To appear in the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore

A CLOSER LOOK AT AUTOENCODERS FOR UNSUPERVISED ANOMALY DETECTION

Oyebade K. Oyedotun and Djamila Aouada

Interdisciplinary Centre for Security, Reliability and Trust (Snt),
University of Luxembourg, L-1855 Luxembourg

ABSTRACT

Unsupervised anomaly detection is a challenging problem, where the aim is to detect irregular data instances. Interestingly, generative models can learn data distribution, and thus have been proposed for anomaly detection. In this direction, the variational autoencoder (VAE) is popular, as it enforces an explicit probabilistic interpretation of the latent space. We note that there are other generative autoencoders (AEs) such as the denoising AE (DAE) and contractive AE (CAE), which also model data generation process without enforcing an explicit probabilistic latent space interpretation. While it is intuitively straightforward to see the benefit of a latent space with explicit probabilistic interpretation for generative tasks, it is unclear how this can be crucial for anomaly detection problems. Consequently, our exposition in this paper is to investigate the extent to which different latent space attributes of AEs impacts their performances for anomaly detection tasks. We take the conventional and deterministic AE that we refer to as plain AE (PAE) as the baseline for performance comparison. Our results obtained using five different datasets reveal that an explicit probabilistic latent space is not necessary for good performance. The best results on most of the datasets are obtained using CAE, which enjoys stable latent representations.

Index Terms— Anomaly detection, autoencoder, variational autoencoder, latent representations

1. INTRODUCTION

Anomaly detection is not a trivial task, considering that anomalies are often rare events in data [1]. Unsupervised anomaly detection, where there are no labelled data for guiding learning is even more challenging. Therefore, considering that deep neural networks (DNNs) are power feature extractors [2], various unsupervised DNN data modeling approaches [1, 3] for anomaly detection can be found in the literature. Autoencoder (AE)-based methods that fall into two main categories in the literature are particularly interesting, since they are easy to understand and simple to train. The first method employs AEs for feature extraction, and then performs clustering using another algorithm, as in [4, 5]. The second method relies on small reconstruction errors for data points that come from the same distribution that the AE was trained on. Other data points are expected to have high reconstruction errors [6, 7].

AEs can be deterministic or generative as discussed in [8, 9, 10, 11]. Deterministic AEs do not learn data generation process, and therefore cannot be sampled from; they essentially learn a compressed representation for data, while retaining the important features. An example of a deterministic AE is the plain AE (PAE) [10]. Generative AEs aim to learn data generation process based on interesting latent representations. In this fashion, the latent representations can be sampled from to generate novel data points. A popular assumption in generative models is that capturing the distribution of the training data facilitates the learning of latent representations, which are more separable [12]. Examples of generative AEs include the denoising AE (DAE) [13], contractive AE (CAE) [13, 14, 15], variational AE (VAE) [10] and β-VAE [11], which is simply a generalization of VAE [10] with $\beta = 1$. It is noteworthy that the work on β-VAE [11] advocated $\beta > 1$. Generative AEs can either provide implicit or explicit probabilistic interpretation of the data distribution learned, as seen in Fig. 1.

On the one hand, the generative characteristics of DAE [13] and CAE [14] that provide implicit probabilistic interpretation are well documented in the literature, but sampling from them is a difficult task [13, 16, 17]. On the other hand, sampling from β-VAE [11] that provide explicit probabilistic interpretation is straightforward [10, 11], however a concrete and unanimous account of its generative learning attributes based on disentangled latent representations is currently lacking in the literature [18, 19, 20, 21]. Another current concern is how the β hyperparameter, which balances reconstruction quality and disentangling attribute impacts the performance of β-VAE [22]. Example, β-VAE with $\beta < 1$ can be seen in [1, 23]. Furthermore, among other works, in [1], PAE (seen as Conv-AE) outperformed VAE (seen as CVAE) on the Yahoo dataset. In [3], PAE (seen as LSTM-AE) agained outperformed VAE (seen as LSTM-VAE) on KPI dataset. This unexpected observation in [1, 3] leaves to imagination whether VAE is clearly superior in performance for anomaly detection.

