Complex & Intelligent Systems (2023) 9:6329–6342
https://doi.org/10.1007/s40747-023-01065-9

ORIGINAL ARTICLE

# A new weighted extreme learning machine based on elastic net regularization embedded exponential regularized discriminative dictionary learning for image classification

Di Wu[1,2] · PinYi Zhao[1] · Qin Wan[1,2]

Received: 29 April 2022 / Accepted: 20 March 2023 / Published online: 5 May 2023


**Abstract**
It is well known that discriminative sparse representation can significantly improve the performance of image classification. However, there remain several tricky issues to be addressed due to the unsatisfied performance and high time consumption. In this paper, a novel classification framework called weighted extreme learning machine exponential regularized discriminative dictionary learning (WELM-ERDDL) is proposed to address these issues. The main contributions of this paper include (1) the WELM is embedded with ERDDL via exponential regularized linear discriminative analysis (ERLDA) for feature mappings while enabling nonlinear and diverse feature representation; (2) in the ELM learning process, the elastic net regularization is utilized to optimize more robust and meaningful output weights; (3) an effective weight update rule is designed for WELM. To verify the effectiveness of the proposed method, several experiments are conducted on real-world image classification databases. The results show that the proposed WELM-ERDDL framework is even more efficient than other state-of-the-art algorithms in general.

**Keywords** Weighted extreme leaning machine (WELM) · Discriminative dictionary learning · Elastic net regularization · Exponential regularized · Image classification

## Introduction

For the attractive growth of image data in the real computer vision field, feature extraction is becoming the central research topic in image classification. Come with the various changes of light condition, background and viewpoint that make the task become more challenge [1]. In the meanwhile, to single out more efficient and robust representative features to deal with such variation is of great important to image classification, in which feature extraction plays a crucial role.

✉ Di Wu
wudi6152007@hnie.edu.cn

PinYi Zhao
258960781@qq.com

Qin Wan
wanqin@hnie.edu.cn

[1] College of Electrical and Information Engineering, Hunan Institute of Engineering, Xiangtan 411104, China

[2] Hunan Province Cooperative Innovation Center for Wind Power Equipment and Energy Conversion, Hunan Institute of Engineering, Xiangtan 411104, China

By extraction means, feature extraction methods can be divided into the ones for handcrafted features and automatically learned features. Among the former feature extraction methods, the successful methods are scale invariant feature transform (SIFT) [2] and histogram of oriented gradient (HOG) [3]. They have always been used with geometric and statistical methods for some special tasks, and their limitations have been exposed in different real-world applications. On the other hand, although the latter feature extraction methods such as sparse representation and deep learning methods have been widely used in past decade and have achieved dramatic progress in the real-world application of computer vision [4], they are still facing some troubles. For spare representation methods, the performance of dictionary learning has an influence on the sparse coding vectors. For the deep learning approaches [5–7], they have always suffered from the complicated parameter tuning process and local minima.

Although sparse representation is susceptible to dictionary learning performance [8], it is very efficient in feature learning given no prior information and it has proven particularly robust in solving image processing problems owing to the following two incomparable merits [9–11]: (1) the receptive

fields of cell can be modeled using sparse coding in the visual cortex; (2) the rich representation can recover the subspace of image patches, leading to sparse representations naturally.

Among all the essential components of high-performing deep learning approaches, feed forward neural networks (FNNs) [12], especially the convolutional neural networks (CNNs) [13] and neural response (NR) [14] have achieved an excellent performance in various real-world tasks, such as face recognition, object tracking, and speech recognition [15], based on the multilayer perceptron. During the parameter optimization and tuning processes, the existing FNNs deeply rely on the backward propagation algorithm and always have lower convergence [16]. Although the tuning time is reduced to some extent, the CNNs are still in need of the tedious weights and bias optimization process, which virtually prolongs the run time and local minimum phenomenon. The performance of the NR methods is influenced by the design and selection of the template.

To overcome these unavoidable drawbacks of sparse representation and deep learning methods, Huang et al. [17] proposed the extreme learning machine (ELM), in which the hidden layer neurons are generated randomly without tuning process and the output weights can be determined analytically. The ELM is wide used in various computer vision applications such as object recognition and image classification. Typically, the original ELM is in a singer-layer structure and can be extended to multilayer framework called multilayer ELM [18] (ML-ELM) to improve its generalization performance. The ML-ELM can be constructed using the ELM-based autoencoders through stacking ideas, but it will lose the full universal approximation merit of ELM. Huang et al. [19] proposed the local receptive fields-based ELM aiming to solve the universal approximation problem, but it has only one feature mapping layer and one pooling layer which fail to extract the sufficient representative features. So further research is counted upon to focus on the topic of improving the classification performance and learning efficiency.

In the learning process of ELM, the output weights are crucial and normally computed using the Moore–Penrose generalized inverse method by minimizing the classification error of the training data. Recent studies have shown this method tends to fail due to data distributions. Several studies have tried to address this unavoidable problem in the real-world application using different technologies. Huang et al. [20] used the $l_2$ norm (Ridge regression) regularization to address the minimum problem. Tang et al. [21] used the $l_1$ norm (Lasso) regularization to derive the sparse solutions in order to restrict the output weights. While in the real application, the hidden nodes usually outnumber the label data so the traditional Lasso method could fail to realize the group selection.

In summarizing the existing learned feature extraction methods, it is found that they all have some drawbacks and suffer from either low classification performance or high time consumption. In this paper, inspired by the sparse representation and ELM, our panel proposes a novel nonlinear feature extraction approach called weighted extreme learning machine exponential regularized discriminative dictionary learning (WELM-ERDDL). The proposed WELM-ERDDL consists of two stages: the WELM-ERDDL feature mapping stage and the WELM learning stage. In the feature mapping stage, the WELM is embedded with exponential regularized discriminative dictionary learning via exponential regularized linear discriminative learning (ERLDA) and sparse coding, so the input features can be transformed via nonlinear feature mapping. In the WELM learning stage, elastic net regularization is used to update the output weights comprising the $l_1$ norm and $l_2$ norm to obtain more compact and meaningful features. Finally, a flexible weight update criterion is designed for the WELM.

Overall, the main contributions of this paper are outlined as follows:

1. A nonlinear WELM embedded feature projection strategy via Exponential Regularized Discriminative Dictionary Learning is given for feature diversity and low computational efficiency.
2. During the ELM learning stage, the output weights are updated through elastic net regularization to enhance its compactness and meaningfulness.
3. An effective adaptive online weight update criterion is designed for the WELM.

The rest of this paper is organized as follows: section "Preliminaries" provides some prior knowledges related to this paper. In section "The proposed WELM-ERDDL", the proposed WELM-ERDDL feature extraction framework is given in detail. In section "The ELM Learning", the efficient ELM learning process is discussed. In section "Classification with the WELM-ERDDL", the classification scheme with the proposed WELM-ERDDL is presented. In section "Experimental results and analysis", the experimental results are shown and analyzed. Finally, in section "Conclusion", the conclusion is shown, and some potential directions of future work are indicated.

## Preliminaries

In this section, a brief introduction of some preliminaries is presented, including the concepts of extreme learning machine (ELM) and weight extreme learning machine (WELM), dictionary discriminative learning (DDL), and elastic net regularization.

