets, having different dynamics from very small datasets found by Bergman & 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 as 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 sports monitoring may indicate injury, disease, or serious issues.

Datasets. We focus on five multivariate time series datasets from the UEA multivariate time series classification archive [2] 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 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-DR	**99.5**±0.1	**93.2**±0.3	**95.7**±0.3	**88.2**±0.5	98.4±0.1

Evaluation Protocol. We evaluate the NAF method on these benchmarks based on two protocols:

- *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 the 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 the classes are treated as anomalies. In this case, the normal data is no longer from one class such that the variability increases significantly.

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 2, 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 Qiu et al. (2021).

Results. Table 2 shows the results of NAF-AD and NAF-DR in comparison to the shallow and deep AD baselines on multiple time series experiments. NAF-DR 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 data refinement. Only on the SAD dataset, our NAF-DR is outperformed by NeuTral with learned transformations which have an advantage over the random transformations, while our data refinement improvement looks marginal because its dataset size is larger than the other experiments. Our NAF-DR 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-DR can well handle both scenarios although it is designed for addressing the specific challenges from small data.

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-DR	89.3±0.2	**81.7**±0.7	**75.8**±0.5	**82.7**±0.3	83.9±0.3

The results of the m-vs-rest tasks are shown in Table 3. In this case, NAF-DR 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-DR shows superior performance on 4 out of 5 experiments. On SAD, NAF-DR 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.

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

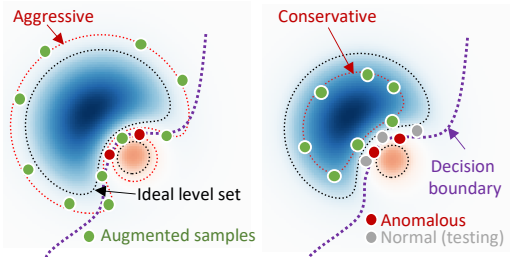

Figure 3: Sample augmentation via NAF-DR: aggressive strategy vs conservative strategy.

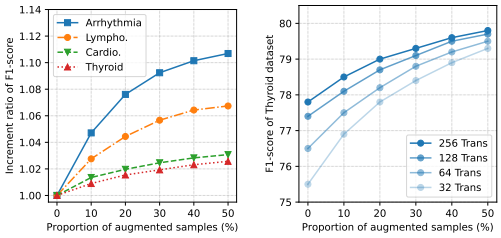

Figure 4: Effect of data refinement and the number of transformations on F1-scores.

4.4 Real-world Application: Failure Detection in Particle Accelerator

Finally, we demonstrate the NAF-DR method on a real-world failure detection problem in the Spallation Neutron Source (SNS) facility which is the world's highest power proton accelerator, see Fig. 5 Blokland et al. (2021). Achieving high availability is a challenging task in beam accelerators because errant beam pulses can cause running failure and damage to the accelerator. The goal of this work is to detect upcoming beam loss using historical high-frequency time series data from monitoring devices and thus stop the accelerator before failure/damage occurs.

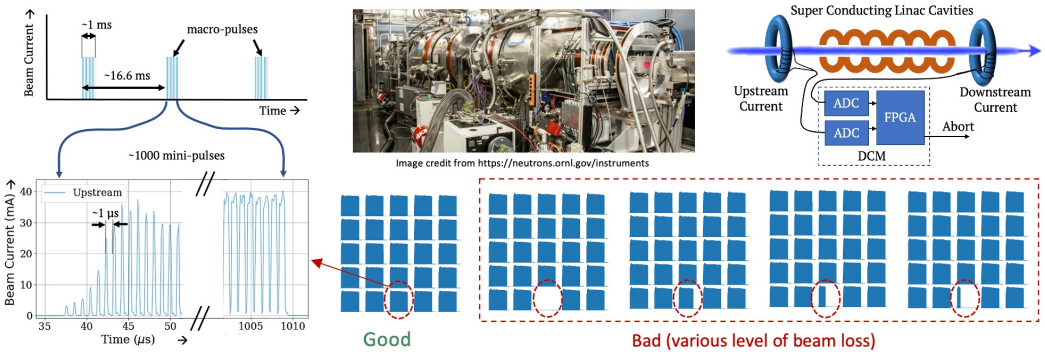

Figure 5: Failure detection to predict errant beam pulses in the Spallation Neutron Source (SNS) accelerator.

