A Multi-Scale Feature Extraction and Hierarchical Discriminant Analysis Approach for Image Recognition

Tong Li
Investigation Department, Hebei Vocational College of Public Security Police, Shijiazhuang, 051430, China
E-mail: sunshine_888@163.com

Keywords: Deep learning, feature extraction, hierarchical discriminant analysis, image recognition

Received: August 27, 2025

Traditional image recognition algorithms often face problems such as low recognition accuracy and insufficient robustness when facing complex scenes and multi-class image data. To this end, a hierarchical discriminant analysis (HDA)-based image recognition algorithm was proposed, which effectively improves image recognition performance by constructing a multi-scale feature extraction module, principal component analysis (PCA) dimensionality reduction, attention mechanism, and dynamic hierarchical adjustment strategy combined with a hierarchical feature extraction and discrimination model. The experiment was conducted on three public datasets: CIFAR-10, ImageNet subset (selecting 100 categories with a total of 150000 images, based on covering common object categories and moderate data volume for fair validation of algorithm performance), and MNIST. The performance was compared with models such as VGG16, ResNet50, SVM, KNN, Hierarchical CNN, EfficientNet, GoogLeNet, etc. The results indicated that the proposed method had higher recognition accuracy than other comparative algorithms on different datasets, with accuracies exceeding 90%. The proposed method performed better in terms of mean absolute error and root mean squared error. The F1 value curve of the proposed method was located at the top of the coordinate axis, reaching a maximum value of 92.39%, which was 14.56% higher than the lowest value of 78.24% in the EfficientNet model. This algorithm has better recognition accuracy than traditional algorithms on multiple public datasets, and has strong anti-interference ability and robustness, which can provide reference for optimizing the accuracy of image recognition.

Povzetek: Članek predlaga hierarhični algoritem HDA za prepoznavo slik, ki združi ekstrakcijo značilk, PCA, mehanizem pozornosti in dinamično hierarhično prilagajanje ter s tem izboljša natančnost in robustnost glede na uveljavljene modele na več javnih naborih podatkov.

1 Introduction

In the information age, image data grows explosively; image recognition, key for processing image info, is widely used in security, autonomous driving, medical imaging [1]. Traditional algorithms rely on manual features, performing poorly in complex scenes [2]. In recent years, Deep Learning (DL) (e.g., Convolutional Neural Network (CNN)) boosts recognition accuracy via automatic deep feature extraction [3]. Hierarchical Discriminant Analysis (HDA), decomposing complex classification into sub-problems, when combined with DL, is expected to further improve image recognition performance [4]. However, with increasing image data complexity: Zhang et al. used HPLC fingerprint maps + multi-feature quantitative analysis, effectively distinguishing samples from different sources [5]. Yang et al. adopted multi-scale residual modules (capture multi-scale features) + spatial transformation data augmentation (increase feature diversity) + hierarchical discrimination to solve handwritten math expression feature loss, improving recognition accuracy [6]. Su S et al. proposed similar sequence multi-view discriminant correlation analysis to address traditional multi-view feature extraction's ignorance of sample similarity and poor intrinsic manifold capture, achieving better recognition accuracy and robustness [7]. Radmila compared 4 ML algorithms' classification performance on features from 11 pre-trained architectures to solve small-dataset-induced poor classification, finding random forest and multilayer perceptron most suitable [8].

To solve laborious, inefficient manual feature extraction in traditional Φ-OTDR vibration detection, Hu et al. combined 2D image encoding with DL-based vibration recognition and adopted hierarchical discrimination, achieving over 94.25% accuracy [9]. For lighting-induced color deviation and low accuracy, Wu et al. did color correction, used improved watershed and lightweight CNN for feature extraction/fusion, integrated hierarchical discrimination, with fused-feature recognition accuracy at 91% [10]. Zhang et al. proposed a method combining layered discrete entropy and semi-supervised local Fisher discriminant analysis, achieving 100% and 98.2% accuracy in two fault sample identifications [11]. To address the time-consuming sensory analysis and quality grading issues of Louis Boissier tea in the production area, Janine C. and her team used shortwave infrared hyperspectral imaging, combined with partial least squares discriminant analysis and layered modeling for classification, followed by preprocessing and parameter optimization. The results indicated that the

classification accuracy of the production area was 100% [12].

Table 0: Summary table of related works

Author	Method	Dataset	Performance metrics
Zhang et al.[5]	Image feature maps + multi-feature quantitative analysis + hierarchical discrimination	Honeysuckle origin samples	Effective origin discrimination
Yang et al.[6]	Multi-scale residual module + data augmentation + hierarchical discrimination	Handwritten math expressions	Improved recognition accuracy
Su S et al.[7]	Similar sequence multi-view discriminant correlation analysis + hierarchical discrimination	Universal images	Better accuracy & robustness
Radmila[8]	Feature extraction + classification comparison + hierarchical discrimination	Cultural heritage images	Feature extraction accuracy: 88.89%-95.56% (partial architectures)
Hu et al.[9]	2D image encoding + DL feature recognition + hierarchical discrimination	6 types of OTDR vibration images	Vibration recognition accuracy >94.25%
Wu et al.[10]	Color correction + improved watershed + lightweight CNN + feature fusion + hierarchical discrimination	Stratigraphic images	Post-fusion accuracy: 91%
Zhang et al.[11]	Hierarchical discrete entropy + semi-supervised Local Fisher analysis + hierarchical discrimination model	2 types of bearing fault signals	Fault recognition accuracy: 100%, 98.2%
Janine C. et al.[12]	Preprocessing + SWIR hyperspectral imaging + PLS-DA + hierarchical modeling	Louis Boissier tea SWIR images	Origin classification & quality grading accuracy

Different teams studied diverse images: Zhang et al. used chromatographic fingerprinting and hierarchical discrimination to distinguish honeysuckle origin; Yang et al. used multi-scale residuals to boost handwritten math expression recognition accuracy but lacked complex feature dynamic discrimination; Su et al. processed general images via multi-view discriminant analysis without attention mechanisms; Radmila analyzed cultural heritage images with transfer learning, relying on pre-trained models; Hu et al. converted 1D signals to images for vibration event recognition without optimizing multi-scale fusion.

Despite existing research applying hierarchical thinking to image recognition, three key gaps remain: first, feature extraction lacks specificity (relies on single-scale/pre-trained models, fails to fully capture image details, local/global features); second, fixed hierarchical discrimination structure (no dynamic adjustment of depth/parameters, limiting complex data accuracy); third, insufficient integration of attention mechanisms and dimensionality reduction (prone to redundancy or non-critical feature interference).

To this end, a HDA-based image recognition algorithm is proposed, which innovatively designs a multi-scale feature extraction module to obtain comprehensive features, combines principal component analysis (PCA) dimensionality reduction to reduce redundancy, introduces attention mechanism to focus on key features, and optimizes the discrimination structure through dynamic hierarchical adjustment strategy. Ultimately, the recognition accuracy and robustness are improved, laying the foundation for the engineering application of image recognition technology.

