Received March 5, 2018, accepted April 3, 2018, date of publication April 6, 2018, date of current version April 25, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2823979

A Greedy Deep Learning Method for Medical Disease Analysis

CHUNXUE WU[1], (Member, IEEE), CHONG LUO[1], (Member, IEEE), NAIXUE XIONG[2], (Senior Member, IEEE), WEI ZHANG[3], AND TAI-HOON KIM[4], (Member, IEEE)

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[2]Department of Mathematics and Computer Science, Northeastern State University, Tahlequah, OK 74464, USA
[3]Department of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310037, China
[4]Department of Convergence Security, Sungshin Women's University, Seongbuk-gu 02844, South Korea

Corresponding author: Naixue Xiong (xiong31@nsuok.edu)

This work was supported in part by the Shanghai Science and Technology Innovation Action Plan Project under Grant 16111107502 and Grant 17511107203 and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY14F020044.

ABSTRACT This paper proposes a new deep learning method, the greedy deep weighted dictionary learning for mobile multimedia for medical diseases analysis. Based on the traditional dictionary learning methods, which neglects the relationship between the sample and the dictionary atom, we propose the weighted mechanism to connect the sample with the dictionary atom in this paper. Meanwhile, the traditional dictionary learning method is prone to cause over-fitting for patient classification of the limited training data set. Therefore, this paper adopts l_2-norm regularization constraint, which realizes the limitation of the model space, and enhances the generalization ability of the model and avoids over-fitting to some extent. Compared with the previous shallow dictionary learning, this paper proposed the greedy deep dictionary learning. We adopt the thinking of layer by layer training to increase the hidden layer, so that the local information between the layer and the layer can be trained to maintain their own characteristics, reduce the risk of over-fitting and make sure that each layer of the network is convergent, which improves the accuracy of training and learning. With the development of Internet of Things and the soundness of healthcare monitoring system, the method proposed have better reliability in the field of mobile multimedia for healthcare. The results show that the learning method has a good effect on the classification of mobile multimedia for medical diseases, and the accuracy, sensitivity, and specificity of the classification have good performance, which may provide guidance for the diagnosis of disease in wisdom medical.

INDEX TERMS Medical big data, machine learning, mobile multimedia, deep learning, dictionary learning, patient classification.

I. INTRODUCTION

With the development of the Internet, big data, cloud computing and artificial intelligence, machine learning is playing an important role in different infrastructures. The combination of machine learning and healthcare are also closely related. Mobile multimedia system for healthcare is important for resource and information management. Meanwhile, Internet of Things (IoT) are now gaining recognition among the health stakeholders as powerful enabling technologies for ubiquitous and widespread healthcare monitoring [1]. Which can make better decisions on patient's diagnoses and lead to overall improvement of healthcare services.

As we all know, there are several major categories of classical machine learning algorithms: neural network, clustering, regression algorithm, decision tree, Bayesian, support vector machine (SVM) and so on [2]. Deep learning is a general term for neural network methods, which is based on learning representations from raw data and contain more than one hidden layer [3]. However, with the rapid development of computer-level, some proposed algorithms remain to be improved, which limited to bottlenecks of the amount of computing and computing capacity at the time. Now machine learning and people's lives are inseparable, for example, it is widely used in the medical diagnosis, recommendation system, weather forecast, environmental supervision and other aspects. With the promotion of wisdom medical care and precision medical, medical big data is closely interrelated with machine learning, which better to drive the development of Internet medical [4].

Nowadays, it is an era of rapid expansion of data, there will be massive data every day in daily life, but it is difficult to filter out the data we need from the hundreds of millions of massive data. In order to deal with massive amounts of data, the technology of big data was born [5]. In fact, the concept of big data really began to slowly popular in 2008. In the special issue of Science, the big data is defined as "represents the progress of human cognitive process, the size of the data set is not in the tolerable time with the current technology, methods and theory to obtain, manage, deal with" [6]. In June 2011, the famous McKinsey & Company issued a study Report: "Big Data: The Next Frontier for Innovation, Competition and Productivity" [7], which announced the arrival of the Big Data Age. The characteristics of big data widely recognized in the traditional sense can be summarized as 4V: Volume, Variety, Velocity and Value [8]. IBM believes that it should also have the characteristics of Veracity [9].

Medical big data is a kind of large number of branches in big data. Accompanied by the wisdom of medical hot, the research of medical data is also in full swing [10]. Medical data for the diagnosis of traditional patients, doctors generally rely on their own empirical knowledge, depending on the clinical symptoms and consultation information to give diagnostic methods, but also brought some diagnostic errors. The development of science and technology has led to the need for rapid and efficient diagnosis and treatment of diseases, while reducing the risk of misdiagnosis caused by clinical diagnosis [11]. This has urge to the study of automated patient classification methods that are conducive to the effective diagnosis and treatment of psychiatric disorders. We have studied the changes in activity in various regions of the brain through blood-oxygen-level-dependent (BOLD) techniques from functional magnetic resonance imaging (fMRI) and the corresponding conditions [12], [13], such as studies of depression and attention deficit hyperactivity disorder (ADHD). IoT provides a new life to the healthcare. One of the better ways is where the doctors are able to certainly and quickly use the relevant patient information through the help of internet of things to take suitable actions [14]. In this paper, the classification of the diagnosis of the patient is studied by combining the knowledge of machine learning and medical big data. The accuracy and error rate of the classification are compared to judge the performance superiority of the algorithm. The application of machine learning in the medical big data will undoubtedly bring a new approach to medical diagnosis.

The paper is organized as follows. We will introduce the relevant research on the relevant situation in Section 2. Then, in Section 3, the proposed GDWDL method is described in the detailed introduction. After that, the verification of the experiment is given and we analyze the result of the experiment in Section 4. Section 5 draws the conclusion and discuss the future work.

II. RELATED WORKS

This paper combines the knowledge of big data and machine learning, and applies the basic algorithms of machine learning to solve the problems of medical big data, such as large amount of data, wide complexity and difficulty in handling. It assists doctors to better treatment for patient through accurate calculation and accurate prediction [15]. In the study of classified medical data from nuclear magnetic resonance imaging, predecessors have extended a series of methods from dictionary learning to make corresponding contributions in this respect.