As such, this paper presents a holistic empirical investigation on how the learning characteristics of aforementioned AEs impact their performances for anomaly detection. Our motivation for this study is that an interesting AE has a latent

Research funded by the National Research Fund (FNR), Luxembourg, under the project reference BRIDGES2020/IS/14755859/MEET-A/Aouada.

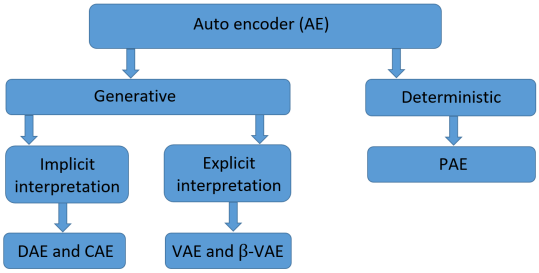

Fig. 1: Classification of autoencoders (AE) studied in this work into generative and deterministic models.

space, which models well the structure of training data. While the desired structure of the latent space for generative tasks is clear [11, 19, 20], the attributes of a good latent space for anomaly detection is tricky. For instance, taking the deterministic latent space of PAE as a baseline, does the generative attributes of AEs, as seen in DAE, CAE and β-VAE, qualify them as more interesting models for anomaly detection? Importantly, does an explicitly interpretable probabilistic latent space in an AE, as seen in β-VAE provide performance improvement for anomaly detection? Namely, our contributions are as follows:

1. Study the extent to which the latent space attributes of popular AEs such as DAE [13], CAE [14] and β-VAE [11] impact performance for unsupervised anomaly detection in comparison to the deterministic PAE [10].
2. Provide extensive empirical results based on F_2-score, along with interesting visualizations of latent representations for understanding decision making in the models.

The remainder of this paper is organized as follows. Section 2 discusses the background and problem statement. Section 3 presents the proposed analysis. In Section 4, we give the experiments. The paper is concluded in Section 5.

2. BACKGROUND AND PROBLEM STATEMENT

2.1. Background

2.1.1. Deterministic Plain Autoencoder

Plain autoencoder (PAE)
The PAE is a model that can extract interesting latent representations in data by learning to reconstruct the input in the output layer. Let the encoder and decoder functions of the AE be parameterized by $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, respectively. Let the model's inputs be $\boldsymbol{x}_i \in \mathbb{R}^d : 1 \leq i \leq N$, and output be $\ddot{\boldsymbol{x}}_i \in \mathbb{R}^d$. As such, $\boldsymbol{h}_i = f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)$ and $\ddot{\boldsymbol{x}}_i = g_{\boldsymbol{\psi}}(f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)) = g_{\boldsymbol{\psi}}(\boldsymbol{h}_i)$. Let $\theta = \{\boldsymbol{\phi}, \boldsymbol{\psi}\}$. Assuming that the output of the AE are real-values, and follow multivariate Gaussian distribution, we can employ the training cost

$$L(\boldsymbol{x}, \ddot{\boldsymbol{x}}; \theta) = \underset{\theta}{\operatorname{argmin}} \Sigma_{i=1}^{N} (\boldsymbol{x}_i - \ddot{\boldsymbol{x}}_i)^2. \quad (1)$$

2.1.2. Generative Autoencoder with Implicit Interpretation

Herein, we discuss generative AEs, DAE and CAE, that do not have an explicit probabilistic interpretation of the generative process [13, 16]. However, novel data points can be generated by sampling from the implicit learnt distribution [13, 16].

Denoising autoencoder (DAE)
Assuming the corruption process $\hat{\boldsymbol{x}}_i \sim \Lambda(\hat{\boldsymbol{x}}_i|\boldsymbol{x}_i)$, the DAE is a regularized AE that tasks the hidden units, $\boldsymbol{h}_i \in \mathbb{R}^s$, to learn the reconstruction of corrupted input data, $\hat{\boldsymbol{x}}_i \in \mathbb{R}^n$, in the output layer as $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^n$. For a DAE with similar parameterization as the PAE, the cost function is

$$L(\boldsymbol{x}, \tilde{\boldsymbol{x}}; \theta) = \underset{\theta}{\operatorname{argmin}} \Sigma_{i=1}^{N} (\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i)^2. \quad (2)$$

For sampling from the DAE, Metropolis-Hastings [16] and Markov chains [13] have been proposed.

