IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

SPECIAL SECTION ON URBAN COMPUTING AND INTELLIGENCE

Received April 22, 2019, accepted May 15, 2019, date of publication May 22, 2019, date of current version July 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918409

A Stacked Multi-Granularity Convolution Denoising Auto-Encoder

YUN YANG[1,2], LIJUAN CAO[1], QING LIU[2], AND PO YANG[1,2,3]

[1]National Pilot School of Software, Yunnan University, Kunming 650091, China
[2]Kunming Key Laboratory of Data Science and Intelligent Computing, Kunming 650500, China
[3]Department of Computer Science, Liverpool John Moores University, Liverpool L3 5UX, U.K.

Corresponding author: Po Yang (poyangcn@gmail.com)

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 61876166 and Grant 61663046, in part by the Yunnan Applied Fundamental Research Project under Grant 2016FB104, in part by the Yunnan Provincial Young Academic and Technical Leaders Reserve Talents under Grant 2017HB005, in part by the Program for Yunnan High Level Overseas Talent Recruitment, and in part by the Program for Excellent Young Talents of Yunnan University.

ABSTRACT With the development of big data, artificial intelligence has provided many intelligent solutions to urban life. For instance, an image-based intelligent technology, such as image classification of diseases, is widely used in daily life. However, the image in real life is mostly unlabeled, so the performance of many image-based intelligent models shows limitations. Therefore, how to use a large amount of unlabeled image data to build an efficient and high-quality model for better urban life has been an urgent research topic. In this paper, we propose an unsupervised image feature extraction method that is referred to as a stacked multi-granularity convolution denoising auto-encoder (SMGCDAE). The algorithm is based on a convolutional neural network (CNN), yet it introduces a multi-granularity kernel. This approach resolved issues with image unicity by extracting a diverse category of high-level features. In addition, the denoising auto-encoder ensures stability and improves the classification accuracy by extracting more robust features. The algorithm was assessed using three image benchmark datasets and a series of meningitis images, achieving higher average accuracy than other methods. These results suggest that the algorithm is capable of extracting more discriminative high-level features and thus offers superior performance compared with the existing methodologies.

INDEX TERMS Unsupervised learning, feature extraction, denoising auto-encoder, convolutional neural network.

I. INTRODUCTION

With the development of big data, cloud computing and the Internet of Things (IoT), artificial intelligence has brought many conveniences to people's lives. Applications such as intelligent transportation systems and smart medical have improved people's quality of life and improved urban living environment. These intelligent applications in cities involve some emerging technologies, like image classification [1]–[3] and speech recognition. For example, image classification plays an important role in both license plate recognition and medical image analysis [4]. Many researchers have proposed many novel intelligent image classification algorithms in recent years, aiming to continuously improve people's living standards by solving problems in industrial production and daily life. However, with the development of information technology, massive unlabeled image data continues to emerge, and the performance of these intelligent algorithms has declined to meet the needs of fast and efficient data processing. Especially, in actual production and life, the acquisition of image data labels still relies on traditional manual work, which is inefficient and wastes resources. And the correctness of labels depends heavily on prior knowledge, so there is a high demand for personnel relevant domain knowledge. In contrast, there is a large amount of unlabeled image in reality, so research on unlabeled data processing has always been a hot field of study. Currently, there are many feature-based image classification algorithms, but many of these techniques have failed to achieve satisfactory results due to the unicity of extracted image features, which only represent certain aspects of an image (e.g., color) and are not conducive to acquiring comprehensive image information.

The associate editor coordinating the review of this manuscript and approving it for publication was Lu Liu.

The increasing availability of massive data collection and storage have promoted the development of machine learning, particularly for image recognition which requires large training sets.

Shao et al. divided image classification into three primary stages: (1) image preprocessing, (2) feature extraction, and (3) classifier selection and design [5]–[7]. Feature extraction plays an especially critical role in this process because it affects algorithm classification performance. Extraction is conventionally conducted using a multi-type global or local descriptor, including techniques such as local binary pattern (LBP) [8], histogram of oriented gradients [9], scale invariant feature transform (SIFT) [10], and independent component analysis (ICA) [11]. Under certain conditions, these approaches have achieved efficient classification for specific applications or data types (e.g., grayscale images). However, manually selected features are difficult to extend and the performance of these algorithms can be inconsistent when applied to other tasks. These models can also struggle with large data sets because of their excessive dependence on a priori knowledge.

Deep learning (DL) algorithms are currently an active area of research within machine learning and can independently learn high-level features from images [12]. Previous studies have shown that deep neural network, a self-learning algorithm, can extract more general purpose features from any image rather than domain adaptive features for specific tasks [13]. The performance of such algorithms relies on learned features and, as such, they are typically less dependent on a priori knowledge than conventional machine learning models. Such as support vector machine (SVM) which is a generalized linear classifier that classifies data binary according to supervised learning [14], logistic regression (LR) [15], and random forest (RF) that refers to a classification algorithm that uses multiple trees to train and predict samples [16]. Further, neural network have a high level of abstraction ability for image and its hierarchical structure can capture hidden features, which is beneficial to improve the accuracy of the classification of image. Common deep learning algorithms include network-in-network [17], deep belief net (DBN) that is a probability generation model [18], and convolution neural network (CNN) [19]. Effective feature extraction is a critical component required for the successful implementation of any DL algorithm. Among these, CNN is perhaps the most common, due to its peculiarity of shared weights and sparse connections, which can significantly reduce model parameters. Despite its applicability in a wide range of fields, this technique exhibits one obvious drawback: it requires a massive labeled data set. These data are required for training the network using a back-propagating (BP) error approach [20]. The acquisition of data tags is difficult in most practical settings, requiring extensive resources, which limits the development and application of CNNs. In contrast, unsupervised feature learning models can extract features from unlabeled data automatically. In other words, it belongs to the category of unsupervised learning [21]. An auto-encoder (AE) is an unsupervised neural network that does not require labeled data in its training process [20]. Stack AE is a specific deep learning algorithm, the performance of which is improved significantly by stacking deep learning layers [22], [23]. However, AE algorithm characteristic full connectivity between layers, which introduces an excessive number of network parameters. In other words, both CNN and AE suffer from common feature extraction limitations. Specifically, extracted features exist singularly and cannot provide a comprehensive image description.

On the one hand, existing feature extraction algorithms have some limitations. On the other hand, researchers know that getting a lot of man-made work to get labeled data is often difficult, expensive, and time consuming, and a large amount of unmarked data can be easily collected. In an effort to resolve these feature extraction issues from unlabeled data, this study proposes a novel stacked multi-granularity convolution denoising auto-encoder (SMGCDAE). This method can effectively adapt to the continuous increase of unlabeled data volume and provide a new image classification solution for smart applications in urban life. Firstly, this technique effectively integrates a CNN and a denoising auto-encoder (DAE) into the same neural network structure. The resulting network inherits the advantages of a CNN, which can extract robust features from unlabeled data with a lower computational learning cost and fewer parameters. In addition, a multi-granularity convolution kernel is introduced using ensemble learning [24]–[26]. The size of this kernel varies, taking advantage of the fact that different convolution kernels can acquire multifarious features. This approach is beneficial because convolution kernels have different receptive fields, allowing them to capture different image features. As a result, the features extracted in this process exhibit a diverse range of attributes, which can be combined to improve generalization performance [27]. Moreover, we draw on the ideas of the predecessors that multiple MGCDAEs were stacked to form the deep network structure, which was then trained using a greedy layer-wise pre-training approach [28]. In summary, our contributions can be highlighted as follows.

