Deep Feature Selection for Anomaly Detection Based on Pretrained Network and Gaussian Discriminative Analysis

JIE LIN, SONG CHEN, ENPING LIN, AND YU YANG (Member, IEEE)
Department of Electronic Science, Xiamen University, Xiamen 361005, China
CORRESPONDING AUTHOR: Y. YANG (e-mail: yuyang15@xmu.edu.cn)

ABSTRACT Deep learning neural network serves as a powerful tool for visual anomaly detection (AD) and fault diagnosis, attributed to its strong abstractive interpretation ability in the representation domain. The deep features from neural networks that are pretrained on the ImageNet classification task have been proved to be useful for AD based on Gaussian discriminant analysis. However, with the ever-increasing complexity of deep learning neural networks, the set of deep features becomes massive where redundancy appears to be inevitable. The redundant features increase computational cost and degrade the performance of the AD method. In this article, we discuss the deep feature selection for the AD task and show how to reduce the redundancy in the representation domain. We propose a horizontal selection (dimensional reduction) method of features with subspace decomposition and a vertical selection to identify the most effective network layer for AD and fault diagnosis. We test the proposed method on two public datasets, one for AD task and the other for fault diagnosis of bearings. We show the significance of different network layers and feature subspaces on AD tasks and prove the effectiveness of the feature selection strategy.

INDEX TERMS Anomaly detection (AD), deep learning, dimensional reduction, fault diagnosis, Gaussian discriminative analysis.

I. INTRODUCTION

An anomaly is defined as an observation with a considerable difference from the normal observations [1], [2]. The anomaly detection (AD) problem extensively exists in a wide variety of research and industrial fields. For example, for the face presentation attack detection in biometrics systems [3], attack accesses can be considered abnormal and should be distinguished from the real-access data. In the medical field, AD might refer to identifying a disease or locating the lesion region in the medical images [4]. In a network intrusion detection system, the anomaly is a behavior with statistically different patterns from the normal network traffic [5]. Specifically, in industrial fields, AD is commonly related to structural health condition monitoring, and fault diagnosis of the machinery aims to identify the unusual behavior of a system [6] or detect the damage in objects [7], [8]. For example, bearing faults detection and diagnosis, which are critical to avoid mechanical failures, can be modeled as AD problems and solved by machine learning methods [9], [10].

Despite extensive research over the past few decades, there remain several great challenges in AD. First, it is almost impossible to collect all normal or abnormal patterns, which generally have a vague boundary and might vary over time [2], [11]. Second, there is a scarcity of available labeled datasets for AD due to the complication of labeling [12]. Third, the definition of abnormal behavior is diverse in different applications, e.g., the same temperature variation value might be considered abnormal if the data comes from human-body health monitoring but normal in weather monitoring.

There are two critical aspects required to be specified in every AD application—the definition of abnormal and the estimation metric of a testing sample’s difference from the normal samples. In different application scenarios, there are different considerations in these two aspects, which lead to...
different strategies for AD. Regarding the image application, the anomaly can be categorized into the low-level sensory anomaly and high-level semantic anomaly [1]. The damage or crack on the object’s surface is a typical example of a low-level anomaly [13], since it is reflected by the pixel-level features. On the contrary, the high-level semantic anomalies are generally assumed to be generated from latent state variables, such as the identity or property of an object [14]. An example is demonstrated in Fig. 1, where the images of flowers are considered as normal. In this case, the kid masquerading as a flower can be regarded as a low-level (sensory) anomaly sample, and the image of tree leaves that are turning red by traffic light is a high-level (semantic) anomaly.

The strategies of AD usually involve two steps: 1) training and 2) testing. During training, we build a generative model to describe the distribution of normal patterns or a discriminative model to determine the decision boundary between normal and abnormal data. In the testing stage, the distances of the testing samples from the normal patterns are estimated according to the trained model and then applied to judge the abnormality.

A large group of AD methods is based on one-class classification [15], [16]. For example, Kim et al. [15] applied a deep learning network to extract features and then adopted a support vector data description (SVDD) classifier to discriminate abnormal data. These classification-based methods can be classified as the supervised learning strategy requiring both normal and abnormal samples during training. However, supervised learning methods are not realistic in many practical applications because of the scarcity of labeled training data and the class imbalance (i.e., the abnormal samples are often more difficult to collect than normal ones).

On the contrary, semi-supervised or unsupervised learning methods only apply normal samples or contaminated datasets in the training stage [17], [18]. Zong et al. [17] used a deep autoencoder for dimension reduction and leveraged an estimated network to obtain the parameters of a Gaussian mixture model (GMM) as a model of the normal pattern distribution. Sabokrou et al. [18] applied a similar strategy as generative adversarial networks (GANs) and combined an autoencoder and a discriminator to obtain the likelihood of the input sample given the target class. These methods learn the deep network from scratch, which might encounter the overfitting issue under insufficient training data. Since data scarcity is a general problem when applying deep learning neural networks for AD or fault diagnosis tasks, there are mainly two strategies to solve this problem. The first one is transfer learning, which involves pretraining a classifier on a large dataset (e.g., ImageNet) and fine-tuning the classifier using a smaller labeled AD dataset related to specific tasks [19]. It can alleviate the data scarcity problem in AD tasks to some extent but might still have the overfitting (over-confident) problem of the softmax classifier as discussed in [20], especially when the AD dataset is small-scaled. The other strategy is to combine the classic discriminator [e.g., Gaussian discriminant analysis (GDA)] and pretrained deep learning features. In 2018, Lee et al. [20] suggested that the deep features extracted by any pretrained softmax neural classifier can be adopted with GDA for AD, which results in a method with even higher accuracy than fine-tuned softmax classifier. Rippel et al. [21], [22] applied the same idea but further employed principal component analysis (PCA) to reduce deep feature dimensions and achieve state-of-the-art performance.

