# Robust Variational Autoencoders and Normalizing Flows for Unsupervised Network Anomaly Detection

Naji Najari, Samuel Berlemont, Grégoire Lefebvre, Stefan Duffner, Christophe Garcia

# Robust Variational Autoencoders and Normalizing Flows for Unsupervised Network Anomaly Detection

Naji Najari[1,2,3], Samuel Berlemont[1], Grégoire Lefebvre[1], Stefan Duffner[2,3], and Christophe Garcia[2,3]

[1] Orange Labs, Meylan, France
{naji.najari, samuel.berlemont, gregoire.lefebvre}@orange.com
[2] LIRIS UMR 5205 CNRS, Villeurbanne, France
{naji.najari, stefan.duffner, christophe.garcia}@liris.cnrs.fr
[3] INSA Lyon, Villeurbanne, France

**Abstract.** In recent years, the integration of connected devices in smart homes has significantly increased, thanks to the advent of the Internet of things (IoT). However, these IoT devices introduce new security challenges, since any anomalous behavior has a serious impact on the whole network. Network anomaly detection has always been of considerable interest for every actor in the network landscape. In this paper, we propose GRAnD, an algorithm for unsupervised anomaly detection. Based on Variational Autoencorders and Normalizing Flows, GRAnD learns from network traffic metadata, a normal profile representing the expected nominal behavior of the network. Then, this model is optimized to detect anomalies. Unlike existing anomaly detectors, our method is robust to the hyperparameter selection, and outliers contaminating the training data. Extensive experiments and sensitivity analyses on public network traffic benchmark datasets demonstrate the effectiveness of our approach in network anomaly detection.

**Keywords:** Unsupervised anomaly detection, robust autoencoders, dynamic outlier filtering. network traffic anomaly detection.

## 1 Introduction

Thanks to the recent advances in Internet of Things (IoT) technologies and the steady growth of IT services, IoT devices have become ubiquitous in multiple domains such as Smart Home, Healthcare, Industry 4.0. Although the IoT has played a key role in the enablement of new services and the development of new business value, there is a growing concern about the security of modern networks. IoT devices have numerous technical limitations such as constrained resources, battery failure, connectivity issues, and are vulnerable to diverse cyber threats. Such failures have serious consequences on the Quality of Service (QoS). Therefore, detecting abnormal events is of paramount importance to mitigate risks, prevent system failures.

Signature-based Intrusion Detection Systems (IDSs) are commonly used to protect IoT devices from cyber threats. They detect network anomalies by comparing the traffic with known attack signatures. Although they are effective to detect already known attacks, these systems are incapable of mitigating non-malicious anomalies or novel attacks, e.g., zero-day attacks [1]. Unsupervised anomaly detection has been a point of interest to mitigate these limitations and develop reliable and secure networks.

Anomaly Detection (AD) is the task of detecting anomalous data points that significantly deviate from expected normal samples [7]. The most common approaches for AD are based on One-Class Classification (OCC) [7]. OCC consists in learning an accurate representation of the norm, relying only on nominal data points. Once the normal data are well-modeled, the algorithm assigns an abnormality score to each test sample. Finally, a threshold criterion separates inliers and outliers. OCC efficacy depends on the availability of anomaly-free training data, and performance may degrade significantly when this assumption is violated. Unfortunately, this violation is likely to occur in real-world applications. For example, in network traffic monitoring, collected network packets may comprise defective data sent by faulty sensors, damaged fiber connectors, or caused by network congestion [12]. Finally, due to data volumes and potentially unknown anomalies, manual labelling of training samples is not feasible.

In this paper, we propose GRAnD, an algorithm for Generative Robust Anomaly Detection. We introduce a training strategy that alternates between filtering outliers contaminating the training dataset and learning a robust representation of the norm. Our training strategy involves little architectural changes and can be integrated with Variational Autoencoders (VAEs) [11] and Normalizing Flows (NFs) [17]. Unlike recent robust generative methods, our approach makes no assumption about the anomaly distribution, or about the fraction of training outliers. Our method comprises three contributions :

- a robust rejection strategy that filters corrupted training samples, based on Extreme Value Theory (EVT). This strategy separates the training data into three disjoint subsets: an inlier subset containing training data deemed nominal, an outlier subset that comprises the "most anomalous" training samples, and a third subset containing critical undetermined instances;
- a training strategy that leverages filtered anomalies to learn a representation where inliers are well reconstructed and outliers are explicitly corrupted;
- an extensive validation on network traffic datasets, which demonstrates that our approach outperforms some state-of-the-art robust methods and robust to the hyperparameter selection.

