HETEROGENEOUS LOSS FUNCTION WITH AGGRESSIVE REJECTION FOR CONTAMINATED DATA IN ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A clean training dataset, which consists of only normal data, is crucial for detecting anomalous data. However, the clean dataset is challenging to produce in practice. Here, a heterogeneous loss function with aggressive rejection is proposed, which strengthens robustness against contamination. Aggressive rejection constrains training on the potential anomalies which is the intersection of normal and abnormal distributions. Heterogeneous loss function applies a mini-batch stochastic choice of an asymptotic polynomial to a generalized robust loss function, which dynamically optimizes the gradient for the intersection further. Through the proposed method, mean square error based models can outperform various robust loss functions and generate comparable performance with robust models for contaminated data.

1 INTRODUCTION

Identifying outliers or abnormalities in data is known as anomaly detection (AD) (Chandola et al., 2009). AD assumes that a model is trained on a clean dataset that consists of only normal data so that the output contains normal features. Autoencoder (AE) (Hinton et al., 2006; Bergmann et al., 2018) is a representative model whose output is generated by an encoder and a decoder. The encoder converts the input into a latent vector, which is then reconstructed by the decoder as the original one. Due to the clean dataset, the output of AE takes on normal features. As a result, the reconstruction error (the difference between input and output) should be close to zero for normal input, whereas high for abnormal input. However, it is challenging to produce clean datasets because of the ambiguity between normal and abnormal data. Moreover, the ground truth depends on the individual. Contaminated data is produced when training data are labeled as normal, although data include both normal and abnormal. It is essential to make the model robust to contamination because it impairs performance.

To address contamination, a few anomalies were postulated and excluded from the training data (Beggel et al., 2019). However, it is challenging to access and eliminate overall anomalies due to the normal-like ones. Both classification and regression have related methods for the problem. Classification approaches utilized ground truth (Xu et al., 2019; Pleiss et al., 2020) or combined cross-entropy loss function and mean absolute error (MAE) (Englesson & Azizpour, 2021; Zhang & Sabuncu, 2018). In the case of regression, robust loss functions were used, such as pseudo-Huber loss (Huber, 1992). The classification approaches, which rely on ground truth or cross-entropy loss, cannot be extended to the field of AD. AD can utilize the regression approaches, but they are ineffective since the training data still contains anomalies enough to interrupt the assumption.

In this paper, a novel robust loss function, heterogeneous loss function with aggressive rejection, is proposed to make detectors robust against contaminated data. Aggressive rejection removes a large amount of data to handle the anomalies as much as possible. Heterogeneous loss function is based on a general and adaptive robust loss function (GA) (Barron, 2019) that generalizes loss function from mean square error (MSE) to Welsch loss (Dennis Jr & Welsch, 1978). The proposed loss function dynamically adjusts the gradient to utilize the normal samples as much as possible while impacting less for abnormal samples.

The experiments showed that heterogeneous loss function with aggressive rejection outperformed the existing robust loss functions and models. In addition, the heterogeneous loss function can be extended to MSE-based AD models, as demonstrated in section 5.3.

2 Related Work

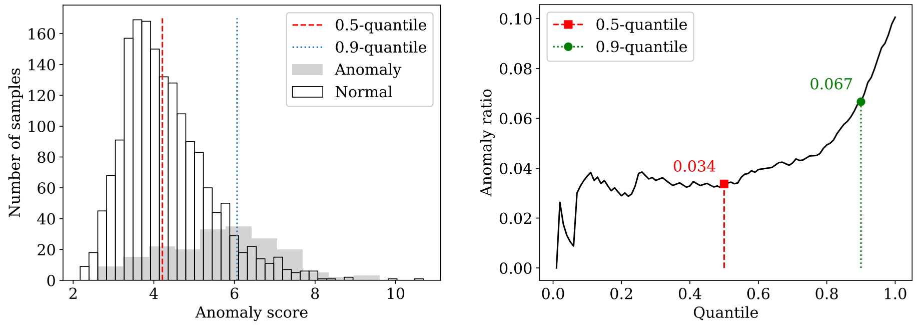
In the field of AD, deep learning has improved performance significantly. Initially, reconstruction error based models such as AE, variational autoencoder (VAE) (Kingma & Welling, 2013), and adversarial autoencoder (AAE) (Makhzani et al., 2015) were used. Since AD assumes a clean dataset, the normal input results in a low reconstruction error, whereas the anomaly results in a high reconstruction error. Memory-augmented autoencoder (MemAE) (Gong et al., 2019) added a memory module to convert the input latent vector to the most relevant latent vector in memory. Deep support vector data description (DSVDD) (Ruff et al., 2018) makes the latent vector of the training data close to the center vector. The score function is measured by the distance between the center vector and the input latent vector. The performance of AD has been increased by using contrastive loss (Tack et al., 2020; Reiss & Hoshen, 2021), outlier exposure (Hendrycks et al., 2018), and a few ground truths (Ruff et al., 2019). These methods are sensitive to contaminated data due to the assumption.

Contaminated data is produced by difficulty or mistake in labeling. The contaminated data makes AD models difficult to detect anomalies by training both normal and abnormal features. The problem has already been addressed in the field of classification and regression. For classification, the mislabeled data are prevented by removing them during or after training or robust loss function. Pleiss et al. (2020) identified the mislabeled data based on the entire predictive results generated in the training process. Englesson & Azizpour (2021) and Zhang & Sabuncu (2018) proposed a loss function that combines MAE, a slow but noise robust loss function, and the opposite cross-entropy loss function. For regression, loss functions such as pseudo-Huber (Huber, 1992), Geman-McClure (Ganan & McClure, 1985), and Cauchy (Black & Anandan, 1996) were employed to reduce the influence of the outliers. A general robust loss function (Barron, 2019) was proposed, generalizing various robust loss functions. The classification methods require ground truth or cross-entropy, and the regression methods still have anomalies in training data. Thus, they are not appropriate for solving the contamination problem.

The previous approaches designed robust models or loss functions with data refinement. Robust variational autoencoder with attention based feature adaptation (RVAE-ABFA) (Gao et al., 2020) is based on deep autoencoding gaussian mixture model (DAGMM) (Zong et al., 2018). They achieved robustness by replacing AE with VAE and adjusting the weight between latent vector and reconstruction error via attention based feature adaption. The strategies for data refinement eliminate samples identified as anomalies (Yoon et al., 2021; Görnitz et al., 2014; Xia et al., 2015). Iterative training set refinement (ITSR) (Beggel et al., 2019) applied one-class support vector machine (OC-SVM) (Schölkopf et al., 1999) to a latent vector of AAE for refinement during training. Normality-calibrated autoencoder (NCAE) (Yu et al., 2021) generated high-confident normal samples in the low entropy space and utilized them for predicting anomalies. Latent outlier exposure (LOE) (Qiu et al., 2022) defined normal and abnormal loss functions. The normal loss function is used for the data discriminated as normal, and the abnormal loss function, such as the reciprocal of the normal loss function, is used for others. They assumed that high differences between normal and abnormal loss functions are anomalies, and 10% of data are treated as anomalies. Pseudo-Huber loss was employed by Liznerski et al. (2020). Since AD takes a clean dataset or a few contaminated data, the refinement strategies set a contamination ratio around 10%, which keeps the normal samples and only removes the high-confident anomalies.