The original dictionary learning was used for signal reconstruction, but later the researchers applied the classification by means of the label information to supervise the learning [16]. The classification methods of dictionary learning are broadly divided into two categories: one is to learn directly dictionary with the recognition, the other is the sparse representation. Sparse representation is widely used in image classification, and Wright *et al.* [17] proposed sparse representation based classification (SRC) to deal with facial recognition problems, which adopts l_1-norm regularization constraint to ensure the sparsity of coding coefficients. The SRC has a strong robustness to noise on the light and shade. However, the dictionary is pre-defined rather than obtains through a training set. As the amount of data increases, sparse coding calculation will increase, which is not conducive to training. Later, there are different people who propose to learn a self-adapted dictionary for each class to improve the previous method [18]–[20]. At the same time, for the study of dictionary learning, the researchers began to learn from another point of view in-depth study. This method is to make the sparse coefficient of identification, which only need to train to learn a whole dictionary and do not require each class separately to learn a corresponding dictionary. Zhang and Li [21] proposed an improved K-SVD (D-KSVD) based on the K-SVD to construct a dictionary to sparse represent the data and achieved a good representation of the dictionary. Yang *et al.* [22] proposed the Fisher Discriminant Dictionary Learning (FDDL) Method that a structured dictionary atom should be able to learn the corresponding class label by training rather than distinguishing between different classes by representing the remainder. This method requires the divergence is small in the class and the divergence of interclass is larger. Vu *et al.* [23] proposed a simple and effective image classification method named discriminant feature-oriented dictionary learning (DFDL). This method emphasizes the similarity of intra-class and the direct differences of inter-class. Wang *et al.* [24] proposed an improved weighted discriminant dictionary learning method that allows better convergence between training samples and dictionary atoms to achieve better classification effects. The literature [25] introduces the depth nonnegative matrix decomposition and a new matrix decomposition method is proposed. Which is suitable for the low dimension representation of the cluster and is superior to other nonnegative matrix decomposition problem.

Although the above series of methods have achieved the corresponding results, they also have some shortcomings. For instance, the dictionary learning methods treat all the samples indiscriminately and ignore the inner relationship between the dictionary atoms and sample. The constraints of the number of the experimental sample data are prone to cause over-fitting phenomenon. At the same time, some algorithms can only find the local optimum solution rather than the global optimal solution. The new approach proposed in this paper will improve these problems, through layer by layer training, optimization, and gradually approximate the optimal solution. This paper proposes a method, which aims to better facilitate the work of physicians and medical staff in providing healthcare by using mobile devices technology.

III. GREEDY DEEP WEIGHTED DICTIONARY LEARNING ALGORITHM

A. DATA SET

We obtained the required data from the Chinese Academy of Sciences Institute of Automation, and used data from depression as training data, then validated the efficiency of our method with a resting-state fMRI database of attention deficit hyperactivity disorder (ADHD). The data were from the ADHD-200 sample for global competition (http://fcon_1000.projects.nitrc.org/indi/adhd200/). In this paper, we choose 30 data sets of depression as a training set. Which get through transforming the pictures that we get from some mobile devices, such as smartphones, PDA, into digital information through a series of tool. In order to rationalize the data distribution and randomness, we select local people and the number of health and patients of each 15. At the same time, the number of health and patients also include eight men and seven women, whose age is random choice at different stages of a random distribution. All the preparations are to eliminate the influence of additional factors on experimental data. Through training of the training set, we use the data set of ADHD to verify our results in detail, to make experimental analysis and draw conclusions.

B. DATA SET PROCESSING

On one hand, the pictures obtained through some mobile multimedia devices are not clear. On the other hand, it might contain some noise that the clinical data we obtained from the Institute of Automation of the Chinese Academy of Sciences, which is not what we want. Meanwhile, the data we get are too complicated, and some interfering data may interfere with the experimental results and affect the results of the tests. In order to ensure data purity and the correctness of the experiment, we have data source processing to exclude irrelevant data. This paper uses a series of processing methods, such as Statistical Parametric Mapping (SPM8, http://www.fil.ion.ucl.ac.uk/spm/software/spm8/), Resting-State fMRI Data Analysis Toolkit (REST, http://restfmri.net/forum/index.php), and Data Processing Assistant for Resting-State fMRI (DPARST, http://www.restfmri.net/forum/taxonomy/term/36), to ensure the effectiveness of the data.

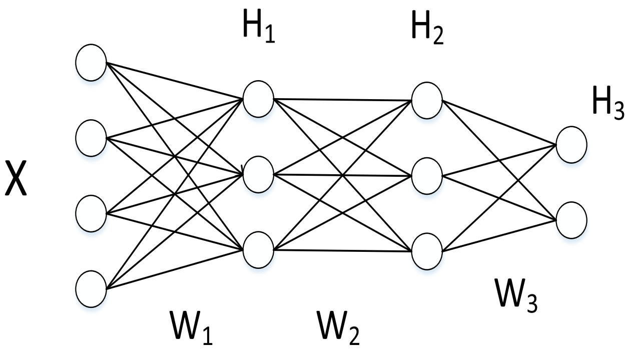


FIGURE 1. Boltzmann machine with three hidden layer.

C. DEEP BOLTZMANN MACHINE

The deep Boltzmann machine (DBM) is developed on the basis of the Restricted Boltzmann machine (RBM). Unlike the RBM with only one hidden layer, DBM has multiple hidden layer structures [26], [27]. As shown in Figure 1, it is the Boltzmann machine structure with the three hidden layer.

RBM is usually an unsupervised learning model, but some researches use class labels to training discriminative Boltzmann machine, and if the hidden layer unit exceeds the corresponding threshold, they will carry on corresponding processing to control the sparseness of learning [28]–[30]. Deep Boltzmann machine is a non-directional learning model, this feedback mechanism is conducive to manage the uncertainty of learning model, which is different from those top-down or bottom-up multi-layer network learning architecture. In the unsupervised way to train a number of restrictions Boltzmann machine, and then reached a good classification effect through the composition of the deep confidence network [31].