(1) A novel multi-granularity convolution kernel is proposed for automated extraction of image features. These features exhibit diverse characteristics, which can be combined to obtain comprehensive key image features.

(2) This novel extraction approach combines the benefits of CNN and DAE, using unsupervised learning to extract robust features and improve classification accuracy with a smaller computational learning cost and fewer parameters.

(3) The proposed algorithm was applied to a real-world data-meningitis data set, which could be used to effectively assist clinicians in diagnosing the disease.

The remainder of this paper is organized as follows. Section II introduces the CNN and AE algorithms utilized in the proposed technique. Section III presents the details of the MGCDAE and its corresponding deep model, SMGCDAE, developed by stacking seven MGCDAEs. Section IV describes the experimental validation process using benchmark and meningitis data sets. Section V presents the corresponding results and analyzes our approach. Finally, conclusions and future work are included in Section VI.

II. RELATED WORK

This section provides an overview of both the convolution neural network approach and the auto-encoder networks used in our proposed algorithm.

A. CONVOLUTION NEURAL NETWORK

CNNs are composed of a convolution layer, a pooling layer, and a full connection layer, which can process data of multiple arrays, for instance, 1D for time series data [29], [30]; 2D for images [31]; and 3D for video. There are three key technique support CNNs that profit from the properties of ideas: shared weights, local connections and use of many layers. The role of the convolutional layer is to identify local feature conjunctions in the previous layer, and the pooling layer then merges semantically similar features [12]. Each unit is connected to local patches, which mapped via kernels in convolution layer. So, different feature maps utilize different kernels in a layer. The architecture ensure that the data such as images can be detected as comprehensive information as possible. The reason is local features of values are highly correlated in array data. CNNs possess unique local connections and shared weights, giving them powerful feature learning capabilities that can significantly reduce model parameters. As a result, CNNs have been widely applied in the field of faces and hands recognition [32], [33]. In 2012, a CNN model proposed by Krizhevsky and Sutskever won first place at the ImageNet competition [34]. Sun et al. used a CNN for multi-instance object recognition [35]. Mattar et al. proposed a two-dimensional locally connected CNN for remote image segmentation [36], and experimental results showed this method had significant potential for remote image recognition. Nooka et al. proposed a hierarchical classification network based on CNN [37]. However, it requires a massive labeled data set and the static convolution kernel captures single image features, which limits further improvements to classification performance. In view of these limitations, this paper combines CNNs and AE into a single network and proposes multi-granularity convolution kernels to extract integrated features via unsupervised learning.

B. AUTO-ENCODER

AEs consist of a three-layer neural network structure: an input layer, a hidden layer, and an output layer. Data transfer from the input layer to the hidden layer is called encoding. Transfer from the hidden layer to the output layer is referred to as decoding. In this process, an input image acquires a potential representation through an encoding operation, which then reconstructs the input image through a decoding step. The potential or distributed feature representation for the input image is learned by minimizing the reconstructed error. For each input vector x^i, the latent features representation α^{i} and the reconstructed vector z^i can be defined as (1) and (2), respectively.

$$\alpha^{\mathrm{i}} = f(W_1 x^i + b_1) \tag{1}$$

where $W_1 \in \mathbb{R}^{c\times n}$ is a weight matrix of encoder and $b_1 \in \mathbb{R}^n$ is encoding bias vector.

$$z^i = g(W_2 \alpha^i + b_2) \tag{2}$$

In the function, $W_2 \in \mathbb{R}^{c\times n}$ is a matrix between hidden layer and output layer, and $b_2 \in \mathbb{R}^c$ is a decoding bias vector. z^i is the reconstruction vector of x^i. The reconstruction error of the loss function in (3):

$$L(x^i, z^i) = \frac{1}{2}\left\| x^i - z^i \right\|^2 \tag{3}$$

However, an AE simply copies the input data. Although the learned feature representation may perfectly reconstruct the original input data, the abstract features are not adequately representative for specific tasks. As a result, AEs include multiple derivative algorithms such as the convolution auto-encoder (CAE) [38], variational auto-encoder (VAE) [39], sparse auto-encoder (SAE) [40], and denoising auto-encoder (DAE) [41]. DAEs, first proposed by Vincent et al. in 2008, are capable of robust feature extraction. This process can effectively improve model generalization performance by setting some input unit values to 0, encoding, and decoding based on the corrupted input data. This process offers powerful data representation and noise-removal features. As such, it is widely used in music denoising [42] and speech recognition [43]. However, DAEs feature full connectivity between layers, which introduces an excessive number of network parameters. As such, we propose a CNN with unique local connections and shared weights to reduce the required parameters.

III. DESCRIPTION OF OUR APPROACH

Image processing has recently entered a new phase with the development of computer vision. However, in practical implementations, feature extraction is susceptible to interference from complicated factors. For example, inaccurate data collection and instrumentation errors can lead to data deviation. As such, this study proposes a stacked multi-granularity convolution denoising auto-encoder (SMGCDAE). This approach can be divided into two parts: (1) construction of a single MGCDAE based on a back-propagation (BP) algorithm and (2) stacking of a multi-MGCDAE to form a deep network (using a greedy layer pre-training method) with powerful non-linear mapping and high-level feature extraction capabilities.

A. A SINGLE MULTI-GRANULARITY CONVOLUTION DENOISING AUTO-ENCODER

Previous studies have focused on achieving sufficient feature learning, specifically based on unsupervised learning algorithms [44]. It is the truth that image features primarily include color, texture, shape, and spatial relationships. Among them, color feature and cultural feature are global features, which describe the surface properties of the scene corresponding to a certain area of the image. The shape feature mainly describes the contour feature of the image and the spatial relationship feature refers to the spatial position or relative direction relationship corresponding to different targets in the image. Thus, all of the features, which only describe certain aspects of the image. Note that, one of the most import sides in measuring the latent high-level feature representation is whether more information can be capture [45]. In general, multi-category features provide a more thorough analysis of available information and can improve image classification performance [46]. As such, comprehensive descriptions require the integration of a diverse range of features. To this end, we propose the concept of multi-granularity convolution kernels to learn high-level features more effectively. In this process, convolution kernels of varying sizes are utilized in the same convolution layer, with each size kernel corresponding to a specific feature of interest. This design ensures the network extracts a diverse group of high-level features, which are then integrated to represent general image information from various mappings. The motivation for using multiple kernel sizes in a single layer is to extract different features in each image. As an added benefit, the resulting model network is sparser than traditional single convolution kernel methods, making it easier to avoid redundant features. In addition to being highly similar to biological nervous systems, this sparse deep network structure is conducive to representing distributions of data [47].

In this study, convolution kernels of dimensions 1 × 1, 3 × 3, and 5 × 5 were selected to design the MGCDAE pipelines. A 5 × 5 kernel can be problematic as it could increase the computational complexity and required runtime. When applied to convolution layers, the 1 × 1 kernel was primarily used to decrease dimensionality, reduce network parameters, and alleviate computational bottleneck. As such, a convolution layer with a kernel size of 1 × 1 preceded the multi-granularity convolution layer. A 1 × 1 CNN pipeline was also included to ensure sufficient sparsity of the network (see Figure 2), which was constructed to be as sparse as possible to allow local connection of each pipeline. The resulting visual convolution network approximately simulates complex image feature distributions.