3 Motivation

The previous robust loss functions try to inhibit training on anomalies. For instance, pseudo-Huber loss (Liznerski et al., 2020) reduces the gradient of high loss, while maintaining the same gradient to MSE for low loss. LOE (Qiu et al., 2022) assumes 10% anomalies in datasets, and trains to maximize the difference between normal and potential anomalies. However, pseudo-Huber loss does not eliminate the potential anomalies and LOE assumes a low contamination ratio. The assumption of a 10% contamination ratio, as shown in Figure 1a, which depicts the distribution of anomaly



(a) Distribution of anomaly score on test data.

(b) Anomaly ratio for different quantiles.

Figure 1: (a) Observation of anomaly scores on 10% contaminated train data. (b) Motivation of aggressive rejection. AE is trained on MNIST dataset (LeCun et al., 2010). The samples over the 0.9-quantile contain more anomalies than normal, but 6.7% of the total data still contain anomalies. Although the samples that are over 0.5-quantile are more normal than the 0.9-quantile case, its anomaly ratio is about 0.034 (3.4%).

scores cannot handle the entire anomalies due to the similar feature between normal and abnormal. However, more anomalies can be managed with a 50% contamination ratio as seen in Figure 1b. Therefore, we introduce a rejection method that assumes 50% of data are potential anomalies, and a heterogeneous loss function to address the problem that eliminates numerous normal samples through gradient adjustment based on anomaly score.

4 Proposed Method

In this section, heterogeneous loss function and aggressive rejection are described. The ambiguous data which intersects the normal and abnormal distributions are rejected. In addition, various loss functions are used for each normal and abnormal sample.

4.1 Rejection

Aggressive Rejection. The previous approaches roughly set 10% contamination ratio (Beggel et al., 2019; Qiu et al., 2022), thus they cannot handle all the anomalies since abnormal distribution overlaps normal distribution as Figure 1a. Moreover, the assumption of contamination ratio limits the performance when the contamination ratio is over 10%. Figure 1b shows that the anomaly ratio decreases as more data are removed. Based on this concept, aggressive rejection removes the anomalies at the expense of a significant amount of normal samples. The formulation of aggressive rejection is given as follows:

$$\begin{gathered} L(x_i, w_i) = w_i L_{MSE}(x_i) \\ w_i = \begin{cases} 0 & \text{if } s_i > s_q \\ 1 & \text{otherwise} \end{cases} \\ L_{MSE}(x_i) = ||x_i - f(x_i)||^2 \end{gathered} \tag{1}$$

where x_i is the training data, w_i is the weight for aggressive rejection, s_i is the i-th anomaly score, s_q is its q-quantile, and $f(\cdot)$ is a model such as AE. Aggressive rejection removes the data in which an anomaly score (defined for each model, e.g., reconstruction error) is higher than s_q. Since q increases monotonically at 0.5 in Figure 1b, q is set to 0.5. The experiments on the q setting can be seen in Appendix A.

Soft Rejection. Although aggressive rejection removes the potential anomalies, it also removes a large amount of normal data, which causes performance degradation, especially on a clean dataset. To address this problem, the rejection weight w_i is adjusted partially by a hyperparameter t_s. As in equation 2, w_i depends on t_s, where $t_s = [0, 1]$. “Hard rejection” in which $t_s = 0$ excludes the rejection target completely, whereas "soft rejection" in which $t_s = (0, 1]$ is trained partially. It is important to set an appropriate value because a low t_s lowers performance on a clean dataset, whereas a high one reduces robustness. Models achieve minimal loss on a clean dataset and robustness on a contaminated dataset when t_s is set to 0.1. (see Appendix B).

$$w_i = \begin{cases} t_s & \text{if } s_i > s_q \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

4.2 Heterogeneous Loss

LOE (Qiu et al., 2022) contains a few normal in the potential anomalies due to the 10% contamination ratio. However, aggressive rejection contains a lot of normal samples because it treats half of the data as potential anomalies. It causes problems for both clean and contaminated datasets.

To address the problem, a new loss function should satisfy two conditions. First, it should consider the anomaly score. The number of normal samples increases as the anomaly score decreases, as shown in Figure 1a. Because there are more normal samples than abnormal samples in low anomaly scores, the loss function should encourage models to be trained well. Second, the loss function should not minimize the abnormal loss function, which is the reciprocal of the normal loss function like LOE, whereas it still produces fast convergence for normal and slow convergence for abnormal. The training is considerably disrupted when the reciprocal of the normal loss function is minimized. As a result, the robust loss function with a lower gradient than MSE should be used as the abnormal loss function. To satisfy these conditions, a novel loss function named heterogeneous loss function is proposed, which adjusts the gradient based on the anomaly score for each potential anomaly. Samples with a high anomaly score use a loss function that is close to MSE, whereas samples with a low anomaly score use a loss function that is close to a robust loss function.

Heterogeneous loss is based on GA loss, a generalized loss function that covers from MSE to Welsch loss by a parameter $\alpha = [-\infty, \infty]$ as in equation equation 3 (the formulation is rewritten slightly since the reconstruction error is used as input). c is the point where various loss functions have similar gradients.

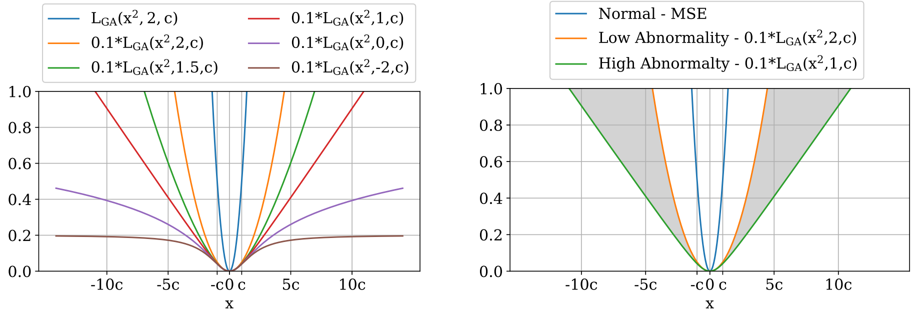
$$L_{GA}(x^2, \alpha, c) = \begin{cases} 0.5x^2/c^2 & \text{if } \alpha = 2 \\ \sqrt{x^2/c^2 + 1} - 1 & \text{if } \alpha = 1 \\ \frac{|\alpha-2|}{\alpha}\left(\left(\frac{x^2/c^2}{|\alpha-2|} + 1\right)^{\alpha/2} - 1\right) & \text{otherwise} \end{cases} \tag{3}$$

Then, modified z-score (Rousseeuw & Croux, 1993) is utilized to project the anomaly score to gradient parameter α. The formulation of the modified z-score is given as follows:

$$\begin{aligned} z_i &= \frac{0.6745(x_i - \hat{x})}{\text{MAD}_i} \\ \text{MAD}_i &= \text{median}_{i \in 1,\ldots,N}(|x_i - \hat{x}|) \\ \hat{x} &= \text{median}_{i \in 1,\ldots,N}(x_i) \end{aligned} \tag{4}$$