In this case, a large amount of data is available from daily monitoring (∼350 data) but most of the data are unlabeled because data labeling needs domain experts' efforts. To demonstrate the applicability of the data refinement scheme with unlabeled data, we extracted normal and anomalous pulses from the archived data from July 2020 and investigate two cases: 4 pulses (2150 good, 148 bad) and 25 pulses (2530 good, 107 bad) with large enough unlabeled data (notice that various levels of beam loss are all labeled as "bad", and the pulse without beam loss is labeled as "good"). Table 4 shows the mean and standard deviation of AUC performance for this real-world time series task. Compared with all baseline methods, the NAF-DR method demonstrates superior performance with better reliability.

Table 4: Mean and standard deviation of AUC for the SNS task

	4 pluses	25 pluses
IForest	68.2±1.4	73.3 ±0.9
LOF	58.8 ±1.2	62.0 ±1.1
OC-SVM	72.1 ±1.3	75.7 ±1.0
DAGMM	74.7±1.1	73.9±0.8
DSVDD	81.9±1.2	82.5±0.9
FB-AE	77.5±0.7	80.1±0.6
GOAD	81.1±1.0	87.3±0.9
LSTM-ED	83.4±0.8	84.9±0.8
NeuTraL	87.3±0.7	88.8±0.6
NAF-AD	89.1±0.6	90.8±0.6
NAF-DR	**91.8**±0.6	**93.4**±0.5

5 Conclusion

We propose a likelihood-based data refinement method for self-supervised anomaly detection on small data beyond images. The key contribution is to develop a new data refinement strategy that benefits from efficient sampling and explicit likelihoods from neural autoregressive flows. We demonstrate the novel method on several tabular and time series benchmarks and one real-world scientific application with superior performance over the state-of-the-art methods.

REFERENCES

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

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.

Willem Blokland, Pradeep Ramuhalli, Charles Peters, Yigit Yucesan, Alexander Zhukov, Malachi Schram, Kishansingh Rajput, and Torri Jeske. Uncertainty aware anomaly detection to predict errant beam pulses in the sns accelerator. *arXiv preprint arXiv:2110.12006*, 2021.

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*, pp. 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*, pp. 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*, pp. 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*, pp. 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*, pp. 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*, pp. 3711–3721. PMLR, 2020.

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.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pp. 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.

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.

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*, pp. 9664–9674, 2021.

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

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*, pp. 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.

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)*, pp. 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. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/1e79596878b2320cac26dd792a6c51c9-Paper.pdf.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 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*, pp. 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*, pp. 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.

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, pp. 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*, pp. 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*, pp. 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*, pp. 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*, pp. 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*, pp. 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, pp. 6845–6852, 2020.

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*, pp. 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.

A APPENDIX

A.1 FLOW-BASED MODEL ARCHITECTURE

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 univariate conditional densities:

$$p_X(x) = p_{X_1}(x_1) \prod_{i=2}^{m} p_{X_i|X_{<i}}(x_i|x_{<i}) \tag{9}$$

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

$$z_i = f_\psi^{-1,(i)}(x_{\leq i}) = t_\psi^{(i)}(x_i, c_\psi^{(i)}(x_{<i})). \tag{10}$$

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 composition of multiple flows due to their simplicity which leads to a limitation on the expressiveness of f_ψ.

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 6.

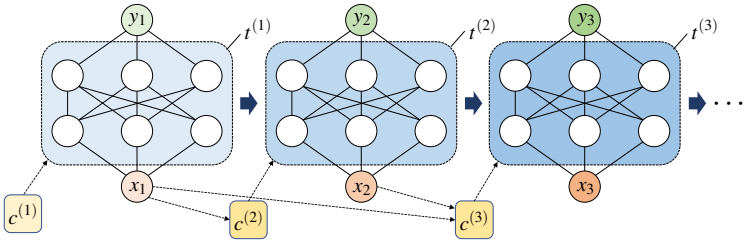

Figure 6: NAF architecture: each $c^{(i)}$ (a neural network) predicts pseudo-parameters for $t^{(i)}$, which in turns processes x_i.

As shown in Fig.6, each conditioner $c^{(i)}$ is an unrestricted function of $x_{<i}$. To parametrize a monotonically increasing transformed network $t^{(i)}$, the outputs of each conditioner $c^{(i)}$ are mapped to the positive real coordinate space by application of an appropriate activation. The result is a flexible transformation with a lower triangular Jacobian whose diagonal elements are positive.

For the efficient computation of all pseudo-parameters, NAF Huang et al. (2018) uses a masked autoregressive network f_θ. The Jacobian of a NAF is computed using the chain rule on f_θ through all its hidden layers. Since f_θ is autoregressive, $\mathbf{J}_{f_\theta}$ is lower triangular and only the diagonal needs to be computed for each i. Therefore, the operation requires only computing the derivatives of each $t^{(i)}$ reducing the time complexity.