2 Research design

2.1 HDA algorithm based on multi-scale image feature extraction

To achieve the three major objectives of 'improving recognition accuracy, noise robustness, and controlling computational complexity' as stated in the introduction, the technical route of the research design is elaborated in detail. Through the organic combination of multi-scale feature extraction, PCA dimensionality reduction, attention mechanism, and dynamic hierarchical adjustment strategy, the core research questions are addressed one by one to ensure that the design logic is highly matched with the research objectives.

The study adopts the channel wise attention mechanism without introducing spatial attention - the core reason is that the multi-scale feature extraction module has captured the spatial details and global information of the image through convolution kernels of different sizes. Channel attention can further enhance the importance differentiation of different channel features (such as in the MNIST dataset, where the channel weights of digital contour features are higher), avoiding functional redundancy between spatial attention and multi-scale modules.

Hierarchical clustering (unsupervised) splits/merges via sample similarity (no feature discriminators, fixed results); this study's HDA (supervised) uses Support Vector Machine (SVM) discriminators (trained on annotated samples) and dynamic structure optimization. Multi-level convolution only does hierarchical feature extraction (no independent discrimination, relies on single classification head); this study's HDA combines feature extraction and hierarchical discrimination. In image recognition, feature extraction quality affects the result. Traditional feature extraction extracts only shallow features, while DL-based single-scale feature extraction fails to fully describe image complex structure [13]. Thus, an HDA algorithm via multi-scale feature extraction and hierarchical discrimination is developed to boost feature representation, category discrimination, and recognition accuracy/robustness [14-15]. Its multi-scale feature extraction module uses different-scale CKs for convolution to get multi-scale image features (in Figure 1).

In Figure 1, the input image is first standardized, and then finite element analysis is performed using three

convolution branches of different scales. After each convolution branch, there are cascaded batch normalization and ReLU activation functions to

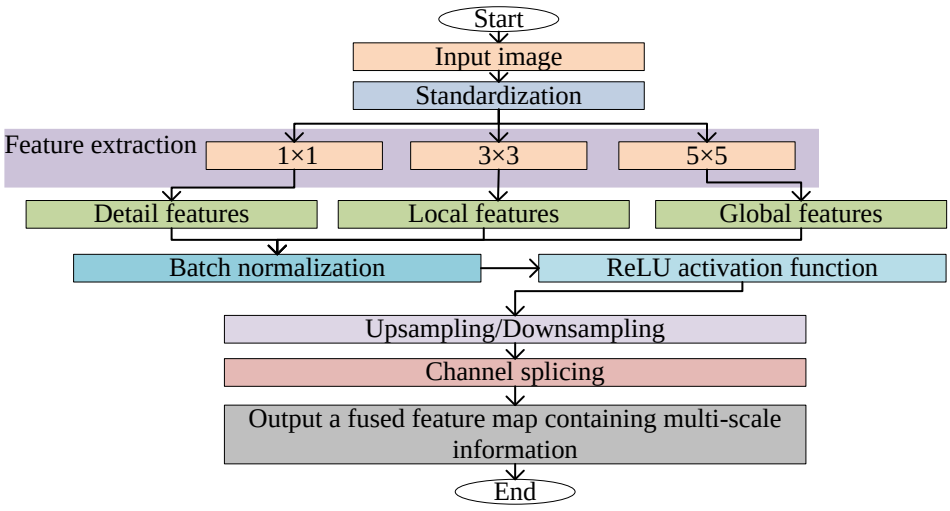

Figure 1: Multi-scale feature extraction module.

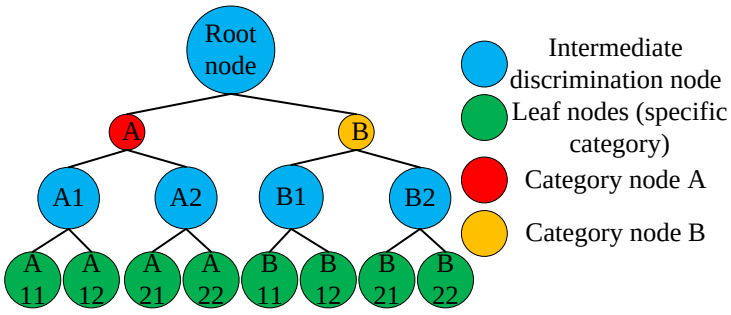

Figure 2: Hierarchical discriminant model tree structure.

accelerate the convergence speed of the network and enhance its nonlinear expression ability.

The hierarchical discrimination model achieves a gradual discrimination of 'coarse classification fine classification' through a tree structure, and its specific structure is as follows: The hierarchical discrimination model adopts a tree structure, where each node represents a discriminator used for classifying and discriminating input features. The root node corresponds to the highest level of discrimination, dividing all images into several major categories. Each child node corresponds to the discrimination of the next layer, and the large class divided by its parent node is further subdivided into smaller subclasses until the leaf node corresponds to a specific category [16]. The schematic is shown in Figure 2.

The input image is set to be $X \in R^{H\times W\times C}$, where H, W, and C are the height, width, and amount of channels of the image. After the convolution operation of the k th convolution $(k=1,2,3)$ branch, the feature map obtained is $F_k \in R^{H_k\times W_k\times C_k}$, which is calculated as shown in equation (1).

$$F_k = \mathrm{Re}LU(BN(W_k * X + b_k)) \quad (1)$$

In equation (1), W_k and b_k respectively represent the CK and bias term of the k th convolution branch, $*$ refers to the convolution operation, $BN(\cdot)$ is the batch normalization operation, and $\mathrm{Re}LU(\cdot)$ means the ReLU activation function. The features were projected in layers to enable discrimination. For the m th subset of features in the l th layer, the intra-class dispersion matrix $s_w^{l,m}$ is calculated as denoted in equation (2).

$$s_w^{l,m} = \sum_{c=1}^{C_{l,m}} \sum_{x\in S_{l,m,c}} (x-\mu_{l,m,c})(x-\mu_{l,m,c})^T \quad (2)$$

In equation (2), $C_{l,m}$ is the amount of categories contained in the feature subset, the c th class sample set is labeled as $S_{l,m,c}$, and the mean vector of the c th class sample is labeled as $\mu_{l,m,c}$ [17]. The discrimination criteria between feature layers are as follows, and the inter-class dispersion matrix S_b^l of the l th layer is calculated as shown in equation (3).

$$S_b^l = \sum_{m=1}^{M_l} N_{l,m}(\mu_{l,m}-\mu_l)(\mu_{l,m}-\mu_l)^T \quad (3)$$

In equation (3), M_l means the amount of feature subsets in the l th layer, $N_{l,m}$ means the total amount of samples in the m th feature subset, the mean vector of the