## Extreme learning machine (ELM)

Suppose the training set $\{x_i, t_i\}_{i=1}^{N}$ is composed of $N$ training samples; the input is $x_i$ whose dimension is $d$; $t_i$ is the label of the output. Then the output of the ELM is [22]:

$$\sum_{j=1}^{L} \beta_j h(x_i) = \sum_{j=1}^{L} \beta_j g\left(x_i \cdot w_j + b_j\right) = t_i, \tag{1}$$

where the parameter $w_j = \left[ w_{j1}, w_{j2}, \ldots w_{jn} \right]$ is the input weight of the $j$th hidden node; $b_j$ is the deviation of the $j$th hidden node; $\beta_j$ is the output weight of the $j$th hidden node. Equation (1) can be simplified into

$$H\beta = T, \tag{2}$$

where $H$ is the hidden layer output matrix:

$$H = \begin{bmatrix} h(x_i) \\ \ldots \\ h(x_N) \end{bmatrix} = \begin{bmatrix} g(x_1 \cdot w_1 + b_1) & \ldots & g(x_1 \cdot w_L + b_L) \\ \vdots & \ddots & \vdots \\ g(x_N \cdot w_1 + b_1) & \ldots & g(x_N \cdot w_L + b_L) \end{bmatrix}. \tag{3}$$

To improve the generalization ability of ELM, a penalty factor $C$ is introduced into (3), and the output weight matrix $\beta$ is

$$\beta = H^T \left( \frac{1}{C} + HH^T \right)^{(-1)} T. \tag{4}$$

ELM aims to minimize the training error and the $l_2$ norm of the output weights, namely

$$\min : L_{\mathrm{ELM}} = \frac{1}{2}\|\beta\|^2 + C \cdot \frac{1}{2} \cdot \sum_{i=1}^{N} \|\xi_i\|^2. \tag{5}$$

Then the output of the extreme learning machine can be expressed as

$$f(x) = h(x)\beta = h(x)H^T \left( \frac{1}{C} + HH^T \right)^{(-1)} T. \tag{6}$$

## Weighted extreme learning machine (WELM)

The main viewpoint of WELM is to assign penalties to different classes, and it can be view as a cost sensitive version of ELM in handling the troublesome imbalanced data problem. In the WELM, the penalty factor $C$ is added to the ELM while the minority class has a greater value of $C$. Then a weighted matrix $W$ is used to regulate $C$, so (5) can be modified as [23]

$$\min : L_{\mathrm{ELM}} = \frac{1}{2}\|\beta\|^2 + \frac{CW}{2} \sum_{i=1}^{N} \|\xi_i\|^2. \tag{7}$$

In (7), the critical problem is to determine the appropriate weight matrix. Zong et al. gave two different versions of computing methods:

$$W_{\mathrm{ELM1}} = \frac{1}{\mathrm{num}(t_i)}, \tag{8}$$

$$W_{\mathrm{ELM2}} = \begin{cases} \frac{0.618}{\mathrm{num}(t_i)} & \text{if } \mathrm{num}(t_i) > \mathrm{AVG}(t_i) \\ \frac{1}{\mathrm{num}(t_i)} & \text{if } \mathrm{num}(t_i) \le \mathrm{AVG}(t_i) \end{cases}, \tag{9}$$

$\mathrm{num}(t_i)$ it is the number of samples belonging to the $i$th class. Finally, (6) can be modified as

$$f(x) = h(x)\beta = h(x)H^T \left( \frac{1}{C} + HWH^T \right)^{(-1)} WT. \tag{10}$$

## Dictionary discriminative learning (DDL)

For the sparse coding problem, a classical dictionary learning problem is shown as follows [24–26]:

$$\underset{D,A}{min} \frac{1}{2}\|X - DA\|^2 + F_{\mathrm{s}}(A)^2 + F_{\mathrm{d}}(A)^2, \tag{11}$$

where $F_{\mathrm{s}}$ stands for the sparsity inducing term; $F_{\mathrm{d}}$ stands for the discriminative term. The research has proven that adding discriminative information to the sparse coding can significantly enhance the classification performance. Liu et al. proposed a specific $l_{1,2}$ norm to learn the discriminative information $F_{\mathrm{d}}$:

$$F_{\mathrm{d}}(A) = \sum_{c=1}^{C} \|A_c\|_{1,2}, \tag{12}$$

where $A_c$ represents the coding vectors from class $c$, while the coding vector from the same class will have the same saprse pattern achieved simultaneously by the sparsity and discriminative encoding.

## Elastic net regularization

Elastic net regularization can solve the variable selection problem effectively by combining the $l_1$ norm and $l_2$ norm to obtain a better solution. Therefore, the elastic net regularization problem can be expressed as follows [27, 28]:

$$P(x;\, \omega) = \omega\|x\|_1 + (1-\omega)\frac{1}{2}\|x\|_2^2, \tag{13}$$

where the parameter $\omega$ controls the proportion between the $l_1$ norm and $l_2$ norm. Evidently, there are two special cases: if $\omega = 0$, the elastic net regularization becomes associated with the $l_2$ norm; if $\omega = 1$, the elastic net regularization becomes associated with the $l_1$ norm.

## The proposed WELM-ERDDL

In this section, the proposed WELM-ERDDL is to be deduced in detail. As known, the whole structure of the proposed WELM-ERDDL consists of two parts: the discriminative projection term and the discriminative sparse regularization term. In the discriminative projection, the exponential regularized linear discriminative analysis (ERLDA) is conducted for the discriminative projection term. Therefore, the proposed approach can not only acquire high-dimensional features for high performance without parameter turning process but also perform dictionary-learning process in a low-dimensional subspace.

Inspired by the dictionary learning problem given in (11), our panel formulates the following objective function of the proposed WELM-ERDDL approach:

$$\min_{D,A} \frac{1}{2}\|\beta HW - DA\|^2 + F_1(\beta)^2 + F_2(A)^2, \tag{14}$$

where $H$ is the nonlinear transform of the original input $X$; $F_1$ stands for the discriminative projection term and $F_2$ stands for the discriminative sparse regularization term, which are regularizations for $\beta$ and $A$, respectively. From (14), one may find that the proposed WELM-ERDDL has two merits:

1. Using the nonlinear transform, the original input is mapped into high-dimensional features $H$ without parameter tuning process so that the universal approximation capability can be guaranteed.
2. The dictionary learning is performed solely in a lower dimensional space.

### The discriminative projection term $F_1$

In this part, the discriminative projection term is to be explained comprehensively. From the objective function given in (14), it can be seen that the discriminative subspace $\beta WH$ is crutial for dictionary learning and its determined by the discriminative projection term $F_1$. Formally, the LDA approach is utilized for regularization, but it is always confronted with small sample size problem. In this paper, the ERLDA is conducted for the discriminative projection term. For the ERLDA approach, the discriminant criterion is given by:

$$J(W, \alpha)_{\text{ERLDA}} = \frac{\left|W^T \exp(S_b) W\right|}{\left|W^T \exp(S_w + \alpha I) W\right|}. \tag{15}$$

Then the orientation matrix is computed by EVD as $\left[\exp(S_w + \alpha I)\right]^{-1}\left[\exp(S_b)\right]$. While the discriminative projection term $F_1$ is computed as follows:

$$F_1(\beta) = \frac{\lambda_1}{2}\text{tr}\Big[\beta(\exp(S_w + \alpha I) - \exp(S_b))\beta^T\Big], \tag{16}$$

where $\lambda_1$ is a hyperparameter; $S_w$ and $S_b$ are the intraclass scatter matrix and the interclass scatter matrix based on the hidden space, respectively. The effect of (16) is to minimize the intraclass scatter and maximize the interclass scatter to separate the classes of features alongside dictionary learning.

