COMPLEMENTARY SET VARIATIONAL AUTOENCODER FOR SUPERVISED ANOMALY DETECTION

Yuta Kawachi, Yuma Koizumi, and Noboru Harada

NTT Media Intelligence Laboratories, Tokyo, Japan

ABSTRACT

Anomalies have broad patterns corresponding to their causes. In industry, anomalies are typically observed as equipment failures. Anomaly detection aims to detect such failures as anomalies. Although this is usually a binary classification task, the potential existence of unseen (unknown) failures makes this task difficult. Conventional supervised approaches are suitable for detecting seen anomalies but not for unseen anomalies. Although, unsupervised neural networks for anomaly detection now detect unseen anomalies well, they cannot utilize anomalous data for detecting seen anomalies even if some data have been made available. Thus, providing an anomaly detector that finds both seen and unseen anomalies well is still a tough problem. In this paper, we introduce a novel probabilistic representation of anomalies to solve this problem. The proposed model defines the normal and anomaly distributions using the analogy between a set and the complementary set. We applied these distributions to an unsupervised variational autoencoder (VAE)-based method and turned it into a supervised VAE-based method. We tested the proposed method with well-known data and real industrial data to show that the proposed method detects seen anomalies better than the conventional unsupervised method without degrading the detection performance for unseen anomalies.

***Index Terms*—** Anomaly detection, variational autoencoder (VAE), neural network

1. INTRODUCTION

Automatic anomaly detection in multimedia has large underlying potential in new business fields. One purpose of anomaly detection is to prevent or detect equipment failures instantly. There is high demand for various practical applications of anomaly detection in fields such as surveillance systems [1, 2, 3], animal husbandry [4, 5, 6], and material and/or equipment inspection [7, 8, 9, 10, 11, 12].

A typical strategy to automatically detect anomalies is an unsupervised approach using only data that are labeled as normal. This approach is often called outlier detection. In outlier detection, an observed sample is identified as an "anomaly" when a kind of dissimilarity from the normal model, which is frequently called the "anomaly score," exceeds a pre-defined threshold. Many conventional unsupervised approaches are interpreted as probabilistic generative models with anomaly scores defined by their likelihoods. Major model choices include regression models[13, 14, 15, 16] and topic models [17, 18].

In practical situations, we may occasionally obtain a part of anomalous data. In such a case, an anomaly detection algorithm may be extended to a supervised binary classification problem, which separates data into normal and anomalous data. However, a simple binary classification algorithm is difficult to be used for anomaly detection because of three main problems:

Imbalanced data: anomalous data are often hard to obtain and are obtained less frequently than normal data. This data imbalance causes over-fitting: the classification results lack generalization [19]. To avoid this phenomenon, imbalance learning techniques are applied to anomaly detection [20].

Labeling cost: basically, human experts judge whether the sample is normal or not. Thus, it is difficult to create a massive amount of labeled data. As a solution, a semi-supervised approach that exploits unlabeled data is used [21]. In this approach, the missing label may be estimated from the small amount of labeled data.

Presence of unseen anomalies: in many practical cases, types of anomalies are diverse. For example, in machine failure cases, we cannot observe all possible failure patterns. There may be many kinds of rare cases in which anomaly data cannot easily be collected. In such cases, conventional supervised and semi-supervised methods with few seen anomaly samples often do not detect unseen anomalies that some unsupervised methods usually do [22]. Although this problem is also important for practical usage and has been tackled in some of the literature [22, 23, 24], it is focused on less than the two above problems.

In this paper, we focus on the last problem: we propose a solution for detecting seen anomalies without degrading detection performance for unseen anomalies in supervised situations. We introduce a novel statistical representation of unseen anomalies into a variational autoencoder (VAE) [25], which is a promising generative neural network that has already been applied to anomaly detection [15, 16, 21]. We turned the VAE into a supervised model by modifying its prior distribution. The key assumption is that anomalies are "not normal." We start by replacing this property with the equivalent relationship between the set expressing the occurrence of normal samples and its complementary set expressing the occurrence of anomalous samples. Then we change the original prior into the prior corresponding to the complementary set. Finally, substituting this to VAE's loss function, we can analytically represent the property of anomalies with few approximations. From an anomaly detection point of view, we show the proposed method includes the unsupervised VAE anomaly detection method [15] as a special case. We also show the proposed method detects both seen and unseen anomalies more accurately than previous unsupervised method in terms of the area under the receiver characteristic curve.