Rather than improving generalization performance with existing techniques, this study proposes a new approach with automated feature learning capabilities. DAE is an unsupervised approach, proposed in 2008, for extracting robust features. Its operational theory, that robust features can be learned from noisy images by contaminating the original image, was used in this study. These robust features can improve generalization performance and ensure stability. There are many ways to add noise [40], where we use random Gaussian noise to destroy the original clean input image.

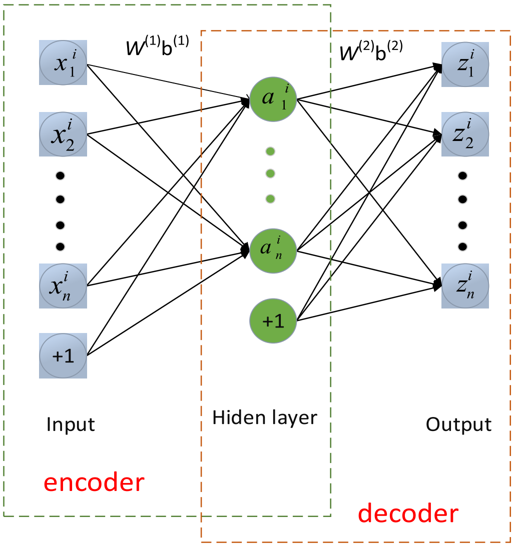


FIGURE 1. The architecture of auto-encoder.

In the encoder shown in Figure 2, a corrupt input vector $\widetilde{x}^i$ is produced from the original input vector x^i by randomly adding Gaussian noise, then enter the nonlinear activation function through linear mapping. As opposed to the AE, the MGCDAE shared weights. The potential representation of the i^{th} feature map was defined for a single-channel input x^i as:

$$\alpha^i = f(W_1 * \widetilde{x}^i + b_1) \tag{4}$$

where W_1 is a weighting matrix for the encoder, b_1 is the encoding bias vector, and $*$ denotes a convolution operation. The term α^i is a latent feature representation and $f(\cdot)$ is an activation function for neurons, which is typically a sigmoid [48] or a Leaky Relu function (as in this study) [49]–[51]. The Leaky Relu function used here can be expressed as:

$$y = \begin{cases} x & x \geq 0 \\ \omega x & x \leq 0 \end{cases} \tag{5}$$

Here, ω is a coefficient. After the addition of stochastic Gaussian noise, the image was input to two hidden layers. It was then transformed to obtain a high-level representation of the hidden layer. These high-level features extracted from different convolution kernels were diverse but each characterized the same image from different perspectives. These features, which had the same dimensions when extracted from a given pipeline, were integrated using a weighted average (see Figure 2). A fusion of features was then performed to acquire comprehensive image descriptions. This fusion approach has been widely used in previous studies to select optimal prognostic parameters [52], [53]. In this paper, feature fusion was achieved by matching the dimensions of the convolution layer. Initially, a feature blending operation fused the features extracted from each previous convolution layer pipeline. Secondly, feature fusion enhanced the contribution of each feature type to the corresponding comprehensive

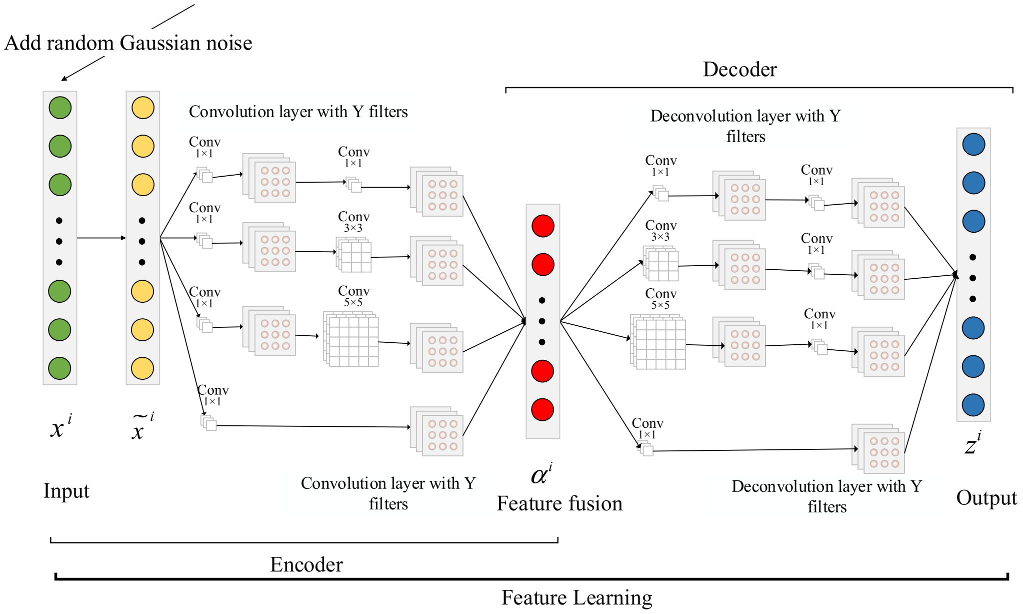


FIGURE 2. The MGCDAE architecture.

Algorithm 1 MGCDAE Training Algorithm, the Training Procedure of Multi-Granularity Convolution Denoising Auto-Encoder

```
Input:
  Training dataset $X$, cost function $J(W, b)$, learning rate $\alpha$, the proportion of noise added $\eta$
Output:
  Parameters $(W_1, b_1)$; Loss value $J(W, b)$.
Randomly set the parameters, $(W_1, b_1)$ $(W_2, b_2)$
Get $\widetilde{x}^i$ by adding stochastic Gaussian noise in $x^i$
For $j = 1$ to $T$ do
    $J(W, b) = \frac{1}{2M} \sum_{i=1}^{M} \left\| x^i - z^i \right\|^2 + \frac{\lambda'}{2} \|W\|_2^2$
    Use the BP algorithm to update $(W_1, b_1)$, $(W_2, b_2)$
end for
return $(W_1, b_1)$ and $J(W, b)$
```

feature, which is beneficial for improving generalization performance [26], [27]. During the decoding step, a potential feature representation α^i (output from the intermediate layer) was used in a nonlinear activation function to output the reconstructed input vector z^i:

$$z^i = g(W_2 * \alpha^i + b_2) \tag{6}$$

In this expression, $g(\cdot)$ is a decoding function, W_2 is a matrix between the hidden and output layers, b_2 is a decoding bias vector, and z^i is a vector reconstruction of x^i. Assuming a given training set $X = \left\{(x_1, y_1), (x_{2,} y_2), \ldots, (x_M, y_M)\right\}$, the overall cost function for the MGCDAE on the data set X can be defined as:

$$J(W, b) = \frac{1}{2M} \sum_{i=1}^{M} \left\| x^i - z^i \right\|^2 \tag{7}$$

A regularization term can then be added:

$$J(W, b) = \frac{1}{2M} \sum_{i=1}^{M} \left\| x^i - z^i \right\|^2 + \frac{\lambda'}{2} \|W\|_2^2 \tag{8}$$

Its role is to prevent overfitting by automatically weakening unimportant feature variables. We considered the encoder to be a feature extractor, which was learned by minimizing the reconstruction error for the cost function in equation (8). In this expression, W and b are the weight matrix and bias vector for the entire MGCDAE network, respectively, and λ' is a regularization coefficient.

