Rethinking Autoencoders for Medical Anomaly Detection from A Theoretical Perspective

Yu Cai, Hao Chen$^{(\boxtimes)}$, and Kwang-Ting Cheng

The Hong Kong University of Science and Technology, Hong Kong, China
`jhc@cse.ust.hk`

Abstract. Medical anomaly detection aims to identify abnormal findings using only normal training data, playing a crucial role in health screening and recognizing rare diseases. Reconstruction-based methods, particularly those utilizing autoencoders (AEs), are dominant in this field. They work under the assumption that AEs trained on only normal data cannot reconstruct unseen abnormal regions well, thereby enabling the anomaly detection based on reconstruction errors. However, this assumption does not always hold due to the mismatch between the reconstruction training objective and the anomaly detection task objective, rendering these methods theoretically unsound. This study focuses on providing a theoretical foundation for AE-based reconstruction methods in anomaly detection. By leveraging information theory, we elucidate the principles of these methods and reveal that the key to improving AE in anomaly detection lies in minimizing the information entropy of latent vectors. Experiments on four datasets with two image modalities validate the effectiveness of our theory. To the best of our knowledge, this is the first effort to theoretically clarify the principles and design philosophy of AE for anomaly detection. The code is available at https://github.com/caiyu6666/AE4AD.

Keywords: Anomaly detection · Autoencoders · Information theory.

1 Introduction

In various medical scenarios, including health screening and rare disease recognition, collecting an adequate number of annotated abnormal images presents a challenge due to the diverse range of diseases and the scarcity of abnormal cases. Consequently, anomaly detection, which only necessitates normal samples during training, has emerged as a promising solution for such scenarios.

Anomaly detection (AD) is a fundamental machine learning problem [9,18] that aims to detect abnormal samples that deviate from expected normal patterns. It is typically formulated as one-class classification (OCC), where models are exclusively trained on normal data. The trained models should effectively detect samples that deviate from normal patterns, eliminating the need for abnormal training data. In the field of medical AD, reconstruction-based methods utilizing AEs have gained popularity as effective strategies [4,21,20,24,10,19].
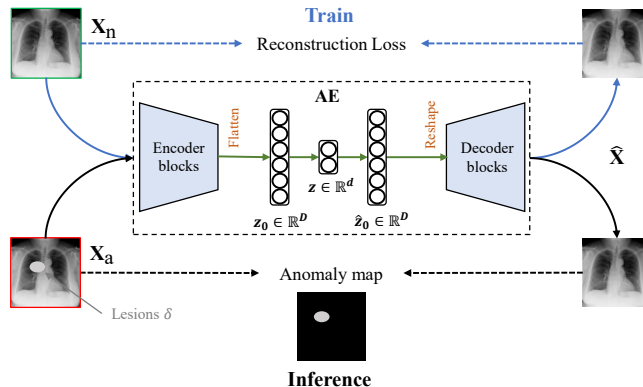
Fig. 1. Overview of the reconstruction AE-based AD. The model is trained to minimize reconstruction loss on normal images $\mathbf{X}_n$. During inference, lesions δ in abnormal images $\mathbf{X}_a$ are assumed unable to be reconstructed by the trained model.

Fig. 1 depicts an overview of AE in AD. The AE is trained to reconstruct normal images, and during inference, the reconstruction error serves as an anomaly score, indicating the extent of deviation from normality. These methods operate on the *assumption* that (i) a model trained solely on normal images can effectively reconstruct normal regions, but (ii) it cannot reconstruct unseen abnormal regions, resulting in high reconstruction errors in abnormal regions.

Considerable efforts have been made for reconstruction-based AD. To enhance the reconstruction quality of normal images (*assumption* (i)), several methods [5,1] introduced adversarial training to AE-based models. Mao et al. [16] proposed an automatic estimation of pixel-level uncertainty in reconstruction to mitigate unexpected reconstruction errors around normal boundaries. Bercea et al. [6] refined the pseudo-healthy reconstruction using estimated dense deformation fields [3]. On the other hand, some approaches focus on preventing the reconstruction of abnormal regions (*assumption* (ii)). Gong et al. [13] proposed to record prototypes of normal training patterns, which are used as inputs to the decoder to facilitate anomaly-free reconstruction. Zimmerer et al. [24] introduced an inpainting task to aid the model in repairing abnormal regions. You et al. [23] proposed a transformer architecture with neighbor masked attention to prevent abnormal information from being reconstructed.