### The discriminative sparse regularization term $F_2$

The traditional $l_{1,2}$ norms in (12) are inspired by multitask learning with similar tasks sharing similar sparse patterns, which means the row sparse structure in (12) must select the same dictionary atoms within the same class. The drawback of this scheme is that it is hard to optimize; meanwhile, it is sensitive to the optimization method. In this section, a simple and novel sparse regulation strategy is to be presented for dictionary learning. Then $F_2$ can be expressed as follows:

$$F_2 = \sum_{c=1}^{C}\Big(\lambda_2 + \lambda_3\|a_{-c}^{(i)}\|_2\Big)\|a_c^{(i)}\|_2, \tag{17}$$

where $a_c^{(i)}$ and $a_{-c}^{(i)}$ are the rows of $A_c$ and $A_{-c}$, respectively, which stand for the coding vectors belonging to and not belonging to the class $c$;$\lambda_2$ and $\lambda_3$ are two hyperparameters. In this part, define $[W_C]_{ii} = \lambda_2 + \lambda_3\|a_{-c}^{(i)}\|_2$, and thus, (17) can be transformed into

$$F_2 = \sum_{c=1}^{C} W_C A_{C1,2}.. \tag{18}$$

### Formulation

In this part, after the discriminative projection term $F_1$ and the discriminative sparse regularization term $F_2$ are determined, the final formula for the proposed WELM-ERDDL is given as follows:

$$\begin{aligned}\min_{D,A} \frac{1}{2}\beta HW - DA^2 + F_1(\beta)^2 + F_2(A)^2 &= \min_{D,A} \frac{1}{2}\beta HW - DA^2 \\ &+ \frac{\lambda_1}{2}\text{tr}\Big[\beta(\exp(S_w + \alpha I) - \exp(S_b))\beta^T\Big] + \sum_{c=1}^{C} W_C A_{C1,2}.\end{aligned} \tag{19}$$

***Remarks*** In (19), $F_1$ and $F_2$ are both designed as discriminative regularization terms, but they are serve for different purposes: the discriminative projection term $F_1$ is used to learn a suitable projection for feature representations, while the discriminative sparse regularization term $F_2$ is designed for discriminative dictionary learning by regularizing the sparse coding vectors.

## The ELM learning

In this section, the parameter learning process of the proposed WELM-ERDDL framework is to be explained in detail. Firstly, the output weights $\beta$ are deduced with a more effective strategy using the elastic net regulation method. Next, a more robust adaptive online learning weight update rule for WELM is given.

### Update of $\beta$

Normally, the output weights $\beta$ are always computed by minimizing the approximation error of the training data, but it suffers from the Moore–Penrose generalized inverse of $H$. In order to solve this troublesome problem, Huang et al., added an $l_2$ norm regularization term in (5), while Tang et al. used the $l_1$ norm to restrict $\beta$ to obtain a more meaningful and sparser value. However, the feature maps may outnumber the training data and the pairwise columns of $H$ may have strong correlations. Fortunately, the elastic net regulation combining the $l_1$ norm and the $l_2$ norm provides an appropriate solution to the valuable selection problem. In this part, $\beta$ is determined through elastic net regularization.

In the output weight learning problem, the elastic net regularization can be expressed as

$$P(\beta;\omega) = \omega\|\beta\|_1 + (1-\omega)\frac{1}{2}\|\beta\|_2^2, \tag{20}$$

Combined with (20), the final formula for the proposed WELM-ERDDL given in (19) can be transformed as follows:

$$\begin{aligned}\min_{D,A}\ & \frac{1}{2}\beta HW - DA^2 + F_1(\beta)^2 + F_2(A)^2 + \lambda_4 P(\beta;\omega)\\ &= \min_{D,A}\frac{1}{2}\beta HW - DA^2\\ &+ \frac{\lambda_1}{2}\mathrm{tr}\Big[\beta(\exp(S_w+\alpha I)-\exp(S_b))\beta^T\Big]\\ &+ \sum_{c=1}^{C} W_C A_{C1,2} + \lambda_4\left[\omega\beta_1 + (1-\omega)\frac{1}{2}\beta_2^2\right],\end{aligned} \tag{21}$$

where $\lambda_4$ is the regularization parameter for elastic net penalty. According to Lagrangian multiplier strategy, assuming that $D$ and $A$ are constant, we have

$$\begin{aligned}\min_{\gamma,\beta,u}\ & \frac{1}{2}\gamma HW - DA^2 + \frac{\lambda_1}{2}\mathrm{tr}\Big[\gamma(\exp(S_w+\alpha I)-\exp(S_b))\gamma^T\Big]\\ &+ \lambda_4\left[\omega\gamma_1 + (1-\omega)\frac{1}{2}\gamma_2^2\right] + \frac{\rho}{2}\gamma - \beta + u^2,\end{aligned} \tag{22}$$

Furthermore, the problem in (22) can be decomposed into three subproblems:

$$\begin{aligned}\gamma^{k+1} = \min_{\gamma}\ & \frac{1}{2}\gamma HW - DA^2 + \frac{\lambda_1}{2}\mathrm{tr}\Big[\gamma(\exp(S_w+\alpha I)-\exp(S_b))\gamma^T\Big]\\ &+ \lambda_4\left[\omega\gamma_1 + (1-\omega)\frac{1}{2}\gamma_2^2\right] + \frac{\rho}{2}\gamma - \beta^k + u_2^{k2},\end{aligned} \tag{23}$$

$$\beta^{k+1} = \mathrm{argmin}\|\gamma^{k+1} - \beta + u^k\|_2^2, \tag{24}$$

$$u^{k+1} = u^k + \gamma^{k+1} - \beta^{k+1}. \tag{25}$$

Consequently, among these three subproblems, the first subproblem in (23) is a sparse coding problem with the Lasso regularization which can be computed by shrinkage function.

The second subproblem in (24) is a quadratic optimization problem whose close form solution is given as followsl:

$$\begin{aligned}\beta^{k+1} = & (2H^T HW + 2\lambda_4(1-\mu)I\\ &+\lambda_1(\exp(S_w+\alpha I)-\exp(S_b))\\ &+\rho I)^{-1}(2H^T HT + \rho\gamma^{k+1} - u^k).\end{aligned} \tag{26}$$

Finally, the optimization problem of (21) can be summarized in Algorithm 1.

**Algorithm 1:** The pseudocode of the output weights $\beta$ update method with the elastic-net regularization

**Inputs:** hidden layer matrix $H$ , label matrix $T$ , $\omega$ and $\lambda_4$

**Initialization:** $\beta$, $\gamma$ , $u$

**For $i = 1$ to $N$ , do**

1. Update $\gamma$ according to (23);
2. Update $\beta$ according to (26);
3. Update $u$ according to (25);
4. **End for**

**Output:** the output weights $\beta$

### Adaptive online weight update for the WELM

This part focuses on the weight setting problem for the novel WELM proposed in this paper. Formally, in the previous work of Zong et al., the weights are computed in terms of the number of samples belonging to each class. The drawback of this method is obvious. According to this strategy, as the labeled samples increase, the weights used to punish the newly added samples tend to decrease sharply, leading to the final classification model being more focused on the previous model irrespective of the newly added samples. So, in this paper, the adaptive online weight learning rule for WELM is given. For the newly added samples, the weights can be taken as follows:

$$w_i = \begin{cases} \frac{N^+}{N^+ + N^-} & \text{if } x_i \text{ belongs to the majority class} \\ \frac{N^-}{N^+ + N^-} & \text{if } x_i \text{ belongs to the minority class} \end{cases}, \quad (27)$$

where $N^+$ and $N^-$ denote the number of samples belonging to the positive class (majority class) and to the negative class (minority class), respectively. During the adaptive online weight update procedure, the weight mainly depends on the ratio$N^+$: $N^-$ (or $N^-$:$N^+$).