Contractive autoencoder (CAE)
The CAE [14] regularizes $\boldsymbol{h}$ in PAE to achieve minimal response for small changes in $\boldsymbol{x}$, $\delta\boldsymbol{x}$. CAE simply penalizes $R(\boldsymbol{h})_{CAE} = \|\delta\boldsymbol{h}/\delta\boldsymbol{x}\|_F^2$, so that the new training cost is

$$L(\boldsymbol{x}, \ddot{\boldsymbol{x}}; \theta) = \underset{\theta}{\operatorname{argmin}} \Sigma_{i=1}^{N} (\boldsymbol{x}_i - \ddot{\boldsymbol{x}}_i)^2 + \lambda R(\boldsymbol{h})_{CAE}, \quad (3)$$

where λ controls the regularization weight. CAE sampling can be achieved using Jacobian-based Gaussian noise [15].

2.1.3. Generative Autoencoder with Explicit Interpretation

Generative AEs such as β-VAE [11] aims to explicitly estimate the conditional probability distribution, $p(\boldsymbol{h}|\boldsymbol{x})$.

$\boldsymbol{\beta}$-Variational autoencoder ($\boldsymbol{\beta}$-VAE)
In $\boldsymbol{\beta}$-VAE [11], computing $p(\boldsymbol{h}|\boldsymbol{x})$ requires estimating $p(\boldsymbol{x}) = \int_{\boldsymbol{h}} p(\boldsymbol{x}|\boldsymbol{h}) d\boldsymbol{h}$, which is intractable. For resolving the problem, a parametric inference model $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{x}_i)$ that is an AE encoder is used for approximating $p(\boldsymbol{h}|\boldsymbol{x})$. Generally, $p(\boldsymbol{h}|\boldsymbol{x})$ is chosen as isotropic unit Gaussian so that the latent space is well-behaved (continuous) in a probabilistic sense, and fosters disentangled representations. Using the KullbackLeibler (KL) divergence, $R(\boldsymbol{h})_{VAE} = KL[q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{x}_i)\|p(\boldsymbol{h}|\boldsymbol{x})]$, the regularized training cost of the $\boldsymbol{\beta}$-VAE is given as

$$L(\boldsymbol{x}, \ddot{\boldsymbol{x}}; \theta) = \underset{\theta}{\operatorname{argmin}} \Sigma_{i=1}^{N} \left[(\boldsymbol{x}_i - \ddot{\boldsymbol{x}}_i)^2 + \boldsymbol{\beta}\, R(\boldsymbol{h})_{VAE}\right], \quad (4)$$

where $\boldsymbol{\beta} \in \mathbb{R}$ weights the regularization for $R(\boldsymbol{h})_{VAE}$. For $\boldsymbol{\beta} = 1$, we have the conventional VAE [10].

2.2. Problem statement

Given an unlabelled dataset $D = \{\boldsymbol{x}_i\}_{i=1}^N$, where $\boldsymbol{x}_i \in n|a$, where n and a denote normal and anomalous data, respectively. Let $D_n = \{\boldsymbol{x}_{n_i}\}_{i=1}^r : \boldsymbol{x}_{n_i} \in n$, $D_a = \{\boldsymbol{x}_{a_i}\}_{i=1}^u : \boldsymbol{x}_{a_i} \in a$, and $N = r + u$ so that $D = D_n \cup D_a$ and $\varnothing = D_n \cap D_a$. Furthermore, let $D_n \sim p_n$ and $D_a \sim p_a$. For unsupervised

anomaly detection, the AE is typically trained on only D_n. Supposing the AE is successfully trained on D_n, the AE learns an encoding function that allows mainly the successful reconstruction of a novel data point $\boldsymbol{x}_j \sim p_n$. It is expected that a data point $\boldsymbol{x}_k \nsim p_n$ will incur a high reconstruction loss. As such, using a suitable reconstruction loss threshold, t_{rec}, it possible to distinguish $\boldsymbol{x}_j$ from $\boldsymbol{x}_k$.