B. A STACKED MULTI-GRANULARITY CONVOLUTION DENOISING AUTO-ENCODER

With only four hidden layers, the non-linear mapping capabilities of the MGCDAE are somewhat limited. Wen et al. found that deep neural networks possess remarkable data abstraction capabilities [54], [55]. Inspired by this, we stacked multiple MGCDAEs in a deep neural network based on a greedy layer-wise pre-training algorithm. The first MGCDAE1 was trained by the BP algorithm. The output of the encoder α_1^i was then used as the input for the second MGCDAE2, which was then trained. The latent feature vector α_2^i was then used as the input to MGCDAE3. This process continued through multiple layers as shown in Figure 3. The N stacked MGCDAEs formed a deep stacked MGCDAE (SMGCDAE). The latent feature representation α_N^i was then calculated as:

$$\alpha_N^i = f(W_1^N * \alpha_{N-1}^i + b_1^N) \tag{9}$$

In this expression, W_1^N and b_1^N are the weight matrix and bias vector of the N^{th} MGCDAEN, respectively. In this way, the function conducts further feature mapping in which the

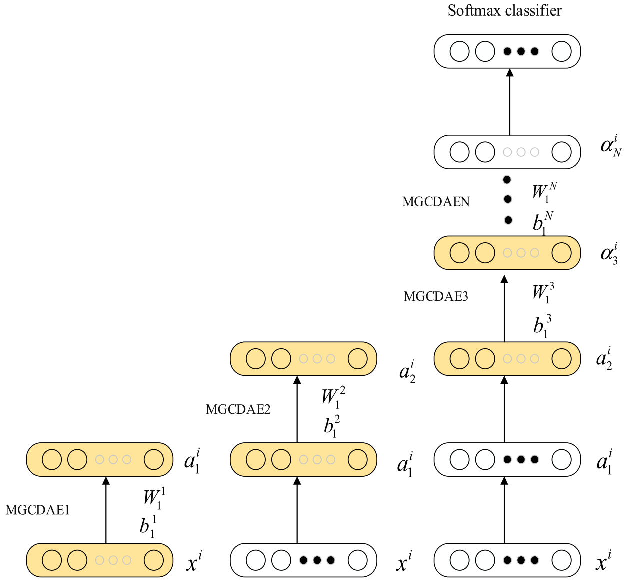


FIGURE 3. The SMGCDAE network architecture.

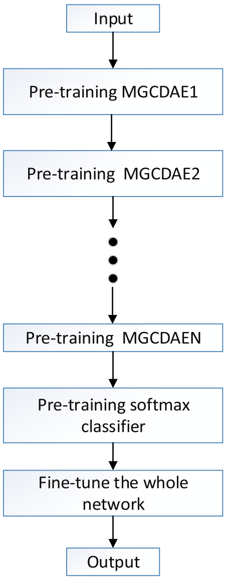


FIGURE 4. The SMGCDAE training process.

output of each hidden layer is an abstract representation of the original image. Therefore, the aim of the deep SMGCDAE is to achieve multiple representation mappings. Feature layers in the initial input image were abstracted layer-by-layer to obtain the final high-level features, which were then input to the classifier after the fusion process. This step completed the final classification task.

Softmax classifiers are typically used for multi-classification in neural networks [4], [56], [57]. In this study, a softmax classifier was used to classify images as follows:

$$f(x) = \frac{1}{\sum_{j=1}^{k} e_j^{\sum_{i=0}^{M} w_i x_i}} \begin{bmatrix} e_1^{\sum_{i=0}^{M} w_i x_i} \\ e_2^{\sum_{i=0}^{M} w_i x_i} \\ \cdots\cdot \\ e_k^{\sum_{i=0}^{M} w_i x_i} \end{bmatrix} \tag{10}$$

where k represents the image category and w_i is the weight of the sample x_i. Training this deep network required a cost function that minimized the reconstruction error, making the output image as similar to the input image as possible. Figure 4 shows the SMGCDE training process, which was divided into two components: pre-training and fine-tuning [58]. In the pre-training phase, the superposition of multiple MGCDAEs was unsupervised in a bottom-to-top fashion. Supervised learning was then used to train the softmax classifier, which had the effect of fine-tuning the entire architecture.

IV. SIMULATION EXPERIMENTS

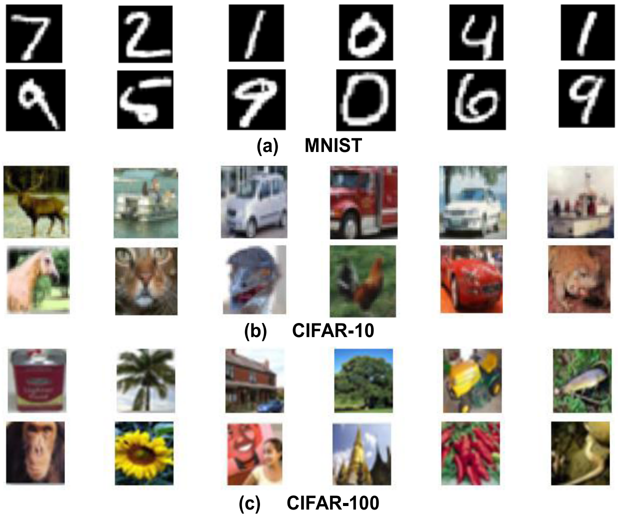
This study primarily investigated a novel method for image feature extraction, based on unsupervised learning, which was then applied to specific image classification tasks. This new approach to feature learning, based on a CNN, has been discussed in detail above. It is necessary to examine the practical efficiency of this model when applied to real data sets. The stability of this approach was verified through a variety of classification tasks, the results of which are reported in this section. Experimental results are analyzed and the advantages of this approach are discussed, along with potential improvements. This verification process was conducted as follows: (1) the fundamental concepts introduced by our approach were compared with existing unsupervised learning algorithms using three benchmark image data sets. (2) Comparative testing was performed with conventional machine learning models using the same data sets. (3) The proposed algorithm was applied to a medical image dataset to verify its stability and effectiveness. Classification accuracy is a commonly adopted metric used to evaluate techniques in the literature [15], [16], [38]. As such, the performance of all comparative approaches was assessed using this criterion:

$$acc(f; D) = \frac{1}{m}\sum_{i=1}^{m} II(f(x_i) = y_i) \tag{11}$$

Here, $II(\cdot)$ represents indicator function, y_i is a true label of x_i and $f(x_i)$ is a predict label produced by a classification algorithm.