D. DICTIONARY LEARNING

Dictionary learning can be simple called sparse coding. We understand the dictionary learning from the perspective of matrix decomposition. In fact, it is equivalent to a given data set X and each column of the matrix X represents a sample. The goal of dictionary learning is to make matrix X decompose into matrix D and matrix K:

$$\mathrm{X} \approx \mathrm{D} * \mathrm{K} \tag{1}$$

Here the coefficient matrix K is as sparse as possible, and each column of D is a normalized vector, D is the dictionary, and each column of D is an atom. In practice, we can directly create the following objective function to carry out the corresponding dictionary learning.

$$\begin{aligned} F = \ & \arg\min_{D,K} \ \|X - DK\|^2 \\ & \text{s.t. } \ \|X\|_0 \leq L \end{aligned} \tag{2}$$

Where L is a constant, which is a sparse constraint parameters. It needs to deal with the corresponding to meet the experimental needs when a certain threshold is not satisfied. For the dictionary learning, there are many other methods of research on the original basis, and there is a detailed description of the expansion thinking in the literature [25], [32], [33].

E. GREEDY DEEP WEIGHTED DICTIONARY LEARNING (GDWDL)

1) OBJECTIVE FUNCTION

We have the clinical data preprocessing and extract the feature of the data information. The extracted data information is expressed as a matrix $\mathrm{X}^{\mathrm{t}} \in \mathbb{R}^{\mathrm{m}\times\mathrm{n}}$, where m represents the number of brain regions, n is the number of code time series. In order to better express the contrast training results of health and disease, we separate training on health group (HG), the matrix expressed as $\mathrm{X} \in \mathbb{R}^{\mathrm{m}\times np}$, the disease group (DG), the matrix is expressed as $\tilde{X} \in \mathbb{R}^{\mathrm{m}\times\mathrm{nq}}$, and p and q are the number of objects each class, respectively. By training we need to find the appropriate dictionary $\mathrm{D}\,(\mathrm{D} = [\mathrm{D}_1, \mathrm{D}_2, \ldots, \mathrm{D}_\mathrm{R}])$ for the sparse representation of the health group HG and the dictionary $\tilde{D}$ used to represent the disease group DG, all of which are $\mathbb{R}^{\mathrm{m}\times\mathrm{R}}$ (where $\mathrm{r} > \mathrm{m}$, $\mathrm{r} \ll \mathrm{np}$, $\mathrm{r} \gg \mathrm{nq}$) with m rows and r columns, R is the number of dictionaries. For simplicity, this paper only discusses and compares the models of the health group, since the patient group and health group are similar in model and can be dealt with in a similar way. The objective function of the health group proposed in this paper is as follows on the basis of dictionary learning:

$$F = \arg\min_{D,K,\tilde{K}} \left(\frac{1}{C} \sum_{i=1}^{C} \frac{W_i}{W} \|x_i - Dk_i\|_2^2 - \frac{\rho}{\tilde{C}} \sum_{j=1}^{\tilde{C}} \frac{\tilde{W}_j}{\tilde{W}} \left\| \tilde{x}_j - D\tilde{k}_j \right\|_2^2 \right)$$
$$\text{s.t.} \ \|d_c\|_2^2 = 1, \quad \|K\|_2 < \varepsilon_1, \quad \left\|\tilde{K}\right\|_2 < \varepsilon_2, \quad c = 1, 2, \ldots r \tag{3}$$

Where $\mathrm{c} = \mathrm{np}$ is the number of columns of X, D is the sparse representation dictionary for the entire training group, $\mathrm{x_i}$ is the column vector of X, $\mathrm{k_i}$ is the coding coefficient of $\mathrm{x_i}$, $\mathrm{W_i}$ is the weight coefficient of $\mathrm{x_i}$, W is the sum of each column vector of the weight ratio. Similarly, $\tilde{C} = \mathrm{nq}$ is the number of columns of $\tilde{X}$, $\tilde{x}_\mathrm{j}$ is the column vector of $\tilde{X}$, $\tilde{k}_\mathrm{j}$ is the coding coefficient of $\tilde{x}_\mathrm{j}$, $\mathrm{W_j}$ is the weight coefficient of $\tilde{x}_\mathrm{j}$, ρ is the regularization parameter. The first expression of the objective function emphasizes that the difference in the internal class is small during the classification process, and the second expression emphasizes the difference between the class and the class. The objective function is set up in order to balance the sparseness of the sample, and to better represent the effect of classification.

It is to improve the accuracy of the classification using weight, to ensure that the various samples fluctuate within their reasonable range, and matches better with the real situation. The weight is defined as follows:

$$W_i = \frac{1}{Z} \exp\left(- \left\| x_i - \bar{d} \right\|_2^2\right) \tag{4}$$

$$\mathrm{W} = \frac{1}{C} \sum_{i=1}^{C} W_i \tag{5}$$

Where Z is the normalized constant, $\bar{d}$ is the mean vector of dictionary atoms in D, and C is the number of column mean vectors. Similarly, the definition for $\tilde{W}_j$ is similar.

2) OPTIMIZATION PROCESS

The matrix decomposition mentioned above is equivalent to a single-layer neural network, and the algorithm proposed in this paper is based on multi-layer dictionary learning, that is, multi-layer matrix decomposition:

$$\mathrm{X} = D_1 * D_2 * \ldots * K \tag{6}$$

The shallow dictionary learning is a non-convex optimization problem. It will make the problem becomes more complex while increasing the hidden layer, and multi-layer dictionary learning to participate in the parameters greatly increased, and sometimes it easily cause the phenomenon of over-fitting in the limited training samples. Therefore, this paper adopts the study thinking of layer-by-layer training based on the previous study, using a similar approach to the method of SAE and DBN [34], [35]. The objective function is modified on the basis of the original. The layer-by-layer training makes ensure that each layer is convergence and the entire training process is perfect and effective. To three-layer decomposition, for instance, the diagram is shown in Figure 2.