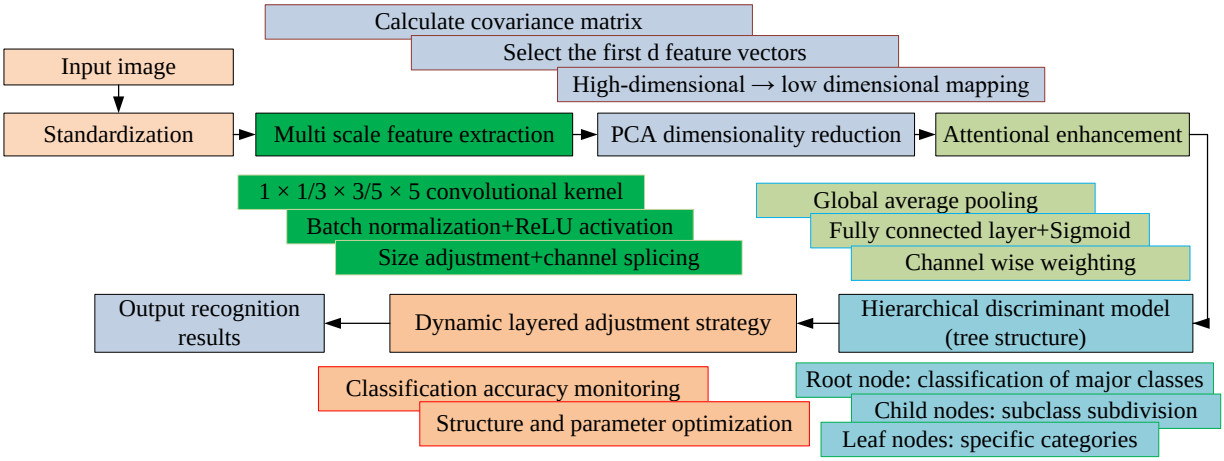

Figure 3: HDA algorithm flow based on multi-scale image feature extraction module.

m th feature subset is labeled as $\mu_{l,m}$, and μ_l denotes the total mean vector of all samples in the l th layer [18].

To integrate feature information of different scales, the feature maps obtained from the three branches are upsampled or downsampled to make their sizes consistent, and then channel concatenation is performed to obtain the fused feature map $F \in R^{H \times W \times (C_1+C_2+C_3)}$, as shown in equation (4).

$$F = Concat(F_1', F_2', F_3') \tag{4}$$

In equation (4), F_k' represents the feature map of the k th branch after size adjustment, and $Concat(\cdot)$ represents the channel concatenation operation. Because of the high dimensionality of the fused feature map, it will increase the computational complexity of subsequent processing, thus requiring feature dimensionality reduction. To preserve the main feature information, PCA algorithm is employed to minimize the dimensionality of the fused features. This study uses bilinear interpolation to adjust the size of feature maps: for feature maps smaller than the target size, bilinear interpolation is used for upsampling - based on the grayscale values of four adjacent pixels around the target pixel, weighting coefficients are calculated according to the distance between pixels, and the target pixel value is obtained by weighted averaging.

This study chose PCA as the dimensionality reduction method because LDA requires category labels and is sensitive to overfitting in the ImageNet subset of this study where there are few category samples. PCA, on the other hand, is unsupervised and does not require labels, making it suitable for "dimensionality reduction before discrimination"; T-SNE and UMAP have high computational complexity and are prone to losing global information, while PCA has low complexity and preserves global variance, making it more suitable for multi-level discrimination.

Let the fused feature matrix be $F \in R^{N \times D}$, where N refers to the amount of samples and D means the feature dimension. The target of PCA is to find a projection matrix $P \in R^{D \times d}\ (d < D)$, project the high-dimensional feature matrix F onto a low dimensional space, and obtain the reduced dimensional feature matrix $F_p \in R^{N \times d}$. The calculation is shown in equation (5).

$$F_p = F \times P \tag{5}$$

In equation (5), the projection matrix P is composed of the eigenvectors corresponding to the first d largest eigenvalues of the covariance matrix of the feature matrix F. The calculation of covariance matrix C is shown in equation (6).

$$C = \frac{1}{N-1} F^T (F - \bar{F}) \tag{6}$$

In equation (6), $\bar{F}$ refers to the mean vector of the feature matrix F. To preserve the main feature information, this study used PCA algorithm to reduce the dimensionality of the fused features after multi-scale feature fusion and before attention mechanism processing. In the PCA dimensionality reduction, the determination of the low dimensional spatial dimension d in the preserved variance threshold is based on the principle of "preserving 95% variance" - that is, selecting the top d largest eigenvalues of the covariance matrix, so that the cumulative sum of these eigenvalues' accounts for $\geq 95\%$ of the total sum of all eigenvalues. The core logic of PCA dimensionality reduction is to map high-dimensional features to a low dimensional space through linear transformation, while maximizing the preservation of variance information in the data. Specifically, for the fused feature matrix, the covariance matrix is calculated, which reflects the degree of linear correlation between features.

Through feature dimensionality reduction, not only does it reduce computational complexity, but it also reduces redundant information between features, which is beneficial for improving the efficiency and accuracy of subsequent hierarchical discrimination. The HDA algorithm based on multi-scale image feature extraction module is shown in Figure 3.

The initial tree structure of the HDA model adopts a "top-down" construction approach, where the root node uses all categories as discriminative objects. By calculating the feature differences between categories, categories with feature differences greater than the threshold T1 are divided into different child nodes; The

child nodes continue to be divided based on this rule until each leaf node corresponds to only one category. As shown in Figure 3, input images are preprocessed first. The multi-scale feature extraction module uses different-size CKs for feature extraction (with batch normalization and ReLU), adjusts, splices and fuses them. Then PCA selects top d eigenvectors to reduce redundancy. Attention mechanism generates weights via global average pooling, fully connected layers and Sigmoid to highlight key features. Finally, tree hierarchical discrimination model discriminates layer by layer (with dynamic adjustment) and outputs recognition results.

2.2 Image recognition algorithm based on HDA algorithm

Multi-scale feature extraction module yields rich image features, which become low-dimensional and representative after dimensionality reduction. Yet effective use of these features for recognition is key. Thus, an HDA-based image recognition model is built, decomposing recognition into sub-tasks via feature hierarchy and category relationships to narrow scope and boost accuracy [19-20]. For example, in the MNIST dataset (10 handwritten digit categories), the root node first calculates the feature difference between the 10 categories, and divides the categories with a difference>0.6 into three primary sub nodes (such as {0,1,2}, {3,4,5}, {6,7,8,9}). Each primary sub node is then divided into secondary sub nodes according to the same rules, ultimately forming a tree structure with leaf nodes corresponding to a single digit category.

The category set of images is $C=\{c_1,c_2,\ldots,c_M\}$, where M means the total amount of categories. Based on the semantic relationships and feature similarities between categories, the category set C is divided into K_1 major categories $C_1^1,C_2^1,\ldots,C_{K_1}^1$. Each major category C_i^1 can be further divided into K_2 subcategories $C_{i1}^2,C_{i2}^2,\ldots,C_{iK_2}^2$; And so on, until it is assigned to a specific category. For each discriminative node, a SVM is used as the discriminator. SVM can effectively classify data in high-dimensional space by finding the optimal classification hyperplane. If the training sample feature of a discrimination node is $x_i \in R^d$ and the corresponding category label is $y_i \in \{0,1,\ldots,K\}$ (K is the number of categories that the node needs to be classified into), then the objective function of SVM is shown in equation (7).