Our problem focus is to investigate to what extent the latent space characteristics of the different AEs discussed in Section 2.1 impact model performance for unsupervised anomaly detection. Specifically, we study in the aforementioned AEs, how the form of learning $\boldsymbol{x}_j \sim p_n$ influences their performances for detecting $\boldsymbol{x}_k \nsim p_n$.

3. PROPOSED ANALYSIS

3.1. Quantitative evaluation

We study the classification performance of the different AEs in Section 2.1 using different unlabelled datasets. The AEs will be trained on only normal (i.e. non-anomalous) data points. Afterwards, the AEs will be tested using arbitrary data points. Given that the AE is trained on D_n, we expect an AE, which successfully learnt $\boldsymbol{x}_j \sim p_n$ to make fewer mistakes in identifying novel data points $\boldsymbol{x}_k \sim p_n$. That is, the AE will have a higher recall than precision. As such, we evaluate performance metrics including $recall = TP/(TP+FN)$ and $precision = TP/(TP+FP)$; where TP, FN and FP refer to true positive, false negative and false positive, respectively. We also evaluate the F_1-score and F_2-score using

$$F_\beta\text{-score} = (1+\beta^2)\frac{Precision \times Recall}{\beta^2 \times Precision + Recall}, \qquad (5)$$

where $\beta = 1$ and $\beta = 2$ for F_1-score and F_2-score, respectively. Note that F_2-score gives more weight to recall than precision, and thus is particularly more useful for anomaly detection as seen in other works [24, 25]. *As such, F_2-score will be the main evaluation metric in this work.*

3.2. Qualitative evaluation

The trained AE models are also evaluated by inspecting their latent representations. For anomaly detection tasks, we expect that the latent representations of good models will form more distinct clusters. For visualization, the t-distributed stochastic neighbor embedding (t-SNE) [26] technique is employed for projecting the latent representations into two-dimensional data.

4. EXPERIMENTS

For experiments, we use five different datasets that include BreastW, Ionosphere, Thyroid, Cardio, and Wisconsin breast Cancer (WBC), which are all obtained from [27]. We note that other works have employed similar datasets as in [28]. Table 1 shows the statistics of the different datasets. The AE models, PAE, DAE, CAE and β-VAE, are trained on the datasets. All AEs have two hidden layers with five and four hidden units, consecutively. For each dataset, we collect all the normal data instances, out of which 80% data instances are randomly selected for training and the other 20% used for testing. All the anomalous data instances are collected and only used for testing. For choosing the model hyperparameters and reconstruction loss threshold, t_{rec}, 15% of the training data is used as validation data. All models are trained for 30 epochs using mini-batch gradient descent with an initial learning of 10^{-3}, which is annealed during training. The DAEs are corrupted with Gaussian noise of zero mean (i.e. $\mu = 0$) and standard deviation of $\sigma = 0.05$ or $\sigma = 0.1$. Contracting penalties of $\lambda = 10^{-4}$ and $\lambda = 10^{-3}$ as in (3) for the CAE model are used. For determining t_{rec}, we use the mean reconstruction loss on the validation data added to its standard deviation. Furthermore, following the work [28], *the experiments are repeated ten times for all the datasets, and the average recall, precision, F_1-score and F_2-score results for the different models are reported.*

The results of the different AEs are given in Tables 2 to 6. Considering the F_2-score as the main performance metric, it is seen that the CAE gives the best performance on most of the datasets, *which can be attributed to its more stable latent representations discussed in Section 2.1.2*. The PAE, DAE and VAE give competitive performance on the datasets. In Table 3, the VAE gives clearly poor results for the ionosphere dataset. We conjecture that the thyroid dataset may be particularly