## Classification with the WELM-ERDDL

In this section, after the learning of $\beta$ and $w$ on the training set, each test sample $x_{\text{test}}$ is set with WELM-ERDDL and the label can be predicted as

$$Y = H\beta^*. \quad (28)$$

Then, the class number $c_i$ of the unlabeled sample data can be determined by finding the maximum in the corresponding row:

$$c_i = \text{argmax } \mathrm{Y}_{ij}. \quad (29)$$

Finally, the whole process of the proposed WELM-ERDDL method is summarized in Algorithm 2.

**Algorithm 2:** The pseudocode of the proposed WELM-ERDDL

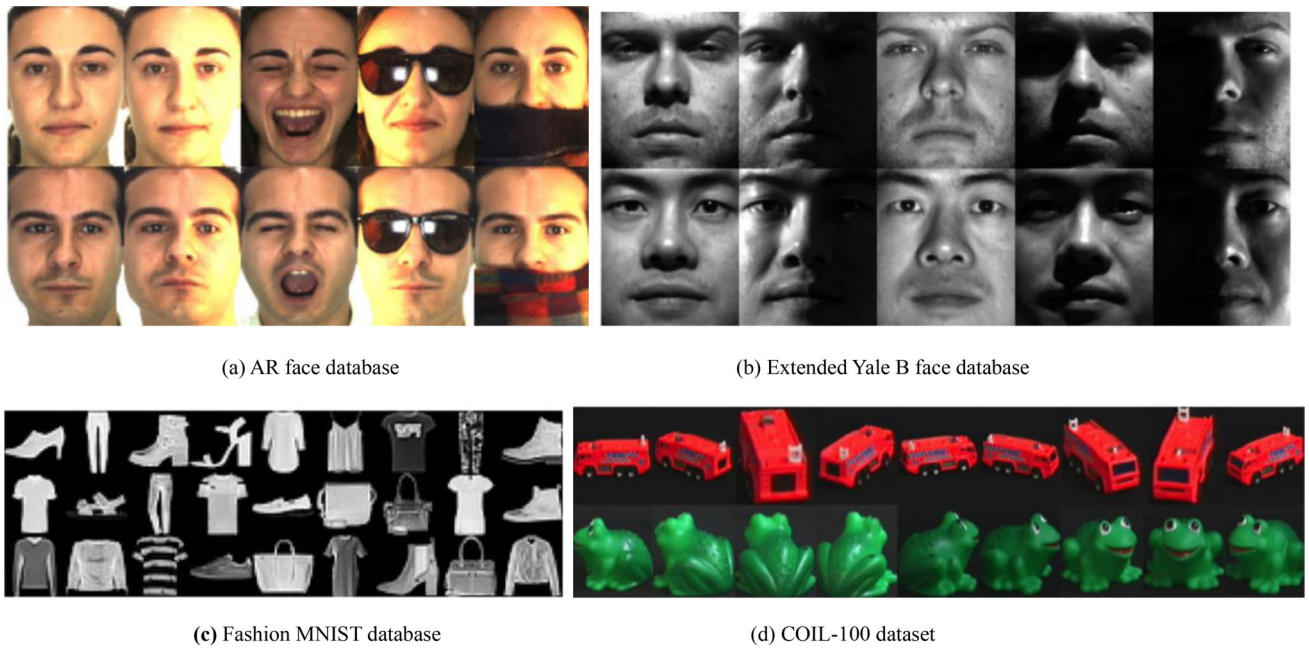
Inputs: input samples $X$, random matrix $W$, output dimension $n$, hyperparameters $\lambda_1$, $\lambda_2$,$\lambda_3$,$\lambda_4$

1. For $i = 1$ to $N$, do
$$h_i = f(Wx_i + b)$$
2. Compute the scatter matrices $S_w$ and $S_b$ ;
3. Update the weight of WELM according to (27);
4. Compute $F_1$ as follows:
$$F_1(\beta) = \frac{\lambda_1}{2} tr[\beta(exp(S_w + \alpha I) - exp(S_b))\beta^T]$$
5. Compute $F_2$ as follows:
$$F_2 = \sum_{c=1}^{C} (\lambda_2 + \lambda_3 \| a_{-c}^{(i)} \|_2) \| a_c^{(i)} \|_2$$
6. Solve the following problem given in (21) using Algorithm 1:
$$\min_{D,A} \frac{1}{2}\|\beta HW - DA\|^2 + F_1(\beta)^2 + F_2(A)^2 + \lambda_4 P(\beta;\omega)$$
$$= \min_{D,A} \frac{1}{2}\|\beta HW - DA\|^2 + \frac{\lambda_1}{2} tr[\beta(exp(S_w + \alpha I) - exp(S_b))\beta^T] + \sum_{c=1}^{C} \|W_C A_C\|_{1,2} + \lambda_4 [\omega\|\beta\|_1 + (1-\omega)\frac{1}{2}\|\beta\|_2^2]$$
7. Obtain the unlabeled data label according (29).

## Experimental results and analysis

In this section, several experiments are provided in diverse ways to demonstrate the effectiveness of the proposed WELM-ERDDL approach.

Each experiment is tested on a PC with Intel Core I7-8700 at 3.40 GHz and 16 GB RAM. The proposed method is implemented using Matlab2013a, with the code for the other models inherited directly from the code published by the respective authors. To verify the effectiveness and robustness of the WELM-ERDDL algorithm, the experiment is divided into four parts:



(a) AR face database

(b) Extended Yale B face database

**(c)** Fashion MNIST database

(d) COIL-100 dataset

**Fig. 1** The four benchmark databases

1. In Section A, the database used in this experiment and corresponding parameter setting are declared.
2. In Section B, the performance of the proposed WELM-ERDDL approach is evaluated using the following four different classical image classification databases: the AR face and the Extended Yale B for face recognition, the Fashion-MNIST dataset and COIL-100 dataset for object classification. Specifically, the proposed new method is compared with SRC, K-SVD, D-KSVD, FDDL, SVGDL, SDDL, ELM, WELM, and HI-DKIELM.
3. In Section C, the learned representation $\beta HW$ is compared with the original data and the classical ELM output $\beta H$ to test its effectiveness.
4. In Section D, the importance of the choice of $\lambda_2$ and $\lambda_3$ of the discriminative sparse regularization term $\boldsymbol{F}_2$ of the given WELM-ERDDL method is validated. In the meantime, the effect of the sparse representation classifier (SRC) is also tested for image classification.
5. In Section E, the performance of the WELM-ERDDL in practical application learning tasks is verified and compared with other baseline methods.