A. BENCHMARK DATASET

Benchmark dataset are widely used in the fields of computer vision and neural networks. In this section, we evaluate the performance of the proposed model using the MNIST [19], CIFAR-10 [59], and CIFAR-100 [59] benchmark dataset for general classification tasks. The MNIST handwritten digit classification dataset has 50,000 training examples and 10,000 testing images which consist of binary images with $28 \times 28 \times 1$. These numbers range from 0 to 9, so there are ten classes. Figure 5(a) shows some examples of MNIST dataset. The CIFAR-10 dataset has 60,000 color natural images,



(a) MNIST

(b) CIFAR-10

(c) CIFAR-100

FIGURE 5. Samples of the three image datasets: (a) MNIST, (b) CIFAR-10, (c) CIFAR-100.

TABLE 1. Baseline dataset partitioning.

Dataset	Class	Dimensionality	Training	Testing
MNIST	10	28×28×1	60,000	10,000
CIFAR-10	10	32×32×3	50,000	10,000
CIFAR-100	100	32×32×3	50,000	10,000

TABLE 2. MGCDAE parameter.

	Filters	Stride	Padding	Activation Function
L2	96	2	same	Leaky Relu
L3	128	1	same	Leaky Relu
L4	128	1	same	Leaky Relu
L5	96	1	same	Leaky Relu
L6	-	2	same	Sigmoid

of which 50,000 are training sets and 10,000 are testing sets. And the dimensions of the image are 32 × 32 × 3 pixels. This dataset contains ten classes: (1) bird; (2) ship; (3) airplane; (4) horse; (5) truck; (6) deer; (7) cat; (8) monkey; (9) car; and (10) dog. These classes are completely mutually exclusive. Figure 5(b) demonstrates some examples of this dataset. The dataset of CIFAR-100 is as same as the CIFAR-10 dataset in format and size, which has 100 classes (i.e., fish, flowers, food, insects, household electrical devices, large man-made outdoor things, medium-sized mammals, non-insect invertebrates, et. al) and distributed in the training and testing sets equally. Figure 5(c) displays samples of CIFAR-100 dataset. In this section, we will conduct two phases of experimentation. Differences in accuracy were first investigated between the prototype model and our proposed method. They were then compared with other existing classification approaches to evaluate the effectiveness of our approach.

Using the model configuration listed in Table 2, we first compared the influence of multi-granularity convolution kernels and single-grained convolution kernels on classification performance. As shown in Table 3, three convolution kernel sizes were used in our approach, requiring a comparison to CAE with a single granular convolution kernel. To improve the generalization performance, we added noise to the image when training the model. In contrast to previous studies (i.e., Vincent), the images were corrupted by adding 20% random Gaussian noise. The addition of this noise further verifies the generalization performance of our approach.

TABLE 3. Classification accuracy on benchmark dataset.

	Dataset		
Approach	MNIST	CIFAR-10	CIFAR-100
CAE (1×1)	93.81%	51.71%	34.22%
CAE (3×3)	97.95%	57.67%	37.16%
CAE (5×5)	98.04%	58.33%	40.01%
MGCAE	98.11%	61.03%	42.76%
CDAE (1×1)	93.99%	51.96%	35.01%
CDAE (3×3)	98.12%	58.01%	38.29%
CDAE (5×5)	98.14%	58.68%	40.26%
MGCDAE	98.22%	61.23%	43.99%

In this section, for each approach, we ran the experiment for 10 times and averaged the value of each experimental result in order to achieve a fair comparison. Table 3 displays the average classification accuracy for each method included in the study. On the one hand, MGCAE achieve the classification accuracy of 98.11% on MNIST dataset, 61.03% on CIFAR-10 dataset and 42.76% on CIFAR-100 dataset, respectively. The experimental results are significantly higher than CAE (1 × 1), CAE (3 × 3) and CAE (5 × 5). It is evident that our approach outperformed the others across all three datasets. It is due to that convolution kernels of varying sizes can capture the comprehensive features of an image and contribute to the improvement of classification accuracy. On the other hand, MGCDAE achieve the classification accuracy of 98.22% on MNIST dataset, 61.23% on CIFAR-10 dataset and 43.99% on CIFAR-100 dataset, respectively. Comparing noise free condition and noise adding, the accuracy of our approach has been increasing by 0.11%, 0.2%, and 1.23% respectively. It demonstrates that adding noise can produce a more robust feature extraction in the original image, effectively improving classification accuracy. The effect of varied noise levels (in the training data set) on classification performance are illustrated in Figure 6, Figure 7, and Figure 8. In the figure, 'MGCDAE' corresponds to one MGCDAE, 'stack-3' represents the superposition of three MGCDAEs, 'stack-5' represents the superposition of five MGCDAEs, and 'stack-7' represents the superposition of seven MGCDAEs. Figure 6, Figure 7, and Figure 8 indicate performance for different proportions of added Gaussian noise in each dataset. It is evident that as the number of layers increases the generalization performance of the model gradually increases. Furthermore, adding noise during the training phase will improve generalization performance, as compared to noise-free conditions.

Different types of traditional machine learning classification approaches were investigated to evaluate the stability and

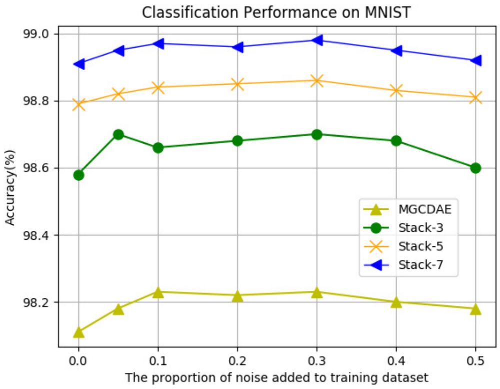


FIGURE 6. Testing accuracy on the MNIST dataset.

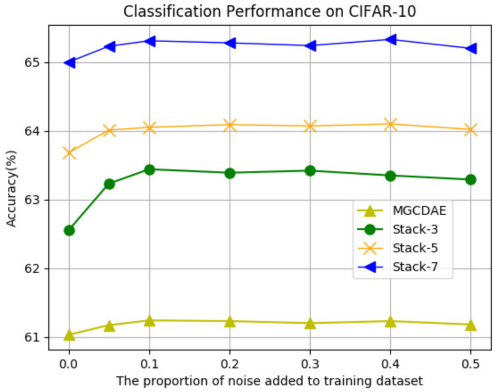


FIGURE 7. Testing accuracy on the CIFAR-10 dataset.

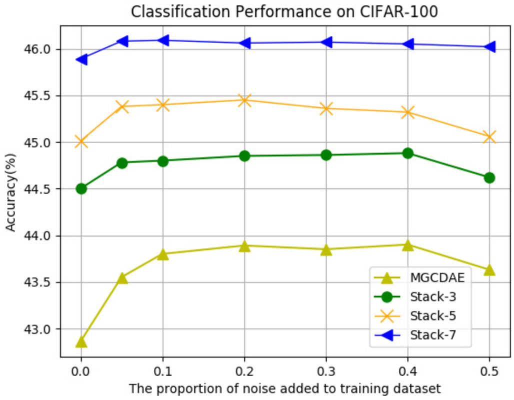


FIGURE 8. Testing accuracy on the CIFAR-100 dataset.

feasibility of our approach (MGCDAE-7), which was compared with Deep Belief Net (DBN), support vector machine (SVM), convolution neural network (CNN), and random forests (RF). Model parameters are shown in Table 2, where the parameters of other techniques were taken from the literature. As shown in Table 4, our approach achieved the highest average classification accuracy of 98.97%, 65.33%, 46.06%, respectively, across all three datasets, demonstrating superior performance. This is likely because our approach not only includes a sparse network structure, which is conducive to representing image distributions, but also can extract robust features. The network layer structure used in our approach could be improved further. While the performance of the completely unsupervised and simple MGCDAE models indicate this to be challenging, the dimensions of the network depth could be leveraged by adding a pooling operation.