However, we notice that *assumption* (ii) does not always hold due to the absence of relevant constraints in the training objectives. While existing methods incorporate various intuitive designs to prevent the reconstruction of anomaly and facilitate *assumption* (ii), they rely on heuristics and lack a solid theoretical foundation, resulting in ambiguous optimal solutions. This lack of a theoretical foundation has led to erroneous arguments, as seen [23,15], that claim the existence of an "identical shortcut" in reconstruction networks, further contributing to confusion. Bercea et al. [6] attempted to tackle this problem by summariz-

ing two desired properties for AE: sufficiency and minimality, which guarantee *assumption* (i) and (ii) respectively. However, their analysis remains qualitative.

This study aims to address this issue by providing a theoretical foundation to explain the workings of reconstruction-based methods and uncover their theoretically optimal solutions. Specifically, we first prove that proper latent dimensions in AE can effectively avoid the "identical shortcut" mentioned in previous works. Additionally, using information theory, we theoretically clarify the principle of AE in AD and establish its optimal solution. Our theory reveals that, apart from the reconstruction loss, another restriction should be imposed on AE to constrain the entropy of its latent space. Ideally, the entropy of the latent space should match that of normal data, enabling effective reconstruction of normal data while preventing the reconstruction of lesions. In experiments, we observe that simple latent dimension reduction is a powerful way to constrain the entropy of the latent space. The AD performance consistently aligns with our theory across varying latent dimensions, with extremely small latent dimensions (e.g., 4) often yielding optimal performance. Our contribution is summarized as follows:

- We prove that with an appropriate latent dimension, AE for AD does not encounter the "identical shortcut" problem.
- Based on information theory, we provide an explanation of how AE operates in AD, and uncover its theoretically optimal solution.
- Experiments on four datasets with two image modalities present consistent trends that align with our proposed theory, validating its effectiveness.

2   The pipeline and limitation of AE in anomaly detection

As shown in Fig. 1, AE ϕ comprises an encoder f_e and a decoder f_d. The encoder compresses the input image $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ into a compact latent vector $\mathbf{Z} = f_e(\mathbf{X}) \in \mathbb{R}^d$, and the decoder maps the latent vector back to the image space $\hat{\mathbf{X}} = f_d(\mathbf{Z}) \in \mathbb{R}^{C \times H \times W}$. We formally denote the normal image as $\mathbf{X}_n$ and the abnormal image as $\mathbf{X}_a$. In the case of typical medical images like chest X-rays and brain MRIs, each abnormal image $\mathbf{X}_a$ can be understood as the corresponding healthy version $\mathbf{X}_n$ with the addition of lesion regions δ, i.e., $\mathbf{X}_a = \mathbf{X}_n + \delta$ [11]. The training objective is to minimize the reconstruction loss on normal images:

$$\min \mathbb{E}[\|\phi(\mathbf{X}_n) - \mathbf{X}_n\|^2], \tag{1}$$

then the well-trained AE is expected to achieve the following ideal objectives:

$$\hat{\mathbf{X}}_n = \phi(\mathbf{X}_n) \to \mathbf{X}_n \tag{2}$$

$$\hat{\mathbf{X}}_a = \phi(\mathbf{X}_a) = \phi(\mathbf{X}_n + \delta) \to \mathbf{X}_n, \tag{3}$$

where $\hat{\mathbf{X}}_n$ and $\hat{\mathbf{X}}_a$ represent the reconstructions generated by AE. Consequently, the reconstruction error, $\mathcal{A}_{rec}(\mathbf{X}) = \|\hat{\mathbf{X}} - \mathbf{X}\|^2$, will be $\mathbf{0}$ if $\mathbf{X} = \mathbf{X}_n$ and δ^2 if $\mathbf{X} = \mathbf{X}_a$, thereby highlighting the abnormal regions.