A general overview shows that the data-driven machine learning strategies, such as deep learning, have superior performance compared to the traditional method on AD tasks. It is widely accepted that the success of the deep learning network relies on its powerful representation ability. The network inputs are transformed into the representation domain and projected into deep features, which are proved to be more effective in many tasks (e.g., classification) than the hand crafted features by traditional methods. After the introduction of the residual network (ResNet) [23] in 2016, with which the gradient vanishing problem within deep neural networks has been well addressed, more complicated neural networks have been proposed with an increasing number of layers and neurons. For example, AlexNet [24] proposed in 2012 as one of the first deep convolutional networks to achieve considerable accuracy, had only five convolutional layers. In 2014, VGGNet [25] with 19 convolutional layers was proposed and soon became one of the most popular image recognition architectures. With the idea of ResNet, an astonishingly high number of layers can be stacked and trained effectively, resulting in complicated networks with more than a hundred layers and millions of neurons. Today, neural networks can have thousands of layers and over 100 million neurons.

Admittedly, a large-scale neural network could have stronger representation ability and extract more kinds of
feature maps with its numerous layers, but do we really need such a massive set of features? In 2019, EfficientNet [26] achieved better performance than ResNet but with a much shallower architecture and fewer parameters. It proved the existence of redundant neurons in a large-scale network and suggested that a leaner network could be more effective.

This article focuses on the low-level AD given only a small-sized training dataset, where abnormal samples might not be available, which is the most common practical situation for AD or fault diagnosis. Motivated by [20], [21], and [22], we choose a pretrained deep learning neural network to extract the features, which are then applied to build a Gaussian distribution model of normal patterns and estimate the likelihood that a testing sample belongs to this normal category. There are two major questions we want to address in this work. First, the representation domain, albeit with a reduced dimension, might be highly redundant for the AD task. How to further reduce or select the features from this representation domain is extremely important, especially when the number of training (normal) samples is less than the number of features. Second, how deep is enough for low-level AD when using the deep neural network for feature extraction?

The contribution of this work is in the following two aspects.

1) We apply a subspace division method to decompose the eigenvectors space of the deep representation domain and analyze the role of each subspace in low-level AD. The subspaces with the most significant contribution to AD are applied for deep feature reduction, which leads to an effective AD method with state-of-the-art performance. This subspace-based dimension reduction method is referred to as the horizontal feature selection in the remainder of this article.

2) We visualize the effect of each network layer, from which the most effective layer for low-level AD tasks can be identified. This procedure can be considered a vertical feature selection. With the visualization results and quantitative analysis, we further reveal the difference between shallow and deep features of the pretrained network and their different effects on AD.

II. PRELIMINARIES

In this article, we adopt a similar generative method as [20], [21], and [22], where the pretrained deep features are fitted with a multivariate Gaussian (MVG).

Given an AD dataset with only normal samples for training, there are three steps in this method. First, a network $\psi_{\theta}(\cdot)$ with pretrained parameters $\theta$, which can be obtained off-the-shelf, needs to be chosen to extract the deep features. Second, all the deep features of the training dataset $X = \{x_i\}$ are obtained as $f_i = \psi_{\theta}(x_i)$, based on which an MVG with parameters $\mu$ (mean) and $\Sigma$ (covariance matrix) is built. Third, during the testing stage, the original data $y$ is mapped to the representation domain and deep features $\psi_{\theta}(y)$ are obtained, then the Mahalanobis distance between $\psi_{\theta}(y)$ and the distribution center $\mu$ is calculated as the confidence score to classify $y$ into normal or abnormal. The framework is shown in Fig. 2(a).

According to the MVG assumption, the features of normal samples $f \in \mathbb{R}^{N_F}$, where $N_F$ is the number of features (feature dimensions), would follow the MVG distribution function:

$$P(f; \mu, \Sigma) = (2\pi)^{-\frac{N_F}{2}}|\Sigma|^{-\frac{1}{2}} \times \exp\left[-\frac{1}{2}(f - \mu)^T\Sigma^{-1}(f - \mu)\right]$$

(1)

where $\mu \in \mathbb{R}^{N_F}$ and $\Sigma \in \mathbb{R}^{N_F \times N_F}$ are the mean vector and covariance matrix of the distribution model. The training process aims to obtain the maximum-likelihood estimation (MLE) of $\mu$ and $\Sigma$:

$$\hat{\mu} = \frac{1}{N_S} \sum_{i=1}^{N_S} f_i$$

(2)

$$\hat{\Sigma} = \frac{1}{N_S - 1} \sum_{i=1}^{N_S} (f_i - \mu)(f_i - \mu)^T$$

(3)

where $N_S$ is the number of training samples.