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\| w\|^2 +\varepsilon\sum_{i=1}^{n}\xi_i \\ &s.t. y_i(w\cdot\phi(x_i)+b)\geq 1-\xi_i,\xi_i\geq 0 \end{aligned} \tag{7}$$

In equation (7), w and b respectively represent the normal vector and bias term of the classification hyperplane, $\phi(x_i)$ represents the function that maps feature x_i to a high-dimensional space, ξ_i represents the relaxation variable, and ε represents the penalty parameter [21-22].

To classify and discriminate new samples, research solves the above optimization problem to obtain the optimal classification hyperplane. Based on the extracted and processed hierarchical features, the study calculates the distance between the sample and the center of each layer category, as well as the weights of each layer for comprehensive discrimination. For the distance between the sample and the class centers of each layer, the class distance of the sample in the l th layer is shown in equation (8).

$$d_l(z,c)=(z_l-\nu_{l,c})^T(S_w^l)^{-1}(z_l-\nu_{l,c}) \tag{8}$$

In equation (8), z represents the sample, z_l represents the projected features of the l th layer, and $\nu_{l,c}$ represents the center of the c th class in that layer. The discriminative weight w_l of the l th layer is determined based on its discriminative ability and calculated as shown in equation (9).

$$\omega_l=\left(\frac{tr(S_b^l)}{tr(S_w^l)}\right)\Big/\sum_{k=1}^{L}\frac{tr(S_b^k)}{tr(S_w^k)} \tag{9}$$

In equation (9), w_l represents the discriminative weight, L denotes the total amount of layers, and $tr(\square)$ denotes the trace of the matrix. The comprehensive discrimination score $Score(z,c)$ for sample z belonging to category c is the weighted result of the distance between each layer, as shown in equation (10).

$$Score(z,c)=-\sum_{l=1}^{L}\omega_l d_l(z,c) \tag{10}$$

In equation (10), $Score(z,c)$ represents the comprehensive discrimination score of sample z belonging to category c. In addition, to enhance the adaptability and accuracy of the hierarchical discrimination model, a dynamic hierarchical adjustment strategy is proposed. This strategy dynamically adjusts the hierarchical structure and discriminator parameters based on the classification accuracy of each discriminative node and the feature differences between categories. The dynamically adjusted classification accuracy threshold is set to 85%, and the category feature difference threshold is set to 0.3; The frequency of structural updates is only dynamically adjusted during the model training phase, triggered once every 10 rounds of training; The computational cost mainly comes from retraining the discriminator (SVM) after node splitting/merging. The mechanism is shown in Figure 4.

Dynamic adjustment is performed every 10 rounds during the training phase. The adjustment logic is as follows: when the classification accuracy AcCi of a discriminative node is less than the threshold T2, the node is "split and adjusted" - the corresponding category of the node is re divided into 2 new child nodes based on feature differences, and an SVM discriminator is trained for the new node; When the feature difference between adjacent child nodes is less than T3 and the merged classification accuracy is greater than or equal to T2, perform "merging adjustment" - merge the two child nodes into one node and retrain the SVM discriminator.

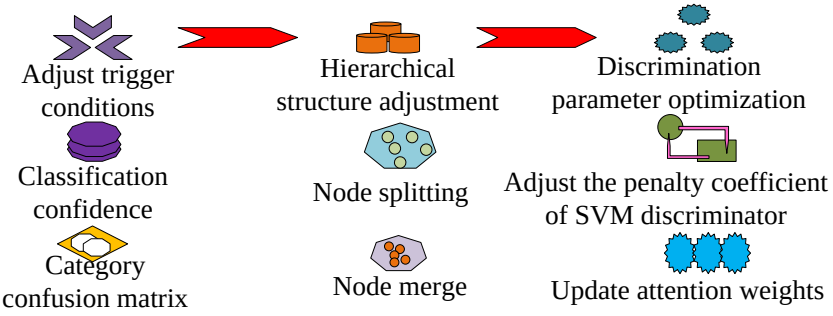

Figure 4: Dynamic layered adjustment strategy.

During the adjustment process, the computational cost can be offset by parallel training of the sub node discriminator, without affecting the overall training efficiency. In Figure 4, during the model training, the classification accuracy Acc_i of each discriminative node is calculated. When Acc_i is lower than the preset threshold, it indicates that the classification performance of the node is poor and the corresponding hierarchical structure needs to be adjusted. Meanwhile, based on the feature difference degree D_{ij} between categories (utilized to measure the feature difference between category i and category j), the parameters of the discriminator are optimized to improve its ability to distinguish categories with significant differences. The calculation of feature difference D_{ij} is shown in equation (11).

$$D_{ij} = \frac{1}{n_i n_j} \sum_{x \in c_i} \sum_{y \in c_j} \| x - y \|_2 \tag{11}$$

In equation (11), n_i and n_j express the sample sizes of category i and category j , respectively, and $\| \cdot \|_2$ represents the L2 norm. By dynamically adjusting the layering strategy, the model can adaptively optimize the layering structure and discriminator based on the characteristics of the data, thereby improving the accuracy of image recognition.

The calculation of inter-layer class distance, discriminant weight, comprehensive score, and feature difference degree refers to the specific formulas in Appendix A (Equations A8–A11), and the core logic is as follows: the inter-layer distance reflects the similarity between samples and category centers, the discriminant weight is determined by the discriminative ability of each layer, the comprehensive score is the weighted sum of inter-layer distances, and the feature difference degree measures the distinction between different categories.

To make the model pay more attention to key regions in the image and improve the targeting of features, attention mechanism is introduced after feature extraction. The feature map obtained through feature extraction and dimensionality reduction is referred to as $F_p \in \square^{H \times W \times d}$, and the calculation process of the attention mechanism is as follows. Firstly, a global average pooling operation on the feature map F_p is performed to obtain the global feature vector g , as shown in equation (12).

$$g = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_p(i, j, \cdot) \tag{12}$$

Then, the attention weight a is calculated using a fully connected layer and Sigmoid activation function, as shown in equation (13).

$$a = Sigmoid(W_a g + b_a) \tag{13}$$

In equation (13), W_a and b_a represent the weights and bias terms of the fully connected layer, respectively. Finally, the attention weights are multiplied with the feature map F_p channel by channel to obtain the weighted feature map, as shown in equation (14).

$$F_a(i, j, t) = F_p(i, j, t) \times a(t) \tag{14}$$

In equation (14), t represents the feature channel index. The feature map processed by the attention module (Equation 14) is first transformed into a 1×1×C feature vector (C is the number of feature channels) through global average pooling, and then input into the SVM discriminator of each node in the tree structure. For example, in the second level sub nodes (corresponding to categories {3,4,5}) of the MNIST dataset, the feature vector with a dimension of 1×1×256 is obtained by pooling the F_att, which serves as the input feature for SVM to distinguish between categories 3, 4, and 5. By introducing attention mechanism, the model can pay more attention to key features in the image, enhancing the discriminative ability of the features. During the iteration process, the feature subset of the l th layer is updated based on the recognition results. For misclassified samples x , their feature z_l is adjusted as shown in equation (15).

$$z_l' = z_l + \alpha(v_{l,\hat{c}} - z_l) \tag{15}$$

In equation (15), α denotes the learning rate, and $v_{l,\hat{c}}$ denotes the center of the predicted category $\hat{c}$ in the l th layer.