The modified z-score is robust against outliers since it is based on the median. In addition, the modified z-score generates a normal distribution and has a relative distance. The modified z-score z is normalized by a maximum between 3.5 and max($|z|$), where 3.5 is the outlier threshold. When m is defined only by max($|z|$), the normalized score ranges from 0 to 1, even if the variance is low. 3.5 is used instead of the low max($|z|$) to make the normalized z close to zero when the variance is low. As the boundary value, 3.5 is utilized to increase the convergence. The normalized z is converted to α by the equation 5, where s_i is the anomaly score, z_i is the modified z-score, and α_r is the parameter for the lowest gradient loss. The minimum of z_i is matched to MSE (normal loss function), and the maximum of z_i is closed to the robust loss function, GA loss with $\alpha = \alpha_r$. Thus, α ranges from α_r to 2. Since α determines the lower boundary of the gradient in the loss function, it is critical to use proper values. In this paper, α is set to 1.5, the median between MSE and pseudo-Huber (see Appendix C). z is applied in the form of a quadratic function because the anomaly ratio increases as a quadratic function (see Figure 1b).

$$\begin{aligned} \alpha_i &= \begin{cases} 2 - (2 - \alpha_r) * (z_i/m)^2 & \text{if } z_i > 0 \\ 2 & \text{otherwise} \end{cases} \\ m &= \max(3.5, \max(|z|)) \end{aligned} \tag{5}$$



(a) Heterogeneous loss function

(b) Example of heterogeneous loss function.

Figure 2: (a) Heterogeneous loss function with soft rejection. It shows loss functions on varying α when $t_s = 0.1$. (b) example of heterogeneous loss function with soft rejection when $t_s = 0.1$, $\alpha_r = 1$ and $c = \sqrt{0.5}$.

Figure 2 illustrates the heterogeneous loss function with soft rejection when $t_s = 0.1$ and $c = \sqrt{0.5}$. The gradient in Figure 2a decreases as the α gets smaller. Figure 2b is an example of heterogeneous loss function with $t_s = 0.1$ and $\alpha_r = 1$. The potential anomalies use the loss function between MSE and pseudo-Huber loss as in the gray area of Figure 2b. The lowest abnormality, where z is close to 0 (0.5-quantile), uses MSE with soft rejection, whereas the largest abnormality, which is at the tail of the z distribution, uses pseudo-Huber loss with soft rejection.

Algorithm 1 summarizes heterogeneous loss function with aggressive rejection. The samples are discriminated as potential anomalies when the anomaly score of model s is larger than s_q, which is q-quantile of s. For the normal samples, w_i is set to 1, whereas the others are set to t_s. The anomaly score in the mini-batch determines the gradient parameters α. The anomaly score is converted to α based on the modified z-score and input α_r. Finally, the model parameters are updated with the loss generated by multiplying soft rejection weight w and loss function L_{GA}.

Algorithm 1 Training with heterogeneous loss function and aggressive rejection

```
Input: Sample $\boldsymbol{X}$, model $f$, hyperparamter $q$, $t_s$, $\alpha_r$, $c$
foreach $Epoch$ do
    foreach $Mini\text{-}batch$ $\boldsymbol{x} \subseteq \mathbf{X}$ do
        $\boldsymbol{s} = ||\boldsymbol{x} - f(\boldsymbol{x})||_2^2$
        $s_q = q$-quantile of $\boldsymbol{s}$
        $\boldsymbol{w} = \begin{cases} t_s & \text{if } s_i > s_q \\ 1 & \text{otherwise} \end{cases}$        ▷ Equation 2
        $\boldsymbol{z} = \text{Modified z-score}(\boldsymbol{s})$        ▷ Equation 4
        $m = \max(3.5, \max(|\boldsymbol{z}|))$
        $\boldsymbol{\alpha} = \begin{cases} 2 - (2 - \alpha_r) * (z_i/m)^2 & \text{if } z_i > 0 \\ 2 & \text{otherwise} \end{cases}$        ▷ Equation 5
        $L = \boldsymbol{w} * L_{GA}(\boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{c})$        ▷ Equation 3
        Update model parameters with $L$
    end for
end for
```

5 Evaluation

This section compares existing robust models and loss functions to heterogeneous loss function with aggressive rejection. Three fundamental image datasets-MNIST (LeCun et al., 2010), FashionMNIST (F-MNIST) (Xiao et al., 2017), and CIFAR-10 (Krizhevsky, 2009)-are used to evaluate the methods. MNIST and F-MNIST consist of 10 classes and 28×28 gray scale images. CIFAR-10

Table 1: AUROC of various robust loss functions

		MNIST		**F-MNIST**		**CIFAR-10**	
Model	**Loss**	**0%**	**20%**	**0%**	**20%**	**0%**	**20%**
AE	MSE	**0.931**	0.796	**0.891**	0.783	0.561	0.540
	pseudo-Huber	0.913	0.783	0.882	0.793	**0.564**	0.544
	GA	0.912	0.779	0.879	0.783	0.561	0.541
	LOE	0.913	0.784	0.881	0.789	0.563	0.542
	Hetero	0.928	**0.849**	0.887	**0.830**	**0.564**	**0.558**
MemAE	MSE	**0.933**	0.770	**0.898**	0.776	0.570	0.547
	pseudo-Huber	0.930	0.815	0.892	0.775	0.572	0.549
	GA	0.929	0.813	0.892	0.770	0.570	0.546
	LOE	0.926	0.819	0.891	0.775	0.572	0.548
	Hetero	0.932	**0.855**	**0.898**	**0.836**	**0.578**	**0.555**
DSVDD	MSE	0.928	0.826	0.916	0.806	0.605	0.566
	pseudo-Huber	**0.931**	0.826	0.917	0.842	0.596	0.567
	GA	**0.931**	0.817	0.917	0.829	0.601	0.567
	LOE	0.930	0.847	**0.918**	0.852	**0.606**	0.568
	Hetero	0.927	**0.870**	0.906	**0.854**	0.601	**0.575**

is made up of 32×32 color images with 10 classes. For general applicability, the experiments on tabular datasets (Rayana, 2016; Dua & Graff, 2017) are reported in Appendix F.