## Database and experiment parameter setting

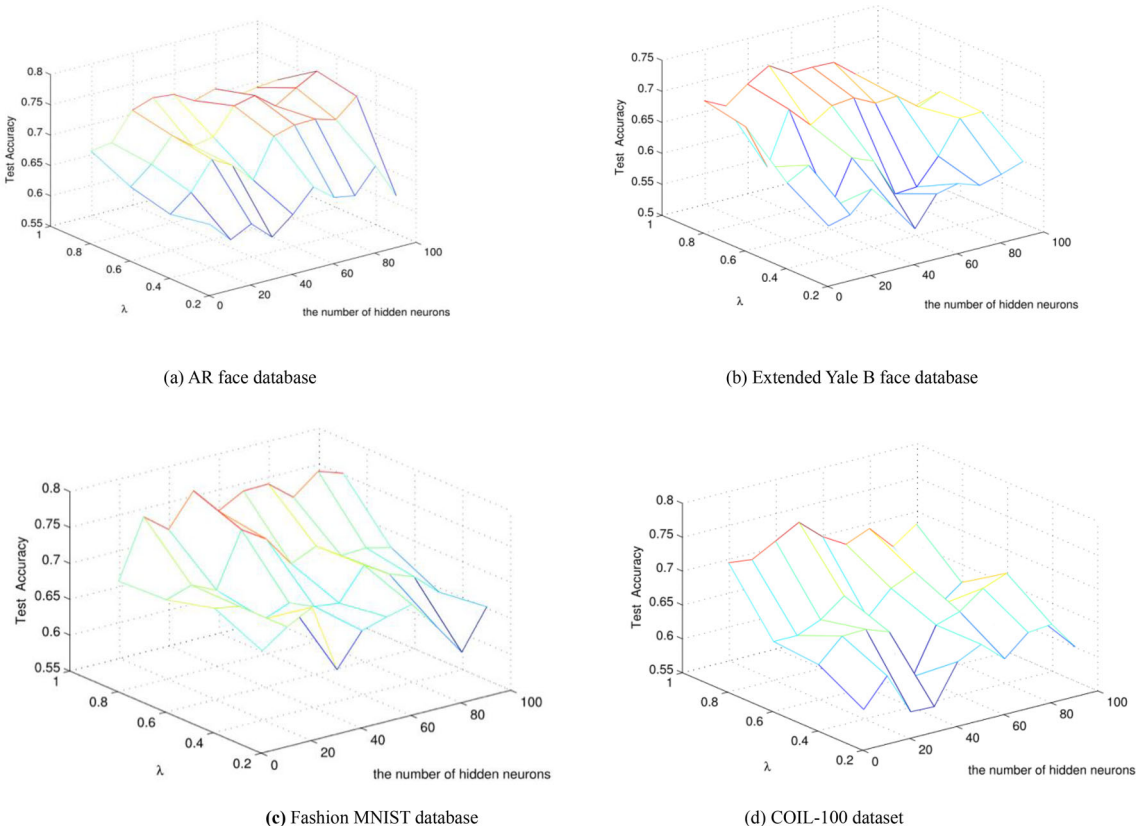
In these experiments, four different classical image classification databases are used: the AR and the Extended Yale B for face recognition, the Fashion- MNIST and COIL-100 for object classification.

The AR face database consists of 4000 color images from 126 people which have high brightness and wide pose variations. The sub-datasets are shown in Fig. 1a. Following the experiment setting, a commonly used subset is used which includes 2600 images from 50 males and 50 females, each of whom has 26 facial images with size $165 \times 120$. 20 images are selected at random for training and the 6 images are remained for testing.

The Extended Yale B face database has 2414 frontal face images collected from 38 people, each of whom has about 64 images. This face database is challengeable for the reason that all the images are captured with different facial expressions, occlusions, and lighting variations. The specific illustrations of the database are shown in Fig. 1b. Following the common experiment parameter setting, each image is normalized into $192 \times 168$ pixels. Half of the images are selected at random for training, with the rest for testing.

The performance of the proposed method on object recognition is also evaluated with the Fashion MNIST database in substitution for the classical MNIST database. This database contains 70,000 images, 60,000 of which serve for training and others for testing. Especially, each image is a $28 \times 28$ gray scale image belonging to each class. The illustrations of the data are shown in Fig. 1c.

Another dataset for object recognition is the COIL-100 dataset which contains almost 7200 images associated with 100 objects, and each image is captured from different views against a clear background. The examples of the dataset can be found in Fig. 1d. Generally, 10 images are selected at




(a) AR face database

(b) Extended Yale B face database

**(c)** Fashion MNIST database

(d) COIL-100 dataset

**Fig. 2** Classification accuracy at varying $\lambda_1$ and number of neurons on the four benchmark databases

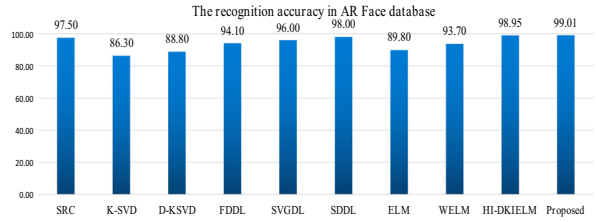
**Table 1** Parameter settings of each database

| Dataset | AR Face | Extended Yale B | Fashion MNIST | COIL-100 |
|---|---|---|---|---|
| Samples | 2600 | 2414 | 70,000 | 7200 |
| Image size | $165 \times 120$ | $192 \times 168$ | $28 \times 28$ | $32 \times 32$ |
| $n$ | 540 | 504 | 512 | 512 |
| $m$ | 540 | 570 | 300 | 1000 |
| $\lambda_1$ | 0.9 | 0.9 | 0.9 | 0.9 |
| $\lambda_2$ | 0.001 | 0.01 | 0.05 | 0.002 |
| $\lambda_3$ | 0.06 | 0.01 | 0.5 | 0.03 |
| $\lambda_4$ | 0.01 | 0.001 | 0.01 | 0.4 |

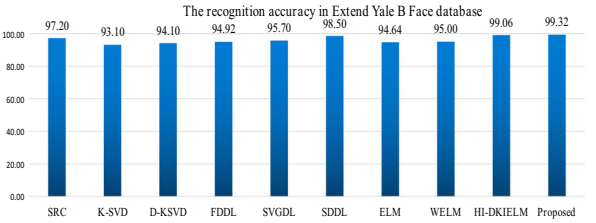
random from each object for training and the rest for testing, and each image is resized to $32 \times 32$ pixels.

For the proposed method, for each database, the optimal parameter is selected using the cross validation method. The hidden dimension of ELM is set as $L = 2000$, while the output dimension is set as $n$. The number of dictionary atoms $m$ and the regularization parameters $\lambda_1 - \lambda_4$ of each dataset are shown in Table 1.

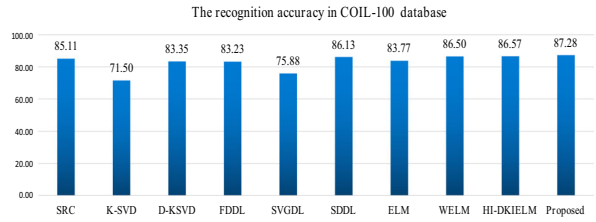
For the ELM, the parameter involves the number of hidden neurons. We select the optimal $\lambda_1$ and the number of hidden neurons through contrast experiment, parameters is set as follows: $\lambda_1 = [0.3, 0.5, 0.7, 0.9]$, the number of hidden neurons between 1 and 100. Figure 2 shows the classify accuracy of different $\lambda_1$ and the number of hidden neurons on four graph data sets. In the following experiments, we choose $\lambda_1 = 0.9$ and 40 neurons when apply our method on PTC, because it has higher accuracy on all the four data sets.