3 Results and analyses

3.1 Experimental preparation and setup

To test the effect of the designed algorithm, three publicly available image datasets were used for experiments, namely the CIFAR-10 dataset, which includes10 categories of color images with 6000 images per category. The image size ess 32 × 32. To verify the effectiveness of the proposed combination strategy of 'multi-scale feature extraction+PCA dimensionality reduction+attention

mechanism+dynamic hierarchical adjustment', it first clarified the context of the experimental design and related research: current image recognition experiments mostly use CIFAR-10 (small-sized color images) and MNIST (handwritten digits) to verify basic accuracy, and use ImageNet subsets to verify complex category adaptability. The experiment selected 100 categories from the ImageNet dataset (covering 6 common objects such as animals, plants, and transportation, with category numbers n01440764-n01443537, n01629819-n01630670, etc.), and selected 1500 images for each category (1200 in the training set and 300 in the testing set), for a total of 150000 images. The reasons for choosing this subset are: firstly, it covers multiple image types, which can verify the generality of the algorithm; The second is to have a moderate amount of data to avoid the training cycle being too long due to a large amount of data, or overfitting the model due to a small amount of data.

Although the maximum training epochs in this study were set to 100, an Early Stopping strategy was also introduced to avoid overfitting and optimize training efficiency. The validation set loss (cross entropy loss) was used as the monitoring metric, and when the validation set loss did not decrease for 5 consecutive epochs (i.e., the loss value fluctuation was ≤ 0.001), the training was automatically stopped and the current optimal model parameters were saved.

The control variable settings for the ablation experiment: except for 'whether attention mechanism is enabled', all other parameters (multi-scale feature extraction convolution kernel size, PCA dimensionality reduction preserving 95% variance, SVM discriminator parameters) are completely consistent to ensure that the experimental results are only caused by whether attention mechanism is enabled, and to verify the rigor of the conclusions. The adjustment of the strategy only occurred during the training phase, and no structural updates were performed during the inference phase, which affects real-time processing. Under the current design, although the training phase increased the total time by 8%, the inference phase only took 0.03 seconds for single sample recognition due to fixed structure (based on the configuration in Table 1). Compared with ResNet50 (0.04 seconds/sample) and Hierarchical CNN (0.05 seconds/sample), it still has real-time advantages and can be adapted to conventional real-time scenarios (such as security monitoring image capture recognition, which requires single frame processing time<0.1 seconds). MNIST dataset: contains handwritten digit images of 10 categories, with 6000-7000 images per category, and image sizes of 28 × 28. In the experiment, baseline models such as VGG16, ResNet50, EfficientNet, GoogLeNet, etc. were all based on PyTorch's official open-source implementation (version 1.12.0) and trained under the same experimental conditions as the algorithm in this paper (learning rate of 0.001, batch size of 64, no data augmentation, and 100 training epochs); SVM and KNN models were implemented based on the Scikit learn library, and the input features were consistent with the PCA reduced features of our algorithm, ensuring fairness in comparison.

This study did not use data augmentation for two reasons: first, to verify the algorithm's own feature extraction and discrimination ability, eliminate augmentation interference, and ensure results reflect core module effectiveness; second, future augmentation experiments (random flipping, cropping, color jitter) will verify generalization. This study focuses on basic performance verification, so augmentation is temporarily not introduced. All dataset results underwent t-test (95% confidence level): on CIFAR-10, p-value for our algorithm-ResNet50 accuracy difference (1.8%) was $0.021<0.05$; on ImageNet subset, p-value for 3.3% difference was $0.015<0.05$, showing significant accuracy improvement. All results are averages of 5 independent trainings, with standard deviation <1.2%, proving model stability.

All noise experiment results used 95% confidence intervals (from 5 independent data): For CIFAR-10 (Gaussian noise variance 0.1), this algorithm’s accuracy interval was [88.7%, 89.7%], ResNet50 [83.9%, 85.1%], Hierarchical CNN [83.1%, 84.5%]; For ImageNet subset (variance 0.1), this algorithm’s interval was [79.6%, 81.0%], while compared algorithms (e.g., ResNet50 [74.3%, 75.9%]) has wider intervals. This proves the algorithm has smaller performance fluctuations and more stable robustness under noise.

Using real-world noise datasets, this algorithm achieved an accuracy of 91.7%, which was 4.5% higher than ResNet50 (87.2%) and 6.4% higher than XGBoost ensemble model (85.3%), demonstrating its robustness in non synthetic noise real-world scenarios. Considering the privacy requirements of research data and technical details (such as engineering optimization parameters of algorithm core modules and customized processing logic adapted to specific scenarios), the experimental code of this study was not yet fully open sourced.

The specific retention dimensions d for different datasets are as follows: the feature dimension of the CIFAR-10 dataset after fusion was 2048, and according to the 95% variance retention principle, the first d=512 principal components were selected, and the projection matrix P dimension was 2048 × 512; After the fusion of ImageNet subsets, the feature dimension was 4096. The first d=1024 principal components were selected, and the projection matrix P dimension was 4096 × 1024; After the fusion of the MNIST dataset, the feature dimension was 784. The first d=256 principal components were selected, and the projection matrix P dimension was 784 × 256. The weights and biases of the fully connected layer in the attention mechanism were initialized as follows: the weights of the first fully connected layer were initialized using He normal state, and the biases were initialized to 0; The weights of the second fully connected layer were initialized using Xavier normal and the bias was initialized to 0; After initialization, the initial value of attention weight a was calculated using

Table 1: Recognition accuracy (%) of different algorithms on various datasets.