Our main contributions are summarized as follows:

- We propose a supervised anomaly detection algorithm as an extension of an unsupervised VAE algorithm [15] with novel statistical representation of anomalies to solve the problem of the presence of unseen anomalies.
- We show the problem can be solved analytically with fewer heuristics and can detect both seen and unseen anomalies.

2. CONVENTIONAL UNSUPERVISED ANOMALY DETECTION USING VAE

2.1. Variational autoencoder (VAE)

Variational autoencoder [25] models transformations between original feature space and simple latent Gaussian distributions. Encoders transform feature space into Gaussian distributions, and decoders transform Gaussian distributions into feature space. Both are composed of neural networks. VAE aims at maximizing marginal likelihood $p(\boldsymbol{x};\boldsymbol{\theta})$, where $\boldsymbol{x}$ is a feature vector and $\boldsymbol{\theta}$ is a vector that includes all parameters in the decoder $p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})$, where $\boldsymbol{z}$ is the latent variable. The marginal likelihood, which is intractable [25], is approximated by the evidence lower bound (ELBO) defined as

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}) = \langle \log p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})\rangle_{q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\phi})} - KL[q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\phi})||p(\boldsymbol{z})], \tag{1}$$

which satisfies $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}) \leq \log p(\boldsymbol{x};\boldsymbol{\theta})$, $q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\phi})$ is the encoder parameterized by $\boldsymbol{\phi}$, $KL[\cdot||\cdot]$ denotes Kullback-Leibler (KL) divergence, and $\langle\cdot\rangle_{p(\cdot)}$ denotes expectation over a distribution $p(\cdot)$. To derive the training algorithm, we approximate the expectation using finite L samples of $\boldsymbol{z}$

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}) \simeq \frac{1}{L}\sum_{l} \log p(\boldsymbol{x}|\boldsymbol{z}^{(l)};\boldsymbol{\theta}) - KL[q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\phi})||p(\boldsymbol{z})], \tag{2}$$

$$\boldsymbol{z}^{(l)} \sim q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\phi}),\ \ l \in \{1,2,\cdots L\}, \tag{3}$$

where $\sim$ means sampling from the right-hand distribution. The standard prior choice is Gaussian: $p(z) = \mathcal{N}(z;0,1)$, and a linear combination constant C is also added to improve performance [26]. Then, the single-dimensional Gaussian prior version of the maximization target is

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}) \simeq \frac{1}{L}\sum_{l} \log p(\boldsymbol{x}|z^{(l)};\boldsymbol{\theta}) - C\cdot\left(-\frac{1}{2} - \log\sigma + \frac{1}{2}\sigma^2 + \frac{1}{2}\mu^2\right). \tag{4}$$

We omitted the notation of dependence between $\boldsymbol{x}$, $\boldsymbol{\theta}$ and μ, σ^2 for simplicity. The first term is the reconstruction loss, and the second is a dissimilarity function between prior distribution of latent variables and the variables emitted from an encoder. The first one tends to fit the reconstructed vector to the original vector, whereas the second keeps the latent variables near to the point of origin.

2.2. Anomaly detection using VAE

VAE has been adapted to anomaly detection in an unsupervised manner [15, 16] and is considered to effectively learn representations of feature vectors [16]. The basic strategy for anomaly detection is to measure the magnitude of reconstruction loss. In the training phase, the VAE is trained with collected samples labeled normal. However, the samples labeled anomalies cannot be used. Given the unlabeled samples, the model can reasonably reconstruct them with low reconstruction loss if the samples are normal. If the samples are anomalies, they cannot be reconstructed sufficiently, which causes reconstruction loss to rise. If we set a pre-defined threshold and define the reconstruction loss as the anomaly score, the samples can be identified as "anomalies" when the anomaly score exceeds the threshold, the same as the conventional outlier detection.