5.1 Datasets and setups

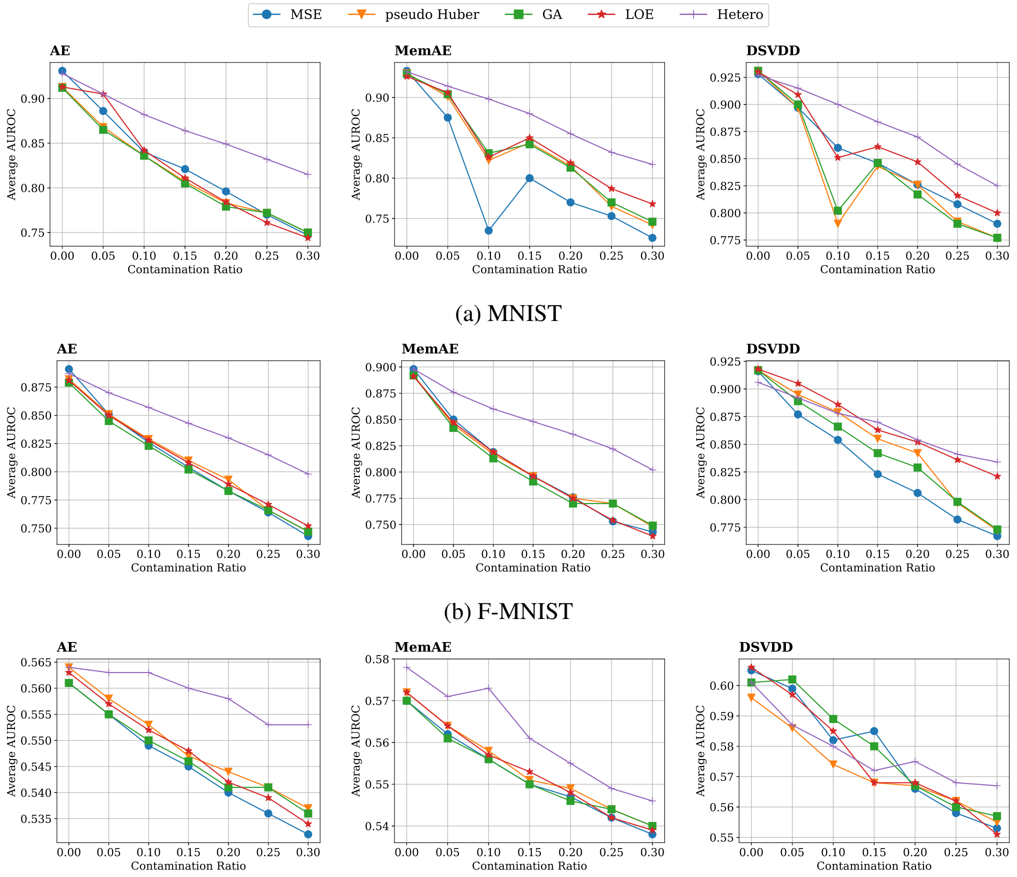
One vs. rest setup which set one class as normal and the other classes as abnormal was used in the experiments. For normal data, training data is twice as much as test data, and 10% of the original training data was used as validation. $\gamma_{ct}/(1-\gamma_{ct}) * N$ number of abnormal data were added, where γ_{ct} was the contamination ratio and N was the number of normal data. 30% of the test data consisted of anomalies. The results without the constraint (using the entire training and test data) are reported in Appendix E. The model with the lowest validation loss was taken as the test model. The validation loss for the proposed method is measured by aggressive and hard rejection. The Area Under Receiver Operating Characteristic (AUROC) is used as a metric. The experiments set each class as normal and measured the average AUROC with three different seeds.

5.2 comparison method

ITSR (Beggel et al., 2019), RVAE-ABFA (Gao et al., 2020), and NCAE (Yu et al., 2021) are employed as the robust models. ITSR utilized OC-SVM and AAE for refinement. RVAE-ABFA developed DAGMM (Zong et al., 2018) by adopting VAE and attention based feature adaption. NCAE utilizes normal samples generated from generative adversarial model (Goodfellow et al., 2020) to refine the dataset. The loss functions are evaluated based on three conventional models-AE (Bergmann et al., 2018), MemAE (Gong et al., 2019), and DSVDD (Ruff et al., 2018). AE is employed since it is a conventional model. DSVDD utilizes MSE but is different from reconstruction error. MemAE is compared due to the additional loss function for memory augmented loss. MSE is substituted by pseudo-Huber (Liznerski et al., 2020), GA loss (Barron, 2019), and LOE (Qiu et al., 2022). GA loss was used to demonstrate how much GA loss affects robustness since it utilizes negative log-likelihood to determine α while the proposed loss function utilizes z distribution. Heterogeneous loss function with aggressive rejection is denoted as Hetero. The details of experiments such as neural network architectures and batch size are described in Appendix D.

5.3 Evaluation with robust loss function

The robust loss functions are used to evaluate the robustness of Hetero loss. Table 1 shows the AUROC on clean datasets ($\gamma_{ct} = 0$) and contaminated datasets ($\gamma_{ct} = 0.2$) (the highest AUROC is in bold). The comparison methods on clean datasets show minimum AUROC loss compared to MSE except for AE on MNIST. Hetero loss on clean datasets shows comparable performance to



(a) MNIST

(b) F-MNIST

(c) CIFAR-10

Figure 3: Evaluation with various loss functions depending on contamination ratio. Three models are evaluated on (a) MNIST, (b) F-MNIST, and (c) CIFAR-10.

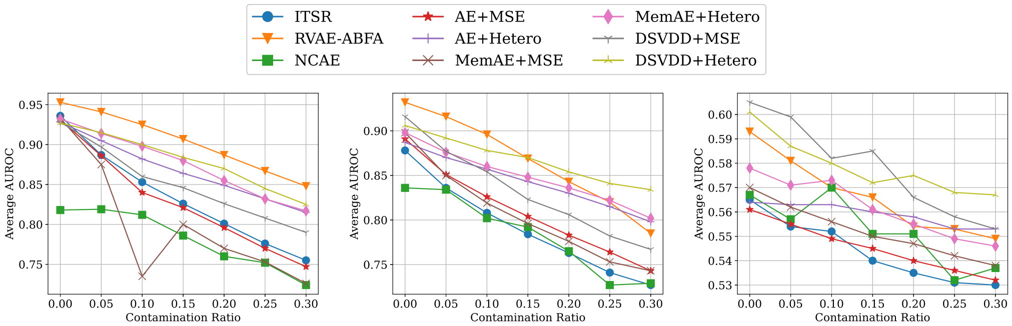
MSE within 0.01. The performance of AE with pseudo-Huber, GA, and LOE does not significantly improve robustness. However, AE with hetero loss achieves over 0.045 robustness. When DSVDD on F-MNIST is trained with robust loss functions, it shows the most effective results. LOE generates a similar performance to Hetero by minimizing the abnormal loss function. Since CIFAR-10 is a hard dataset compared to MNIST or F-MNIST, the AUROC on the clean dataset is low. Therefore, robustness has not improved significantly. Hetero loss surpasses 0.002 0.084 more robustness compared to the overall robust loss functions on 20% contaminated data.

Figure 3 visualizes the AUROC of robust loss functions depending on contamination ratio γ_{ct}. Hetero loss achieves the most robustness in the case of AE and MemAE. In the case of DSVDD on F-MNIST, LOE outperforms hetero when the contamination ratio is under 10%, but hetero loss outperforms LOE when the contamination ratio is over 20%. According to Qiu et al. (2022), LOE performs best when the assumed γ_{ct} is equal to the actual γ_{ct}. LOE achieves robustness by minimizing the abnormal loss function, but the $\gamma_{ct} = 0.1$ assumption limits the robustness. Since DSVDD is not a reconstruction error based model, MSE and GA loss show robustness on CIFAR-10 when the contamination ratio is under 15%. However, hetero loss outperforms the other loss functions when the contamination ratio is 20% and shows a large difference as the contamination ratio increases.