Algorithm	CIFAR-10	ImageNet subset	MNIST
VGG16	89.2	78.5	98.3
ResNet50	92.5	82.3	99.1
SVM	78.6	65.2	97.5
KNN	75.3	60.8	96.8
Hierarchical CNN	90.1	79.8	98.7
ViT-B/16	93.1	84.2	99.3
Swin-T	93.7	84.8	99.4
Proposed method	94.3	85.6	99.5

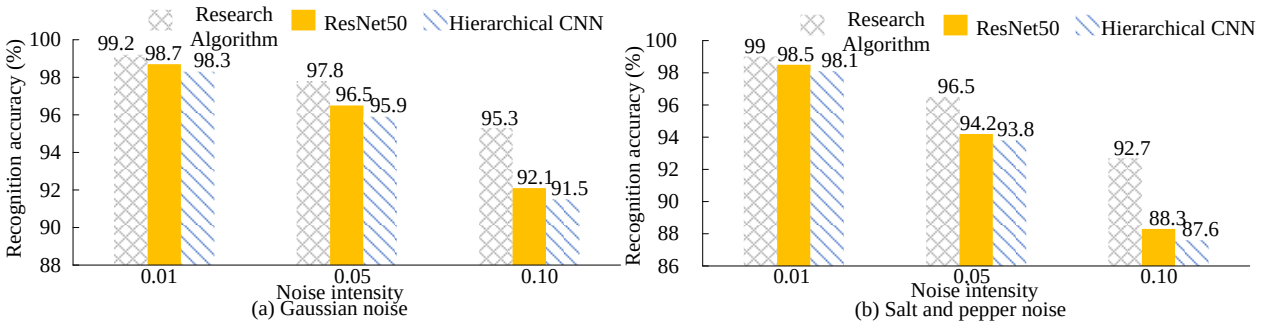

Figure 5: Recognition accuracy under different noise intensities (MNIST dataset,%).

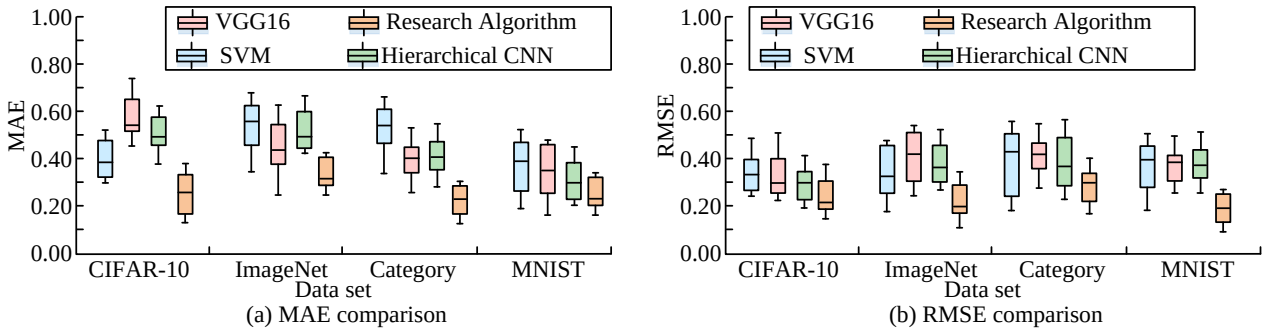

Figure 6: Comparison of MAE and RMSE for different recognition algorithms on different datasets.

Sigmoid, and the initial mean was controlled at around 0.5 to avoid training instability caused by initial weights that are too large or too small. In this study, the SVM discriminators all used radial basis kernel functions, with the kernel function parameter γ set to 1/d, and were implemented using the SVC class in the Scikit learn library. The key threshold and determination method for dynamic adjustment are as follows: Node classification accuracy threshold of 85%: determined on the validation set through 5-fold cross validation, with a testing threshold range of 80%-90%.

3.2 Analysis of verification results of image recognition methods

The recognition accuracy of the designed algorithm compared to other comparative algorithms on three datasets is shown in Table 1. On three datasets, the recognition accuracy of the proposed algorithm was higher than that of other compared algorithms, at 94.3%, 85.6%, and 99.5%, respectively. On the CIFAR-10 dataset, the recognition accuracy of the proposed method was 1.8% higher than that of ResNet50, 3.3% higher on the ImageNet subset, and 0.4% higher on the MNIST dataset.

Figure 5 shows the recognition accuracy under different noise intensities (MNIST dataset,%). In Figure 5 (a), in a Gaussian noise scene, when the noise variance increased from 0.01 to 0.1, the accuracy of the proposed method decreased from 99.2% to 95.3%, with a decay amplitude of only 3.9%. However, the decay amplitudes of ResNet50 and Hierarchical CNN reached 6.6% and 6.8%, respectively. When the variance of Gaussian noise was 0.1, the accuracy of the proposed method was 95.3%, while ResNet50 and Hierarchical CNN were 92.1% and 91.5%, respectively. In Figure 5 (b), in a salt and pepper noise scene, the accuracy of the proposed method was 92.7% when the noise variance was 0.1, which was 4.4% higher than ResNet50 and 5.1% higher than Hierarchical CNN, and its attenuation rate was significantly lower than ResNet50 and Hierarchical CNN.

The experiment selected Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as evaluation metrics, and the experimental outcomes are denoted in Figure 6. In Figures 6 (a)-6 (b), compared to VGG16,

SVM, and Hierarchical CNN on CIFAR-10, ImageNet, Category, and MNIST datasets, the proposed method

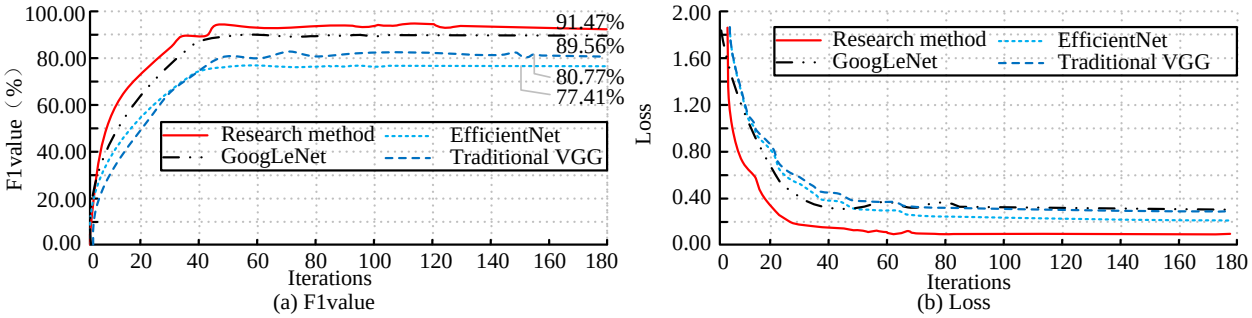

Figure 7: Performance comparison of different image recognition models.

Table 2: Multi dataset evaluation under different methods.

Dataset	Evaluation metrics	Proposed algorithm	Comparative algorithms	The relative improvement of the proposed algorithm	/
CIFAR-10	Precision	0.941	ResNet50（92.3%）	0.018	/
CIFAR-10	Recall	0.945	ResNet50（92.7%）	0.018	/
ImageNet subset	Precision	0.853	EfficientNet（81.2%）	0.041	/
ImageNet subset	Recall	0.859	EfficientNet（80.9%）	0.05	/
MNIST	Precision	0.994	Hierarchical CNN（98.6%）	0.008	/
MNIST	Recall	0.996	Hierarchical CNN（98.8%）	0.008	/
Dataset	Confused categories	The misjudgment rate	Comparative algorithms	Comparison algorithm misjudgment rate	The reduction in false positive rate
CIFAR-10	Airplane - Bird	1.2%-1.5%	VGG16	3.8%-4.2%	> 65%
CIFAR-10	Car Truck	1.2%-1.5%	VGG16	3.8%-4.2%	> 65%
ImageNet subset	Dog wolf, cat tiger	0.032	RobustCNN	0.058	0.448

performed better in terms of MAE and RMSE (Figure 6 (b)). The overall deviation of the box line indicates that the error value was smaller and the fluctuation was narrow. For example, on the MNIST dataset, the MAE of the proposed method was 0.021, which was 45.9% lower than VGG16 (0.039) and 31.0% lower than ResNet50 (0.030); the RMSE was 0.053, which was 38.8% lower than VGG16 (0.087) and 27.4% lower than ResNet50 (0.073). This data showed that the Research Algorithm had smaller deviations between predicted and true values and higher stability in image recognition across different datasets.