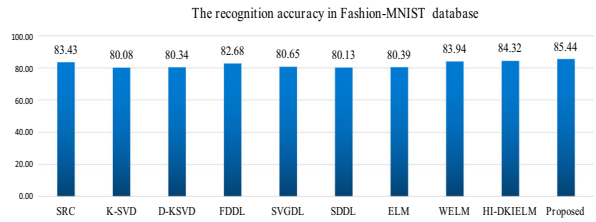

**(a) AR Face database**




**(b) Extended Yale B database**




**(c) COIL-100 database**




**(d) Fashion-MNIST database**

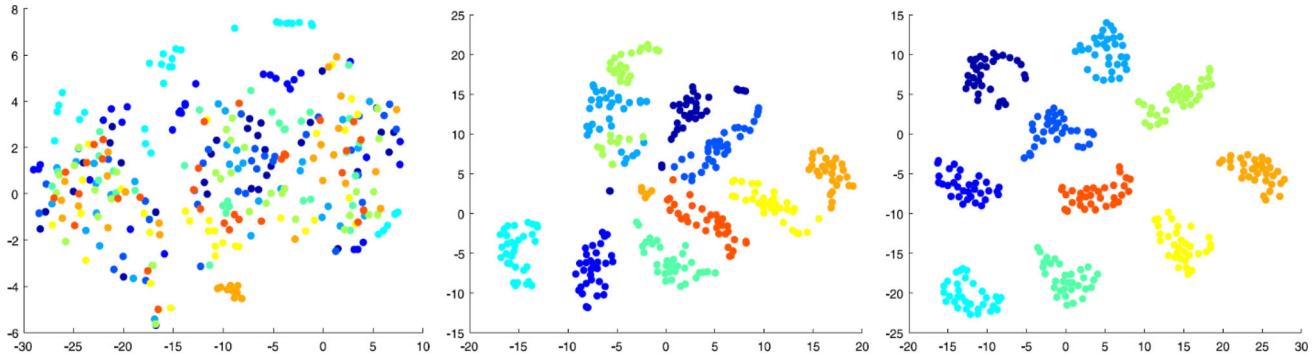
**Fig. 3** Recognition accuracy of the four benchmark databases

## Evaluation of the performance of the proposed WELM-ERDDL

In this part, the focus is laid on evaluating the performance of the proposed WELM-ERDDL method on the four benchmark databases. Specifically, the proposed new method is compared with SRC, K-SVD, D-KSVD, FDDL, SVGDL, SDDL, ELM, WELM, and HI-DKIELM. Among these compared methods, SRC is nonparametric with the training samples used as the dictionary directly; D-KSVD is based on the K-SVD method with different discriminative regularizations; FDDL employs Fisher discrimination while SVGDL further extends this method using support vector formulation, and thus SDDL applies the $l_{12}$ norms in the regularizations to constrain the supports of the coding vectors; WELM is the weighted version of ELM with the weights computed in terms of the number of samples belonging to each class; HI-DKIELM is covered in our former works given in [29]. The experimental results of the proposed WELM-ERDDL as well as those of the compared methods are summarized in Fig. 3. For each method, all experimental results are measured with the optimal hyperparameter settings and averaged almost 10 runs. In the meanwhile, the detailed analysis is made with respect to each database.

The detailed analysis on the experimental results is performed in the following:

1. For the Extended Yale B database, it can be deemed that the proposed WELM-ERDDL method achieves the best recognition result among all the compared methods. A scrutiny into the data reveals that the margin between the proposed method and SDDL is not large, almost 0.82%. The proposed approach avoids $l_0$ minimization and is more stable for multiple runs with std of 0.24% over 10 runs. However, for the SDDL method, for the reason that it suppresses the overlapping support of different classes via $l_0$ minimization which is approximate to the $l_2$ minimization, it is unstable and sensitive to the regulation factor. On the other hand, it can be seen that SDDL, SVGDL, and HI-DKIELM achieve high accuracy compared with other methods. Moreover, our method has over 2.11% accuracy gain compared with SRC, which means the proposed method can achieve effective discriminative sparse representation.
2. For the AR face database, the proposed WELM-ERDDL method outperforms all other compared methods, while the SDDL method achieves large margins than do other dictionary learning methods, showing the better performance in support discrimination.
3. For the COIL-100 database, our method also outperforms the other compared methods. From the results in Fig. 2, the K-SVD method has the worst recognition performance with recognition rate being only 71.50%; meanwhile, the D-KSVD method exhibits a significantly superior performance that demonstrates the importance of discrimination for dictionary learning. From the detailed results in Fig. 3, it is also found that the SRC method has a superior performance to SVD-based methods probably because the COIL-100 database is set up against a clear background and because of the regular structure in which the images can be constructed nicely.
4. For the Fashion-MNIST database, specifically, the performance is tested under a restricted experimental environment with 600 images selected at random for training and the whole test set for testing evaluation. From the



(a) The original data points. (b) The ELM output. (c) The sparse coding vectors.

**Fig. 4** Visualizations of learned representations using t-SNE. Data points of the same color belong to the same class. **a** Corresponds to the original data points. **b** Corresponds to the ELM outputs. **c** Corresponds to the sparse coding vectors

experimental results in Fig. 3, our proposed method demonstrates the best accuracy. Specifically, the SDDL method fails to yield a superior result to other compared methods on previous three databases for the reason that the support size is not large enough to describe the test partition.

## Testing of the effectiveness of the learned representation

In this part, the learned representations $\beta HW$ are visualized to evaluate the effectiveness of the proposed WELM-ERDDL method using the Extended Yale B database and compare its results with the output results $\beta H$ of traditional ELM. While the output weights $\beta$ are learned via (26) and original data points, the weights $W$ are learned via (27). Specifically, the learned representations $\beta HW$ given in this paper are further used to compute the sparse representations $A$, while the original samples, the traditional output results $\beta H$ of ELM, and the sparse coding vectors obtained in this paper are embedded into two dimensions using t-SNE, with the experimental results shown in Fig. 4.

From the results in Fig. 4a, it can be clearly found that the original data points are cluttered with mixed structure. However, when projected with learned output weights $\beta$ via (26), the representations become more separable. This phenomenon shows that the learned projection is more able to generate discriminate outputs. But Fig. 4b reminds that there are still some data points that are far apart and that some classes are even mixed with points of other classes. While the outputs of sparse coding remedy these drawbacks and exhibit clear clusters with better separate representations as shown in Fig. 4c, which means the class weights learned via (27) can effectively improve the discriminative learning and empirically justify the designed WELM-ERDDL scheme in this paper is meritorious for image classification.

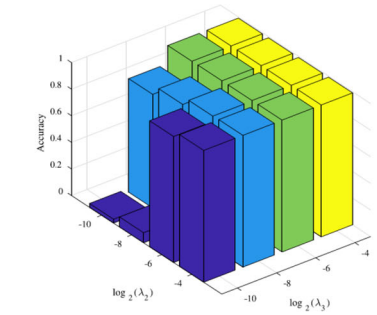
## Testing of the effectiveness of $\lambda_2$, $\lambda_3$, and SRC

This part begins with a test on the importance of the proper choice of $\lambda_2$ and $\lambda_3$ which control the sparsity of the coding vectors for the proposed WELM-ERDDL method. The similar simple illustrations in the log scale for the four databases are shown in Fig. 5.