TABLE 4. Classification accuracy for traditional machine learning approach.

	Dataset		
Approach	MNIST	CIFAR-10	CIFAR-100
RF	96.80%	50.17%	37.25%
SVM	98.60%	16.32%	28.53%
DBN	98.75%	53.94%	40.37%
CNN	98.81%	63.15%	42.29%
Our Approach	98.97%	65.33%	46.06%

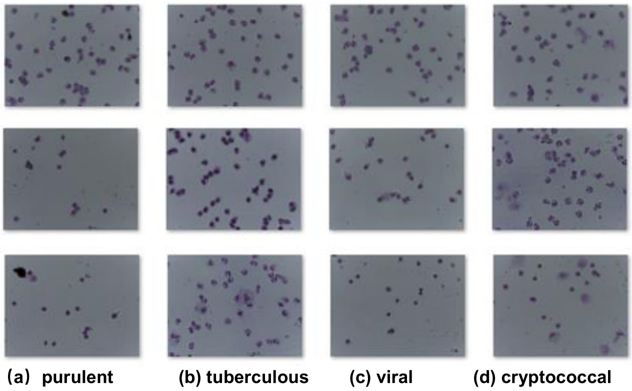

(a) purulent (b) tuberculous (c) viral (d) cryptococcal

FIGURE 9. Image of different type meningitis: (a) purulent, (b) tuberculous, (c) viral, (d) cryptococcal.

B. MENINGITIS DATABASE

In general, Medical image datasets are have two apparent characteristics [60]:

1) The visual characteristics are not always easy-to distinguish, some are visually different while others may be slightly similar.

2) Inherent complex situation in medical image data [61], such as high-dimensionality and the presence of noise.

Thus, the proposed method was applied to a data set of meningitis images acquired from a hospital in Kunming, Yunnan Province, China. This test provided a useful assessment of the approach, as medical images are notably complex compared to other image types. Microscopy images were grouped into four classes corresponding to the four types of meningitis: (a) purulent meningitis, (b) tuberculous meningitis, (c) viral meningitis, and (d) cryptococcal meningitis as Figure 9 shows. The dataset consisted of 1320 RGB cerebrospinal fluid images with dimensions of 2048 1356 3. In the dataset, there are 360 purulent meningitis samples, 290 tuberculous meningitis samples, 368 viral meningitis samples and 332 cryptococcal meningitis samples. By seeing from

Figure 9, in these meningitis images, the cerebrospinal fluid corresponding to each disease type was highly similar when viewed under a microscope, with little difference between classes. Doctors observing this fluid are biased by personal experience, previous knowledge, and other conditions that can affect the final diagnosis. The proposed algorithm could be utilized to remove this bias and improve meningitis classification and treatment.

1) DATA PREPROCESSING

Deep neural networks possess extraordinary abstraction capabilities. However, training a competitive depth model requires a significant amount of data. Due to its limited size, the meningitis image set did not meet the requirements for accurately training a model of this type. Chawla et al. proposed an oversampling data enhancement method known as the Smote algorithm, which synthesizes new samples from a few types [62]. The synthetic strategy randomly selects a sample x_j from its nearest neighbors for each minority class sample x_i. It then randomly selects a point on the line between x_i and x_j as the new class of synthetic samples. This method was employed in our study to process the images and produce 5280 additional samples. However, the size of the images was still relatively large, which increased the complexity of the algorithm and the required runtime. As such, bilinear interpolation with OpenCV was used to process the images and reduce their size. Multiple experiments suggested an optimal input image size of 256 × 256 × 3. After optimizing image quantities and dimensions, the data set was standardized to prevent oversensitivity to different indicators. Upon completion of this process, each indicator is on the same order of magnitude, making it suitable for a comprehensive comparative evaluation. Specifically, we introduced the MAX–MIN scaling method to standardize data according to the following equation:

$$f(x_i) = \frac{x_i - \min(X)}{\max(X) - \min(X)} \tag{12}$$

where $\max(X)$ is the maximum value of X and $\min(X)$ is the minimum value of X. The data set was divided using a random 80/20 division for training and testing, respectively. So it has a training dataset (4224 samples) and a test dataset (1056 samples).

2) EXPERIMENT

The performance of the proposed model in classifying the meningitis data was evaluated by comparing its accuracy with other models. This included prototype methods such as stacked CAE (SCAE), stacked CDAE (SCDAE), and traditional machine learning approaches such as RF, SVM, DBN and the state of the art CNN. Classification was conducted ten times using each model and the average classification accuracy was recorded (see Table 5). The classification accuracy of our approach has achieve 84.02%, which obviously higher than both prototype approach and traditional classification approach. It is evident that our approach was comparable to existing classification models. At the same time, this result demonstrates that our approach achieves high classification accuracy for real-world data set. In practice, image data often include artifacts caused by various factors in the collection, storage, or retrieval processes. The stability of our approach in the presence of noise was investigated using the meningitis data. Gaussian noise was added to the deep model training set to ensure extraction of robust high-level features from the image. Gaussian noise was also included in the testing set to simulate common image artifacts. This experiment was repeated 10 times and the average classification accuracy was recorded for various noise ratios. Experimental results showed that our approach achieved good average classification accuracy (see Figure 10). The reduction is smooth at the noise level, which suggests the method is insensitive to noise and performs well in practical applications. This again validates the effectiveness and robustness of our model.

TABLE 5. Classification accuracy for meningitis dataset.

Approach	Prototype Approach	SCAE (3×3)	57.42%
		SCAE (5×5)	58.89%
		SCDAE (3×3)	57.68%
		SCDAE (5×5)	60.01%
	Traditional Classification Approach	SVM	65.37%
		RF	58.34%
		DBN	72.8%
		CNN	83.52%
	Our Approach	Stack-7	84.02%

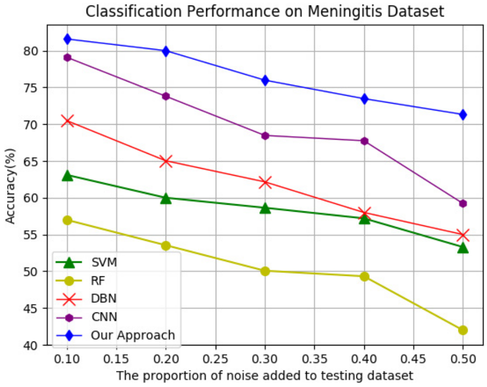


FIGURE 10. Classification performance with noise in the meningitis dataset.