However, the performance of AE is suboptimal due to a mismatch between the training objective (Eq. 1) and the ideal task objectives (Eq. 2 and 3). While the training objective encourages AE to generate reconstructions that are identical to the network input, the expected reconstructions of abnormal images differ from the input during inference. This discrepancy can lead to false negatives, as the AE may successfully reconstruct some abnormal regions due to the generalization ability of deep neural networks. In the extreme case, if AE learns a function $\phi(\mathbf{X}) = \mathbf{X}, \forall \mathbf{X}$, it perfectly satisfies the training objective, but cannot detect any anomalies. This phenomenon is called "identical shortcut" [23,15]. To better understand this issue and seek potential solutions, we conduct a thorough theoretical analysis of properties of AE in the following.

3   Theoretical analysis of AE in anomaly detection

This section theoretically analyzes AE in AD. We firstly elucidate its fundamental property of reconstruction to establish that an AE with an appropriate latent dimension effectively avoids the "identical shortcuts", and then leverage information theory to derive the optimal solution for AE.

3.1   An inherent property of AE

While previous papers [23,15] attribute the failure of *assumption* (ii) to identical shortcut, we contend that such a statement is erroneous as it overlooks the changes in the latent dimension. To rectify this misconception, we present theoretical evidence in Proposition 1 to substantiate our argument.

Proposition 1. *Given an AE (Fig. 1), let $\mathbf{Z}_0 \in \mathbb{R}^D$ be the feature vector before the latent vector $\mathbf{Z} \in \mathbb{R}^d$, $\hat{\mathbf{Z}}_0 \in \mathbb{R}^D$ be the feature vector after $\mathbf{Z}$. Then, if $d < \frac{D}{2}$, the AE cannot learn identical mapping.*

Proof. According to *data processing inequality* [12], post-processing cannot increase information. It suggests that, to achieve identical mapping, the bottleneck of AE must preserve all information content for the input. Otherwise, the decoder would be unable to reconstruct the lost information. Therefore, we simplify the analysis by studying identical mapping in the bottleneck.

As shown in Fig. 1, we have $\mathbf{Z} = g_1(\mathbf{Z}_0)$ and $\hat{\mathbf{Z}}_0 = g_2(\mathbf{Z})$, where g_1 and g_2 are FC layers in the bottleneck. Assume $\hat{\mathbf{Z}}_0 \equiv \mathbf{Z}_0$. Considering $\hat{\mathbf{Z}}_0 = g_2(g_1(\mathbf{Z}_0)) = (\mathbf{Z}_0 W^{(1)} + b^{(1)})W^{(2)} + b^{(2)} = \mathbf{Z}_0 W^{(1)} W^{(2)} + (b^{(1)} W^{(2)} + b^{(2)})$, we have:

$$W^{(1)} W^{(2)} = \mathbf{I}_{D \times D}, \tag{4}$$

$$b^{(1)} = \mathbf{0}_{1 \times d}, b^{(2)} = \mathbf{0}_{1 \times D}. \tag{5}$$

Here $W^{(1)} \in \mathbb{R}^{D \times d}, W^{(2)} \in \mathbb{R}^{d \times D}, d^{(1)} \in \mathbb{R}^d, d^{(2)} \in \mathbb{R}^D$ are learnable weights of g_1 and g_2. The solution of Eq. 5 definitely exits. If Eq. 4 has at least one solution, the number of independent scalar equations (D^2) should not exceed the number
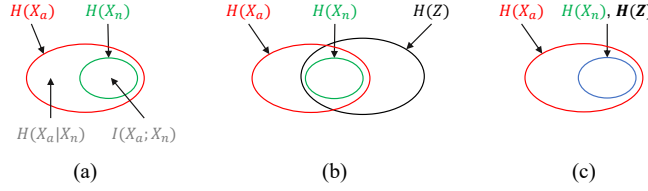
Fig. 2. Venn diagram of $H(\mathbf{X}_n), H(\mathbf{X}_a), H(\mathbf{Z})$.[1] (a) Relationship between $H(\mathbf{X}_n)$ and $H(\mathbf{X}_a)$; (b) $H(\mathbf{Z})$ of an AE trained with Eq. 1; (c) $H(\mathbf{Z})$ of an optimal AE.