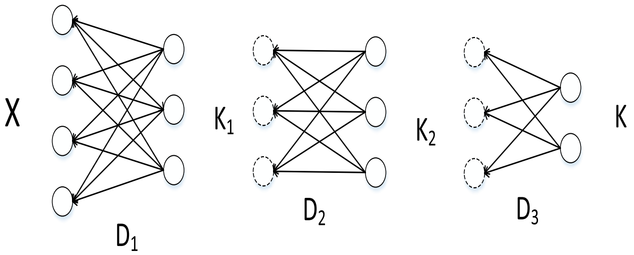


FIGURE 2. **The decomposition process of layer-by-layer learning.**

In fact, the process of deep learning is the first training to learn the feature K_1 and the weighted dictionary D_1 of the first layer.

$$\mathrm{X} = D_1 * K_1 \tag{7}$$

Then learn we treat the feature K_1 learned from training of the first layer as input of the second layer to learn, and then get the second feature K_2 and the weighted dictionary D_2

$$K_1 = D_2 * K_2 \tag{8}$$

And so on, you can achieve deeper dictionary learning. In the whole process of training, in order to ensure more accurate classification, reduce the coupling between class and class, increase the cohesion within the class, better reflect the authenticity of the data, avoid cause problem of the deep neural network error accumulated too long and reduce the error, we hope to get a dictionary each layer can satisfy:

$$D_i * K_i = X_i \tag{9}$$

$$D_j * K_i = 0 (i \neq j) \tag{10}$$

So we increase the constraint of coefficient recognition force, the corresponding constraints on each layer of the dictionary and optimize the coefficient with F-norm regularization constraint. η is a regularization parameter to prevent over-fitting,

$$A_i = \|x_i - Dk_i\|_2^2 + \|x_i - D_i k_i\|_2^2 \tag{11}$$

$$B_j = \left\|\tilde{x}_j - Dk_j\right\|_2^2 + \left\|\tilde{x}_j - D_j k_j\right\|_2^2 \tag{12}$$

The objective function is modified as follows:

$$F^* = arg \min_{D,K,\tilde{K}} (\frac{1}{C}\sum_{i=1}^{C}\frac{W_i}{W}A_i - \frac{\rho}{\tilde{C}}\sum_{j=1}^{\tilde{C}}\frac{\tilde{W}_j}{\tilde{W}}B_j + \eta\sum_{i\neq j}\left\|D_j K_i\right\|_2^2 + \phi\,\|K\|_F^2) \tag{13}$$

While training each layer can be optimized through two ways, one is the dictionary matrix D and $\tilde{D}$ remain unchanged to update the coefficient encoding K and $\tilde{K}$, the other is to maintain the coefficient encoding K and $\tilde{K}$ unchanged to update the dictionary matrix D and $\tilde{D}$. The objective function is solved when the dictionary matrix remains unchanged. Similarly, the two classes are treated in the same way:

$$k_i^* = arg \min_{k_i} (\frac{W_i}{W}\,\|x_i - Dk_i\|_2^2 + \lambda\,\|k_i\|_2^2) \tag{14}$$

Where λ is the regularization parameter.

By solving the objective function with the least squares method, k_i is derived from the above equation, and the derivative is zero. The optimal solution is obtained as follows:

$$\frac{\partial k_i^*}{\partial k_i} = arg \min_{k_i} (2(\frac{W_i}{W}\left(D^T D + I\right) k_i - D^T x_i)) \tag{15}$$

Where I is the unit matrix, $D^T D + I$ and $D^T x_i$ are calculated in advance at each stage. After each layer is updated with the coefficient matrix k_i, $\frac{W_i}{W}\left(D^T D + I\right) k_i$ is used to calculate the best sparse matrix.

When the coding coefficients K and $\tilde{K}$ remain unchanged, the dictionary matrices D and $\tilde{D}$ can be updated by layer-by-layer training and then we update D ($D = [D_1, D_2, \ldots, D_R]$) by setting K unchanged. We update $D_i = [d_1, d_2, \ldots, d_c]$ by training on each layer, when updating D_i, all D_j ($j \neq i$) are kept constant and the objective function can be simplified as:

$$F^{**} = \arg\min_{D} (\frac{1}{C}\sum_{i=1}^{C}\frac{W_i}{W}A_i - \frac{\rho}{\tilde{C}}\sum_{j=1}^{\tilde{C}}\frac{\tilde{W}_j}{\tilde{W}}B_j + \eta\sum_{i\neq j}\left\|D_j K_i\right\|_2^2 + \phi\,\|K\|_F^2)$$
$$\text{s.t. } \|d_c\|_2^2 = 1, \quad c = 1, 2, \ldots r \tag{16}$$

Then we can simplify the solution of the equivalent expression:

$$F = \arg\min_{D_i} (\frac{1}{C}\sum_{i=1}^{C}\frac{W_i}{W}(\|X - D_i K_i\|_F^2 + \|X_i - D_i\|_F^2) + \eta\sum_{i\neq j}\left\|D_j K_i\right\|_F^2)$$
$$\text{s.t. } \|d_c\|_2 = 1, \mathrm{c} = 1, 2, \ldots, \mathrm{R} \tag{17}$$

When $\hat{X} = \mathrm{X} - \sum_{\mathrm{j=1,j\neq i}}^{\mathrm{c}} \mathrm{D_j K_j}$ and $\mathrm{K_i}$ are the representation matrix of X on $\mathrm{D_i}$, the above equation can be written as:

$$F = \arg\min_{D_i} \frac{1}{C}\sum_{i=1}^{C}\frac{W_i}{W}\,\|\Gamma_i - D_i\Psi_i\|_F^2$$
$$\text{s.t. } \|d_c\|_2 = 1, \mathrm{c} = 1, 2, \ldots, \mathrm{R} \tag{18}$$

Where, $\Gamma_i = \left[\hat{X} X_i 0 \ldots 0 0 \ldots 0\right]$, $\Psi_i = [K K_i K_1 \ldots K_{i-1} K_{i+1} \ldots K_c]$, 0 is a zero matrix selected based on the context. The optimal solution of the objective function is solved by updating each dictionary atom and referencing the algorithm described in [23] and [36].

$$z_c = \frac{1}{S'_{c,c}}\left(h_c - Ds'_c\right) + d_c \tag{19}$$

$$d_c = \frac{z_c}{\|z_c\|_2} \tag{20}$$

Where $S'_{c,c}$ represents the value of S' at position (c, c), h_c represents the cth column of matrix X, and s'_c represents the cth column value of S'.

The algorithm proposed in this paper is shown in Table 1.

IV. APPLICATION TEST AND DATA ANALYSIS

In this section, we apply the data validation to the algorithm proposed in this paper to test the performance of the algorithm, such as the accuracy, sensitivity and the mean error rate. Compared differences with the previous algorithms in these performances.