## 2 Related Work

AD is an active research field that has always been a point of interest in different applications such as network intrusion detection, fraud detection, fault diagnosis, and predictive maintenance. Four families of approaches were proposed: *probabilistic, neighbor, domain and reconstruction-based methods* [7]. *Statistical and*

*probabilistic-based methods* typically model the inlier distribution by learning the parameters of a parametric function. Samples that have low likelihood under this model are considered anomalies. This category includes Gaussian mixture models [22], and kernel density estimators [8]. *Neighbor-based methods*, a.k.a. proximity-based methods, assume that outliers are far from their nearest neighbors, while inliers are close to each other. Well-known proximity-based method include Local Outlier Factor (LOF) [6] and Angle-Based Outlier Detection (ABOD) [13]. *Domain-based methods* estimate a boundary that separates the inlier domain from the rest. Anomalies are samples outside this inlier boundary. One-class SVM (OC-SVM) [19] and Support Vector Data Description (SVDD) [21] are two popular domain-based algorithms. *Reconstruction-based anomaly detection* assumes that, unlike outliers, inliers can be projected into a low-dimensional subspace. The reconstruction error represents a score of data abnormality, as the reconstruction errors of anomalies are higher than inliers. Particularly, AutoEncoders (AEs) have been trained to map nominal input data into a compact latent space, to learn a non-linear representation of the nominal class [7]. Besides, generative models have been profusely proposed for anomaly detection [4]. Furthermore, numerous studies explored Generative Adversarial Networks (GANs) [7] for AD.

The above methods have been applied to detect network traffic anomaly detection [9]. Although they show good results when trained with anomaly-free data, their performance drastically decreases when the training data is contaminated with outliers. In a real-world environment, there is no guarantee that the collected training data are entirely clean. Atypical abnormal traffic may be hidden in the collected data, due to adversarial attacks, or packet collisions. Consequently, it is advocated to develop robust unsupervised anomaly detectors, insensitive to training contaminants [18].

Zhou and Paffenroth [23] proposed Robust Deep Autoencoders (RDAs) to filter sparse corrupted samples from the input data matrix. Robust Subspace Recovery (RSR) [15] is another line of work in robust anomaly detection. RSR assumes that inliers can be projected into a linear low-dimensional subspace, while outliers are not well modeled in this subspace. Lai et al. [14] introduced Robust Subspace Recovery AutoEncoder (RSRAE), where they integrated an RSR-layer in a classical autoencoder. Regarding robust generative autoencoders, Akrami et al. [3] proposed a Robust VAE (RVAE). Their approach uses the robust $\beta$-divergence instead of the standard Kullback-Leibler (KL) divergence. Minimizing the $\beta$-divergence involves reweighting each sample likelihood gradient with its probability density.

Recently, Kotani et al. [12] used RDA for network flow intrusion detection. Although these approaches proved to reduce the number of false positives on real-world traffic datasets, they involve an explicit regularization, defined by one or many critical hyperparameters. Prior knowledge about the outlier ratio and additional assumptions either on the inlier, the outlier class, or both, are required to select the optimal hyperparameters. Generally, such hyperparameters are empirically tuned with a dedicated validation subset containing manually

ground-truth-labeled data. In the context of anomaly detection, labeled outliers are too scarce to form a balanced validation subset. Also, in most situations, the ratio of training outliers is not known. Therefore, hyperparameter selection is prone to misspecification. However, the methods above-mentioned are all sensitive to their hyperparameters, since slightly changing them can drastically degrade their anomaly detection performances.

In contrast, our approach does not make any assumptions about outlier distribution. We propose a robust training strategy that jointly performs two tasks. This strategy filters training outliers using EVT. Then, training outliers are leveraged to infer a better representation that can be generalized to unseen anomalies. This strategy can be incorporated with VAEs and NFs, and involves minimal architectural changes.