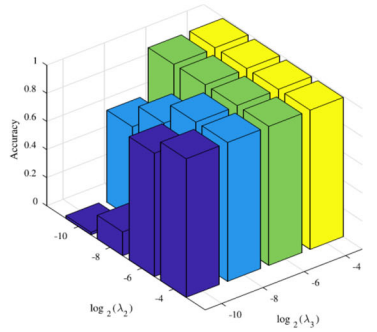
From the results in Fig. 5, it can be clearly seen that the algorithm is more sensitive to $\lambda_2$ which controls the intraclass sparsity. On the other hand, when a suitable value of $\lambda_3$ is selected, the proposed method should be able to exhibit a reasonable performance. It can also be noted that for Extended Yale B, AR Face, and COIL-100, the parameters are more stable to parameter choices, while for Fashion-MNIST the parameters should be selected with greater care.

Finally, the effect of the sparse representation classifier (SRC) for image classification is evaluated. In the experiments, the SRC method is chosen since it uses the nonparametric coding vectors from the training set learned with the class specific weighted $l_{1,2}$ norm while computing the test code using the $l_1$ norm. In the experiments, it is compared in other three cases: using no ELM embedding (also called "*w.o.ELM*" for short), using no MMC, (abbreviated as " *w.o.MMC*"), and using a linear predictor for SRC during test (abbreviated as "*Lin Pred*"). The detailed results are shown in Fig. 6.

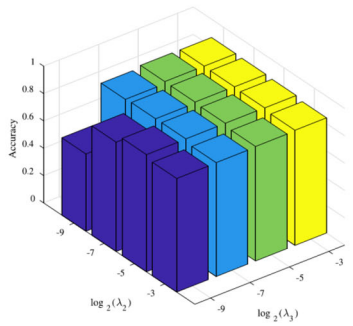
From the results in Fig. 6, the use of SRC improves performance. For AR Face and COIL-100, SRC does not give significant boost (only about 0.1%). On the other hand, SRC is critical to the Fashion-MNIST database, which improves the performance by almost 2.9%.



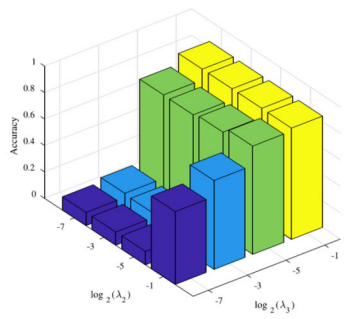

(a) Extended Yale B database




(b) AR Face database




(c) COIL-100 database




(d) Fashion-MNIST database

**Fig. 5** Average accuracy of WELM-ERDDL for different choices of $\lambda_2$ and $\lambda_3$ for each database

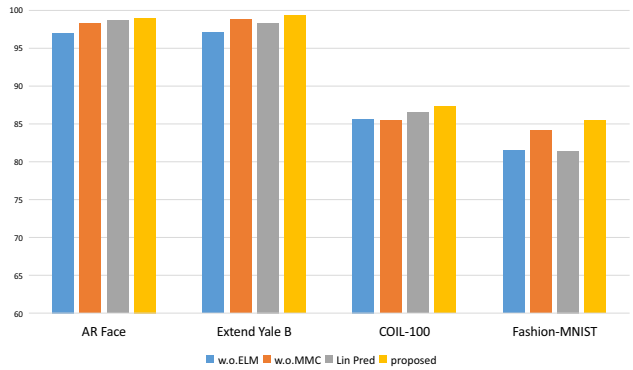



**Fig. 6** Average test accuracy of WELM-ERDDL under different experimental settings

## Real-world application learning tasks

In this section, the performance of the proposed WELM-ERDDL is to be evaluated in the real-world application: image classification task.

The original data are color images from the Corel dataset. Each image is segmented, using the Blob world system, into fragments that represent instances. The fragments containing specific visual contents (e.g., elephant) are labeled positive, while the remaining fragments are labeled negative. Therefore, the fragments (i.e., instances) from the same kind of images (e.g., elephant) create a binary learning problem. Given the five different image datasets: Tiger, Elephant, Fox, Bikes, and Cars, the number of instances is 1096, 1259, 1474, 5215, and 5600, respectively. The instances in the datasets Tiger, Elephant, and Fox are described by a 230-dimensional feature vector which represents the color, texture, and shape of the region, while the instances in the datasets Bikes and Cars are represented by a 90-dimensional feature vector.

Visual content-based image retrieval is an important application of image classification, for example, finding pictures containing an elephant from a dataset. In this subsection, sample images from the benchmark datasets are shown in Fig.7.

The detailed experimental results are shown in Tables 2 and 3 for image classification tasks in terms of classification accuracy (ACC) and area under the curve (AUC), respectively. In both tables, the proposed WELM-ERDDL method achieves the best ACC and AUC for all image datasets, indicating that it is superior to other methods in performing content-based image retrieval tasks. The extraordinary performance is owing to many local approximations created by this proposed method. The results show that the naive Bayes (NB) method on the datasets Tiger, Elephant, Bikes, and Cars has the worst ACC and AUC performance. However, the SVM method has the worst performance for the dataset Fox. For other baselines, more detailed experimental results can be found in Tables 2 and 3.

**Fig. 7** Example images used in the experiments from the COREL image categorization database

**Table 2** Experimental results on image classification datasets concerning classification accuracy (ACC) % and runtimes(S)

| Dataset | Proposed | HI-DKILEM | ELM | EN-ELM | W-ELM | SVM | KNN | NB |
|---|---|---|---|---|---|---|---|---|
| Tiger | 82.94 | 80.51 | 76.78 | 78.05 | 76.91 | 67.67 | 77.31 | 60.74 |
| Elephant | 85.23 | 82.29 | 79.03 | 80.42 | 78.87 | 73.85 | 78.56 | 65.55 |
| Fox | 68.87 | 67.85 | 64.46 | 65.18 | 64.11 | 52.19 | 64.83 | 57.15 |
| Bikes | 79.86 | 77.00 | 76.29 | 76.49 | 75.65 | 67.78 | 71.77 | 61.63 |
| Cars | 66.56 | 63.53 | 62.36 | 62.14 | 61.77 | 57.36 | 61.47 | 55.17 |
| Runtime (s) | 17.78 | 15.89 | 6.07 | 40.86 | 37.69 | 8.13 | 28.31 | 15.79 |

**Table 3** Experimental results on image classification datasets concerning the area under the curve (AUC) of ROC (%) and runtimes (S).

| Dataset | Proposed | HI-DKILEM | ELM | EN-ELM | W-ELM | SVM | KNN | NB |
|---|---|---|---|---|---|---|---|---|
| Tiger | 92.62 | 88.43 | 84.00 | 86.59 | 84.46 | 65.07 | 77.15 | 61.05 |
| Elephant | 93.34 | 90.33 | 87.12 | 89.08 | 87.05 | 71.83 | 78.44 | 66.23 |
| Fox | 78.85 | 74.87 | 70.05 | 71.51 | 69.74 | 53.56 | 64.87 | 57.44 |
| Bikes | 87.55 | 83.65 | 82.59 | 82.43 | 82.35 | 67.85 | 71.73 | 61.64 |
| Cars | 73.55 | 69.20 | 67.41 | 67.26 | 66.83 | 56.23 | 61.26 | 55.93 |
| Runtime (s) | 15.62 | 13.76 | 7.29 | 36.51 | 34.74 | 7.86 | 24.15 | 16.23 |

## Conclusion

In this paper, a novel nonlinear feature extraction approach called Weighted Extreme Learning Machine Exponential Regularized Discriminative Dictionary Learning (WELM-ERDDL) has been proposed. Evaluations on common benchmark datasets have shown that the proposed method has achieved better results than state-of-the-art dictionary learning algorithms. The proposed method has several distinct features from those of existing ELM-based methods.