of variables $(2Dd)$, i.e., $D^2 \leq 2Dd \Leftrightarrow d \geq \frac{D}{2}$. As a contraposition, if $d < \frac{D}{2}$, Eq. 4 has no solution. Thus, if $d < \frac{D}{2}$, the AE cannot learn identical mapping. □

Proposition 1 reveals that, AE with appropriate latent dimensions can effectively circumvent the undesired identical shortcut. Consequently, we contend that there is no need to introduce more complex modules to address this problem.

3.2 The optimal solution for AE

Although AE with $d < \frac{D}{2}$ does not suffer from identical shortcut, we observe that some undesirable abnormal regions are reconstructed due to the generalization ability of the network. This motivates us to theoretically analyze its reason and ideally, find the optimal solution for guiding the design process. Our argument with theoretical evidence is presented in Proposition 2.

Proposition 2. *Given an AE for anomaly detection, let $\mathbf{X}_n$ be the normal image, $\mathbf{X}_a$ be the abnormal image, and $\mathbf{Z}$ be the latent vector of the AE. An optimal AE should satisfy: (1) $I(\mathbf{X}_n; \mathbf{Z}) = H(\mathbf{X}_n)$, and (2) $I(\mathbf{X}_a; \mathbf{Z}) = H(\mathbf{X}_n)$.*[1]

Proof. Medical abnormal images typically consist of the healthy version with an addition of lesion regions, as shown in Fig. 1 ($\mathbf{X}_a$). Therefore, all possible normal patterns could appear in abnormal images. This can be represented as

$$I(\mathbf{X}_n; \mathbf{X}_a) = H(\mathbf{X}_n) \tag{6}$$

As shown in Fig. 2(a), this reveals that the information content of abnormal data $H(\mathbf{X}_a)$ comprises the information content of normal data $H(\mathbf{X}_n)$ and the information content specific to lesions $H(\mathbf{X}_a|\mathbf{X}_n)$.

The reconstruction process can be formulated as a Markov chain: $\mathbf{X} \to \mathbf{Z} \to \hat{\mathbf{X}}$. According to *data processing inequality* [12], we have:

$$I(\mathbf{X}; \mathbf{Z}) \geq I(\mathbf{X}; \hat{\mathbf{X}}). \tag{7}$$

(1) For a normal input $\mathbf{X}_n$, according to Eq. 2, the optimal AE should satisfy $\hat{\mathbf{X}}_n = \mathbf{X}_n$. Combining Eq. 7, we have $I(\mathbf{X}_n; \mathbf{Z}) \geq I(\mathbf{X}_n; \hat{\mathbf{X}}_n) = I(\mathbf{X}_n; \mathbf{X}_n) =$

---

[1] I indicates the mutual information and H indicates the information entropy [22].

$H(\mathbf{X}_n)$. Using the property of mutual information and joint entropy [12], we have $I(\mathbf{X}_n; \mathbf{Z}) = H(\mathbf{X}_n) + H(\mathbf{Z}) - H(\mathbf{X}_n, \mathbf{Z}) \leq H(\mathbf{X}_n)$. Hence, $I(\mathbf{X}_n; \mathbf{Z}) = H(\mathbf{X}_n)$. (2) For an abnormal input $\mathbf{X}_a = \mathbf{X}_n + \delta$, according to Eq. 3, the optimal AE should satisfy $\hat{\mathbf{X}}_a = \mathbf{X}_n$. To prevent anomaly δ from being reconstructed, $\mathbf{Z}$ should not contain abnormal information $H(\mathbf{X}_a|\mathbf{X}_n)$, so $I(\mathbf{Z}; \mathbf{X}_a|\mathbf{X}_n) = 0$. Combining the property of mutual information [12], we have $I(\mathbf{Z}; \mathbf{X}_a) = I(\mathbf{Z}; \mathbf{X}_a|\mathbf{X}_n) + I(\mathbf{Z}; \mathbf{X}_n; \mathbf{X}_a) = I(\mathbf{Z}; \mathbf{X}_n; \mathbf{X}_a)$. Considering $I(\mathbf{Z}; \mathbf{X}_n) = H(\mathbf{X}_n)$ and $I(\mathbf{X}_a; \mathbf{X}_n) = H(\mathbf{X}_n)$, we simplify that $I(\mathbf{Z}; \mathbf{X}_a) = I(\mathbf{Z}; \mathbf{X}_n; \mathbf{X}_a) = I(\mathbf{X}_n; \mathbf{X}_a) = H(\mathbf{X}_n)$.   □