## 3 Background

### 3.1 Generative AEs

We consider the task of unsupervised AD under the standard variational inference setting. Generative models aim to find the optimal parameters $\theta$ that maximize the likelihood $p_\theta(x) = \mathbb{E}_{p(z)}[p_\theta(x|z)]$, where $z$ is the model latent variable and $p(z)$ is a predefined prior. However, this likelihood is intractable because of the marginalization over the latent variable $z$. Variational inference aims to approximate the posterior probability $p(z|x)$ with a parametric distribution $q_\phi(z|x)$, parameterized by $\phi$. Regardless of the choice of this distribution, we can reformulate the log-likelihood as follows:

$$\log p_\theta(x) \geq \mathbb{E}_q[\log p_\theta(x|z)] - \mathbb{D}_{KL}[q_\phi(z|x)||p(z)] = -\mathcal{F}(x), \tag{1}$$

where $q_\phi(z|x)$ is the approximate posterior distribution for the latent variables, and $\mathcal{F}$ is the negative free energy, a.k.a., the evidence lower bound (ELBO). This energy comprises two terms. The first term is the reconstruction error, and the second one represents the KL divergence between the approximate distribution and the prior distribution. A common choice of the approximate distribution family is the multivariate Gaussian distribution with a diagonal covariance matrix. Recently, NFs have been used to provide a richer parametric family of approximate posterior to capture complex structures of the latent space. NFs transform an initial simple density function to a more sophisticated one, by applying a sequence of invertible transformations.

### 3.2 EVT

The objective of EVT is to quantify the probability of occurrence of extreme values in a distribution function. Recently, EVT has been applied to detect anomalies in many applications including network traffic data streams [20]. The Peaks-Over-Threshold (POT) is a typical approach used to model the extreme

values of samples that exceed a specific high threshold. This approach is a result of the Picakands-Balkema-de-Han theorem of EVT [5].

Let $(X_1, X_2, \ldots, X_n)$ be $n$ independent and identically distributed (iid) random variables. Let $F_u$ be their conditional excess distribution function, i.e., $F_u(x) = P(X - u > x | X > u)$, where $u$ is a high threshold. The POT method models the extreme values that exceed the threshold $u$, using the Generalized Pareto Distribution (GPD) parametrized by two parameters, $\xi$ and $\sigma$:

$$F_u(x) \to 1 - G_{\xi,\sigma}(x), \text{ as } u \to \infty \quad \text{where} \quad \begin{cases} G_{\xi,\sigma}(x) = 1 - (1 + \frac{\xi x}{\sigma})^{-\frac{1}{\xi}}, & if\ \xi \neq 0 \\ G_{\xi,\sigma}(x) = 1 - e^{-\frac{x}{\sigma}}, & if\ \xi = 0. \end{cases} \quad (2)$$

In practice, the two parameters of the GPD are empirically estimated by fitting the GPD to the data. The maximum likelihood estimation is typically used to find these optimal parameters $\tilde{\xi}$ and $\tilde{\sigma}$ . Once the extreme values are modeled with the optimal GPD, $G_{\tilde{\xi},\tilde{\sigma}}$, we can identify rare extreme samples that have very low probability [20]. Given a small probability $q$, we can compute the threshold $t_q$ such that, $P(X > t_q) < q$.

$$P(X - u > t_q | X > u) = \tilde{F}_u(t_q) \sim 1 - G_{\tilde{\xi},\tilde{\sigma}}(t_q). \quad (3)$$

$$\text{If } \xi \neq 0, \quad t_q \simeq u + \frac{\tilde{\xi}}{\tilde{\sigma}}((\frac{nq}{N})^{\tilde{\xi}} - 1), \quad (4)$$

where $n$ is the total number of observations, and $N$ is the number of $X_i$ exceeding the threshold $u$, $X_i > u$. A key question arises as to how to choose the threshold $u$. Siffer et al. [20] state that "the value of $u$ is not paramount except that it must be high enough." In practice, $u$ is generally selected as a high empirical quantile of the data, e.g., 90% quantile.

## 4 Contributions

This paper focuses on unsupervised anomaly detection where the unlabeled training data may contain both inliers and outliers, with an imbalanced class distribution. We assume that the majority of the training instances are nominal, along with a small ratio of "contaminants", i.e. outliers. The ratio of these contaminants, which we call $\gamma_p$, is not known in advance. In the following, we introduce GRAnD, an algorithm for Generative Robust Anomaly Detection. Our contribution alternates between filtering training outliers and learns a robust distribution of the norm. In the following, we will first explain the rejection strategy that isolates training contaminants. Then, we will detail the objective function to optimize.