It should be noted that when the number of samples is small with respect to the number of features (e.g., $N_S$ is less than or close to $N_F$), the above estimations (2) and (3) would produce a large deviation from the ideal values. In particular, the estimated covariance matrix $\hat{\Sigma}$ would contain zero eigenvalues if $N_S$ is not larger than $N_F$, and causes trouble in calculating the probability in (1) since $\hat{\Sigma}$ is invertible. It is a common situation for a relatively small training dataset. To address this issue, [21] applies Ledoit–Wolf shrinkage method [27] to obtain the inverse of $\hat{\Sigma}$.
In the testing stage, the Mahalanobis distance can be calculated given \( \hat{\mu} \) and \( \hat{\Sigma}^{-1} \)

\[
d(g) = \sqrt{(g - \hat{\mu})^T \hat{\Sigma}^{-1} (g - \hat{\mu})} \tag{4}
\]

where \( g \) is the features of the testing sample. If the Mahalanobis distance is large, then the probability of the corresponding sample coming from the distribution (1) is low. In other words, this Mahalanobis distance is a measure of the difference between the testing sample and the normal patterns. Then, a testing sample can be predicted abnormal if the corresponding Mahalanobis distance is above the threshold. Apparently, the Euclidean distance is a special case of Mahalanobis distance, with a stronger assumption that the covariance matrix \( \Sigma \) of the normal features is an identity matrix, which does not hold in the general case. Therefore, using the Euclidean distance would give more biased results than the Mahalanobis distance, which has been proved in [21] and [22].

For choosing an effective threshold shown in Fig. 2(c), a method based on the acceptable false-positive rate (FPR) and \( \chi^2 \)-distribution assumption of \( d(g) \) was suggested in [21] and [22], which could be effective given a sufficiently large training dataset but fail with a relatively small training dataset. In that case, i.e., when the number of training samples is comparative or smaller than the number of features, using the maximum Mahalanobis distance of the validation samples as the threshold might be a more convenient choice.

In [21] and [22], this MVG-based AD strategy is applied to the MVTec AD dataset [28], where the abnormal samples contain defects, such as scratches, dents, contaminations, and some structural changes. The experimental results reveal that the detection accuracy can be further improved by reducing the dimension of deep features \( \psi_\theta(y) \). For the dimensional reduction, PCA is first performed on the features \( \psi_\theta(X) \) of all the normal samples, then the eigenvectors corresponding to the most significant eigenvalues are discarded, which results in a space named NPCA with much lower dimensions. The deep features are projected in this NPCA space to reduce the redundancy in the representation domain. This subspace method was previously applied in the pixel space for face recognition [29], and we find it is also effective in the latent feature space.

First, PCA is performed on the estimated covariance matrix in (3) to obtain eigenvectors and eigenvalues

\[
\hat{\Sigma} = \frac{1}{N_S - 1} (F - \mu)(F - \mu)^T = VD^T \tag{5}
\]

where \( F = [f_1 \cdots f_{N_S}] \in \mathbb{R}^{N_F \times N_S} \) is the matrix form of all the deep features given by the pretrained network. \( V \) is the matrix of eigenvectors, and \( D \) is a diagonal matrix with the eigenvalues as the diagonal elements, i.e.,

\[
D = \text{diag}(\lambda_1 \lambda_2 \cdots \lambda_{N_F}).
\]

Here, we assume that the eigenvalues are sorted in descending order: \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N_F} \). When \( N_S \) is insufficiently large or features information is too simple and uniform, \( \hat{\Sigma} \) is semi-positive and invertible.

Then, we divide the eigenvectors and eigenvalues into three groups. The first group \( [\lambda_i, V_i]_{i=1}^{N_S} \) contains the largest eigenvalues and the corresponding eigenvectors, representing the feature components with the strongest commonality.
among all the sample features. In the AD task, this group contains more global contextual information, and \( \Phi_1 = \{V_i\}_{i=1}^{m_1} \) is thus named the strong signal subspace. The second group \( \{\lambda_i, V_i\}_{i=m_1+1}^{m_2} \) includes components with weaker commonality, and its corresponding subspace \( \Phi_2 = \{V_i\}_{i=m_1+1}^{m_2} \) is then called the weak signal subspace. The third group \( \{\lambda_i, V_i\}_{i=m_2+1}^{m_N} \) is with negligible eigenvalue \( \lambda_i \approx 0 \), and the corresponding \( \Phi_3 = \{V_i\}_{i=m_2+1}^{m_N} \) is called the orthogonal subspace, which is orthogonal to the signal subspace.

These three subspaces have different roles in AD tasks. The strong signal subspace \( \Phi_1 \) contains dominating information within the features, which should have the most significant contribution to a data reconstruction task. However, when dealing with AD problems, if the anomaly information is subtle (e.g., a low-level texture anomaly), \( \Phi_1 \) would not be helpful. That is the reason why the removal of the most significant eigenvectors would result in an improved performance in [21]. The weak signal subspace \( \Phi_2 \) contains some subtle details in the features, which might be critical for AD since the difference between low-level anomalies and normal patterns is often local and fine-scaled. The orthogonal subspace \( \Phi_3 \) includes feature components that do not exist in the training dataset. Therefore, if the projection of a testing sample on \( \Phi_3 \) is not zero, it might contain features different from the normal patterns, which is a strong indicator of anomaly.