Proposition 2 demonstrates the conditions that the latent vector of an optimal AE should satisfy: (1) it should provide all information content of normal data, but (2) should not contain any information content of abnormal information, as depicted in Fig. 2(c). Previous AE trained with only Eq. 1 is optimized to achieve $\hat{\mathbf{X}}_n = \mathbf{X}_n$, then it has $I(\mathbf{X}_n; \mathbf{Z}) = H(\mathbf{X}_n), I(\mathbf{X}_a; \mathbf{Z}) \geq H(\mathbf{X}_n)$. Therefore, this AE satisfies Proposition 2(1), but fails to fulfill (2). The Venn diagram in Fig. 2(b) depicts this situation, indicating that its $H(\mathbf{Z})$ extends beyond the scope of $H(\mathbf{X}_n)$ and inadvertently provides information about lesions $H(\mathbf{X}_a|\mathbf{X}_n)$, leading to false negatives. This is an intractable issue since abnormal images are unavailable during training. To address this problem and achieve Proposition 2(2), it is ideal for $H(\mathbf{Z})$ to be minimized and approach the same level as $H(\mathbf{X}_n)$, transforming the scope of $H(\mathbf{Z})$ from Fig. 2(b) to (c). This requirement can be formulated as a regularization:

$$\min H(\mathbf{Z}), \text{if } H(\mathbf{Z}) > H(\mathbf{X}_n). \tag{8}$$

Combining Eq. 1 with 8, $H(\mathbf{Z})$ no longer contains abnormal information but retains normal information, thereby guaranteeing the reconstruction to be normal.

In summary, our theory suggests that in AD, AE tends to benefit from minimizing the entropy of the latent space, aiming to approach the entropy of normal data. This ensures that anomalies cannot be represented and reconstructed by the model. Meanwhile, for more complex datasets with a higher information content, it is required to increase the entropy of the latent space to match that of the normal data. In practice, this can be achieved explicitly through latent dimension adjustment, or implicitly by enforcing latent space restrictions [14,13].

4   Experiments

4.1   Datasets and implementation details

We conduct experiments on four datasets, including RSNA[1], VinDr-CXR[2] [17], Brain Tumor[3], and BraTS2021 [2] for evaluation. We follow [7,8] to build the first three datasets and implement AE. Additionally, we reorganize BraTS2021

---

[1] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

[2] https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection

[3] https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset
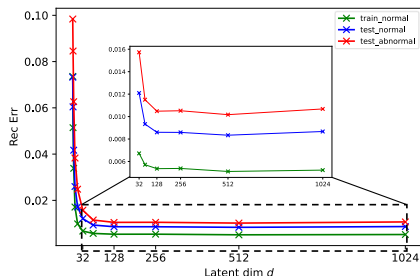
Fig. 3. Reconstruction errors on RSNA dataset w.r.t. the latent dimension.

[2] to evaluate pixel-level AD. The image-level performance is measured using the area under the ROC curve (AUC) and average precision (AP), while at the pixel level, it is assessed by pixel-level AP (AP_{pix}) and the best possible Dice score ($\lceil Dice \rceil$). Details of data and implementation are in the Supplementary.