Dataset	# Instances	# Dimension	Anomalous ratio
BreastW	683	9	35%
Ionosphere	351	33	36%
Thyroid	3772	6	2.5%
WBC	278	30	5.6%
Cardio	1831	21	9.6%

Table 1: Datasets statistics

Model	Recall	Precision	F_1-score	F_2-score
PAE	94.64%	94.90%	94.74%	94.67%
DAE ($\sigma = 0.05$)	99.00%	96.42%	97.66%	98.45%
DAE ($\sigma = 0.1$)	98.79%	96.47%	97.61%	98.31%
CAE ($\lambda = 10^{-4}$)	99.16%	95.66%	97.37%	98.44%
CAE ($\lambda = 10^{-3}$)	99.41%	99.45%	97.90%	**98.80%**
β-VAE (β=0.1)	89.67%	95.87%	92.41%	90.69%
β-VAE (β=1.0)	92.97%	96.09%	94.48%	93.56%
β-VAE (β=5.0)	92.59%	94.52%	93.49%	92.94%

Table 2: Anomaly detection results on breastW dataset

Model	Recall	Precision	F_1-score	F_2-score
PAE	95.71%	93.28%	94.41%	95.18%
DAE ($\sigma = 0.05$)	96.27%	93.68%	94.93%	95.72%
DAE ($\sigma = 0.1$)	96.03%	95.02%	95.45%	95.78%
CAE ($\lambda = 10^{-4}$)	96.59%	95.72%	96.13%	96.40%
CAE ($\lambda = 10^{-3}$)	97.38%	94.23%	95.75%	**96.72%**
β-VAE (β=0.1)	54.92%	88.72%	67.65%	59.37%
β-VAE (β=1.0)	54.68%	87.18%	66.96%	58.98%
β-VAE (β=5.0)	56.51%	88.52%	68.70%	60.79%

Table 3: Anomaly detection results on ionosphere dataset

Model	Recall	Precision	F_1-score	F_2-score
PAE	53.94%	79.61%	63.77%	80.87%
DAE ($\sigma = 0.05$)	93.01%	53.57%	67.86%	80.94%
DAE ($\sigma = 0.1$)	95.70%	52.83%	67.95%	82.19%
CAE ($\lambda = 10^{-4}$)	93.66%	50.90%	65.71%	80.00%
CAE ($\lambda = 10^{-3}$)	93.33%	55.70%	69.46%	81.91%
β-VAE (β=0.1)	95.16%	55.36%	69.68%	**82.85%**
β-VAE (β=1.0)	94.95%	51.99%	67.13%	81.41%
β-VAE (β=5.0)	92.80%	52.71%	67.09%	80.39%

Table 4: Anomaly detection results on thyroid dataset

Model	Recall	Precision	F_1-score	F_2-score
PAE	82.39%	75.85%	78.87%	80.92%
DAE ($\sigma = 0.05$)	85.74%	75.41%	80.14%	83.36%
DAE ($\sigma = 0.1$)	85.23%	72.55%	78.16%	82.19%
CAE ($\lambda = 10^{-4}$)	86.35%	75.10%	80.18%	83.68%
CAE ($\lambda = 10^{-3}$)	86.48%	75.15%	80.27%	**83.84%**
β-VAE (β=0.1)	85.23%	77.21%	80.72%	83.29%
β-VAE (β=1.0)	83.92%	76.05%	79.48%	82.02%
β-VAE (β=5.0)	85.23%	77.25%	80.85%	83.40%

Table 5: Anomaly detection results on cardio dataset

Model	Recall	Precision	F_1-score	F_2-score
PAE	78.10%	67.03%	71.82%	75.36%
DAE ($\sigma = 0.05$)	77.62%	68.00%	72.04%	75.17%
DAE ($\sigma = 0.1$)	78.10%	69.32%	72.75%	75.69%
CAE ($\lambda = 10^{-4}$)	80.48%	60.83%	69.03%	75.37%
CAE ($\lambda = 10^{-3}$)	79.52%	70.55%	74.01%	**77.00%**
β-VAE (β=0.1)	77.14%	69.85%	72.09%	74.64%
β-VAE (β=1.0)	76.67%	66.95%	70.66%	73.92%
β-VAE (β=5.0)	78.10%	65.43%	70.73%	74.82%