For the above consideration, we combine \( \Phi_2 \) and \( \Phi_3 \) to construct a transformation matrix

\[
V_S = [\Phi_2, \Phi_3] \in \mathbb{R}^{N_F \times (N_F - m_1)}
\]

for the dimensional reduction of each feature

\[
\tilde{f}_i = V_S^T f_i \in \mathbb{R}^{(N_F - m_1)}.
\]

Then, the mean vector and covariance matrix of the reduced-dimensional features \( \tilde{f} \) would be

\[
\tilde{\mu} = \frac{1}{N_S} \sum_{i=1}^{N_S} \tilde{f}_i = V_S^T \hat{\mu}
\]

\[
\tilde{\Sigma} = V_S^T \Sigma V_S = \text{diag}(\lambda_{m_1+1}, \lambda_{m_1+2}, \ldots, \lambda_{N_F})
\]

From (8), we can see that \( \tilde{\Sigma} \) is still invertible since it contains zero eigenvalues. Therefore, we also perform the Ledoit–Wolf shrinkage method to estimate \( \tilde{\Sigma}^{-1} \). The detailed procedure of subspace decomposition is presented in Table 1. Please refer to [29] for more comprehensive descriptions.

It should be noted that the transformation matrix \( V_S \) could also be chosen as \( \Phi_3 \), depending on the specific feature maps \( f_i \) and the redundancy within that representation domain.

Now the MVG model of the normal pattern distribution would follow the same dimension reduction (horizontal selection) as (6):

\[
\tilde{g} = V_S^T g.
\]

And the corresponding Mahalanobis distance is

\[
d(\tilde{g}) = \sqrt{(\tilde{g} - \tilde{\mu})^T \tilde{\Sigma}^{-1} (\tilde{g} - \tilde{\mu})}.
\]

If \( d(\tilde{g}) \) is greater than the threshold, then this sample \( y \) is considered as abnormal.

The working point (threshold) cannot be easily set according to a target FPR, especially when a sum of the Mahalanobis distances of multiple stages is applied [21], [22]. Some empirical approaches might also be applied, e.g., based on a validation dataset, to choose the threshold.

| TABLE 1. Procedure of subspace decomposition. |
|-----------------------------------------------|
| **Input:** eigenvalues \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N_F} \) of the covariance matrix, and the number of training samples \( N_y \). |
| **Main:** Estimate the cut-off point \( m_1 \) and \( m_2 \). |
| 1) Determine the rank of the covariance matrix \( m_y \), which is normally set as \( \min(N_y-1, N_y) \). |
| 2) Calculate the median of \( \{\lambda_1, \lambda_2, \ldots, \lambda_{m_y}\} \): |
| \( \lambda_{med} = \text{median}(\lambda_i | 1 \leq i \leq m_y) \). |
| 3) The starting point of the weak signal subspace is then: |
| \( m_i = 1 + \arg\max \{\lambda_i | \lambda_i < \lambda_{med} + \mu(\lambda_{med} - \lambda_{m_1})\} \). |
| where \( \mu \) is an empirical constant and is simply set to 1 in the following experiments. |
| 4) Estimate the cut-off point \( m_2 \) between weak signal subspace \( \Phi_2 \) and orthogonal subspace \( \Phi_3 \) using the eigenratio-spectrum: |
| \( m_y = \arg\max_{1 \leq i \leq m_2} \frac{\lambda_{m_i}}{\lambda_{m_{i+1}}} \). |
| since at \( m_y \), the neighborhood eigenvalues should have a relatively significant difference. |
| **Output:** Three subspace transformation matrices \( \{\Phi_1, \Phi_2, \Phi_3\} \). |

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

The algorithm’s performance is strongly influenced by the pretrained model. Following the same experimental settings as [21], we employ ResNet [23] and EfficientNet [26] as the feature extractors. These networks have been pretrained on the ImageNet classification task and are applied in Gaussian discriminative analysis for AD without any fine-tuning.

We test the proposed AD method on the MVTec AD [28] dataset, which contains 15 different categories with a total of 5354 high-resolution color images. In the MVTec AD dataset, the training data are all normal images, and the testing data include normal images and abnormal images containing over 70 different types of defects, including scratches, dent marks, contamination, and various structural changes. No data augmentation has been performed in our experiments.

To evaluate the performance of the method, we apply the area under the receiver operating characteristic curve (AUROC), which is a threshold-independent metric. A random classifier has an AUROC close to 0.5, while an ideal classifier should have an AUROC of 1. Since there are 15 categories of objects in the dataset, we perform AD for each category independently, then calculate the mean and standard...
error of the mean (SEM) of the AUROC values of all 15 categories. According to our experiments, EfficientNet-B4, as a variant of EfficientNet, performs better than other models considering the efficiency and accuracy of the detection.