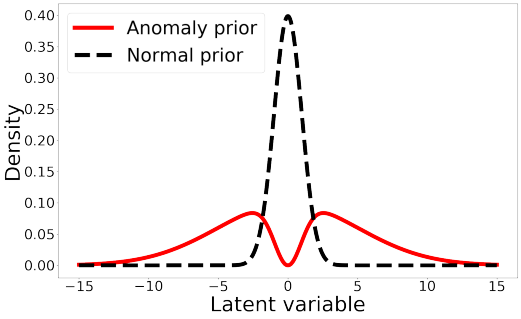


Fig. 1: Visualization of proposed anomaly prior.

3. SUPERVISED ANOMALY DETECTION BY PRIOR MODIFICATION OF VAE (PROPOSED)

3.1. Prior distribution and KL divergence for anomalies

Our assumption is that anomalies are "not normal." In other words, anomalies are regarded as the complementary set of the normal set; the normal region and the anomalous region are both mutually exclusive and collectively exhaustive. Using a probabilistic density function for normal $p_{\mathrm{n}}(\boldsymbol{z})$ and anomalous $p_{\mathrm{a}}(\boldsymbol{z})$, we initially formulated this relationship as

$$p_{\mathrm{a}}(\boldsymbol{z}) \equiv \frac{1}{Y'}\left(\max_{\boldsymbol{z}'} p_{\mathrm{n}}(\boldsymbol{z}') - p_{\mathrm{n}}(\boldsymbol{z})\right), \tag{5}$$

where the constant Y' aims to satisfy this equation as a probabilistic distribution. If $p_{\mathrm{n}}(\boldsymbol{z})$ is a uniform distribution, this definition satisfies the property of the complementary set. The main problem is actually Y' is infinity because the mass of the distribution explodes. To ensure $p_{\mathrm{a}}(\boldsymbol{z})$ is a probabilistic distribution, we multiply a distribution $p_{\mathrm{w}}(\boldsymbol{z})$ that is wide enough for each dimension. Then the distribution is

$$p_{\mathrm{a}}(\boldsymbol{z}) \equiv \frac{1}{Y}p_{\mathrm{w}}(\boldsymbol{z})\left(\max_{\boldsymbol{z}'} p_{\mathrm{n}}(\boldsymbol{z}') - p_{\mathrm{n}}(\boldsymbol{z})\right), \tag{6}$$

where Y is the finite normalizing constant.

Using this as a prior distribution, we expand the conventional unsupervised VAE into a supervised one to distinguish anomalies in the latent space. Since we set the conventional VAE as the representation for normal samples and it uses a Gaussian distribution as a prior, we also set the Gaussian distribution to $p_{\mathrm{n}}(\boldsymbol{z})$ and $p_{\mathrm{w}}(\boldsymbol{z})$ to make (6) the representation for anomaly samples. Then, the single-dimensional version of $p_{\mathrm{a}}(\boldsymbol{z})$ shown in Fig. 1 can be written as

$$p_{\mathrm{a}}(z) \equiv \frac{1}{Y}\mathcal{N}(z;0,s^2)\{\max_{z'}\mathcal{N}(z';0,1) - \mathcal{N}(z;0,1)\}, \tag{7}$$

where the constants in this equation are described as

$$\max_{z'}\mathcal{N}(z';0,1) = \frac{1}{\sqrt{2\pi}}, \tag{8}$$

$$Y = \int_{-\infty}^{\infty} p_{\mathrm{a}}(z)dz = \frac{1}{\sqrt{2\pi}}\{1 - \frac{1}{\sqrt{s^2+1}}\}, \tag{9}$$

and a hyper-parameter s^2 that determines the width of the distribution. We defined the multi-dimensional version as a product of each dimension composed of the single-dimensional version. Therefore, we derive the single-dimensional version of KL divergence below.

By substituting $p_{\mathrm{a}}(z)$ to the KL divergence in (2), the KL divergence can be rewritten as

$$KL[q(z|x;\phi)||p_{\mathrm{a}}(z)] = \int_{-\infty}^{\infty} \mathcal{N}(z;\mu,\sigma^2) \log \frac{\mathcal{N}(z;\mu,\sigma^2)}{\frac{1}{Y}\mathcal{N}(z;0,s^2)\{\frac{1}{\sqrt{2\pi}} - \mathcal{N}(z;0,1)\}} dz. \tag{10}$$