A.2 DATASET DETAILS

A.2.1 TABULAR DATASETS

The details of tabular datasets are provided as follows:

- **Arrhythmia**: A cardiology dataset from the UCI repository Asuncion & Newman (2007) containing attributes related to the diagnosis of cardiac arrhythmia in patients. The datasets consist

of 16 classes: class 1 are normal patients, 2-15 contains different arrhythmia conditions, and class 16 contains undiagnosed cases. Following the data preparation of previous works, only 274 continuous attributes are considered. The abnormal classes include 3, 4, 5, 7, 8, 9, 14, and 15. The rest classes are considered normal.

- **Cardiotocography**: The Cardiotocography dataset Asuncion & Newman (2007) is obtained from the ODDS repository, with each sample having 21 features. The dataset comprises 1831 samples, including 176 anomalies (9.6% contamination).
- **Lymphography**: The Lymphography dataset Asuncion & Newman (2007) is obtained from the ODDS repository, with each sample having 18 features. The dataset comprises 148 samples, including 6 anomalies (4.0% contamination).
- **Thyroid**: A medical dataset from the UCI repository Asuncion & Newman (2007), containing attributes related to whether a patient is hyperthyroid. Following the data preparation of previous works, only 6 continuous attributes are considered. The hyperfunction class is treated as abnormal, and the rest of the 2 classes are considered normal.
- **KDD**: The KDD Intrusion Detection dataset was created by an extensive simulation of a US Air Force LAN network. The dataset consists of the normal and 4 simulated attack types: denial of service, unauthorized access from a remote machine, unauthorized access from a local superuser, and probing. The dataset consists of around 5 million TCP connection records. Following the evaluation protocol in Zong et al. (2018), we use the UCI KDD 10% dataset, which is a subsampled version of the original dataset. The dataset contains 41 different attributes. 34 are continuous and 7 are categorical. Following Zong et al. (2018), we encode the categorical attributes using 1-hot encoding.
- **KDDCUP**: The KDDCUP99 10 percent dataset from the UCI repository contains 34 continuous attributes and 7 categorical attributes. Following the data preparation of previous works, 7 categorical attributes are represented by one-hot vectors. Eventually, the data has 120 dimensions. The attack samples are considered normal, and the non-attack samples are considered as abnormal.
- **KDDCUP-Rev**: It is derived from the KDDCUP99 10 percent dataset. The non-attack samples are considered normal, and attack samples are considered as abnormal. Following the data preparation of previous works, attack data is sub-sampled to consist of 25% of the number of non-attack samples.

The statics of the tabular data is shown in Table 5.

Table 5: 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

A.2.2 TIME SERIES DATASETS

We provide more details and information about the time series data, which are from the UEA multivariate time series classification archive Bagnall et al. (2018).

- **SAD**: The full name is 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.
- **NATOPS**: The full name is Naval air training and operating procedures standardization. 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.

- **CT**: The full name is 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.
- **EPSY**: The full name is Epilepsy. 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.
- **RS**: The full name is Racket Sports. 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.

The statistical information of the time series data is summarized in Table 6

Table 6: 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

A.3 Further Discussion

A.3.1 How does the data refinement work with NAF?

Figure 3 in the main context shows the data augmentation scheme via neural autoregressive refinement 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 data refinement, 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. In 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 data refinement from being too aggressive or too conservative.

A.3.2 How does NAF-DR improve the AD performance?

Figure 4 in main context 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 the F1-score via five data refinement 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 the training cost limitation. Figure 4 in the main context (left) shows the effect of transformations on data refinement. Although a smaller number of transformations increases the classification error (also reported by Bergman & Hoshen (2020)), our NAF-DR can decrease the error via sample augmentation and achieve an equivalent accuracy by using fewer transformations.

A.4 LIMITATIONS AND SOCIAL IMPACT OF THE WORK

The authors are aware of the limitations of the presented case studies focusing on small-scale AD problems with the tabular and time-series type of data formats. The computational challenge in sequential retrain and update is also a concern, specifically for generative models. As part of future work, the authors plan to investigate large-scale tabular and time-series problems but also balance the computational cost and sampling efficiency trade-off.

The presented work falls into the basic research category. As such, the authors are not aware of any potential direct negative societal impact of the proposed work. On the contrary, the authors of this paper believe that the presented theory is a minor contribution towards general knowledge, which accumulation has been historically proved to inherently benefit all humanity.