B. COMPARISONS WITH STATE-OF-THE-ART METHODS

The results are shown in Table 2 with a comparison of several state-of-the-art methods, i.e., Reconstruction-by-inpainting-based AD (RIAD) [30], Patched SVDD [31], DifferNet [32], Triplet Networks [33], and Gaussian AD (GAD) [21]. The proposed method has the same framework as GAD, while we adopt a more sophisticated feature reduction method. All these methods adopt neural networks but with different strategies.

Table 2 clearly shows the superiority of the GDA based on deep features as both GAD and our method surpass the other methods. The improvement in the AUROC metric proves the effectiveness of the proposed feature reduction method.

In Table 2, we use all feature maps output from the nine blocks of EfficientNet and average the Mahalanobis distances as described in [21] for both GAD and our method.

C. VISUALIZE THE CLUSTERING EFFECT OF DEEP FEATURES

The effectiveness of our method is attributed to the deep features extracted from the representation domain and the further feature reduction using subspaces $\Phi_2$ and $\Phi_3$. To demonstrate this intuitively, we apply t-SNE [34] to visualize the clustering of anomaly and normal samples in the original pixel space, the shallow feature space with a dimension reduction on $\Phi_1$, and the deep feature space with dimension reduction on $[\Phi_2, \Phi_3]$. According to the previous discussions, the feature space can be divided into the strong signal subspace $\Phi_1$, weak signal subspace $\Phi_2$, and the orthogonal subspace $\Phi_3$, where $\Phi_2$ and $\Phi_3$ play significant roles in AD tasks. On the other hand, different layers extract features with different levels of abstraction. The feature maps extracted from the nine stages of EfficientNet are denoted as $s_1, s_2, \ldots, s_9$. Here, we select $s_1$ and $s_7$ to show the effect of features with different depths.

As shown in Fig. 4(a) and (b), the normal and abnormal samples are clustered together without a clear boundary in the original pixel space and the shallow feature space. On the contrary, normal and abnormal samples can be discerned visually in the deep feature space, as shown in Fig. 4(c), since there is strong intraclass similarity and high interclass variance. The result also shows that $\Phi_2$ and $\Phi_3$ contain key information for AD.

V. DISCUSSIONS

This section further discusses the effects of vertical feature selection (the depth of the network) and horizontal feature selection (dimensional reduction with feature subspace).

A. EFFECT OF HORIZONTAL FEATURE SELECTION

In this section, the vertical selection of features is not performed. We apply the feature maps from all the network levels to study the impact of horizontal feature selection (or dimensional reduction of deep features) on AD. Three variants of ResNet and three variants of EfficientNet, all of which have been pretrained on the ImageNet dataset, are applied. The results of feature reduction are shown in Table 3.

For ResNet, ResNet34 outperforms ResNet18 and ResNet50, whereas B4 outperforms the others for EfficientNet. It suggests that a more complicated model does not necessarily lead to better performance for the AD task. At the same time, we notice that the performance of EfficientNet is better than that of ResNet, which verifies from the experimental point of view that the classification accuracy of the deep model on the ImageNet dataset has a strong correlation with the role it can play in the AD task. In other words, if the pretrained deep features are effective for classification, they can also be beneficial for AD.

Different columns of Table 3 represent the results of different feature reduction operations. When no dimensional reduction is performed, the AUROC values are roughly between 0.8 and 0.9. When dimensional reduction [by (6)] is performed with the transformation matrix $V_S$ chosen to be the strong signal subspace $\Phi_1$, performance is significantly degraded. On the contrary, dimensional reduction with the subspaces $\Phi_2$ or/and $\Phi_3$ can significantly improve

### Table 2. AUROC values of different algorithms on the MVTec AD dataset.

| Approach        | AUROC ± SEM (%) |
|-----------------|-----------------|
| RIAD[30]        | 91.7 ± 1.8      |
| Patch SVDD[31]  | 92.1 ± 1.7      |
| DifferNet[32]   | 94.7 ± 1.3      |
| Triplet Networks[33] | 94.9 ± 1.2 |
| GAD[21]         | 95.8 ± 1.2      |

The proposed method based on EfficientNet-B4:
- horizontal selection: 96.1 ± 1.1
- horizontal & vertical selection: 96.6 ± 0.9
TABLE 3. AUROC values (± SEM %) of deep feature-based GDA for AD on the MVTec AD dataset with different network structures and different subspace reductions. The bold numbers indicate the best results for each network structure (each row), and the red bolded number is the best one among all these results.

| Horizontal Model | No Reduction | \(\Phi_3\) | \(\Phi_2\) | \(\Phi_1\) | \([\Phi_2, \Phi_3]\) |
|------------------|--------------|------------|------------|------------|-----------------|
| ResNet18         | 84.20 ± 4.45 | 77.55 ± 5.36 | 84.30 ± 3.82 | 87.35 ± 3.63 | 87.09 ± 3.70 |
| ResNet34         | 86.61 ± 4.43 | 81.03 ± 5.59 | 85.45 ± 3.84 | 89.23 ± 3.43 | 88.82 ± 3.50 |
| ResNet50         | 84.98 ± 4.58 | 80.68 ± 5.10 | 86.23 ± 3.89 | 85.00 ± 4.15 | 85.77 ± 4.09 |
| EfficientNet-B2  | 92.91 ± 2.26 | 87.41 ± 3.64 | 93.26 ± 2.03 | 94.97 ± 1.61 | 94.88 ± 1.63 |
| EfficientNet-B4  | 94.06 ± 1.74 | 88.91 ± 3.43 | 94.67 ± 1.44 | 96.13 ± 1.06 | 95.80 ± 1.15 |
| EfficientNet-B6  | 93.52 ± 1.52 | 87.64 ± 2.88 | 94.14 ± 1.22 | 95.95 ± 1.00 | 95.56 ± 1.04 |