### 4.1 Robust Rejection Strategy

The objective of this rejection strategy is to separate nominal training data points from anomalies. The main idea consists in setting a relevant threshold

to segment the reconstruction scores assigned to training samples, in order to reject outliers having extreme scores.

We hypothesize that, early in the training phase, contaminants have larger free energy (cf. Equation 1), compared to inliers. Consequently, we propose to isolate these extreme values by thresholding the energy with the POT approach, described in Section 3.2. The POT approach requires the selection of two parameters: the initial threshold $u$, and the risk parameter $q$. In our experiments, we define $u$ as follows :

$$u = Q_3(\mathcal{F}) + \alpha * IQR(\mathcal{F}) \tag{5}$$

where $\mathcal{F}$ is the free energy of the training instances, $Q_3$ is the third quartile, and $IQR$ is the Inter-Quartile Range, which is defined as the difference between the third and the first quartiles. $\alpha$ controls the scale of the decision rule. In all our experiments, we fixed $\alpha = 1.5$ and $q = 0.001$. In Section 5.5, we study the sensitivity of our contribution with respect to $\alpha$ and $q$.

Using the POT parameters, we propose to split the input data into three subsets $\mathbb{X} = \mathbb{L} \cup \mathbb{S} \cup \mathbb{U}$, as illustrated in Figure 1. The subset $\mathbb{L}$ contains nominal training samples, having energy lower than the initial threshold $u$ of the POT method. $\mathbb{S}$ contains anomalous data points, with an energy higher than $t_q$, computed using Equation 4. $\mathbb{U}$ comprises the remaining critical samples, with an energy higher than $u$ and lower than $t_q$. These sample energies are neither very low to be considered nominal, nor high enough to be rejected as anomalies.

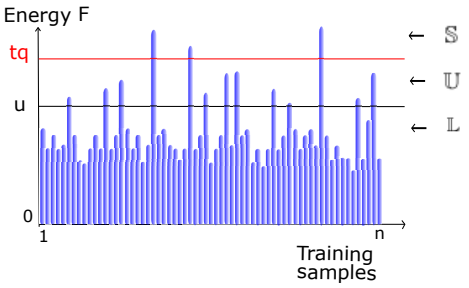



Fig. 1: Illustration of the rejection strategy using the POT approach.

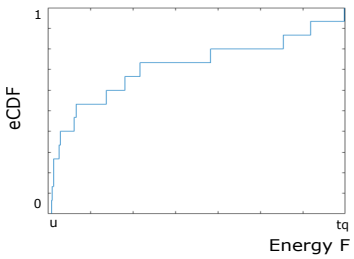



Fig. 2: Empirical cumulative distribution function of $\mathbb{U}$ samples

### 4.2 Training Loss

The rejection strategy splits the training data into three subsets $\mathbb{L}$, $\mathbb{S}$, and $\mathbb{U}$. We train the autoencoder to jointly perform three tasks: (i) minimize $\mathbb{L}$ sample energy, (ii) badly reconstruct $\mathbb{S}$ samples by maximizing their energy, (iii) maximize a weighted energy function of $\mathbb{U}$ instances, which takes into account the probability of anomalous of these instances. The idea is to associate to each critical instance in $\mathbb{U}$ a weight in $[0, 1]$ that quantifies whether the instance is anomalous or not.

Let $\mathbb{U} = \{X_1, X_2, \ldots, X_n\}$ contain a sequence of $n$ iid instances. We firstly sort these instances in increasing order according to their free-energies $(\mathcal{F}(X_1), \mathcal{F}(X_2), \ldots, \mathcal{F}(X_n))$. We use the empirical Cumulative Distribution Function (eCDF) to define the anomalousness weight of a $X_i \in \mathbb{U}$.

$$P(X_i \in \mathbb{U} \text{ is anomalous}) = eCDF_n(\mathcal{F}(X_i)) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\mathcal{F}(X_j) \leq \mathcal{F}(X_i)} \quad (6)$$

where $\mathbb{1}$ is the indicator function. As illustrated in Figure 2, $\mathbb{U}$ samples with energy close to the threshold $u$ have a small probability close to 0. Conversely, samples with high scores, i.e., close to $t_q$, have probabilities close to 1.