Pseudo-Huber and GA shows insufficient robustness because they simply mitigate the effects of large losses rather than assuming the anomalies and avoiding the training of anomalies. LOE assumes the 10% contamination ratio and minimizing the abnormal loss function, which improves the robustness, but loses the performance on clean dataset. It can exceed the hetero loss in some cases like CIFAR-10 with DSVDD, but it can also cause a significant performance decline on clean

Table 2: AUROC of various robust AD models

Model	MNIST		F-MNIST		CIFAR-10	
	0%	**20%**	**0%**	**20%**	**0%**	**20%**
ITSR	0.936	0.801	0.878	0.763	0.565	0.535
RVAE-ABFA	**0.953**	**0.887**	**0.932**	0.843	0.593	0.554
NCAE	0.818	0.760	0.836	0.765	0.567	0.551
AE	0.931	0.796	0.891	0.783	0.561	0.540
AE + Hetero	0.928	0.849	0.887	0.830	0.564	0.558
MemAE	0.933	0.770	0.898	0.776	0.570	0.547
MemAE + Hetero	0.932	0.855	0.898	0.836	0.578	0.555
DSVDD	0.928	0.826	0.916	0.806	**0.605**	0.566
DSVDD + Hetero	0.927	0.870	0.906	**0.854**	0.601	**0.575**



(a) MNIST (b) F-MNIST (c) CIFAR-10

Figure 4: Comparison with robust models. The models are trained on three datasets-(a) MNIST, (b) F-MNIST, and (c) CIFAR-10.

datasets such as AE on MNIST. Moreover, when the contamination ratio is above 10%, LOE performs worse. Hetero loss outperforms the other robust loss functions by handling a lot of anomalies through aggressive rejection and a mini-batch distribution based gradient adjustment. The experiment demonstrates that hetero loss can be used with a various MSE-based AD models to generate high robustness on contaminated datasets and minimal AUROC loss on clean datasets.

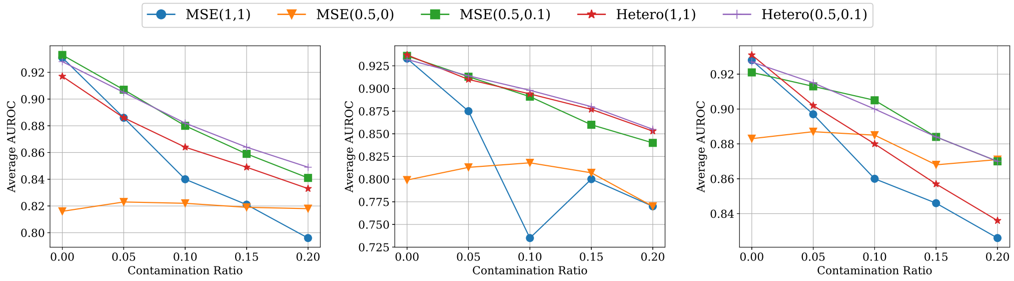
5.4 Evaluation with robust models

The standard models with Hetero loss are compared to robust models. The AUROC of comparison on different contamination ratios is shown in Table 2. NCAE has a low decrease in the AUROC of contamination datasets compared to clean datasets but lacks performance. ITSR and RVAE-ABFA perform more robustness than AE and MemAE. However, hetero loss function improves the robustness of AE and MemAE by about 0.05 compared to ITSR. DSVDD with Hetero loss shows comparable results toward RVAE-ABFA, which shows the most robust results on MNIST. Although RVAE-ABFA shows the highest AUROC compare to the other methods on F-MNIST, DSVDD with hetero loss shows 0.012 higher robustness. Hetero loss achieves the highest robustness on CIFAR-10. It demonstrates that models with hetero loss can achieve comparable results to robust models.

Figure 4 illustrates the robustness of models on varying contamination ratios. On MNIST dataset, RVAE-ABFA has the most robustness, but hetero exhibits the most comparable outcomes. When the contamination ratio is above 20% on F-MNIST dataset, DSVDD with hetero loss outperforms RVAE-ABFA. In the case of CIFAR-10, DSVDD outperforms the RVAE-ABFA, and hetero makes MemAE comparable to RVAE-ABFA. Compared to the existing robust models, Hetero loss achieves high robustness without increasing inference time or modifying the architecture.

Table 3: Ablation Study of Hetero loss

Loss	q	t_s	AE		MemAE		DSVDD	
			0%	20%	0%	20%	0%	20%
MSE	1	1	0.931	0.796	0.933	0.770	0.928	0.826
MSE	0.5	0	0.816	0.818	0.799	0.804	0.883	**0.871**
MSE	0.5	0.1	**0.933**	0.841	0.936	0.840	0.921	0.870
Hetero	1	1	0.917	0.833	**0.937**	0.853	**0.931**	0.836
Hetero	0.5	0.1	0.928	**0.849**	0.932	**0.855**	0.927	0.870



(a) AE

(b) MemAE

(c) DSVDD

Figure 5: Ablation Study of Hetero loss on MNIST dataset-(a) AE, (b) MemAE, and (c) DSVDD.

5.5 Ablation study

In this experiment, each component was removed to measure its effect. The performance on MNIST with clean and contaminated datasets is shown in Table 3 and Figure 5. q and t_s are the key components that determine the effect of soft or hard rejection. The loss function and values are denoted as $L(q, t_s)$, where L is the loss function, q is the q-quantile, and t_s is the degree of rejection.

Effect of aggressive rejection. Hard rejection-MSE(0.5,0) is more robust than the standard loss-MSE(1,1) on the contaminated data. However, hard rejection on a clean dataset shows significant performance degradation of AE (0.115), MemAE (0.134), and DSVDD (0.045). Since it excludes many normal samples, it has an adverse effect under clean datasets. Soft rejection-MSE(0.5,0.1) complements the decline. It shows comparable AUROC to MSE on clean datasets. Since the normal samples in the exclusion are used for training in part, soft rejection improves robustness for AE, MemAE, and DSVDD by 0.045, 0.07, and 0.044, respectively.

Effect of heterogeneous loss. Heterogeneous loss function without aggressive rejection is referred to as Hetero(1,1). It performs better than the baseline MSE loss and demonstrates greater robustness than MemAE with MSE(0.5,0.1). Moreover, Hetero(0.5, 0.1), hetero loss with aggressive rejection, shows 0.008 and 0.15 improvements for AE and MemAE compared to MSE(0.5, 0.1). The experiment demonstrates that the difference between normal and abnormal data can be produced by the gradient adaptation based on the mini-batch distribution.

6 Conclusion

In this paper, heterogeneous loss function with aggressive rejection is introduced. Numerous normal data are involved in the rejection target to handle the overall anomalies and excluded partially via soft rejection. In addition, the mini-batch distribution is transformed into the gradient of general robust loss to suppress the convergence of suspicious outliers. The loss function outperforms the previous robust loss function and shows comparable results to robust models. The experiments proved that the MSE-based models could strengthen robustness against contaminated data by the proposed loss function. In the future, we will utilize the minimization of abnormal loss function to promote the difference between normal and abnormal samples. To complement the minimization, we will apply the dynamic q function based on the abnormality in distribution.

REFERENCES

Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4331–4339, 2019.

Laura Beggel, Michael Pfeiffer, and Bernd Bischl. Robust anomaly detection in images using adversarial autoencoders. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 206–222. Springer, 2019.

Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

Michael J Black and Paul Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding*, 63(1):75–104, 1996.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

John E Dennis Jr and Roy E Welsch. Techniques for nonlinear least squares and robust regression. *Communications in Statistics-simulation and Computation*, 7(4):345–359, 1978.

Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297, 2021.

Stuart Ganan and D McClure. Bayesian image analysis: An application to single photon emission tomography. *Amer. Statist. Assoc*, pp. 12–18, 1985.