The performance. It confirms the effectiveness of subspace decomposition and dimensional reduction. Furthermore, when vertical feature selection is not performed (i.e., both deep and shallow features extracted from the network layers with different depths are applied), dimension reduction with \(\Phi_3\) would be better than that with \([\Phi_2, \Phi_3]\). The reason is that redundant information (i.e., information not directly related to the AD tasks) in the shallow features is more than that in the deep features. As a result, in this case, condensing the features to a more considerable extent is important for the method to suppress the interference induced by the redundant information.

B. EFFECT OF VERTICAL FEATURE SELECTION

Here, we use EfficientNet-B4, which has the best performance in Table 3, as an example to show the effect of vertical feature selection for AD. There are nine different stages in EfficientNet-B4, with feature maps denoted by \(s_1, s_2, \ldots, s_9\). The first and last stages are simple convolutional structures, and the other seven stages are called MBConvBlocks, which are composed of multiple network layers. It is commonly known that each layer in a deep neural network captures different kinds of information in the image. Anomalies in the image, such as scratches, dents, defects, etc., can be summarized as texture anomalies. When the features extracted by a specific layer in the network can focus on the image pixels related to abnormal textures, they are very likely to work well on the AD task.

Deep learning is generally considered a black box that works in an end-to-end way. Researchers have proposed many interesting techniques to open the black box and interpret the mechanism of deep learning, such as class activation mapping (CAM) [35], Grad-CAM [36], and others. Considering AD as a binary classification problem, we use Grad-CAM to visualize the classification criteria of different layers in the pretrained network layers, which can serve as an instruction for selecting the most effective deep features for AD tasks.

The heat map \(H(u, v)\) generated by Grad-CAM is termed GC-Feature, where the spatial information in the original image is kept, and the intensities represent the contribution of the original image pixel at \((u, v)\) on the target feature maps. In other words, considering a target network layer as a decision maker on a specific task, its GC-Feature presents the distribution of attention it focuses on the original image.

We demonstrate an example in Fig. 5, where the heat maps of stages 2–8 are obtained with a broken bottle image as the input of EfficientNet-B4. The GC-Features from some shallow network layers (e.g., \(s_2\)–\(s_4\)) contain low-level information like the overall profile of the object, as shown in Fig. 5(b)–(d), where the edges are sharp and clear, but the texture and defective region are not evident. On the other hand, the GC-Features from the deeper network layers (e.g., \(s_7\) and \(s_8\)) clearly highlight the defective region, which might be due to the difference in the local distribution with regard to the whole image. It is strong evidence of the effectiveness of deep features for AD, even if the feature maps are extracted from a network trained for tasks other than AD (e.g., multiclass classification).

Moreover, it can be observed by comparing Fig. 5(g) and (h) that feature maps from a deeper layer might be too abstractive where some subtle defects might be overlooked. Therefore, deeper features might not work better for the proposed AD method. For a quantitative analysis of this issue, we use feature maps from different stages of EfficientNet-B4 and present the experimental results in Table 4, from which three conclusions can be drawn.

1) For AD tasks, shallow features (e.g., \(s_1\)) are always less useful since the redundancy in these feature maps would interfere with the detection results. An extreme
example is that the $\Phi_1$ subspace of the feature map $s_1$ from the first stage (which is supposed to contain more redundancy than anomaly information) would result in an AUROC of 57.5%, which is close to the random classifier.

2) Deeper features are not always more effective. For example, neither $s_8$ nor $s_9$ could offer more useful information for the AD task than $s_7$, consistent with the intuitive observation in Fig. 5. Therefore, Grad-CAM visualization could be a good indication of vertical feature extraction, and it could be done even when only a few ($\geq 1$) abnormal samples are available.

3) A proper dimensional reduction (or horizontal selection) of the feature would be extremely helpful for shallow feature maps since it can reduce the redundant information. Even when no vertical feature extraction is performed (i.e., both shallow and deep feature maps are applied), the horizontal selection would significantly improve the performance of this method. It is suggested to use $\Phi_3$ as the transformation matrix if too many shallow features are involved, and use $[\Phi_2, \Phi_3]$ if only deep features (e.g., $s_7$) are applied.