4.2   Results and analysis

Validation of Proposition 1. Fig. 3 presents reconstruction errors w.r.t. the latent dimension on RSNA dataset, which shows trends that align with our theory. Firstly, we observe that when d is small, an increase in d results in a decrease of reconstruction errors. While $d > \frac{D}{2} = 512$, an increase in d does not lead to smaller errors. This validates that AE with a small d does not encounter identical shortcut, while for $d > \frac{D}{2}$, the bottleneck becomes saturated. Secondly, even when $d > \frac{D}{2}$, errors of normal training data are smaller than those of normal testing data, which, in turn, are smaller than errors of abnormal testing data. This suggests that even if the bottleneck is saturated, identical mapping does not occur in AE due to the limited capacity of the network.

Validation of Proposition 2. We validate Proposition 2 by controlling $H(\mathbf{Z})$ via latent dimension adjustment. Tab. 1 presents the performance of AE with different d on AD, which aligns with our proposition. Firstly, reducing d from

Table 1. Performance of AE with different values of latent dimension d.

Latent dim d	RSNA		VinDr-CXR		Brain Tumor		BraTS2021			
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AP_{pix}	$\lceil Dice \rceil$
128	$60.5_{\pm0.6}$	$60.3_{\pm0.4}$	$48.6_{\pm0.2}$	$51.9_{\pm0.3}$	$86.5_{\pm0.3}$	$76.9_{\pm0.4}$	$80.5_{\pm0.9}$	$90.5_{\pm0.7}$	$25.6_{\pm2.9}$	$32.7_{\pm2.3}$
64	$60.7_{\pm0.6}$	$60.7_{\pm0.6}$	$49.0_{\pm0.2}$	$52.0_{\pm0.3}$	$86.8_{\pm0.1}$	$77.2_{\pm0.2}$	$81.3_{\pm0.2}$	$91.0_{\pm0.2}$	$20.8_{\pm3.1}$	$28.8_{\pm2.8}$
32	$62.8_{\pm0.4}$	$62.5_{\pm0.2}$	$52.5_{\pm0.4}$	$55.6_{\pm0.5}$	$\mathbf{87.1_{\pm0.2}}$	$\mathbf{77.8_{\pm0.2}}$	$82.1_{\pm0.3}$	$91.6_{\pm0.1}$	$25.2_{\pm3.8}$	$32.7_{\pm3.3}$
16	$67.5_{\pm0.9}$	$66.7_{\pm0.5}$	$56.4_{\pm0.4}$	$60.2_{\pm0.5}$	$85.9_{\pm0.3}$	$76.9_{\pm0.1}$	$\mathbf{82.6_{\pm0.1}}$	$\mathbf{92.0_{\pm0.1}}$	$33.2_{\pm2.2}$	$39.2_{\pm1.7}$
8	$69.4_{\pm1.2}$	$68.3_{\pm0.8}$	$57.4_{\pm0.3}$	$61.7_{\pm0.3}$	$84.1_{\pm0.6}$	$75.5_{\pm0.8}$	$81.0_{\pm0.5}$	$90.9_{\pm0.3}$	$\mathbf{36.1_{\pm7.9}}$	$\mathbf{42.3_{\pm5.9}}$
4	$\mathbf{72.9_{\pm2.1}}$	$\mathbf{70.3_{\pm0.9}}$	$\mathbf{61.3_{\pm0.6}}$	$\mathbf{64.1_{\pm0.8}}$	$70.8_{\pm1.4}$	$64.1_{\pm1.0}$	$79.4_{\pm0.1}$	$89.6_{\pm0.3}$	$33.4_{\pm5.4}$	$40.0_{\pm4.2}$
2	$67.5_{\pm0.8}$	$65.4_{\pm1.0}$	$60.5_{\pm0.3}$	$59.7_{\pm0.2}$	$47.3_{\pm1.7}$	$48.3_{\pm1.0}$	$77.2_{\pm0.3}$	$87.5_{\pm0.1}$	$30.0_{\pm2.5}$	$38.0_{\pm1.7}$
1	$66.5_{\pm0.4}$	$63.6_{\pm0.9}$	$59.4_{\pm1.6}$	$58.6_{\pm1.6}$	$36.6_{\pm0.3}$	$42.4_{\pm0.6}$	$72.8_{\pm0.8}$	$84.5_{\pm0.7}$	$35.9_{\pm5.1}$	$40.8_{\pm3.4}$

Table 2. Comparison of typical AD methods. We set $d = 16$ by default. $d_{optimal} = 4, 4, 32, 16$ for RSNA, VinDr-CXR, Brain Tumor, and BraTS2021 datasets, respectively.