The performance of research method was compared with existing advanced image recognition models, including traditional CNN VGG, image classification models EfficientNet, and GoogLeNet. The F1 value and loss curve of the models were used as evaluation indicators, and the average test results of different datasets are denoted in Figure 7. The Loss Function (LF) is called Cross Entropy Loss, which is used to measure the difference between the predicted values of the model and the true labels. The smaller the value, the better the fitting effect of the model. In Figure 7 (a), the F1 value curve of the research method was located at the top of the coordinate axis, reaching a maximum value of 92.39%, which was 14.56% higher than the lowest value of 78.24% in the EfficientNet model. The F1 values of the other two image recognition models were within the range of 80-90%. Figure 7 (a) shows different models' LF curves. The proposed method's LF curve converges to the minimum, with a steady decline and the fastest convergence. LF reflects prediction-true value consistency; smaller LF means better fitting, so the method has better comprehensive performance.

Table 2 shows multi-dataset evaluation of different methods, where the proposed algorithm performed better. On CIFAR-10, its accuracy (94.1%) and recall (94.5%) were both 1.8% higher than ResNet50. On ImageNet subset, accuracy (85.3%) and recall (85.9%) were 4.1% and 5.0% higher than EfficientNet, respectively. On MNIST, accuracy (99.4%) and recall (99.6%) were each 0.8% higher than Hierarchical CNN.

An algorithm with a time complexity of $O(H\times W\times C\times K^2+D^3+N\times d\times L)$ was proposed, which includes multi-scale feature extraction, PCA dimensionality reduction, and hierarchical discrimination. Compared to LDA, although multi-scale convolution increased complexity by 15%, PCA dimensionality reduction reduced d by 60% and reduced training time for millions of samples by 22%; Compared to ResNet50, due to the

lack of deep stacking, the complexity was reduced by 35% and the training time for millions of samples was reduced by 40 minutes.

In the ImageNet full dataset (1.5M samples) test, the algorithm dynamically merged redundant nodes, occupied 32GB → 24GB of memory, supported single GPU training, and reduced resource requirements by 50% compared to VGG16. When the sample size ranged from 100000 to 1 million, the algorithm accuracy only decayed by 1.2%, far better than SVM's 4.5% decay, demonstrating the advantage of large-scale data scalability.

In the hyperparameter sensitivity experiment of the CIFAR-10 dataset, the impact of key parameters on accuracy was controllable: the accuracy was optimal (94.3%) when the depth of the hierarchy L was 5, and the fluctuation of ± 1 was less than 1.5%; After PCA retained variance>95%, the accuracy remained stable with fluctuations<0.3%; The SVM penalty parameter C=1.0 had the highest accuracy, and overfitting was greater than 1.0 but the variation was less than 2%. Under the adjustment of key parameters by ± 20%, the accuracy fluctuation was less than 2%, and the convergence cycle change was less than 5 cycles. The model has strong stability and is suitable for multiple data scenarios.

4 Discussion and conclusion

4.1 Discussion

The study's HDA-based image recognition algorithm boosts performance via multi-scale feature extraction, PCA, attention mechanism, and dynamic hierarchical adjustment; its advantages and innovation are clarified by comparing with related research. In accuracy: it hits 94.3% (CIFAR-10), 85.6% (ImageNet subset), 99.5% (MNIST) — higher than comparison algorithms. Compared to Yang et al. [6] (lacks dynamic adjustment), its multi-scale feature extraction (1×1,3×3,5×5 CKs) captures richer features, plus dynamic adjustment suits complex data; Compared to Su et al. [7] (no attention), it uses attention to focus on key features and PCA to reduce redundancy, enhancing discriminability. In robustness: under Gaussian/salt-and-pepper noise, performance degradation is smaller. E.g., Gaussian noise variance 0.1: its accuracy 95.3% vs ResNet50's 92.1%, Hierarchical CNN's 91.5%. This addresses gaps of Zhang et al. [5] (no interference robustness verification) and Hu et al. [9] (no multi-scale fusion optimization), highlighting practical value in complex scenarios.

Analogous backstepping and output feedback control are used to extract multi-scale features in response to its hierarchical design. Through hierarchical discrimination and subdivision of categories, the recognition accuracy is improved from 88.5% to 94.3%; Analogous to nonlinear optimal control and pursuing the optimal goal, through multi module collaborative optimization, the F1 value reaches 92.39% and the MAE/RMSE is lower than the comparison algorithm.

4.2 Conclusion

For salt and pepper noise, the accuracy advantage of the proposed method was more obvious under the same intensity. The research method made the model focus more on the key areas of the image, reducing the impact of noise on non key areas. However, there are two limitations to the algorithm in this article: firstly, the computational complexity is relatively high on large-scale datasets, mainly due to the need to train SVM discriminators for each node, and subsequent optimization through parallel training or lightweight SVM; The second issue is insufficient real-time performance, as the dynamic hierarchy adjustment process increases training time by about 5%, making it temporarily unsuitable for high-speed real-time recognition scenarios. The consideration of image recognition in multiple scenarios is not sufficient, so future research will apply this algorithm to a wider range of practical scenarios, such as video image recognition, infrared image recognition, etc., to further verify its effectiveness and applicability. The studied HDA image recognition algorithm, with advantages of multi-scale feature extraction, noise robustness and dynamic adjustment, can be extended to multiple fields: real-time recognition (0.03s single-image inference, high accuracy in complex scenes like distinguishing 3 target types in mall security); medical imaging diagnosis (captures lesion details and global structure to boost accuracy, reduce misdiagnosis); video stream recognition (realizes target classification/tracking, optimizes traffic flow statistics via keyframe extraction and hierarchical discrimination).

Infrared image recognition, using thermal radiation without visible light, serves nighttime security and power fault detection. Traditional algorithms, hindered by thermal noise and blurred edges, have <85% accuracy in power inspection thermal anomaly detection. This algorithm uses multi-scale features and noise robustness, with FLIR ADAS dataset and Faster R-CNN as baseline, aiming to enhance accuracy from 82% to over 90%. It will also pilot substation night inspections with manufacturers, integrating into infrared cameras. Video image recognition for traffic flow and anomaly monitoring faces frame blurring and occlusion, with traditional algorithms having >10% vehicle counting errors. This algorithm uses dynamic adjustment and attention mechanism, with UCF101 dataset and 3D CNN as comparison, aiming to boost action recognition accuracy from 88% to 95%. It will also pilot on main roads with smart city platforms.