Table 6: Anomaly detection results on WBC dataset

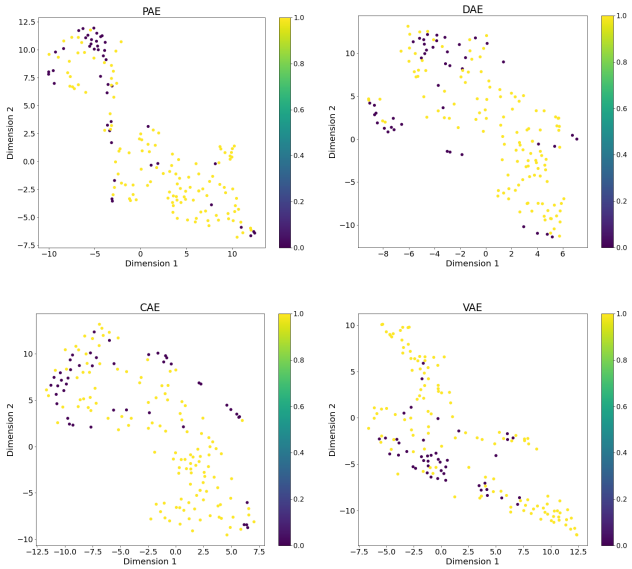

Fig. 2: Latent representations for the AE models trained on ionosphere dataset using t-SNE projection

amenable to the Gaussian prior on the VAE latent structure, and hence the interesting result. Fig. 2 and Fig. 3 show the

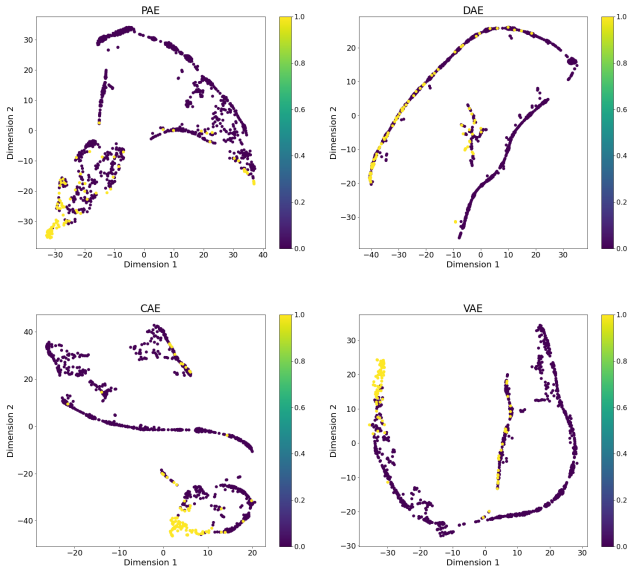

Fig. 3: Latent representations for the AE models trained on thyroid dataset using t-SNE projection

latent representations of the different AEs on the ionosphere and thyroid datasets, respectively. It is seen in Fig. 2 that for the PAE, DAE and CAE, the anomalous latent representations (i.e. purple points) concentrate on the extreme ends on the latent space, and thus facilitate anomaly detection. In contrast, for the VAE, the anomalous latent representations spread out in the centre of the latent space, and thus make anomaly detection difficult. We hypothesize that such failure of the VAE could be related to the restriction of latent space continuity imposed by the Gaussian prior distribution; see Section 2.1.3. In Fig. 3, the latent representations of all the AEs have decent latent spaces, though there is some overlap of the normal (i.e. yellow points) and anomalous latent representations.

5. CONCLUSION

Autoencoders (AEs) with different formulations, and thus learning attributes are popular for anomaly detection. The variational autoencoder (VAE) with a continuous latent space and explicit probabilistic interpretation has attracted attention for anomaly detection. Interestingly, unclarity for the operation and performance of VAE can be found in the literature. As such, this paper investigates the extent to which VAE compares with other AEs such as the plain autoencoder (PAE), denoising autoencoder (DAE) and contractive autoencoder (CAE) using five different datasets. We find that the explicit probabilistic interpretation and continuous latent space attributes of VAE does not often translate to better performance in comparison to PAE, DAE and CAE. Instead, the CAE with more stable latent representations gives the best performance on most of the datasets, while VAE clearly fails on one of the datasets.