Yuda Gao, Bin Shi, Bo Dong, Yan Chen, Lingyun Mi, Zhiping Huang, and Yuanyuan Shi. Rvae-abfa: robust anomaly detection for highdimensional data using variational autoencoder. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 334–339. IEEE, 2020.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Nico Görnitz, Anne Porbadnigk, Alexander Binder, Claudia Sannelli, Mikio Braun, Klaus-Robert Müller, and Marius Kloft. Learning and evaluation in presence of non-iid label noise. In *Artificial Intelligence and Statistics*, pp. 293–302. PMLR, 2014.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pp. 492–518. Springer, 1992.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. *arXiv preprint arXiv:2007.01760*, 2020.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056, 2020.

Chen Qiu, Aodong Li, Marius Kloft, Maja Rudolph, and Stephan Mandt. Latent outlier exposure for anomaly detection with contaminated data. *arXiv preprint arXiv:2202.08088*, 2022.

Shebuti Rayana. Odds library, 2016. URL `http://odds.cs.stonybrook.edu`.

Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. *arXiv preprint arXiv:2106.03844*, 2021.

Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

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. *arXiv preprint arXiv:1906.02694*, 2019.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999.

Tom Shenkar and Lior Wolf. Anomaly detection for tabular data with internal contrastive learning. In *International Conference on Learning Representations*, 2021.

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.

Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1511–1519, 2015.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL `http://arxiv.org/abs/1708.07747`.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. *Advances in neural information processing systems*, 32, 2019.

Jinsung Yoon, Kihyuk Sohn, Chun-Liang Li, Sercan O Arik, Chen-Yu Lee, and Tomas Pfister. Self-trained one-class classification for unsupervised anomaly detection. *arXiv preprint arXiv:2106.06115*, 2021.

Jongmin Yu, Hyeontaek Oh, Minkyung Kim, and Junsik Kim. Normality-calibrated autoencoder for unsupervised anomaly detection on data contamination. *arXiv preprint arXiv:2110.14825*, 2021.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

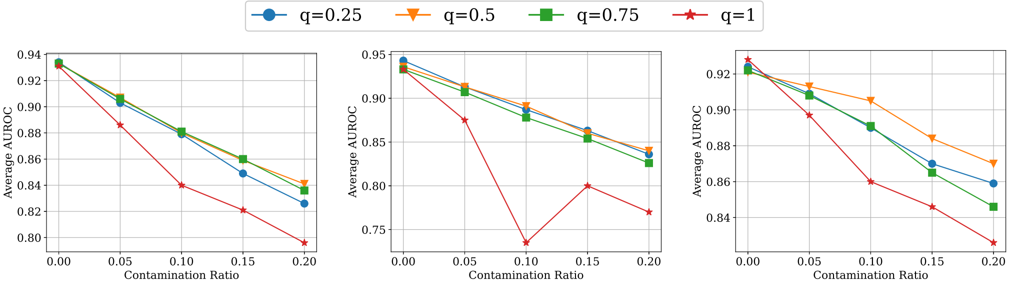
Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.

A QUANTILE IN AGGRESSIVE REJECTION

Table 4 and Figure 6 show the effect of q-quantile on MNIST dataset. Since hetero loss is based on the z-score and is optimum when $q = 0.5$, MSE with soft rejection ($t_s = 0.1$) is employed instead of hetero loss. 0.25-quantile performs better than 0.75-quantile except for AE. Although $q = 0.25$ includes more anomalies in the exclusion, 0.5-quantile rejection performs the best. The tendency is maintained even the contamination ratio increases (see Figure 6). It indicates that over-elimination is not always ensured robustness.

Table 4: Effect of q-quantile on 20% contaminated MNIST dataset.

q	**AE**	**MemAE**	**DSVDD**
0.75	0.836	0.822	0.846
0.5	**0.841**	**0.840**	**0.870**
0.25	0.828	0.836	0.848



(a) AE

(b) MemAE

(c) DSVDD

Figure 6: Ablation Study of Hetero loss on MNIST dataset-(a) AE, (b) MemAE, and (c) DSVDD.

B Studies of soft rejection

Sensitivity Figure 5 illustrates the sensitivity of MSE and hetero loss with soft rejection. As t_s increases, their performance decreases inversely. However, hetero loss reduces the sensitivity of t_s. When $t_s = 1$, the soft rejection becomes a non-robust loss function MSE. Therefore, soft rejection becomes less effective as t_s increases. In contrast to MSE, hetero loss which is still robust due to gradient adaptation exhibits less difference between $t_s = 0.1$ and $t_s = 0.5$ on 20% contaminated data.

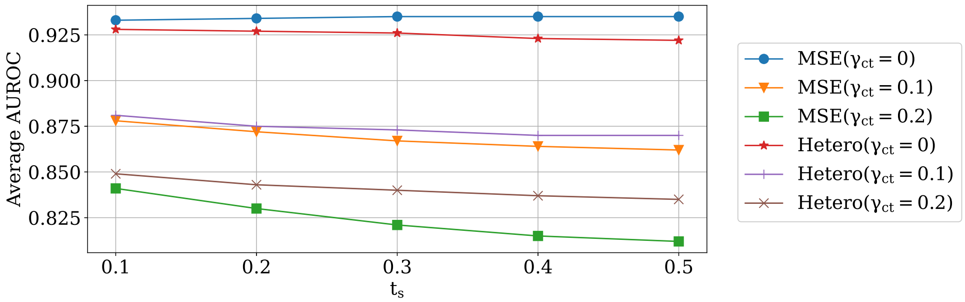
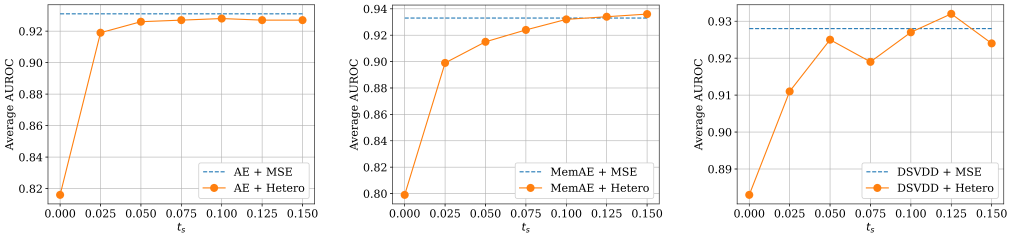


Figure 7: Sensitivity of hetero loss and soft rejection. AE is trained with MSE and hetero loss on MNIST. It describes AUROC for each contamination ratio on varying t_s.

Optimizing AE, MemAE, and DSVDD with hetero loss are trained on clean MNIST ($\gamma_{ct} = 0$) to optimize the t_s. Except for t_s, the parameters are utilized as previously ($q = 0.5$, $\alpha_r = 1.5$ and $c = \sqrt{0.5}$). The dashed line means the standard model with only MSE. AUROC is saturated from 0.05 for AE and 0.1 for MemAE. DSVDD is unstable but similar to MSE when t_s is 0.1. As a result, t_s is set to 0.1 generally, considering minimum performance loss and lowest value.