### C. FAULT DIAGNOSIS APPLICATIONS

The proposed AD method can be applied to the multi-class classification problem, such as fault diagnosis. When categorizing samples into $C$ different classes, a Gaussian discriminator can be built for each class. We test this method on the Case Western Reserve University (CWRU) bearing fault dataset [37], which contains vibration signals of both normal and faulty bearings under four operational conditions with different loads. Under each operational condition, there are one health state and nine different fault categories with different fault diameters and locations. For this experiment, we only choose the vibration signals under a working load of 3 hp. The time-domain signals are first cropped into nonoverlapped segments, each of which contains 1024 points. These segments are then transformed into images by continuous wavelet transform (CWT), following the same settings as [10], and resized into $300 \times 300$ pixels. Since the spectrum in the CWT domain has only one channel, we use “Jet” colormap to expand the CWT spectrum into a three-channel image with pseudocolor, which can then serve as the input of the deep learning neural networks designed for image processing tasks (e.g., EfficientNet and ResNet). Examples of the resulted images representing the normal state and nine fault states are shown in Fig. 6. After the preprocessing, there are about 118–120 images for each state and 1189 images in the whole dataset. We divide the whole dataset into training (50%) and testing (50%) subsets.

We can see the effect of the deep learning neural network by the t-SNE plots of the deep features from different network stages. In the first stage, the network extracts shallow features from the input data, and the clustering quality of the whole set is not ideal, where samples from some groups (fault-2, fault-3, fault-5, and fault-8) are mixed. On stage-4 and stage-6, the features have significantly better separation. Some previous research, such as [10], applied transfer learning for fault diagnosis on the CWRU dataset. For transfer-learning-based approaches, the adopted deep learning network is pretrained on a large image dataset (e.g., ImageNet), which is similar to our method, and then fine-tuned on the target fault diagnosis dataset. The success of the transfer learning method can be rationalized by Fig. 7, which shows that the pretrained network can map the input signal into a feature space to achieve better separation among different categories. Therefore, only the classification layers in the network are required to be trained to distinguish different fault patterns on the target dataset during the fine-tuning stage of the transfer learning approach. This fine-tuning stage is replaced by the GAD in our method.

The performance of the proposed method on this fault diagnosis case is also evaluated using AUROC. Table 5 shows that stages 4–6 (especially stage 6) provide better features for fault diagnosis than other stages, consistent with
FIGURE 6. CWT images of the preprocessed bearing data in the CWRU dataset. (a) Normal (health) state. (b) Fault 1 (ball bearing fault with 7 mils diameter). (c) Fault 2 (ball bearing fault with 14 mils diameter). (d) Fault 3 (ball bearing fault with 21 mils diameter). (e) Fault 4 (inner-race bearing fault with 7 mils diameter). (f) Fault 5 (inner-race bearing fault with 14 mils diameter). (g) Fault 6 (inner-race bearing fault with 21 mils diameter). (h) Fault 7 (outer-race bearing fault with 7 mils diameter). (i) Fault 8 (outer-race bearing fault with 14 mils diameter). (j) Fault 9 (outer-race bearing fault with 21 mils diameter).

FIGURE 7. Visualization of the clustering by t-SNE dimensional reduction on the CWRU dataset in the representation spaces of (a) stage-1, (b) stage-4, (c) stage-6, and (d) stage-8. The black stars denote the normal samples, and the colored dots denote faulty samples.

the t-SNE analysis shown in Fig. 7. Besides, if all features from all the stages are applied, then dimension reduction (horizontal reduction) on the features would be beneficial to the diagnosis, as shown in the last row of Table 5.

However, it should also be noted that the overall effect of feature selection is quite limited for this case. Without horizontal selection, the classification results (the first column marked as “all features”) are close to or sometimes even better than those with horizontal selection. And the results with vertical selection (stages 1–9) have no significant difference from the ones with features from all the stages. The reason is that the difference between the CWT images of different categories is relatively apparent, as shown in Fig. 6. It does not require too sophisticated feature extraction and dimension reduction algorithms to distinguish samples of different categories. Compared to the AD tasks such as the MVTec AD, where the abnormal features are generally subtle and can only be detected from a small area of the input image, the abnormal features in the CWT images for the CWRU fault diagnosis task are more eye-catching. In other words, although the MVTec AD and the CWRU fault diagnosis can both be considered low-level sensory AD tasks, the CWRU task is apparently at an even lower level than the MVTec AD task.

VI. CONCLUSION

This article aims to provide more theoretical guidance for the GDA for AD based on pretrained deep features, which is a simple but intriguing method. We propose a feature selection strategy and achieve state-of-the-art performance for AD. The feature selection includes vertical and horizontal selections, where the vertical selection is to decide which network layer contributes the most to the AD task and the horizontal selection is to reduce the dimensions of the features by subspace projection. The vertical selection can be assisted with Grad-CAM visualization, while the horizontal selection can be completed with a subspace decomposition procedure based on PCA.

It should be noted that this AD method based on pretrained deep features might not be as effective for high-level semantic AD. It is almost impossible for a network trained on the ImageNet classification dataset to generate useful features for some specific semantic AD problem (e.g., distinguishing the red leaves from a group of flowers as shown in Fig. 1) without any fine-tuning operations. In other words, the deep features learned from an ImageNet classification task can be generalized to low-level AD tasks but might fail on high-level AD tasks. In the future, we will extend this...
study to the semantic anomaly and focus on the difference between the two tasks, which might shed further light on the deep learning mechanism.

**ACKNOWLEDGMENT**

Appreciation is given to Xiaoliang Liu (School of Foreign Languages, Sun Yat-sen University, China) for proofreading this manuscript.