**GRAnD Objective Function** Given the three subsets of data $\mathbb{L}$, $\mathbb{S}$, and $\mathbb{U}$, respectively generated from the three distributions, $D_L$, $D_S$, and $D_U$, GRAnD optimizes the following objective function:

$$\mathcal{L}(x) = \mathbb{E}_{x\sim D_L}[\mathcal{F}_{\mathcal{L}}(x)] + |m - \mathbb{E}_{x\sim D_S}[\mathcal{F}_{\mathcal{S}}(x)]| + eCDF_m(\mathcal{F}_{\mathcal{U}}(x))\,|m - \mathbb{E}_{x\sim D_U}[\mathcal{F}_{\mathcal{U}}(x)]| \quad (7)$$

The objective function comprises three components:

- $\mathbb{E}_{x\sim D_L}[\mathcal{F}_{\mathcal{L}}(x)]$ is the expectation of the free energy function of $\mathbb{L}$ samples, defined in equation 1. This first component aims to minimize the energy of $\mathbb{L}$ samples.
- $\mathbb{E}_{x\sim D_S}[\mathcal{F}_{\mathcal{S}}(x)]$ is the expectation of the free energy function of $\mathbb{S}$ samples. $|.|$ is the absolute distance, and $m \in \mathbb{R}^+$ is a margin value. By maximizing this energy, we train the autoencoder to badly reconstruct the potential training contaminants. Since this energy function is positive and unbounded, we propose to fix an upper bound $m$, to prevent it from diverging in the training. In all our experiments, we fix $m = 100$.
- $\mathbb{E}_{x\sim D_U}[\mathcal{F}_{\mathcal{U}}(x)]$ is the expectation of the free-energy function of $\mathbb{U}$ samples. We weight the objective function of $\mathbb{U}$ instances according to their anomalousness probability, computed with the eCDF function. These weights account for the uncertainty of the classification of $\mathbb{U}$ instances.

## 5 Experiments

### 5.1 Dataset Description

**NSL-KDD Dataset** Firstly, we conduct experiments using the NSL-KDD dataset [16], which is one of the most popular datasets used to evaluate network Intrusion Detection Systems (IDSs). Two distinct subsets are provided: the training subset contains 125 973 records and the test subset has 22 544 records. Each data point is represented by 41 features extracted from the network traffic, e.g. the duration of the flow, the TCP flags; and labeled as normal or anomalous. 39

types of network attacks are present in this dataset, ranging from Denial of Services (DoS), to Probe attacks. To investigate algorithm sensitivity with respect to the ratio of anomaly contamination, we vary the anomaly percentage contaminating the training data. We prepare four training subsets, containing 0%, 5%, 10%, and 15% of outliers. These anomalies are selected randomly from all NSL-KDD anomalous training instances. Then, we rescale numerical features to be in the range [0,1], using the min-max normalization method, and categorical features are one-hot encoded.

**MedBIoT Dataset** The MedBIoT dataset [10] is a recent public dataset that contains the network traffic collected from a large network containing 83 real and emulated IoT devices. These devices belong to four categories: switches, light bulbs, locks, and fans. To generate malicious traffic, the authors executed three prominent malware attacks: Mirai, Bashlite, Torii attacks, and labeled the collected training data accordingly. In overall, 17 million network packets were collected: 30% of this traffic is anomalous and the remaining 70% is benign. 61 flow-based features are extracted from the traffic, e.g., flow duration, number and length of packets per flow. A detailed description of each extracted feature is available in [2]. We randomly split the benign data into 60% for the training, 20% for the validation, and 20% for testing. Similarly to NSL-KDD experiments, we prepare four training datasets with different contamination ratios 0%, 5%, 10%, and 15%. Finally, categorical features are encoded using Count Encoder and numerical features are normalized using the Min-Max normalization method.

### 5.2 Competing Methods

We compare our approach, GRAnD, against unsupervised AD methods frequently used in the literature: OCSVM with a Gaussian kernel, Isolation Forest (IF), vanilla VAE, vanilla Planar Flow (vanilla PF), Deep Autoencoding Gaussian Mixture Model (DAGMM) [24], and RVAE [3]. In line with prior works, performances are assessed using the Area Under the Curve of the Receiver Operating Characteristics (AUROC).