(a) AE

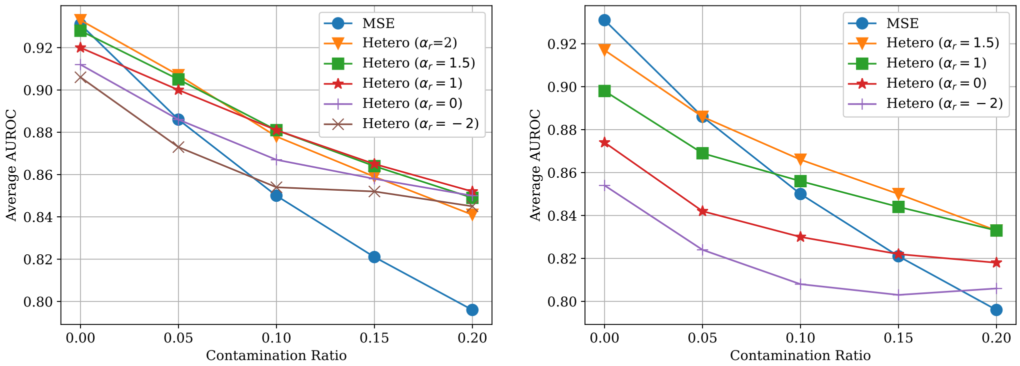
(b) MemAE

(c) DSVDD

Figure 8: Minimum t_s to ensure the AUROC of MSE for each model. Figure shows average AUROC of (a) AE, (b) MemAE and (c) DSVDD.

C Optimizing a Parameter of Heterogeneous Loss

A variety of robust loss functions can be used as the lowest gradient loss function in hetero loss. The pseudo-Huber ($\alpha = 1$), Cauchy ($\alpha = 0$), and Geman-McClure ($\alpha = -2$) are the representative loss functions. Figure 9 illustrates the effect of α_r in hetero loss. The loss function should enhance robustness on contaminated data while minimizing performance degradation on a clean dataset. When α is adjusted below 1, hetero loss degrades on the clean dataset. $\alpha = 1$ shows insufficient performance on the clean dataset, whereas $\alpha = 2$ shows lower robustness. Therefore, the intermediate point $\alpha = 1.5$ is used since it achieves comparable performance on clean and contaminated data.



(a) Loss function with aggressive rejection.

(b) Loss function without aggressive rejection.

Figure 9: Figure shows robustness on varying α_r. (a) and (b) show robustness for each lowest gradient loss function with or without aggressive rejection ($t_s = 0.1$).

D DETAILS ON EXPERIMENTS

Setup Batch size and epochs are set to 100 and 300, respectively, except for ITSR, NCAE, and DSVDD. ITSR refines data for every 10 epochs after the first 100 epochs and is trained on refined data with 100 epochs. NCAE uses 150 epochs and DSVDD uses 150 epochs for pre-training and 100 epochs for the rest. The parameters are updated by Adam optimizer (Kingma & Ba, 2014) with 0.0001 learning rate and 10^{-6} weight decay. As mentioned before, the parameter q, α_r, and c in Hetero loss are set to 0.5, 1.5, and $\sqrt{0.5}$.

AE/MemAE/ITSR The architectures of AE are based on the report in Gong et al. (2019). On MNIST and FashionMNIST datasets, the encoder consist of three convolution modules that consists of convolution, batch normalization (Ioffe & Szegedy, 2015), and leaky ReLU activation (Xu et al., 2015) with 16-32-64 filters that kernel and stride size are 3 and 2. On CIFAR-10 dataset, the encoder consists of four convolution modules, 64-128-128-256 filters that kernel and stride size are 3 and 2.

DSVDD/RVAE-ABFA/NCAE The architectures of autoencoder in DSVDD, RVAE-ABFA, and NCAE are based on the report in Ruff et al. (2018). On MNIST and FashionMNIST datasets, the encoder consists of two convolutions (8×5×5-filters and 4×5×5-filters) and a final fully-connected layers of 32 units. The batch normalization, leaky ReLU, and (2×2)-max-pooling are followed by the convolutions. On CIFAR-10 dataset, the encoder consists of three convolutions (32×5×5-filters, 64×5×5-filters, and 128×5×5-filters) and a final fully-connected layer of 128 units (except for RVAE-ABFA). RVAE-ABFA uses only 32 dimensions due to the computation error. For RVAE-ABFA, the encoder has an additional fully-connected layer for mean and variance of the latent distribution. The bias of layers in DSVDD and NCAE is eliminated due to the trivial solutions as reported in Ruff et al. (2018).

The decoder is symmetric to the encoder, in which convolution is substituted by deconvolutions, and max-pooling is substituted by up-sampling. The last deconvolution has no additional operations such as batch normalization.

E OVERALL DATA RESULTS

The results on MNIST and F-MNIST with overall data can be seen in Table 5. All normal data in training data are used. 90% of them are used for training and the remainder for validation. As mentioned previously, $\gamma_{ct}/(1-\gamma_{ct}) * N$ number of anomalies are added, where γ_{ct} is the contamination ratio and N is the number of normal data. The evaluation makes use of all 10,000 test data. For NCAE, both the reported results by Yu et al. (2021) and the results reproduced using the author's code are shown in the table. The loss function with the highest AUROC is underlined, and the highest AUROC in the comparison is in bold on every dataset.

With the exception of MemAE on MNIST, the existing robust loss functions are insufficient to improve the robustness of AE and MemAE. Comparing DSVDD with hetero loss to MSE, the robustness is increased by 0.043 and 0.048 on each dataset. In the case of DSVDD, LOE decreases performance on clean MNIST and exhibits robustness values that are higher than hetero loss on F-MNIST. RVAE-ABFA and reproduced NCAE achieve greater robustness than in section 5.4 with the entire MNIST data usage. RVAE-ABFA exhibits the best robustness on MNIST (0.918), whereas NCAE, according to reports, exhibits the highest AUROC (0.889) on the 20% contaminated F-MNIST. Among AE, MemeAE, and DSVDD with robust loss functions, DSVDD with hetero loss generally performs best. DSVDD with hetero loss has a large difference (0.06) for MNIST and a low difference (0.028) for F-MNIST when compared to RVAE-ABFA and NCAE, which have the best performance on contaminated data. Since LOE performs better on F-MNIST, the minimization of the abnormal loss function can improve hetero loss when it does not impair the proposed loss function.

Table 5: AUROC with the entire data usage.

		MNIST		**F-MNIST**	
Model	**Loss**	**0%**	**20%**	**0%**	**20%**
ITSR	-	0.939	0.789	0.882	0.754
RVAE-ABFA	-	**0.951**	**0.918**	**0.925**	0.832
NCAE reproduced	-	0.871	0.829	0.806	0.720
NCAE as reported	-	0.940	0.898	0.915	**0.889**
AE	MSE	0.913	0.789	0.889	0.767
	pseudo-Huber	0.911	0.779	0.884	0.773
	GA	0.916	0.783	0.883	0.777
	LOE	0.910	0.788	0.883	0.778
	Hetero	0.919	0.844	0.897	0.818
MemAE	MSE	0.855	0.725	0.895	0.773
	pseudo-Huber	0.913	0.760	0.897	0.785
	GA	0.900	0.773	0.894	0.784
	LOE	0.910	0.798	0.896	0.783
	Hetero	0.906	0.829	0.905	0.828
DSVDD	MSE	0.921	0.815	0.920	0.814
	pseudo-Huber	0.804	0.788	0.918	0.812
	GA	0.800	0.788	0.916	0.815
	LOE	0.821	0.837	0.916	0.870
	Hetero	0.905	0.858	0.921	0.861

F Outlier Detection Datasets

F.1 Setups

Table 6: Information of ODDS. It shows dimension, the number of data (N) and anomalies (A) for each dataset.