A. ENVIRONMENT AND PARAMETER SETTINGS

We verify validation of test data with a win7 system 4-core and 8G memory, and draw the relevant conclusions through the comparison test results. The normalized parameters $\rho = 0.001$, $\lambda = 0.2$ and $\eta = 0.3$ in the experiment. In order to avoid interference from other factors, all parameters are chose by the same method. We randomly perform 10 experiments to

TABLE 1. Algorithm process.

```
Algorithm : GDWDL
1.  Parameters Initialization
    Initialize the atoms of $D_i$ as the feature vectors of $X_i$.
2.  Input
    Matrix $X$ and $\tilde{X}$, dictionary size r, regularization parameter $\rho$, $\lambda$ and $\eta$
3.  Process Core
    1) while the algorithm is not converged and the training layer isn't the last one do
    2) update the weight W and $\widetilde{W}$ by solving Eq. (4) and (5)
    3) Fix D, update K and $\tilde{K}$ by solving Eq.( 14)
    4) Fix K and $\tilde{K}$ , all $d_l, l \neq i$ update $d_i$ by solving Eq.( 20)
       Update all $d_i$ and hence the whole dictionary $D_i$ is updated
4.  Output
    Return to step 3 until the objective function values achieves the optimal solution or the maximum number of iterations is reached. Then output D and K.
```

verify for each data set we tested, and finally use the average of 10 times to measure the final result.

B. CLASSIFICATION EVALUATION INDICATORS

True positive (TP) refers to the number of samples that the actual sample is abnormal lesions and is detected with lesions. False positive (FP) refers to the number of samples that have been detected without abnormal disease but the actual sample is abnormal disease. True negative (TN) refers to the number of samples that are actually no symptoms and are detected without disease. False negative (FN) refers to the number of abnormal disease but the samples are detected as disease-free [37]. True positive rate (TPR), False positive rate (FPR), Accuracy (Acc), Specificity (Spe) are defined as follows:

$$\mathrm{Sen} = \mathrm{TPR} = \frac{TP}{TP + FN} \tag{21}$$

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

$$\mathrm{Spe} = \frac{TN}{TN + FP} \tag{23}$$

$$\mathrm{FPR} = \frac{FP}{TN + FP} \tag{24}$$

At the same time, we introduce the receiver operating characteristic curve (ROC) to intuitively observe the recognition ability of the disease classification, and judge the superiority of the algorithm proposed by the curve.

C. RESULTS ANALYSES

In this section, we analyze the test results in detail. Compared the differences between the algorithms, we make the chart for an intuitive comparison through many tests and recording the corresponding experimental data. We verify the superiority of the algorithm proposed in the accuracy, sensitivity, specificity and error rate in detail.

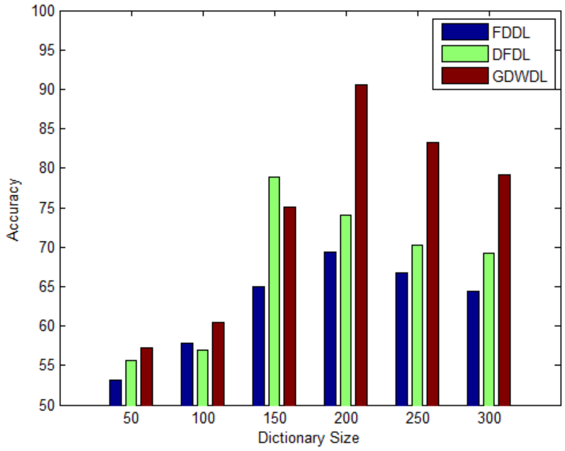


FIGURE 3. Comparison of the accuracy of algorithms with different dictionary size.

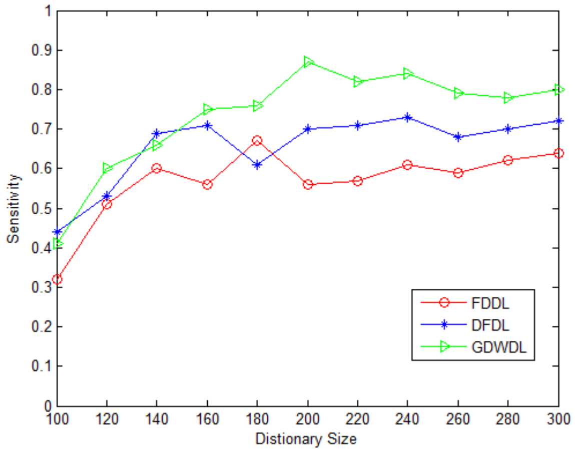


FIGURE 4. Comparison of the sensitivity of algorithms with different dictionary size.

We set up different dictionary sizes to train and verify the accuracy of these algorithms in the data set of depression. As shown in Figure 3, the histogram clearly compares the accuracy differences of these algorithms with different dictionary sizes. We can see that the accuracy of training for different algorithms under different dictionary sizes is different in understanding the overall trend of the algorithm. The GDWDL algorithm proposed in this paper is higher than the other two algorithms from the overall trend, and we learn that it is more appropriate in the dictionary size of 200.

As shown in Figure 4, it presents the sensitivity trends of the algorithms in different dictionary sizes. At the beginning, it is difficult to judge which algorithm is better, because the training rate of data is not high and the learning level is low. However, with the dictionary size increases, the trend is clear when the number of dictionaries reaches 180, after then the

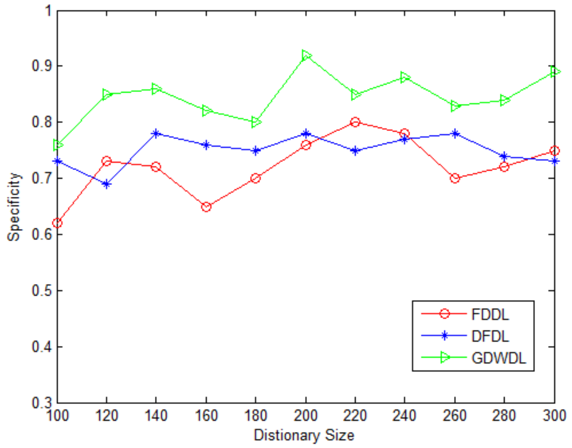


FIGURE 5. Comparison of the specificity of algorithms with different dictionary size.