V. DISCUSSION

As reported above experiments, our approach achieved high quality generalization performance with simple implementation technique yet competitive for general classification tasks. By combining a CNN with a DAE, we introduced a promising implementation technique for general classification tasks. A series of simulations provided a comprehensive investigation of the experimental results. The advantages of our proposed algorithm include the following:

First, in order to simplicity and efficiency extract global information of the image, drawing on the idea of ensemble learning, we proposed an approach to automatically extract the abstract high-level features of images from images based on CNN. That is, setting various convolution kernels in the same layer. Different convolution kernels focus on different features, which means that different features can be extracted. These features were then fused, strengthening the contribution of each category to the final classification accuracy. In addition, this construction provided other technical advantages. For example, the multi-granularity deep neural network construction is structured as sparse as possible, which can approximately simulate the complex data distribution characteristics. The conclusion has been tested on a variety of datasets, covering the three benchmarks image dataset. As listed in Table 3, the simulation result demonstrate that our approach provides better classification performance in comparison with its prototype approach.

Secondly, the unsupervised feature extraction method combines the strengths of CNN and DAE by constructing a new symmetric MGCDAE network and avoiding their limitations in classification tasks. On the one hand, based on the convolution operation, we use the unsupervised learning method to effectively learn the high-level abstract features hidden in the unlabeled image. On the other hand, when training the model, Gaussian noise is added to the original input to obtain the corrupted image, then the model is forced to learn robust features from the noise image with a smaller computational learning cost and fewer parameters. According to the results in Table 3, Figures 6-8, the generalization performance of the model is obviously improved by adding noise in the training stage. The other experimental results, Table 4 that evaluate on benchmark datasets, and Figure 8 implement on real-world dataset consistently demonstrate that our approach not only achieve better classification performance than traditional machine learning approaches but also exceed the state-of-the-art classification method like CNN. Moreover, on real-world meningitis dataset, Figure 9 has shown that our approach possesses some robustness when deal with noisy image data compare with other methods. We analyzed that it is due to our special training mechanism, which to some extent increases the model insensitivity to noise data. During all the experiments, we find that the following are particularly indispensable in our approach:

(1) Multi-granularity Convolution Kernels. These convolution kernels not only can extract various high-level feature from images but also make the network as sparse as possible, which is good for simulate the distribution of images.

(2) Adding Noise to Corrupt the Data. In practice, clean data is inevitably corrupted. Then, when we train the model, adding noise makes the model extract more robust features and improve its robustness and generalization performance.

There are still something worth mentioning here. In our work, we chose 1 × 1, 3 × 3, 5 × 5 three type convolution kernels, which just was more convenience. In our framework, other sizes of convolution kernels are also allowed. Moreover, the type of noise added is not limited to Gaussian noise, and other noise is also possible [63]. Further literature [64] proposed a classification approach called denoising and spare auto-encoder by combining DAE and SAE. In their research, they increase the sparseness of the network by adding SAE restrictions in the DAE, making it as sparse as possible to better represent the distribution of data. In contrast, our approach adopt multi-granularity convolution kernels in same layer to make our network similar to human brain. Such design significantly improves not only self-learning ability but also the computing efficiency. This in turn leads to high-quality classification analysis and much faster learning.

Although the experimental results show that our approach reaches a competitive generalization performance on different datasets, there still remain some issues that could impact its practicability. Similar to other deep learning methods, our novel approach is also need a number of data. Deep learning researchers all know that deep learning algorithm is a kind of data-driven methods, which means that the performance of the algorithm depends on the amount of data. So our approach cannot deal with the problem of small sample. Furthermore, while our approach preset possesses three size convolution kernels, how to automatic select the type of convolution kernel for different data types is still a problem to be further studied.

VI. CONCLUSION

In this paper, we proposed a simple but efficient unsupervised approach to image feature extraction (SMGCDA), which combined DAE and CNN to learn and extract high-level features from unlabeled images. This approach improved on existing algorithms by introducing the concept of multi-granularity convolution kernels. Experimental results demonstrated the effectiveness and robustness of this technique. The nonlinear mapping capabilities of a single MGCDAE were improved using a greedy layer-by-layer pre-training method to stack multiple MGCDAEs, forming a deep neural network. Simulation experiments were performed with handwritten digits images, natural images, and microscopy slide images. The results demonstrated the effectiveness and practicability of this approach. In future research, we will further explore a general model and apply it to other images (e.g., color ultrasound of the heart) to solve the practical problems of heart disease recognition.

REFERENCES

[1] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[2] C. Garcia and M. Delakis, "Convolutional face finder: A neural architecture for fast and robust face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, pp. 1408–1423, Nov. 2004.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 770–778.

[4] J. A. Koziol, E. M. Tan, L. Dai, P. Ren, and J. Y. Zhang, "Restricted Boltzmann machines for classification of hepatocellular carcinoma," *Comput. Biol. J.*, vol. 2014, Apr. 2014, Art. no. 418069.

[5] L. Shao, L. Liu, and X. Li, "Feature learning for image classification via multiobjective genetic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1359–1371, Jul. 2014.

[6] H. Yin, X. Jiao, Y. Chai, and B. Fang, "Scene classification based on single-layer SAE and SVM," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3368–3380, May 2015.

[7] Y. Luo, Y. Wen, D. Tao, J. Gui, and C. Xu, "Large margin multi-modal multi-task feature extraction for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 414–427, Jan. 2015.

[8] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 469–481.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[11] Y. Yang, A. Wiliem, A. Alavi, and P. Hobson, "Classification of human epithelial type 2 cell images using independent component analysis," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2014, pp. 733–737.

[12] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[13] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw*, vol. 61, pp. 85–117, Jan. 2015.

[14] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.

[15] G. King and L. Zeng, "Logistic regression in rare events data," *Political Anal.*, vol. 9, no. 2, pp. 137–163, 2001.

[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[17] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400

[18] G. E. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, May 2009.

[19] Y. LeCun , L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Readings Cognit. Sci.*, vol. 323, pp. 399–421, Oct. 1986.

[21] B. H. Barlow, "Unsupervised learning," *Neural Comput.*, vol. 1, no. 3, pp. 295–311, Mar. 1989.

[22] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The difficulty of training deep architectures and the effect of unsupervised pretraining," in *Proc. Artif. Intell. Statist.*, vol. 5, Apr. 2009, pp. 153–160.

[23] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[24] T. G. Dietterich, "Ensemble methods in machine learning," *Proc Int. Workshgp Multiple Classifier Syst.*, Dec. 2000, pp. 1–15.

[25] Y. Yang, *Temporal Data Mining via Unsupervised Ensemble Learning*. Atlanta, GA, USA: Elsevier, 2017.

[26] Y. Yang and J. Jiang, "Hybrid sampling-based clustering ensemble with global and local constitutions," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 952–965, May 2016.

[27] L. Zhao and X. Wang, "A deep feature optimization fusion method for extracting bearing degradation features," *IEEE Access*, vol. 6, pp. 19640–19653, 2018.

[28] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological*, vol. 59, nos. 4–5, pp. 291–294, Sep. 1988.

[29] Y. Yang and K. Chen, "Time series clustering via RPCL network ensemble with different representations," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev*, vol. 41, no. 2, pp. 190–199, Mar. 2011.

[30] Y. Yang and K. Chen, "Temporal data clustering via weighted clustering ensemble with different representations," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 2, pp. 307–320, Feb. 2010.

[31] L. Jin, Y. Yun, X. Wang, Z. Zhao, and T. Li, "A novel parallel distance metric-based approach for diversified ranking on large graphs," *Future Gener. Comput. Syst.*, vol. 88, pp. 79–91, Nov. 2018.