This equation is decomposed as

$$\begin{aligned} KL[q(z|x;\phi)||p_{\mathrm{a}}(z)] = & \\ & \int_{-\infty}^{\infty} \mathcal{N}(z;\mu,\sigma^2) \log \mathcal{N}(z;\mu,\sigma^2) dz \\ & + \log Y \\ & - \int_{-\infty}^{\infty} \mathcal{N}(z;\mu,\sigma^2) \log \mathcal{N}(z;0,s^2) dz \\ & - \int_{-\infty}^{\infty} \mathcal{N}(z;\mu,\sigma^2) \log\{\frac{1}{\sqrt{2\pi}} - \mathcal{N}(z;0,1)\} dz. \end{aligned} \tag{11}$$

By applying Taylor series expansion of the logarithmic function and linear approximation defined as $\log(x + \frac{1}{2\pi}) \simeq -\log 2\pi + 2\pi x$ to the fourth term of (11), the KL divergence can be approximately calculated as

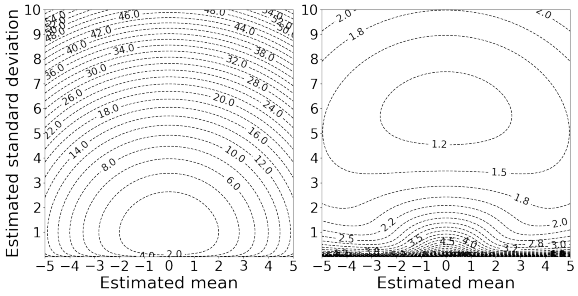
$$\begin{aligned} & KL[q(z|x;\phi)||p_{\mathrm{a}}(z)] \\ & \simeq \sqrt{\frac{2\pi}{\sigma^2+1}} \exp\left(\frac{-\mu^2}{2\,(\sigma^2+1)}\right) \\ & + \frac{\mu^2+\sigma^2}{2\,s^2} - \log\sigma + \log s + \log\left(\sqrt{s^2+1} - 1\right) \\ & - \frac{\log\left(s^2+1\right)}{2} + \frac{\log(2\pi)-1}{2}. \end{aligned} \tag{12}$$

Fig. 2 shows the KL divergences of normal and anomalous cases. Whereas the conventional VAE for normal cases works as a regularizer that forces latent variables to be located near to the point of origin, the proposed VAE for anomalous cases forces latent variables to be located far from the point of origin when the estimated standard deviations are small.

3.2. Implementation of proposed method

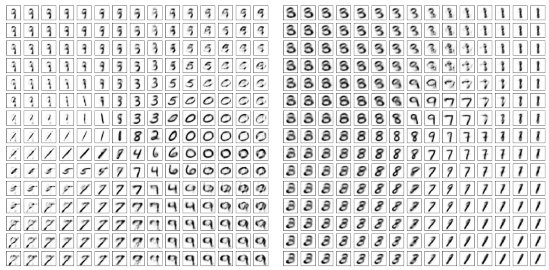
The anomaly score for the proposed model should be the KL divergence between the normal case prior $\mathcal{N}(z;0,1)$ and the encoded distribution $\mathcal{N}(z;\mu,\sigma^2)$ since its latent space represents normal and anomaly data by using dissimilarity from $\mathcal{N}(z;0,1)$. To avoid the curse of dimensionality [27], an anomaly score calculated in low-dimensional latent space such as KL divergence may be preferable. We also use evidence lower bound and conventional reconstruction loss for comparison. The evidence lower bound is the sum of the reconstruction loss and the KL divergence.

We adopt an alternating training procedure. For each epoch of VAE training, we inserted an epoch using anomalous training data under the proposed anomaly prior. If we have no anomalies to train, this procedure is equivalent to the conventional VAE anomaly detection [15]. Balancing the weight of reconstruction loss and KL divergence term has been found to be important for practical use. We set $C = 10$ only in the training procedure and when the proposed anomaly prior is used.



(a) Normal prior case **(b)** Anomaly prior case

Fig. 2: Visualization of KL divergences for normal and anomaly priors. Latent variables near zero mean have high KL divergence in anomaly prior when they have small variance.



(a) 9 vs $\bar{9}$ condition. **(b)** Case 3 condition.

Fig. 3: Latent space visualization of proposed CS-VAE model using its decoder trained by MNIST. Ranges shown are square of $[-10, 10]$. Normal patterns are encoded at center, and anomaly patterns surround them.