### 5.3 Training Parameter Settings

In all experiments, we use the standard Feedforward Neural Network (FNN) architectures for all autoencoders. In NSL-KDD experiments, the autoencoders are composed of a 3-layer MLP with 122-8-122 units. In MedBIoT experiments, the autoencoders are a 5-layer MLP with 61-32-16-32-61 units. All latent layers are followed by ReLU activation function. The last layer of the decoder is followed by a sigmoid function. We use an adaptive learning rate: initially, we use a learning rate of 0.001, which is divided by two if the training loss does not decrease after 20 consecutive epochs. We stop the training when the learning rate is lower than $10^{-6}$ or the number of epochs becomes higher than 500 epochs. We use a batch size of 256 in all experiments. We initialize model parameters

randomly. To limit the impact of random parameter initialization, we repeat each experiment five times and average the results over these five runs.

Our approach comprises three specific hyperparameters: the rejection parameter $\alpha$ that controls the initial threshold $u$, the risk parameter $q$, and the margin $m$. In all experiments, $q$ is fixed to 0.001, $\alpha$ to 1.5, and $m$ to 100. We conduct a sensitivity analysis experiment in Section 5.5, to assess our approach robustness regarding the hyperparameters. We use grid search to select competing methods optimal hyperparameters, which maximize their AUROC on the validation subset. The experiments were run on a laptop equipped with a 12-core Intel i7-9850H CPU clocked at 2.6GHz and with NVIDIA Quadro P2000 GPU.

### 5.4 Experimental Results and Discussion

**NSL-KDD Experimental Results** We show in Figure 3 the results of the comparison between GRAnD and other competing methods on the NSL-KDD. When the training data are contaminated with anomalies, our approach significantly outperforms competing methods. While the performance of competing methods decreases with higher pollution ratios $\gamma_p$, our approach is more stable, with an average AUROC around 94% and very little deviation, for the three contamination ratios 5%, 10%, and 15%. These results mainly highlight the benefit of the robust rejection strategy, where no prior knowledge about the outlier ratio is required in advance.

When the training data are anomaly-free, GRAnD performance slightly degrades, with an AUROC of 92.6% with a standard deviation of 0.8%. This observation can be explained by the fact that GRAnD leverages training outliers to learn a robust projection, where inliers are well reconstructed, while outliers are poorly reconstructed. When training data do not contain anomalies, GRAnD-PF and GRAnD-VAE performances are very similar to vanilla-PF and vanilla-VAE, respectively. Despite this slight increase, GRAnD remains very competitive, with around 6% points better AUROC than IF. Finally, for all contamination ratios, GRAnD-PF slightly outperforms GRAnD-VAE.

**MedBIoT Experimental Results** We present the MedBIoT results in Figure 4. As mentioned previously, we train an anomaly detector for each device type. We obtain similar results for the four device types. Due to space constraints, we report the most representative results in Figure 4. For the four device types, and for all contamination ratios, GRAnD-PF, GRAnD-VAE, and RVAE outperform other anomaly detectors, with an AUROC of $99.9 \pm 0.1\%$. In particular, we highlight the robustness of our contribution compared to vanilla-VAE and vanilla-PF. While the latter performances are considerably impacted when the contamination ratio is higher than 10%, GRAnD yields stable results. For example, for $\gamma_p = 10\%$, GRAnD-PF and GRAnD-VAE exceed vanilla-VAE and vanilla-PF AUROCs by 19% and 23%, on average. Consequently, the robustness of GRAnD for IoT network traffic anomaly detection is validated on this dataset. Although GRAnD and RVAE yield close results, we will show in the next section that, unlike RVAE, GRAnD is robust to the hyperparameter selection.

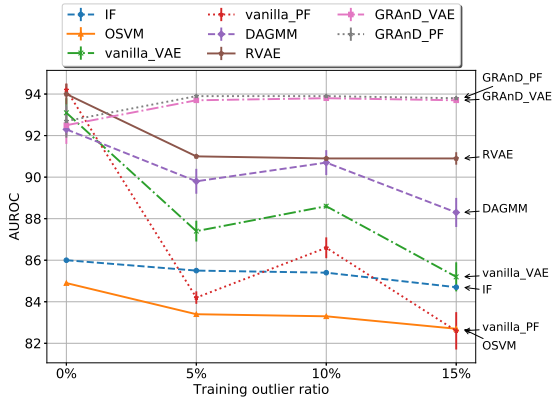
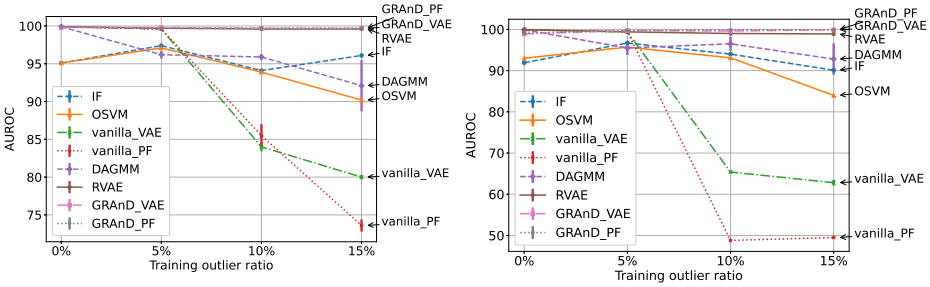