References

[1] M. Memarzadeh, B. Matthews, and I. Avrekh, "Unsupervised anomaly detection in flight data using convolutional variational auto-encoder," *Aerospace*, vol. 7, no. 8, p. 115, 2020.

[2] M. Bianchini and F. Scarselli, "On the complexity of neural network classifiers: A comparison between shallow and deep architectures," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 8, pp. 1553–1565, 2014.

[3] Z. Niu, K. Yu, and X. Wu, "Lstm-based vae-gan for time-series anomaly detection," *Sensors*, vol. 20, no. 13, p. 3738, 2020.

[4] X. Zhao and M. Jia, "A novel unsupervised deep learning network for intelligent fault diagnosis of rotating machinery," *Structural Health Monitoring*, vol. 19, no. 6, pp. 1745–1763, 2020.

[5] L. Seydoux, R. Balestriero, P. Poli, M. De Hoop, M. Campillo, and R. Baraniuk, "Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning," *Nature communications*, vol. 11, no. 1, pp. 1–12, 2020.

[6] E. Principi, D. Rossetti, S. Squartini, and F. Piazza, "Unsupervised electric motor fault detection by using deep autoencoders," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 2, pp. 441–451, 2019.

[7] K. D. Garcia, C. R. de Sá, M. Poel, T. Carvalho, J. Mendes-Moreira, J. M. Cardoso, A. C. de Carvalho, and J. N. Kok, "An ensemble of autonomous auto-encoders for human activity recognition," *Neurocomputing*, vol. 439, pp. 271–280, 2021.

[8] T. Jebara, *Machine learning: discriminative and generative*. Springer Science & Business Media, 2012, vol. 755.

[9] G. Harshvardhan, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "A comprehensive survey and analysis of generative models in machine learning," *Computer Science Review*, vol. 38, p. 100285, 2020.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014, pp. 1–14.

[11] 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 *International Conference on Learning Representations*, 2017, pp. 1–22.

[12] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[13] Y. Bengio, L. Yao, G. Alain, and P. Vincent, "Generalized denoising auto-encoders as generative models," in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[14] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: explicit invariance during feature extraction," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011, pp. 833–840.

[15] S. Rifai, Y. Bengio, Y. N. Dauphin, and P. Vincent, "A generative process for sampling contractive auto-encoders," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012, pp. 1811–1818.

[16] G. Alain and Y. Bengio, "What regularized auto-encoders learn from the data-generating distribution," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3563–3593, 2014.

[17] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[18] H. Sikka, W. Zhong, J. Yin, and C. Pehlevant, "A closer look at disentangling in β-vae," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 888–895.

[19] R. T. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in vaes," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 2615–2625.

[20] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh, "Disentangling disentanglement in variational autoencoders," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4402–4412.

[21] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," in *International Conference on Learning Representations*, 2017, pp. 1–17.

[22] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning*, 2018, pp. 2649–2658.

[23] A. Kumar and B. Poole, "On implicit regularization in β-vae β-vaes," in *International Conference on Machine Learning*, 2020, pp. 5480–5490.

[24] R. Zhang, G. Zhang, L. Liu, C. Wang, and S. Wan, "Anomaly detection in bitcoin information networks with multi-constrained meta path," *Journal of Systems Architecture*, vol. 110, p. 101829, 2020.

[25] R. Mayer, M. Hittmeir, and A. Ekelhart, "Privacy-preserving anomaly detection using synthetic data," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2020, pp. 195–207.

[26] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[27] S. Rayana, "Outlier detection datasets (odds)," 2016. [Online]. Available: http://odds.cs.stonybrook.edu/

[28] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International conference on learning representations*, 2018.