4. EXPERIMENTS

4.1. MNIST

The MNIST [28] dataset is publicly available and consists of handwritten images of digits 0-9. We constructed two kinds of evaluation tasks. Task 1: N vs. $\bar{N}$. In this task, we let one digit be a seen anomaly, the other digits be normal, and a uniform noise be an unseen anomaly. The seen anomaly detection task is similar to that of An and Cho [15]. Task 2: three groups of digits. The digits are tagged as normal, seen anomalies, or unseen anomalies. The tagging patterns are based on numerically ascending order, and digits are grouped into three categories: 1, 2 and 3; 4, 5, and 6; and 7, 8, and 9. The details are shown in Table 1.

In the following experiments, we used Adam with batch size 100 and elapsed 200 epochs to train the network. The encoder and decoder are composed of three-layer perceptrons with 500 hidden units. Two latent variables are used. The number of samples L to approximate integration is set to 1 in both training and evaluation in all experiments. The hyper-parameter s^2 is set to 400. We used the

Table 1: Experimental conditions: $U[0,1]$ is a multidimensional uniform distribution ranging $[0,1]$. In the training procedure, we used samples labeled "normal" for all models, "seen" for only the proposed model, and "unseen" for no models.

task	normal	seen	unseen
1 (N vs $\bar{N}$)	$0, \cdots, 9$ w/o N	N	$U[0,1]$
2 (Case 1)	$1,2,3$	$4,5,6$	$7,8,9$
2 (Case 2)	$4,5,6$	$7,8,9$	$1,2,3$
2 (Case 3)	$7,8,9$	$1,2,3$	$4,5,6$
Air conditioner	usual sound	failure sound	$U[0,1]$

area under the receiver characteristic curve (AUROC) to evaluate detection performance. If we can distinguish normal and anomaly samples completely, the AUROC value is 1. If we cannot, the AUROC value is 0.5. If the scores are inverted between normal and anomaly samples completely, the AUROC value is 0. The results are shown in Table 2. The VAE with reconstruction loss (RL) is regarded as a single sampling version of the existing work [15].

From Table 2, we can observe the proposed complementary set VAE (CS-VAE) detects anomalies better than the unsupervised conventional VAE especially in the seen condition and equal to or better than the unsupervised one in the unseen condition.

In the unseen condition in Task 1, both the conventional and proposed models detect unseen uniform noise anomalies well. However, the proposed algorithm using only the KL divergence score degrades performance. This may be because the latent variable may occur at the center when the model encounters samples not contained in the training data. However, reconstruction loss produced by a decoder may still be large because the decoder will decode as if unseen anomaly samples were normal samples. Therefore, if we use ELBO, which is the sum of reconstruction loss and KL divergence, we may avoid this effect.

In the seen condition in Task 1, we can observe the conventional VAE is not good at detecting 1, 7, and 9 as anomalies. This is a similar tendency to that in the prior work [15]. On the other hand, the proposed model detects such anomalies well by using ELBO and KL scores. The proposed model using ELBO scores is especially good at stably detecting anomalies, whereas the proposed model using only KL scores is not good at detecting 8 as an anomaly.

In Task 2, the proposed model using ELBO scores also detects anomalies better than the conventional method regardless of whether they are seen or unseen. Thus, the proposed method is also good at detecting anomalies when those unseen anomalies look similar to the seen ones. Fig. 3 shows that the normal and anomaly samples are separated in the latent space while they are transforming continuously.

4.2. Air conditioner failure sound

We also tried detecting real-world industrial anomalies. This experiment had 10 epochs. A 40-dimensional mel-filterbank acoustic feature for each 10 ms and its Δ and $\Delta\Delta$ are used. The reconstruction loss is changed to mean squared error. We defined the usual operation sound as normal and the sound after failure as anomalous. The detailed conditions and results are shown in Tables 1 and 2.

We can see both the conventional and proposed models detect anomalies perfectly when using reconstruction loss and ELBO. However, the performances of the proposed method using KL scores are greatly improved. As we observed that ELBO or KL scores work better than reconstruction loss score in the previous MNIST experiment, we can expect the proposed model to work stably regardless of task difficulties under real-world conditions.