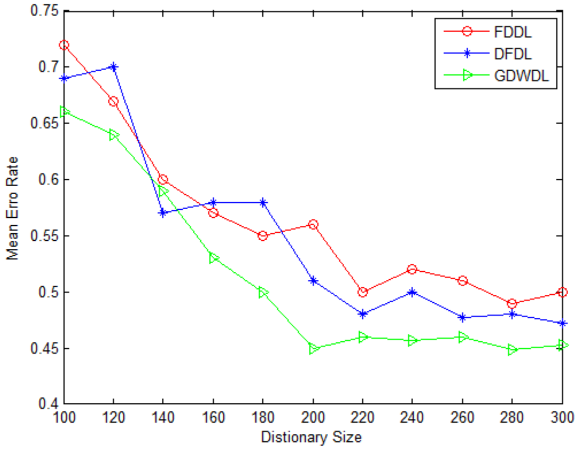


FIGURE 6. Comparison of the mean error rate of algorithms with different dictionary size.

growth trend is apparent slow, and the algorithm proposed in this paper is superior to the other two algorithms in sensitivity. Therefore, as a whole, the GDWDL algorithm proposed in this paper has better sensitivity than the other two algorithms, and it can better the patient classification.

As shown in Figure 5, it presents the specificity trends of the algorithms in different dictionary sizes. Throughout the whole process, the GDWDL algorithm proposed in this paper is higher than the other two algorithms in specificity, which explains that the algorithm proposed in this paper can effectively identify actual disease-free is really negative from the experimental data and accurately improve the classification efficiency of the algorithm.

We train the algorithm in different dictionary sizes and record the error rate data of the experiment several times. The data analysis based on the average value is shown in Figure 6. We have analyzed the error rate of training has been reduced with the increase of the size of the dictionary. However, it does not mean that the dictionary size is bigger the better, and after a period of training, the dictionary size reaches a bottleneck, which is not the main factor on the impact of the error rate, so the late error rate dropped significantly slow. The dictionary size reaches a certain threshold, and then increase dictionary size could not have the average error reduce but will rise. For the algorithm proposed in this paper, the minimum error rate distribution is just about 200 in the dictionary size, indicating that the dictionary size is more reasonable in this point. The graph data shows that the error rate of the algorithm proposed in this paper is generally lower than other algorithms, which further proves that the algorithm is superior to other algorithms.

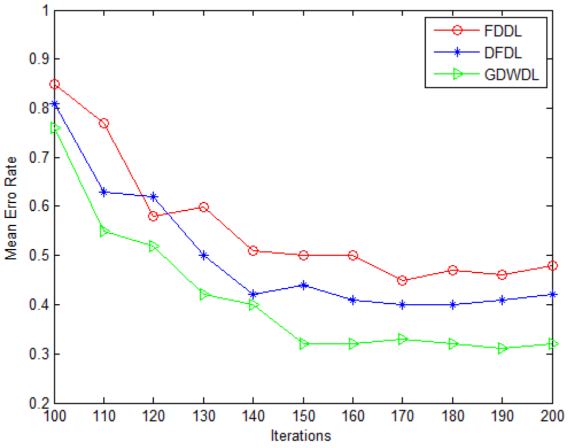


FIGURE 7. Comparison of the average error rates for different iterations.

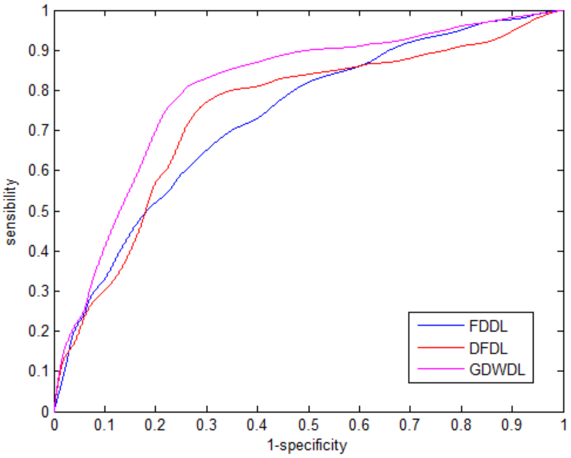


FIGURE 8. ROC curve.

Through the above test, we selected the dictionary size of 200 is more reasonable, and when setting the test sample training, we change the number of iterations to observe and record the mean error rate of the algorithm for the classification. As shown in Figure 7, the overall error rate is decreased as the number of iterations increases, but it is not always reduced, and the latter tends to be smooth and maintained in a specific value neighborhood. The overall error rate of the GDWDL algorithm proposed in this paper is lower than the other two algorithms, the rate of error reduction is relatively large, and it is stable earlier, which has better advantage.

TABLE 2. Time complexity analysis of algorithm.

Algorithm	Complexity	Running Time
FDDL	$O(ckm(h + 2cpk) + c^2mn)$	627.60s
DFDL	$O(ckm(h + p)(2n + L^2))$	314.11s
GDWDL	$O(ck^2m(h + p) + 2kmn)$	187.92s

TABLE 3. Comparison of the algorithms performance on the ADHD dataset.

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)
SVM[38]	62.73	33.33	87.50
SRC[17]	72.73	66.67	75.00
FDDL[39]	68.82	64.87	80.41
DFDL[23]	77.64	66.72	78.67
GDWDL	90.67	87.64	92.76

Figure 8 shows the comparison of the ROC curve of these algorithms. We can see that the GDWDL algorithm proposed does improve the area of the ROC curve to a certain extent by observing the trend and the concavity and convexity of each curve, which is better than the other two algorithms to verify the effectiveness of the algorithm.

D. TIME COMPLEXITY ANALYSIS

In this section, we compare the time complexity for the proposed GDWDL and competing dictionary learning methods: FDDL and DFDL. The time complexity analysis quantitatively describes the running time of the algorithm, indicating the computational workload. As we all know, in most of the dictionary learning methods, complexity mainly depends on the sparse coding step. From Table 2, it is clear that the proposed GDWDL is the least expensive computationally on the ADHD dataset. The parameters are as follows: $c = 2$ (classes), $k = 200$ (atoms per class), $m = 158$ (columns of sample), $n = 103$ (rows of sample), $h = q = 30$ (samples per class), and L is the sparsity level. The running time for these methods is shown in the final column of Table 2. We can find that GDWDL is faster than other algorithms.