Fig. 3: NSL-KDD experimental results: comparison of AD methods based on average AUROCs and deviations over five runs for multiple contamination ratios.




(a) Results for the device category fan

(b) Results for the device category lock

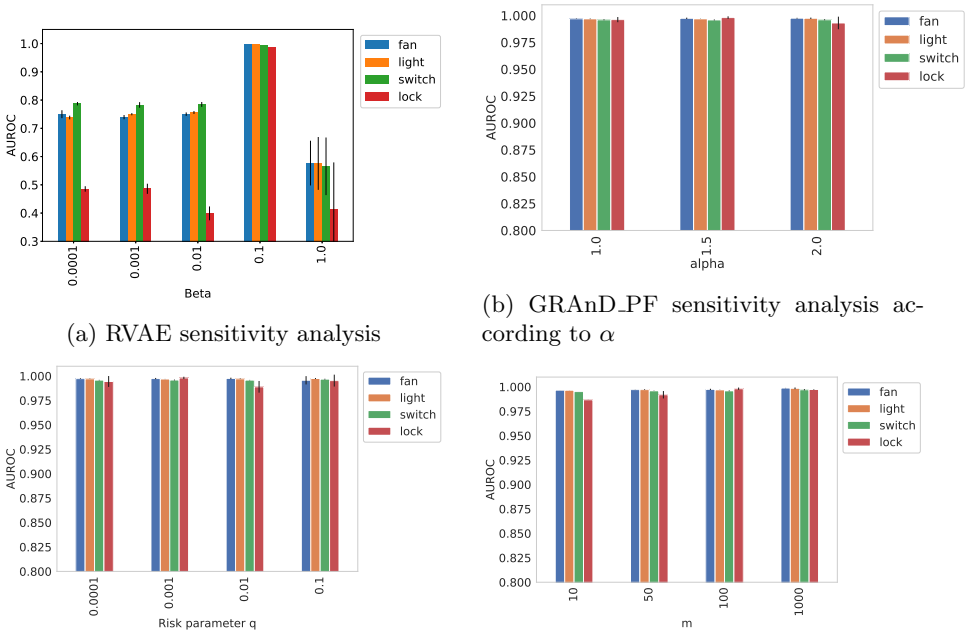
Fig. 4: The MedBIoT experimental results, for the the device categories fan and lock. We report the avgerage AUROC with the standard variation over five runs.

### 5.5 Sensitivity Analysis

As mentioned in numerous works in the anomaly detection community, it is advocated to develop robust anomaly detectors that do not depend on user-defined parameters. The sensitivity to hyperparameters is problematic in unsupervised AD, since outlier labels are scarce, and the selection of the optimal hyperparameters is not guaranteed. We conduct further experiments to assess the sensitivity of our approach regarding its hyperparameters. We train different models with distinct hyperparameters to study the variation of the performance on the same test subset. Due to space constraints, we report in Figure 5 the results of the sensibility analysis of RVAE and GRAnD-PF on the MedBIoT dataset, with $\gamma_p = 10\%$.

In Figure 5a, we show RVAE performance for different $\beta \in \{0.0001, 0.001, 0.01, 0.1, 1\}$. Since GRAnD is defined using three hyperparameters, $m$, $q$, and $\alpha$, we run three experiments, where we only vary one hyperparameter and we keep the remaining ones fixed. Figure 5a shows that RVAE is sensitive to the hyperpa-

rameter $\beta$. For all device types, RVAE AUROC drastically decreases, when $\beta$ changes. In contrast, GRAnD-PF performance is not impacted by the variation of its hyperparameters, and the AUROC is stable around $99.8 \pm 0.1\%$.