**REFERENCES**

[1] L. Ruff et al., “A unifying review of deep and shallow anomaly detection,” Proc. IEEE, vol. 109, no. 5, pp. 756–795, May 2021, doi: 10.1109/JPROC.2021.3052449.

[2] R. Chalapathy and S. Chawla, “Deep learning for anomaly detection: A survey,” Jan. 2019, arXiv:1901.03407.

[3] S. Fatemifar, S. R. Arashloo, M. Awais, and J. Kittler, “Client-specific anomaly detection for face presentation attack detection,” Pattern Recognit., vol. 112, Apr. 2021, Art. no. 107696, doi: 10.1016/j.patcog.2020.107696.

[4] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” in Proc. Inf. Process. Med. Imag., 2017, pp. 146–157, doi: 10.1007/978-3-319-59050-9_12.

[5] C.-H. Tsang, S. Kwong, and H. Wang, “Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection,” Pattern Recognit., vol. 40, no. 9, pp. 2373–2391, Sep. 2007, doi: 10.1016/j.patcog.2006.12.009.

[6] S. Khan and T. Yairi, “A review on the application of deep learning in system health management,” Mech. Syst. Signal Process., vol. 107, pp. 241–265, Jul. 2018, doi: 10.1016/j.ymssp.2017.11.024.

[7] E. Favarelli and A. Giorgetti, “Machine learning for automatic processing of modal analysis in damage detection of bridges,” IEEE Trans. Instrum. Meas., vol. 70, Nov. 2021, Art. no. 2504013, doi: 10.1109/TIM.2020.3038288.

[8] R. Pandit, D. Infeld, and T. Dodwell, “Operational variables for improving industrial wind turbine yaw misalignment early fault detection capabilities using data-driven techniques,” IEEE Trans. Instrum. Meas., vol. 70, Apr. 2021, Art. no. 2508108, doi: 10.1109/TIM.2020.3073698.

[9] Y. Xu, Z. Li, S. Wang, W. Li, T. Sarkodie-Gyan, and S. Feng, “A hybrid deep-learning model for fault diagnosis of rolling bearings,” Measurement, vol. 169, Feb. 2021, Art. no. 108502, doi: 10.1016/j.measurement.2020.108502.

[10] S. Shao, S. McAleer, R. Yan, and P. Baldi, “Highly accurate machine fault diagnosis using deep transfer learning,” IEEE Trans. Ind. Informat., vol. 15, no. 4, pp. 2446–2455, Apr. 2019, doi: 10.1109/TII.2018.2864759.
[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 25, 2012, pp. 1–9.

[25] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015, arXiv:1409.1556.

[26] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. 36th Int. Conf. Mach. Learn., May 2019, pp. 6105–6114.

[27] O. Ledoit and M. Wolf, “A well-conditioned estimator for large-dimensional covariance matrices,” J. Multivariate Anal., vol. 88, no. 2, pp. 365–411, Feb. 2004, doi: 10.1016/S0047-259X(03)00096-4.

[28] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, “MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 9584–9592, doi: 10.1109/CVPR.2019.00982.

[29] B. Mandal, X. Jiang, H.-L. Eng, and A. Kot, “Prediction of eigenvalues and regularization of eigenfeatures for human face verification,” Pattern Recognit. Lett., vol. 31, no. 8, pp. 717–724, Jun. 2010, doi: 10.1016/j.patrec.2009.10.006.

[30] V. Zavrtanik, M. Kristan, and D. Skočaj, “Reconstruction by inpainting for visual anomaly detection,” Pattern Recognit., vol. 112, Apr. 2021, Art. no. 107706, doi: 10.1016/j.patcog.2020.107706.

[31] J. Yi and S. Yoon, “Patch SVDD: Patch-level SVDD for anomaly detection and segmentation,” 2020, arXiv:2006.16067.

[32] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. 86, pp. 2579–2605, 2008.

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 2921–2929, doi: 10.1109/CVPR.2016.319.

[34] K. Loparo. “Case western reserve university bearing data center Website,” 2012. Accessed Jul. 4, 2022. [Online]. Available: https://engineering.case.edu/bearingdatacenter/welcome

JIE LIN received the B.S. degree in electronic science from Xiamen University, Xiamen, China, where he is currently pursuing the M.S. degree. His research interests include anomaly detection and machine learning.

SONG CHEN received the B.S. degree in electronic science from Xiamen University, Xiamen, China, where he is currently pursuing the M.S. degree. His research interests include signal processing and machine learning.

ENPING LIN received the B.S. degree in electronic science from Fuzhou University, Fuzhou, China, in 2017. He is currently pursuing the Ph.D. degree in electronic science with Xiamen University, Xiamen, China. His research interests include signal processing, convex optimization, and machine learning algorithms.

YU YANG (Member, IEEE) received the B.S. degree in acoustics from Nanjing University, Nanjing, China, in 2004, and the Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2009. From 2009 to 2011, she was a Research Assistant with the Electrical Engineering Department, Tsinghua University. She was a Postdoctoral Researcher with Iowa State University, Ames, IA, USA, and the Massachusetts Institute of Technology, Cambridge, MA, USA, from 2012 to 2015. Since 2015, she has been an Assistant Professor with the Electronic Science Department, Xiamen University, Xiamen, China. Her research interests include signal processing, machine learning, and optimization algorithms.