1. A nonlinear WELM embedded feature projection strategy via Exponential Regularized Discriminative Dictionary Learning has been given to achieve feature diversity and low computational efficiency.
2. During the ELM learning stage, the output weights have been updated through elastic net regularization to enhance their compactness and meaningfulness.
3. An effective adaptive online weight update criterion has been designed for the WELM.

In future work, there is development space for exploring the proposed method. First, it is still challengeable to bring more insights into ELM to explore its deep learning capability. Second, the dropout technique or locality encoding method may be considered to further improve the performance of the algorithm, more effective approaches will be needed to cope with large-scale image classification problems.

**Acknowledgements** The authors would like to thank the anonymous reviewers for their constructive comments.

**Funding** This work was supported by National Natural Science Foundation of China (Grant no. 62006075), National Natural Science Foundation of Hunan Province of China (Grant no. 2022JJ30198), Hunan Provincial Science and Technology Department (CN) (Grant no. 21A0460), the science and technology innovation program of Hunan Province (Grant no. 2020RC5019) and Postgraduate Scientific Research Innovation Project of Hunan Province.

**Availability of Data and Materials** The data for this work are available upon request. Please direct all inquiries to the relevant author.

## Declarations

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.



## References

1. Zhang R, Lan Y, Huang GB et al (2012) Universal approximation of extreme learning machine with adaptive growth of hidden nodes. IEEE Trans Neural Netw Learn Syst 23:365–371
2. Cui D, Huang G, Kasun LLC, Zhang G, Han W (2017) Elm-net: feature learning using extreme learning machines. In 2017 IEEE international conference on image processing (ICIP) (pp 1857–1861)
3. Foroughi H, Ray N, Zhang H (2018) Object classification with joint projection and low-rank dictionary learning. IEEE Trans Image Process 27(2):806–821
4. Wang S, Zhu E, Yin J, Porikli F (2018) Video anomaly detection and localization by local motion based joint video representation and OCELM. Neurocomputing 277:161–175
5. Sun K, Mou S, Qiu J, Wang T, Gao H (2018) Adaptive fuzzy control for non-triangular structural stochastic switched nonlinear systems with full state constraints. IEEE Trans Fuzzy Syste 2018:1
6. Zhang W, Zhanga Z, Wang L et al (2019) Extreme learning machines with expectation kernels. IEEE Trans Image Process 96:1–13
7. Lekamalage CL et al (2016) Dimension reduction with extreme learning machine. IEEE Trans Image Process 25(8):3906–3918
8. Yang Y, Wu QMJ (2016) Extreme leaning machine with subnetwork hidden nodes for regression and classification. IEEE Trans Cybern 46(12):2885–2898
9. Wu D, Qu ZS, Guo FJ, Zhu XL, Wan Q (2019) Hybrid intelligent deep kernel incremental extreme learning machine based on differential evolution and multiple population gray wolf optimization methods. Automatika 60(1):48–57
10. D Wu, ZS Qu, FJ Guo , Q Wan (2019) Hybrid Multilayer Incremental Hybrid Cost Sensitive Extreme Learning Machine with Multiple Hidden Output Matrix and Subnetwork Hidden Nodes . IEEE ACCESS 7:118422–118434.
11. Scardapane S, Comminiello D, Scarpiniti M, Uncini A (2015) Online sequential extreme learning machine with kernels. IEEE Trans Neural Netw Learn Syst 26(9):2214–2220
12. Li S, Song S, Huang G, Wu C (2019) Cross-domain extreme learning machines for domain adaptation. IEEE Trans Syst Man Cybern Syst 49(6):1–14
13. Yu K, Liang J, Qu B, Luo Y, Yue C (2021) Dynamic selection preference-assisted constrained multiobjective differential evolution. IEEE Trans Syst Man Cybern Syst. https://doi.org/10.1109/TSMC.2021.3061698
14. Lakhdhar W, Mzid R, Khalgui M, Li Z, Frey G, Al-Ahmari A (2019) Multiobjective optimization approach for a portable development of reconfigurable real-time systems: from specification to implementation. IEEE Trans Syst Man Cybern Syst 49(3):623–637
15. Chen Q, Ding J, Yang S, Chai T (2020) A novel evolutionary algorithm for dynamic constrained multiobjective optimization problems. IEEE Trans Evol Comput 24(4):792–806
16. Zhou Y, Xiang Y, He X (2021) Constrained multi-objective optimization: test problem construction and performance evaluations. IEEE Trans Evol Comput 25(1):172–186. https://doi.org/10.1109/TEVC.2020.3011829
17. Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1):489–501

18. Huang G-B, Chen L, Siew C-K (2006) Universal approximation using incremental constructive feed forward networks with random hidden nodes. IEEE Trans Neural Netw 17(4):879–892
19. Zeng Y, Li Y, Chen J et al (2020) ELM embedded discriminative dictionary learning for image classification. Neural Netw 123:331–342
20. Wang W, Zhang R (2014) Improved convex incremental extreme learning machine based on enhanced random search. Electr Eng Electron Engg 238:2033–2040
21. Tang J, Deng C, Huang G-B (2016) Extreme learning machine for multilayer perceptron. IEEE Trans Neural Netw Learn Syst 27(4):809–821
22. Dasgupa S, Stevens CF (2017) A neural algorithm for a fundamental computing problem. Science 358(6364):793–796
23. Kunjie Y, Jing L, Boyang Q, Caitong Y (2021) Purpose-directed two-phase multiobjective differential evolution for constrained multiobjective optimization. Swarm Evol Comput 60:100799
24. Kunjie Y, Xu C, Xin W, Zhenlei W (2017) Parameters identification of photovoltaic models using self-adaptive teaching-learning-based optimization. Energy Convers Manage 145:233–246
25. Pierezan J, Coelho LDS (2018) Coyote optimization algorithm: a new metaheuristic for global optimization problems. In: 2018 IEEE Congress on Evolutionary Computation (CEC). Rio de Janeiro, Brazil, USA: IEEE
26. Ling T, Zhi-Hui Z, Wang YX, Wang ZJ, Yu WJ, Zhang J (2018) Competitive swarm optimizer with dynamic grouping for large scale optimization. In: Proc. IEEE Congr. Evol. Comput. (CEC 2018), Rio de Janeiro, Brazil, pp 2655–2660
27. Rahman CM, Rashid TA (2020) A new evolutionary algorithm: learner performance based behavior algorithm. Egypt Inf J. https://doi.org/10.1016/j.eij.2020.08.003
28. Abdullah JM, Ahmed T (2019) Fitness dependent optimizer: inspired by the bee swarming reproductive process. IEEE Access 7:43473–43486. https://doi.org/10.1109/ACCESS.2019.2907012
29. Di Wu, Li T, Wan Q (2021) A hybrid deep kernel incremental extreme learning machine based on improved coyote and beetle swarm optimization methods. Compl Intell Syst 7:3015–3032