Dataset	Dimension	N	A (%)
Arrhythmia	274	452	66 (15%)
Thyroid	36	3772	93 (2.5%)
KDDCUP	120	494,021	97,278 (20%)
KDDCUP-rev	120	121,597	24,319 (20%)

Four tabular datasets are used for general applicability. The information about the datasets is shown in Table 6. The following provides configuration details.

Arrhythmia Arrhythmia dataset can be obtained from Rayana (2016). Minority classes (3, 4, 5, 7, 8, 9, 14, and 15) are defined as abnormal class, whereas the others are defined as normal class. Three fully connected layers with 128-64-32 units are employed for autoencoder architecture. The layers except for the final layer in encoder and decoder are followed by batch normalization and leaky ReLU activation. The batch size is set to 64.

Thyroid Thyroid dataset also can be obtained from Rayana (2016). There are three classes and a minority class (hyperfunction) is defined as an abnormal class. The architecture for Thyroid has a similar architecture that was used on Arrhythmia dataset but has 32-16-4 units. The batch size is set to 512.

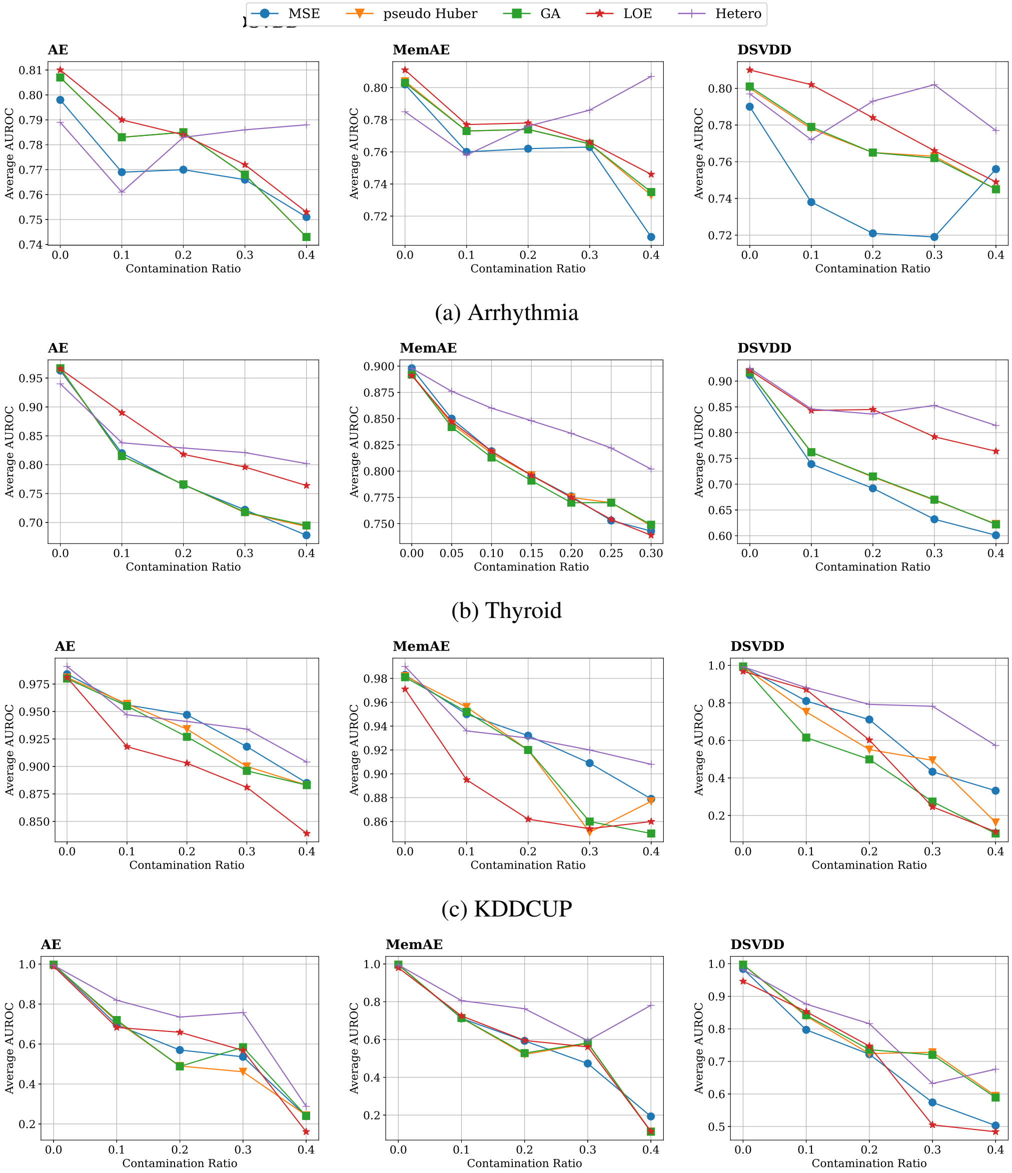
KDDCUP KDDCUP99 10 percent dataset is taken from Dua & Graff (2017). Originally, there were 41 dimensions, where 34 of which were continuous and the rest were categorical. Since the categorical features are transformed by one-hot encoding, the final dimensions become 120. Although 20% of the data is labeled as “normal”, the “normal” class is defined as anomalies since the “normal” class is the minority class. The architecture for KDDCUP has three fully connected layers (32-16-8 units) with batch normalization and ReLU activation. The batch size is set to 4096, and AE and MemAE are trained with 100 epochs.

KDDCUP-Rev KDDCUP-Rev is the reverse version of KDDCUP. “normal” class is defined as normal and the others are defined as anomalies. The dataset contains all of the normal samples, and anomalies are randomly sampled with a 4:1 ratio of normal to abnormal samples. The configurations of KDDCUP-Rev were the same as those of KDDCUP.

Half of the normal samples are used for training. Artificial anomalies for contaminated training data are constructed by adding zero-mean Gaussian noise to test anomalies, as in Shenkar & Wolf (2021) and Qiu et al. (2022) since the datasets have a few anomalies. The standard deviations are derived from test anomalies. The last trained models are used in the evaluation. MSE, pseudo-Huber, GA loss, and LOE are employed to evaluate hetero loss. The rest of the configurations such as learning rate are the same as those used on image datasets.

F.2 Evaluation

Figure 10 reports the average AUROC with 10 different seeds on each dataset. Although hetero loss is less robust than the other losses on the Arrhythmia and KDDCUP datasets, it outperforms when the contamination ratio is greater than 20%. In the case of Thyroid dataset, AE and DSVDD had a similar tendency to Arrhythmia and KDDCUP. Regardless of the contamination ratio, MemAE delivers the best performance. Hetero loss resulted in the best performance on KDDCUP-Rev except for DSVDD with a 30% contamination ratio. With a low contamination ratio, heterogeneous loss has a comparable AUROC to other robust losses, which describes that the loss can effectively make a difference in training speed between normal and abnormal. With a high contamination ratio, hetero loss achieves significant improvements through aggressive rejection which half of the training data handled as potential anomalies. The experiments show that hetero loss can be employed broadly and in other domains.



(a) Arrhythmia

(b) Thyroid

(c) KDDCUP

(d) KDDCUP-Rev

Figure 10: Evaluation with various loss functions depending on contamination ratio. Three models are evaluated on (a) Arrhythmia, (b) Thyroid, (c) KDDCUP, and (d) KDDCUP-Rev.