Table 2: Comparison of AUROCs[%] (higher the better). Best scores before rounding in bold. Proposed method is good at detecting anomalies that conventional method is not (underlined).

model	VAE (Conventional)			CS-VAE (Proposed)		
score	RL[15]	ELBO	KL	RL	ELBO	KL
MNIST Task 1 (unseen anomaly detection)						
0 vs $\bar{0}$	**100.0**	**100.0**	39.7	**100.0**	**100.0**	30.1
1 vs $\bar{1}$	**100.0**	100.0	59.8	100.0	100.0	73.7
2 vs $\bar{2}$	**100.0**	**100.0**	58.6	**100.0**	100.0	0.2
3 vs $\bar{3}$	100.0	100.0	37.5	**100.0**	100.0	2.8
4 vs $\bar{4}$	**100.0**	**100.0**	24.7	**100.0**	**100.0**	29.8
5 vs $\bar{5}$	**100.0**	**100.0**	40.8	**100.0**	**100.0**	9.6
6 vs $\bar{6}$	**100.0**	**100.0**	54.1	100.0	100.0	1.0
7 vs $\bar{7}$	100.0	100.0	6.2	100.0	**100.0**	4.4
8 vs $\bar{8}$	**100.0**	**100.0**	49.0	**100.0**	100.0	1.2
9 vs $\bar{9}$	**100.0**	**100.0**	54.9	**100.0**	**100.0**	4.3
MNIST Task 1 (seen anomaly detection)						
0 vs $\bar{0}$	94.6	94.5	50.2	85.7	97.6	**98.9**
1 vs $\bar{1}$	58.6	58.4	38.7	2.8	93.2	**<u>99.0</u>**
2 vs $\bar{2}$	95.5	95.5	41.3	92.9	**98.7**	95.1
3 vs $\bar{3}$	85.1	84.9	38.6	95.7	**97.2**	91.8
4 vs $\bar{4}$	74.7	74.7	40.1	89.4	**97.0**	94.5
5 vs $\bar{5}$	85.4	85.4	46.7	95.6	**98.3**	86.0
6 vs $\bar{6}$	93.0	93.0	41.9	61.9	97.0	**97.5**
7 vs $\bar{7}$	66.6	66.5	41.7	48.9	94.5	**<u>96.9</u>**
8 vs $\bar{8}$	85.7	85.3	29.7	86.8	**91.9**	68.5
9 vs $\bar{9}$	59.7	59.2	24.8	65.6	**<u>81.9</u>**	70.4
MNIST Task 2 (unseen anomaly detection)						
Case 1	93.1	93.1	50.6	83.0	**93.2**	74.6
Case 2	80.5	80.5	43.8	87.2	**96.3**	83.9
Case 3	88.1	88.0	57.6	89.6	**90.0**	71.3
MNIST Task 2 (seen anomaly detection)						
Case 1	95.0	95.0	52.0	72.2	98.7	**98.7**
Case 2	84.9	84.6	42.5	63.1	97.2	**97.6**
Case 3	72.8	72.6	56.7	56.6	98.0	**98.2**
Air conditioner failure sound detection						
unseen	**100.0**	**100.0**	50.0	**100.0**	**100.0**	**100.0**
seen	**100.0**	**100.0**	98.3	100.0	100.0	99.8

5. CONCLUSION

We presented a supervised anomaly detection method based on a variational autoencoder (VAE). By defining the property of unseen anomalies by using a complementary set and representing it as a prior, we analytically solved the KL divergence term of VAE for anomalies with fewer heuristics. We tested the proposed algorithm on MNIST (Tasks 1 and 2) and in a real-world air conditioner failure sound detection task. The results revealed that the proposed method detects seen anomalies better than and unseen anomalies as well as the conventional unsupervised VAE. We also found that the proposed model using an evidence lower bound anomaly score detected anomalies stably in all the tested tasks.

6. REFERENCES

[1] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Probabilistic novelty detection for acoustic surveillance under real-world conditions," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 713–719, 2011.

[2] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks," in *Proc. ICASSP*. IEEE, 2015, pp. 1996–2000.

[3] T. Komatsu and R. Kondo, "Detection of anomaly acoustic scenes based on a temporal dissimilarity model," in *Proc. ICASSP*. IEEE, 2017, pp. 376–380.