(a) RVAE sensitivity analysis

(b) GRAnD_PF sensitivity analysis according to $\alpha$

(c) GRAnD_PF sensitivity analysis according to $q$

(d) GRAnD_PF sensitivity analysis according to $m$

Fig. 5: The sensitivity analysis of RVAE and GRAnD_PF on MedBIoT dataset. We report the average AUROC with the standard variation over five runs.

## 6 Conclusion and Future Work

In this paper, we proposed GRAnD, a robust generative method for unsupervised anomaly detection. Our approach uses Extreme Value Theory to filter out outliers contaminating the data, and learns a robust representation, where inliers can be accurately reconstructed, while outlier reconstructions are corrupted. Extensive experiments were conducted on benchmark datasets, and showed that our approach outperforms classical anomaly detection methods, all the while showing an outstanding robustness to hyperparameter selection. In the future, we will extend GRAnD to detect anomalies in time-series and sequential data. We will adapt our rejection strategy to detect contextual and collective sequential anomalies. Finally, since our contribution involves a minimal change to the underlying model architecture, future studies could fruitfully explore other generative models, such as adversarial autoencoders and GANs.

## References


1. Malware and network attacks in 2019, https://www.helpnetsecurity.com/2019/12/13/network-attacks-2019/
2. NFStream - a Network Data Analysis Framework., https://nfstream.org/
3. Akrami, H., Joshi, A.A., Li, J., Aydore, S., Leahy, R.M.: Robust Variational Autoencoder. CoRR (2020)
4. An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability (2015)
5. Balkema, A.A., Haan, L.d.: Residual Life Time at Great Age. The Annals of Probability **2**(5), 792 – 804 (1974)
6. Breunig, M., Kriegel, H., Ng, R., Sander, J.: Lof: Identifying density-based local outliers (2000)
7. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys **41**, 1–58 (2009)
8. Desforges, M.J., Jacob, P.J., Cooper, J.E.: Applications of probability density estimation to the detection of abnormal conditions in engineering. Proceedings of the Institution of Mechanical Engineers **212**(8), 687–703 (1998)
9. Fernandes, G., Rodrigues, J.J.P.C., Carvalho, L.F., Al-Muhtadi, J.F., Proença, M.L.: A comprehensive survey on network anomaly detection. Telecommunication Systems **70**(3), 447–489 (Mar 2019)
10. Guerra-Manzanares, A., Medina-Galindo, J., Bahsi, H., Nõmm, S.: MedBIoT: Generation of an IoT Botnet Dataset in a Medium-sized IoT Network:. In: ICISSP. pp. 207–218. Valletta, Malta (2020)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. CoRR (2014)
12. Kotani, G., Sekiya, Y.: Unsupervised Scanning Behavior Detection Based on Distribution of Network Traffic Features Using Robust Autoencoders. In: (ICDMW). pp. 35–38 (2018)
13. Kriegel, H.P., Schubert, M., Zimek, A.: Angle-based outlier detection in high-dimensional data. In: ACM SIGKDD. p. 444–452. KDD '08 (2008)
14. Lai, C.H., Zou, D., Lerman, G.: Robust subspace recovery layer for unsupervised anomaly detection. In: ICLR (2020)
15. Lerman, G., Maunu, T.: An overview of robust subspace recovery (2018)
16. NSL-KDD: Canadian Institute for Cybersecurity | UNB
17. Rezende, D.J., Mohamed, S.: Variational inference with normalizing flows. In: ICML (2015)
18. Ringberg, H., Soule, A., Rexford, J., Diot, C.: Sensitivity of pca for traffic anomaly detection. In: ACM SIGMETRICS. p. 109–120 (2007)
19. Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C.: Support Vector Method for Novelty Detection p. 7
20. Siffer, A., Fouque, P.A., Termier, A., Largouet, C.: Anomaly Detection in Streams with Extreme Value Theory. In: ACM SIGKDD. pp. 1067–1075 (2017)
21. Tax, D.M., Duin, R.P.: Support Vector Data Description. Machine Learning **54**, 45–66 (2004)
22. Yang, X., Latecki, L.J., Pokrajac, D.: Outlier Detection with Globally Optimal Exemplar-Based GMM. In: SDM. pp. 145–154 (2009)
23. Zhou, C., Paffenroth, R.C.: Anomaly Detection with Robust Deep Autoencoders. In: ACM SIGKDD - KDD '17 (2017)
24. Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep Autoencoding Gaussian Mixture Model For Unsupervised Anomaly Detection. ICLR p. 19 (2018)