[32] R. Vaillant, C. Monrocq, and Y. L. Cun, "Original approach for the localisation of objects in images," *IEE Proc.-Vision, Image Signal Process.*, vol. 141, no. 4, pp. 245–250, Aug. 1994.

[33] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," in *Proc. Eur. Conf. Comput. Vis.*, Jun. 2005, pp. 43–58.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[35] M. Sun, T. X. Han, M.-C. Liu, and A. Khodayari-Rostamabad, "Multiple instance learning convolutional neural networks for object recognition," in *Proc. Int. Conf. Pattern Recognit.*, Dec. 2017, pp. 3270–3275.

[36] E. A. Mattar and K. M. Al-Rewihi, "Remote image segmentation and understanding via artificial convolution neural network," in *Proc. Middle East Int. GIS Conf. Workshops*, Oct. 1998.

[37] S. P. Nooka, S. Chennupati, K. Veerabhadra, S. Sah, and R. Ptucha, "Adaptive hierarchical classification networks," in *Proc. Int. Conf. Pattern Recognit.*, Dec. 2017, pp. 3578–3583.

[38] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Int. Conf. Artif. Neural Netw.*, 2011, pp. 52–59.

[39] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 16–23.

[40] Y. Bengio, "Learning Deep Architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Nov. 2009.

[41] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2008, pp. 1096–1103.

[42] M. Zhao, D. Wang, Z. Zhang, and X. Zhang, "Music removal by convolutional denoising autoencoder in speech recognition," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Dec. 2016, pp. 338–341.

[43] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder," in *Proc. Interspeech*, Aug. 2013, pp. 3512–3516.

[44] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[45] Y. Qi, C. Shen, D. Wang, J. Shi, X. Jiang, and Z. Zhu, "Stacked sparse autoencoder-based deep network for fault diagnosis of rotating machinery," *IEEE Access*, vol. 5, pp. 15066–15079, 2017.

[46] J. van de Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," *IEEE Trans Image Process*, vol. 16, no. 9, pp. 2207–2214, Sep. 2010.

[47] S. Arora, A. Bhaskara, R. Ge, and T. Ma, "Provable bounds for learning some deep representations," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2014, pp. 584–592.

[48] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, May 1993.

[49] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, Jun. 2013, p. 3.

[50] H. H. Aghdam, E. J. Heravi, and D. Puig,"Recognizing traffic signs using a practical deep neural network," in *Proc. 2nd Iberian Robot. Conf.*, Dec. 2015, pp. 399–410.

[51] C. Zhang and P. C. Woodland, "DNN speaker adaptation using parameterised sigmoid and ReLU hidden activation functions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5300–5304.

[52] J. Coble, P. Ramuhalli, L. Bond, J. W. Hines, and B. Upadhyaya, "A review of prognostics and health management applications in nuclear power plants," *Int. J. Prognostics Health Manage.*, vol. 6, no. 1, p. 016, 2016.

[53] P. Baraldi, G. Bonfanti, and E. Zio, "Differential evolution-based multi-objective optimization for the definition of a health indicator for fault diagnostics and prognostics," *Mech. Syst. Signal Process.*, vol. 102, pp. 382–400, Mar. 2018.

[54] T. Wen and Z. Zhang, "Deep convolution neural network and autoencoders-based unsupervised feature learning of EEG signals," *IEEE Access*, vol. 6, pp. 25399–25410, 2018.

[55] Z. H. Zhou and J. Feng, "Deep forest: Towards an alternative to deep neural networks," 2017, *arXiv:1702.08835*. [Online]. Available: https://arxiv.org/abs/1702.08835

[56] R. Salakhutdinov and G. E. Hinton, "Replicated softmax: An undirected topic model," in *Proc. Adv. Neural Int. Process. Syst.*, 2009, pp. 1607–1614.

[57] B. Chen, W. Deng, and J. Du, "Noisy softmax: Improving the generalization ability of DCNN via postponing the early softmax saturation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4021–4030.

[58] D. Caromel and M. Leyton, "Fine tuning algorithmic skeletons," in *Proc. Int. Euro-Par Conf. Parallel Process.*, 2007, pp. 72–81.

[59] A. Krizhevsky, ''Learning Multiple Layers of Features from Tiny Images,'' vol. 1, 2009.

[60] P. Cao, J. Yang, W. Li, D. Zhao, and O. Zaiane, ''Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD,'' *Comput. Med. Imag. Graph.*, vol. 38, no. 3, pp. 137–150, Apr. 2014.

[61] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, ''Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance,'' *Neural Netw.*, vol. 21, nos. 2–3, pp. 427–436, Mar./Apr. 2008.

[62] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ''SMOTE: Synthetic minority over-sampling technique,'' *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[63] J. H. Liu, W. Q. Zheng, and Y. X. Zou, ''A robust acoustic feature extraction approach based on stacked denoising autoencoder,'' in *Proc. IEEE Int. Conf. Multimedia Big Data*, Apr. 2015, pp. 124–127.

[64] A. Moussavi-Khalkhali, M. Jamshidi, and S. Wijemanne, ''Feature fusion for denoising and sparse autoencoders: Application to neuroimaging data,'' in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, Dec. 2016, pp. 605–610.

YUN YANG received the B.Sc. degree (Hons.) in information technology and telecommunication from Lancaster University, Lancaster, U.K., in 2004, the M.Sc. degree in advanced computing from Bristol University, Bristol, U.K., in 2005, and the M.Phil. degree in informatics and the Ph.D. degree in computer science from The University of Manchester, Manchester, U.K., in 2006 and 2011, respectively. He was a Research Fellow with the University of Surrey, Surrey, U.K., from 2012 to 2013. He is currently a Full Professor of machine learning with the National Pilot School of Software, Yunnan University, Kunming, China, the Director of the Key Laboratory of Data Science and Intelligent Computing, Yunnan Education Department, and the Director of the Kunming Key Laboratory of Data Science and Intelligent Computing. His current research interests include machine learning, data mining, pattern recognition, and temporal data process and analysis. He serves as an Associate Editor for the *Journal of Yunnan University (Natural Sciences Edition)*.

LIJUAN CAO is currently pursuing the master's degree with Yunnan University. Her current research interests include deep learning, and medical data process and analysis.

QING LIU received the B.Sc. and M.Sc. degrees in computer science from Yunnan University, Kunming, China, in 1985 and 1993, respectively. He was a Visiting Scholar with the University of Illinois, in 1999. He is currently a Full Professor of machine learning software engineering and the Associate Dean of the National Pilot School of Software, Yunnan University. His current research interests include software engineering, data science, and intelligent computing.

PO YANG received the B.Sc. degree (Hons.) in computer science from Wuhan University, China, in 2004, the M.Sc. degree in computer science from the University of Bristol, in 2006, and the Ph.D. degree in electronic engineering from the University of Staffordshire, in 2010. He is currently a Senior Lecturer in computer science with Liverpool John Moores University. Since 2006, he has been generating over 70 international journal and conference papers in the fields of pervasive healthcare, image processing, parallel computing, and RFID-related Internet of Things (IoT) applications. His research interests include the Internet of Things, pervasive healthcare, image processing, and parallel computing. He serves as an Associate Editor for the IEEE Journal of Translational Engineering in Health and Medicine and IEEE Access.