E. VERIFICATION OF OTHER DATA SETS

We have training experiments on the depressive data set and get the above series of results, and then we will further validate the superiority of our proposed method on the ADHD dataset. We use leave-one-out cross-validation (LOOCV) to verify the effectiveness of the algorithm and compare the accuracy, sensitivity and specificity of the algorithm in the ADHD data set in order to evaluate the performance of the classification. Table 3 shows the relevant comparative data.

V. CONCLUSIONS AND FUTURE WORK

This paper proposes a novel method of deep learning, GDWDL, which applied to the classification of mobile multimedia for medical diseases. With the help of mobile multimedia technology, we timely follow-up observation to patients and exchange the collected information into data information. It can better and more effectively classify the patients, taking into account the large amount of data on the basis of large data accumulation Complex and difficult. Using the weight method to measure intrinsic relation between the sample and the dictionary atom, we deal with the over-fitting phenomenon for the limited training set through l_2-norm regularization constraint. At the same time, we introduce the model of the deep network learning, and make local information between the layers train to ensure that each layer is convergence, layer by layer to promote their own characteristics, in order to achieve the best classification effect. The combination of healthcare and machine learning can play a great role in machine learning, and better improve efficiency in the field of mobile multimedia for healthcare. In this paper, we can easily see that the algorithm proposed in this paper is superior to other algorithms in the performance of patient classification by processing the collected medical data and testing the data.

Although the proposed algorithm is superior to other algorithms such as FDDL and DFDL, this is only the result of experimental verification on depression and ADHD data sets. There are some shortcomings to be further study in the future. The focus of the latter research will be on the other data set to verify its classification performance. At the same time, our training sample data is too small, it is necessary to increase more data sets to test better. We choose the dictionary size is less than 300, the number of iterations in 200, this algorithm is only to verify the superiority in this range, but the latter also need to increase the dictionary size and iteration times to experiment. With the new European General Data Protection Regulations (GDPR), make black-box approaches difficult to use [40], we will strengthen research in this area later. In addition, we need to verify the stability of the algorithm and compare against more classification methods in the future.

ACKNOWLEDGMENT

The authors would like to appreciate all anonymous reviewers for their insightful comments and constructive suggestions to polish this paper in high quality.

AUTHOR CONTRIBUTIONS

C. Luo, C.X. Wu and N.X. Xiong conceived the idea, designed the experiments and analyzed the data; W. Zhang and C.X. Wu performed the experiments and conducted the analyses; N.X. Xiong collected and processed the data; T. Kim interpreted the results and drew the conclusions; C. Luo wrote the paper. All authors agree with the above contribution details.

CONFLICTS OF INTEREST

All authors declare no conflict of interest.

REFERENCES

[1] M. Healy and P. Walsh, "Detecting demeanor for healthcare with machine learning," in *Proc. IEEE Int. Conf. Bioinformat. Biomed.*, Nov. 2017, pp. 2015–2019.

[2] Z. H. Zhou, *Machine Learning*. Beijing, China: Tsinghua Univ. Press, 2015.

[3] D. Singh *et al.*, "Human activity recognition using recurrent neural networks," in *Machine Learning and Knowledge Extraction* (Lecture Notes in Computer Science), vol. 10410. Cham, Switzerland: Springer, 2017, pp. 267–274.

[4] R. W. Liu, L. Shi, S. C. H. Yu, N. Xiong, and D. Wang, "Reconstruction of undersampled big dynamic MRI data using non-convex low-rank and sparsity constraints," *Sensor*, vol. 17, no. 3, p. 509, 2017, doi: 10.3390/s17030509.

[5] A. Ali, J. Qadir, R. ur Rasool, A. Sathiaseelan, A. Zwitter, and J. Crowcroft, "Big data for development: Applications and techniques," *Big Data Anal.*, vol. 1, p. 2, Jul. 2016, doi: 10.1186/s41044-016-0002-4.

[6] D. Graham-Rowe *et al.*, "Big data: Science in the petabyte era," *Nature*, vol. 455, no. 7209, pp. 8–9, 2008.

[7] J. Manyika *et al.* (May 2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. [Online]. Available: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

[8] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Proc. Int. Conf. Collaboration Technol. Syst.*, May 2013, pp. 42–47.

[9] IBM. (Oct. 2, 2012). *What is Big Data*. [Online]. Available: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[10] S. Siuly and Y. Zhang, "Medical big data: Neurological diseases diagnosis through medical data analysis," *Data Sci. Eng.*, vol. 1, no. 2, pp. 54–64, 2016, doi: 10.1007/s41019-016-0011-3.

[11] N. X. Xiong *et al.*, "Comparative analysis of quality of service and memory usage for adaptive failure detectors in healthcare systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 4, pp. 495–509, May 2009.

[12] H.-I. Suk, S.-W. Lee, and D. Shen, "Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis," *Brain Struct., Function*, vol. 221, no. 5, pp. 2569–2587, 2016.

[13] S. Zhang, X. Li, J. Lv, X. Jiang, L. Guo, and T. Liu, "Characterizing and differentiating task-based and resting state fMRI signals via two-stage sparse representations," *Brain Imag. Behavior*, vol. 10, no. 1, pp. 21–32, 2016.

[14] A. Divya *et al.*, "Secured smart healthcare monitoring system based on IOT," *Asian J. Appl. Sci. Technol.*, vol. 1, no. 2, pp. 53–56, Mar. 2017.

[15] H. Wang, Y. Chunfeng, H. Weiming, and S. Changyin, "Supervised class-specific dictionary learning for sparse modeling in action recognition," *Pattern Recognit.*, vol. 45, no. 11, pp. 3902–3911, 2012.

[16] M. Şimşek and E. Polat, "The effect of dictionary learning algorithms on super-resolution hyperspectral reconstruction," in *Proc. 25th Int. Conf. Inf., Commun. Autom. Technologies.*, Oct. 2015, pp. 1–5.

[17] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[18] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 1601–1604.

[19] W. D. Yu, P. Ray, and T. Motoc, "A RFID technology based wireless mobile multimedia system in healthcare," in *Proc. IEEE Int. Conf. E-Health Netw., Appl. Services (Healthcom)*, Aug. 2006, pp. 1–8.

[20] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3501–3508.

[21] Q. Zhang and B. X. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2691–2698.

[22] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2012, pp. 543–550.