Method	RSNA		VinDr-CXR		Brain Tumor		BraTS2021			
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AP_{pix}	$\lceil$Dice$\rceil$
AE	$67.5_{\pm0.9}$	$66.7_{\pm0.5}$	$56.4_{\pm0.4}$	$60.2_{\pm0.5}$	$85.9_{\pm0.3}$	$76.9_{\pm0.1}$	$82.6_{\pm0.1}$	$92.0_{\pm0.1}$	$33.2_{\pm2.2}$	$39.2_{\pm1.7}$
VAE [14]	$67.9_{\pm0.8}$	$67.0_{\pm0.8}$	$56.4_{\pm0.4}$	$60.1_{\pm0.5}$	$85.3_{\pm0.4}$	$76.5_{\pm0.3}$	$80.6_{\pm0.8}$	$90.9_{\pm0.5}$	$44.0_{\pm5.3}$	$47.1_{\pm3.4}$
MemAE [13]	$69.6_{\pm0.4}$	$68.1_{\pm0.4}$	$56.9_{\pm0.9}$	$60.9_{\pm0.9}$	$82.6_{\pm0.6}$	$74.3_{\pm0.2}$	$63.9_{\pm7.7}$	$78.5_{\pm6.9}$	$22.8_{\pm5.6}$	$30.8_{\pm4.8}$
CeAE [24]	$68.0_{\pm0.5}$	$66.8_{\pm0.6}$	$56.1_{\pm0.2}$	$60.1_{\pm0.3}$	$85.3_{\pm1.5}$	$76.0_{\pm1.5}$	$81.7_{\pm1.5}$	$91.5_{\pm0.7}$	$28.6_{\pm3.2}$	$35.8_{\pm2.4}$
AE[$d_{optimal}$]	$\mathbf{72.9}_{\pm2.1}$	$\mathbf{70.3}_{\pm0.9}$	$\mathbf{61.3}_{\pm0.6}$	$\mathbf{64.1}_{\pm0.8}$	$\mathbf{87.1}_{\pm0.2}$	$\mathbf{77.8}_{\pm0.2}$	$\mathbf{82.6}_{\pm0.1}$	$\mathbf{92.0}_{\pm0.1}$	$33.2_{\pm2.2}$	$39.2_{\pm1.7}$

128 to 1 initially improves the performance and then leads to deterioration, with the optimal d typically being quite small. Notably, the performance of $d = 4$ surpasses that of $d = 128$ by more than 10% AUC on both RSNA and VinDr-CXR datasets, indicating that normal information $H(\mathbf{X}_n)$ can be represented by a compact vector. On the other hand, a too large value of d may result in generalization to abnormal samples, while too small value cannot sufficiently represent normal information, leading to performance deterioration.

Secondly, the optimal d varies across different image modalities, reflecting differences in $H(\mathbf{X}_n)$. For RSNA and VinDr-CXR datasets, the optimal d is 4, whereas for Brain Tumor and BraTS2021 datasets, it is 32 and 16, respectively. This disparity can be attributed to the fact that MRIs offer more information content compared to X-rays. MRIs are volumetric scans that capture detailed tissue information, exhibiting greater variations among healthy subjects and encompassing variations among axial slices. These characteristics enable MRIs to surpass the information content of X-rays, necessitating a larger d to effectively expand $H(\mathbf{Z})$ for MRI datasets.