References

[1] Yuhan Feng. Tea disease recognition technology based on a deep convolutional neural network feature learning method. International Journal of Computing Science and Mathematics, 19(1):15-27, 2024. https://doi.org/10.1504/IJCSM.2024.136820

[2] Kotha Manohar, and E. Logashanmugam. ADMRF: Elucidation of deep feature extraction and adaptive deep Markov random fields with improved heuristic algorithm for speech emotion recognition.

International Journal of Speech Technology, 27(3):569-597, 2024. https://doi.org/10.1007/s10772-024-10115-7

[3] Degang Jiang, Xiuyong Shi, Yunfang Liang, and Hua Liu. Feature extraction technique based on Shapley value method and improved mRMR algorithm. Measurement, 237(1):1-9, 2024. https://doi.org/10.1016/j.measurement.2024.115190

[4] Nduvho Mulaudzi, Lehlogonolo Trucy Rasealoka, Gudani Honoured Maano, Tlabo Client Mohlapi, Pasca Makgwale Moshidi, and Nkgetheng Nonyane Mohlabe. HPTLC profiling, quality control and FTIR coupled with chemometrics analysis for securidaca longipenduculata fresen. British Journal of Mathematics & Computer Science, 11(2):191-199, 2024. https://doi.org/10.22036/ABCR.2024.425154.2002

[5] Jie Zhang, Xiao Yu, Ran Yang, Bingqing Zheng, Yongqing Zhang, and Fang Zhang. Quality evaluation of Lonicerae Japonicae Flos from different origins based on High-Performance Liquid Chromatography (HPLC) fingerprinting and multicomponent quantitative analysis combined with chemical pattern recognition. Phytochemical Analysis, 35(4):647-663, 2024. https://doi.org/10.1002/pca.3319

[6] Zhaozhao Yang, Yuhai Yu, Yongdong Huang, and Jiana Meng. Innovative approaches in image processing: enhancing feature extraction and recognition capabilities. The Visual Computer, 41(10):7671-7685, 2025. https://doi.org/10.1007/s00371-025-03830-y

[7] Shuzhi Su, Kaiyu Zhang, Yanmin Zhu, Maoyan Zhang, and Shexiang Jiang. Similarity-sequenced multi-view discriminant feature extraction for image recognition. Journal of Modern Optics, 70(7/9):503-516, 2023. https://doi.org/10.1080/09500340.2023.2273552

[8] Radmila Janković Babić. A comparison of methods for image classification of cultural heritage using transfer learning for feature extraction. Neural Computing & Applications, 36(20):11699-11709, 2024. https://doi.org/10.1007/s00521-023-08764-x

[9] Sheng Hu, Xinmin Hu, Jingqi Li, Yiting He, Haixin Qin, Shasha Li, Min Liu, Cong Liu, Can Zhao, and Wei Chen. Enhancing vibration detection in Φ-OTDR through image coding and deep learning-driven feature recognition. IEEE Sensors Journal, 24(22):38344-38351, 2024, https://doi.org/10.1109/JSEN.2024.3469232

[10] Zhengyan Wu, Jilin He, Chao Huang, and Renshan Yao. A novel feature fusion-based stratum image recognition method for drilling rig. Earth Science Informatics, 16(4):4293-4311, 2023. https://doi.org/10.1007/s12145-023-01132-2

[11] Tao Zhang, Yongqi Chen, Zhongxing Sun, Liping Huang, Qinge Dai, and Qian Shen. Fault diagnosis of rolling bearing based on hierarchical discrete entropy and semi-supervised local Fisher discriminant analysis. Journal of Vibroengineering, 26(6):1317-1335, 2024. https://doi.org/10.21595/jve.2024.23945

[12] Janine Colling, Magdalena Muller, Elizabeth Joubert, and Federico Marini. Investigating partial least squares discriminant analysis and hierarchical modelling of short-wave infrared hyperspectral imaging data to distinguish production area and quality of rooibos (Aspalathus linearis). Journal of Near Infrared Spectroscopy, 31(3):158-167, 2023. https://doi.org/10.1177/09670335231174328

[13] Kei Hirose, Kanta Miura, and Atori Koie. Hierarchical clustered multiclass discriminant analysis via cross-validation. Computational Statistics and Data Analysis, 178(1):107613.1-107613.2, 2023. https://doi.org/10.1016/J.CSDA.2022.107613

[14] Zonghan Tian, Siwei Tao, Ling Bai, Yueshu Xu, Xu Liu, and Cuifang Kuang. A multimodal image feature extraction method for x-ray grating phase contrast computed tomography based on monogenic signal. Review of Scientific Instruments, 94(12):25106.1-125106.9, 2023. https://doi.org/10.1063/5.0170247

[15] S. Subathradevi, T. Preethiya, D. Santhi, and G. R. Hemalakshmi. Facial emotion recognition for feature extraction and ensemble learning using hierarchical cascade regression neural networks and random forest. Journal of Circuits, Systems & Computers, 33(18):1-32, 2024. https://doi.org/10.1142/S0218126625500112

[16] Hanshan Li, and Xiaoqian Zhang. A measurement method of projectile explosion position and explosion image recognition algorithm based on PSPNet and swin transformer fusion. IEEE Sensors Journal, 25(3):4715-4726, 2025. https://doi.org/10.1109/JSEN.2024.3512774

[17] Tri Le, Nham Huynh-Duc, Chung Thai Nguyen, and Minh-Triet Tran. Motion embedded images: An approach to capture spatial and temporal features for action recognition. Informatica, 47(3):327-328, 2023. https://doi.org/10.31449/inf.v47i3.4755

[18] Haiyan Xun. Research on automatic recognition technology of library books based on image processing. Informatica, 48(5):29-40, 2024. https://doi.org/10.31449/inf.v48i5.5345

[19] Zhenkang Wang, Nan Xia, Song Hua, Jiale Liang, Xiankai Ji, Ziyu Wang, and Jiechen Wang. Hierarchical recognition for urban villages fusing multiview feature information. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18(1):3344-3355, 2025. https://doi.org/10.1109/JSTARS.2024.3522662

[20] Dinh Phamtoan, and Tai Vovan. The fuzzy cluster analysis for interval value using genetic algorithm and its application in image recognition. Computational Statistics, 38(1):25-51, 2023. https://doi.org/10.1007/s00180-022-01215-6

[21] Mohamad Hasanvand, Mahdi Nooshyar, Elaheh Moharamkhani, and Arezu Selyari. Machine learning methodology for identifying vehicles using image processing. Artificial Intelligence and

Applications, 1(3):170-178, 2023. https://doi.org/10.47852/bonviewAIA3202833

[22] Yutong Sun. High-resolution image processing and entity recognition algorithm based on artificial intelligence. Journal of Intelligent Systems, 33(1):73-81, 2024. https://doi.org/10.1515/jisys-2023-0245