[23] T. H. Vu, H. S. Mousavi, V. Monga, G. Rao, and U. K. A. Rao, "Histopathological image classification using discriminative feature-oriented dictionary learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 3, pp. 738–751, Mar. 2016.

[24] X. Wang, Y. Ren, Y. Yang, W. Zhang, and N. X. Xiong, "A weighted discriminative dictionary learning method for depression disorder classification using fMRI data," in *Proc. IEEE Int. Conf. Big Data Cloud Comput.*, Oct. 2016, pp. 618–623.

[25] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. W. Schuller, "A deep matrix factorization method for learning attribute representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 417–429, Mar. 2017.

[26] N. Agarwalla, D. Panda, and M. K. Modi, "Deep learning using restricted boltzmann machines," *Int. J. Comput. Sci., Inf. Secur.*, vol. 7, no. 3, pp. 1552–1556, 2016.

[27] M. R. Alam, M. Bennamoun, R. Togneri, and F. Sohel, "A joint deep boltzmann machine (jDBM) model for person identification using mobile phone data," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 317–326, Feb. 2017.

[28] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *Proc. Int. Conf. BDLP*, 2008, pp. 536–543.

[29] M. J. Gangeh *et al.* (Feb. 2015). *Supervised Dictionary Learning and Sparse Representation—A Review*. [Online]. Available: https://arxiv.org/abs/1502.05928

[30] Z. Cui, S. S. Ge, Z. Cao, J. Yang, and H. Ren, "Analysis of different sparsity methods in constrained RBM for sparse representation in cognitive robotic perception," *J. Intell. Robot Syst.*, vol. 80, pp. 121–132, Dec. 2015.

[31] B. U. Pedroni *et al.*, "Neuromorphic adaptations of restricted Boltzmann machines and deep belief networks," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2013, pp. 1–6.

[32] N. Akhtar, F. Shafait, and A. Mian, "Discriminative Bayesian dictionary learning for classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2374–2388, Dec. 2016.

[33] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.

[34] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[35] H. Lee *et al.*, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1096–1104.

[36] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010.

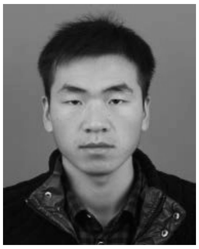
[37] M. W. Miller *et al.*, "False-negative sentinel lymph node biopsy in head and neck melanoma," *Otolaryngol.-Head Neck Surgery, Official J. Amer. Acad. Otolaryngol.-Head Neck Surgery*, vol. 145, no. 4, pp. 606–611, 2011.

[38] K. Ota *et al.*, "Effects of imaging modalities, brain atlases and feature selection on prediction of Alzheimer's disease," *J. Neurosci. Methods*, vol. 256, pp. 168–183, Dec. 2015.

[39] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based Fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 209–232, Sep. 2014.

[40] A. Holzinger *et al.* (Aug. 2017). "A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop." [Online]. Available: https://arxiv.org/abs/1708.01104

CHUNXUE WU received the Ph.D. degree in control theory and control engineering from the China University of Mining and Technology, Beijing, China, in 2006. He is currently a Professor with the Computer Science and Engineering and Software Engineering Division, School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China. His research interests include wireless sensor networks, distributed and embedded systems, wireless and mobile systems, and networked control systems.



CHONG LUO is currently pursuing the master's degree in computer technology with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China. His research interests include networks communication, big data, and machine learning.

NAIXUE XIONG received the Ph.D. degree from Wuhan University in sensor system engineering and the Japan Advanced Institute of Science and Technology in dependable sensor networks, respectively. He is currently an Associate Professor with the Department of Mathematics and Computer Science, Northeastern State University, Tahlequah, OK, USA. Before he attended Northeastern State University, he was with Georgia State University, Wentworth Technology Institution, and Colorado Technical University about 10 years. His research interests include cloud computing, security and dependability, parallel and distributed computing, networks, and optimization theory.

He published over 280 international journal papers and over 120 international conference papers. Some of his works were published in the IEEE JSAC, the IEEE or ACM transactions, ACM Sigcomm workshop, the IEEE INFOCOM, ICDCS, and IPDPS. He was a recipient of the Best Paper Award at the 10th IEEE International Conference on High Performance Computing and Communications and the Best student Paper Award at the 28th North American Fuzzy Information Processing Society Annual Conference. He has been a General Chair, Program Chair, Publicity Chair, PC member, and OC member of over 100 international conferences, and as a Reviewer of about 100 international journals, including the IEEE JSAC, the IEEE SMC (Park: A/B/C), the IEEE Transactions on Communications, the IEEE Transactions on Mobile Computing, the IEEE Transactions on Parallel and Distributed Systems. He is serving as an Editor-in-Chief, Associate Editor, or Editor member for over 10 international journals (including an Associate Editor for the IEEE Transactions on Systems, Man andCybernetics: Systems, an Associate Editor for *Information Science*, an Editor-in-Chief of the *Journal of Internet Technology*, and an Editor-in-Chief of the *Journal of Parallel and Cloud Computing*), and a Guest Editor for over 10 international journals, including *Sensor* Journal, WINET, and MONET.

Dr. Xiong is the Senior Member of the IEEE Computer Society. He is the Chair of Trusted Cloud Computing Task Force, the IEEE Computational Intelligence Society, and the Industry System Applications Technical Committee.

WEI ZHANG received the B.E. degree from the School of Information Science and Engineering, Wuhan University of Science and Technology, China, in 2000, and the M.Ec. and Ph.D. degrees from the Computer School, Wuhan University, China, in 2004 and 2008, respectively. He is currently an Associate Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, China. His research interests include wireless sensor network and intelligent computing. He is a member of Association for Computing Machinery and China Computer Federation.

TAI-HOON KIM received the M.S. and Ph.D. degrees in electrics and electronics and computer engineering from Sungkyunkwan University, South Korea. He is working for Department of Convergence Security, Sungshin Women's University. He wrote 16 books about the software development, OS such as Linux and Windows 2000, and computer hacking and security. And he published about 100 papers by 2006. He was a Chair and program committee of international conferences and workshops. He was a Guest Editor of AJIT and FGCS Journal, and now he is an Editor-in-Chief of JSE and IJSIA Journal. He researched security engineering, the evaluation of information security products or systems with Common Criteria and the process improvement for security enhancement. In these days, he researches also some approaches and methods making IT systems more secure.