Comparison with other methods. Tab. 2 compares the performance of typical AD methods with AE using the optimal latent dimension. Among these methods, MemAE [13] and VAE [14] incorporate specific designs for latent space restriction, CeAE [24] utilizes an inpainting task to aid in repairing abnormal regions. Following [7], we set $d = 16$ by default for reconstruction methods.

The results further supports Proposition 2. Especially, we observe that directly adjusting the latent dimension in AE (AE[$d_{optimal}$]) yields superior performance compared to more complex approaches that enforce latent space restrictions. AE[$d_{optimal}$] outperforms VAE, MemAE, and CeAE on all four datasets, demonstrating the superiority of the simple and explicit latent dimension adjustment over the more complex implicit restrictions.

5   Conclusion and Discussion

This paper investigated a theoretical analysis of AE in anomaly detection. We prove that an appropriate latent dimension can avoid "identical shortcut" in AE. By leveraging information theory, the optimal solution of AE is uncovered. Our

findings indicate that, apart from the reconstruction loss, imposing a constraint on the entropy of the latent space is crucial for preventing the reconstruction of abnormal information. Experiments validate our theoretical framework, and highlight the efficacy of simple latent dimension reduction in constraining the entropy and achieving significant performance improvements. Overall, this paper provides a theoretical foundation for guiding the design of AE in anomaly detection, facilitating the development of more effective and reliable anomaly detection methods.

However, the current approach for adjusting the latent dimension relies on evaluation results to find the optimal value, which is undesirable. To overcome this limitation, our future work aims to quantify the information entropy of normal training data, $H(\mathbf{X}_n)$, and develop self-adaptive methods that dynamically constrain $H(\mathbf{Z})$ to approach $H(\mathbf{X}_n)$ on different datasets. This approach would eliminate the need for manual selection of the latent dimension and enhance the adaptability of AE in various anomaly detection scenarios.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: Ganomaly: Semi-supervised anomaly detection via adversarial training. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 622–637. Springer (2019)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
4. Baur, C., Denner, S., Wiestler, B., Navab, N., Albarqouni, S.: Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. Medical Image Analysis **69**, 101952 (2021)
5. Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4. pp. 161–169. Springer (2019)
6. Bercea, C.I., Rueckert, D., Schnabel, J.A.: What do aes learn? challenging common assumptions in unsupervised anomaly detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 304–314. Springer (2023)

7. Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.T.: Dual-distribution discrepancy for anomaly detection in chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 584–593. Springer (2022)

8. Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.T.: Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. Medical Image Analysis **86**, 102794 (2023)

9. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys (CSUR) **41**(3), 1–58 (2009)

10. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 (2018)

11. Chen, X., You, S., Tezcan, K.C., Konukoglu, E.: Unsupervised lesion detection via image restoration with a normative prior. Medical image analysis **64**, 101713 (2020)

12. Cover, T.M.: Elements of information theory. John Wiley & Sons (1999)

13. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: 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)

14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

15. Lu, R., Wu, Y., Tian, L., Wang, D., Chen, B., Liu, X., Hu, R.: Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. Advances in Neural Information Processing Systems **36**, 8487–8500 (2023)

16. Mao, Y., Xue, F.F., Wang, R., Zhang, J., Zheng, W.S., Liu, H.: Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 529–538. Springer (2020)

17. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(1), 1–7 (2022)

18. Pang, G., Shen, C., Cao, L., Hengel, A.V.D.: Deep learning for anomaly detection: A review. ACM computing surveys (CSUR) **54**(2), 1–38 (2021)

19. Pinaya, W.H., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. Medical Image Analysis **79**, 102475 (2022)

20. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis **54**, 30–44 (2019)

21. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International conference on information processing in medical imaging. pp. 146–157. Springer (2017)

22. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)

23. You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., Le, X.: A unified model for multi-class anomaly detection. Advances in Neural Information Processing Systems **35**, 4571–4584 (2022)

24. Zimmerer, D., Kohl, S.A., Petersen, J., Isensee, F., Maier-Hein, K.H.: Context-encoding variational autoencoder for unsupervised anomaly detection. arXiv preprint arXiv:1812.05941 (2018)