[4] P. Coucke, B. De Ketelaere, and J. De Baerdemaeker, "Experimental analysis of the dynamic, mechanical behaviour of a chicken egg," *Journal of sound and vibration*, vol. 266, no. 3, pp. 711–721, 2003.

[5] Y. Chung, S. Oh, J. Lee, D. Park, H. H. Chang, and S. Kim, "Automatic detection and recognition of pig wasting diseases using sound data in audio surveillance systems," *Sensors*, vol. 13, no. 10, pp. 12929–12942, 2013.

[6] Y. Chung, J. Lee, S. Oh, D. Park, H. H. Chang, and S. Kim, "Automatic detection of cow ' s oestrus in audio surveillance system," *Asian-Australasian journal of animal sciences*, vol. 26, no. 7, pp. 1030, 2013.

[7] Y. Ono, Y. Onishi, T. Koshinaka, S. Takata, and O. Hoshuyama, "Anomaly detection of motors with feature emphasis using only normal sounds," in *Proc. ICASSP*. IEEE, 2013, pp. 2800–2804.

[8] J. Ye, M. Iwata, K. Takumi, M. Murakawa, H. Tetsuya, Y. Kubota, T. Yui, and K. Mori, "Statistical impact-echo analysis based on Grassmann manifold learning: Its preliminary results for concrete condition assessment," in *Proc. EWSHM*, 2014.

[9] T. Tagawa, Y. Tadokoro, and T. Yairi, "Structured denoising autoencoder for fault detection and analysis," in *Proc. ACML*, 2015, pp. 96–111.

[10] H. Fujii, A. Yamashita, and H. Asama, "Defect detection with estimation of material condition using ensemble learning for hammering test," in *Proc. ICRA*. IEEE, 2016, pp. 3847–3854.

[11] P. A. Delgado-Arredondo, D. Morinigo-Sotelo, R. A. Osornio-Rios, J. G. Avina-Cervantes, H. Rostro-Gonzalez, and R. de Jesus Romero-Troncoso, "Methodology for fault detection in induction motors via sound and vibration signals," *Mechanical Systems and Signal Processing*, vol. 83, pp. 568–589, 2017.

[12] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma," in *Proc. EUSIPCO*. EURASIP, 2017.

[13] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based encoder-decoder for multi-sensor anomaly detection," in *Proc. ICML*, 2016.

[14] D. Cheboli, "Anomaly detection of time series," M.S. thesis, The University of Minnesota, 2010.

[15] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," 2015.

[16] S. Suh, D. H. Chae, H. G. Kang, and S. Choi, "Echo-state conditional variational autoencoder for anomaly detection," in *Proc. IJCNN*. IEEE, 2016, pp. 1015–1022.

[17] O. Isupova, D. Kuzin, and L. Mihaylova, "Anomaly detection in video with Bayesian nonparametrics," in *Proc. ICML*, 2016.

[18] H. Soleimani and D. J. Miller, "ATD: Anomalous topic discovery in text documents," in *Proc. ICML*, 2016.

[19] S. Wang, L. L. Minku, and X. Yao, "Resampling-based ensemble methods for online class imbalance learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1356–1368, 2015.

[20] C. K. Maurya, D. Toshniwal, and G. V. Venkoparao, "Online anomaly detection via class-imbalance learning," in *Proc. IC3*. IEEE, 2015, pp. 30–35.

[21] G. Osada, K. Omote, and T. Nishide, "Network intrusion detection based on semi-supervised variational auto-encoder," in *Proc. ESORICS*. Springer, 2017, pp. 344–361.

[22] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward supervised anomaly detection," *Journal of Artificial Intelligence Research*, vol. 46, no. 1, pp. 235–262, Jan. 2013.

[23] D. M. J. Tax, *One-class classification*, Ph.D. thesis, Delft University of Technology, 2001.

[24] B. Du and L. Zhang, "A discriminative metric learning based anomaly detection method," *IEEE Trans. Geoscience and Remote Sensing*, vol. 52, no. 11, pp. 6844–6857, 2014.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[26] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. ICLR*, 2017.

[27] A. Zimek, E. Schubert, and H. P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 5, no. 5, pp. 363–387, 2012.

[28] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.