

A study on rapid identification of medical vector organisms based on improved Transformer

Yan Zhou^{1,†}, Jie Zhong^{1,*†}, Xin Fang^{1,†}, Juan Huang^{1,†} and Lingyu Yan^{2,*†}

¹ Hubei International Travel Healthcare Center (Wuhan Customs Port Outpatient Department), Wuhan

² School of Computer Science, Hubei University of Technology, Wuhan.

Abstract

The identification of port and imported medical vector species is an extremely important issue for customs work, which is related to public health safety. The traditional methods of vector identification and identification mainly rely on manual labor. In order to reduce personnel workload, this paper proposes a new lightweight image classification network model for vector biometric recognition. The backbone network of the model is a network architecture based on the combination of convolutional neural networks and Transformers, a convolutional neural network model constructed with smaller parameters. Firstly, a fusion module is introduced to replace the depthwise separable convolution in shuffleNet with a regular conv3 × 3, and h-swish is used to replace the ReLU activation function, which reduces costs while better reducing the number of model parameters. Then, this article uses a network architecture that combines convolution and improved Transformer to input images in sequence into the Transformer to obtain global feature information, ensuring that the network model can extract local and global features of the image. At the same time, the Tranformer is improved to reduce the number of model parameters..

Keywords

Biometric identification, Image classification, Convolutional neural networks, Transformer

1. Introduction

Accurate and timely identification of medical vector organisms (hereinafter referred to as "vector organisms") at ports of entry and exit is an important issue in the work of the Customs and Excise Department.

Among them, flies, mosquitoes and cockroaches are important parts of vector organisms. Therefore, it is a meaningful task to conduct research on the identification of

CITI'2024: 2nd International Workshop on Computer Information Technologies in Industry 4.0, June 12–14, 2024, Ternopil, Ukraine

* Corresponding author.

† These authors contributed equally.

✉ zhouyan@163.com (Y. Zhou); zhongjie@163.com (J. Zhong); fangxin@163.com (X. Fang); huangjuan@163.com (J. Huang); yanranyaya2024@163.com (L. Yan)

ORCID: 0000-0003-3412-1639 (Y. Zhou); 0000-0002-5385-5761 (J. Zhong); 0000-0002-9421-8566 (X. Fang); 0009-0006-6171-7293 (J. Huang); 0000-0002-8434-5473 (L. Yan)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

vector organisms at ports of entry and exit. The traditional method of vector organism identification [1-2] mainly relies on the experience accumulated by the staff in the actual identification work, which requires the staff to match the physical and morphological characteristics of the vector organisms distinguished by the naked eye (or with the help of microscopic magnification equipment) with the descriptive text in the books on species classification, which is very much dependent on the practical experience of the staff. However, due to the large number of vector organisms, some vector organisms have similar apparent characteristics, and the staff are not all experts with rich practical experience in identification, relying only on the staff's experience to determine the types of vector organisms, especially in the initial screening stage of the sample size is large, there is a slower speed, poorer real-time, and subjective so that the impact of the fresh sample morphology identification of the detection of the problem of the initial screening, making this method is very time-consuming and labor-intensive.

At present, applying deep learning methods to vector organisms rapid identification is the main trend of intelligence [3-5], but most deep learning models are characterized by many model parameters and large model volume, which require high memory and computational resources and consume too much time for inference and computation, and are difficult to be applied to small mobile devices, which restricts the popularization and application of network models. The traditional solution is to deploy the model on the server, perform the computation on the server, and then feedback the results to the cell phone, but this method is too dependent on the network, and the recognition efficiency will be greatly affected by the network latency and the absence of the network, etc. Therefore, the lightweight convolutional model has gradually emerged. Therefore gradually appeared lightweight convolutional neural network^[6-8], lightweight network compared to the mainstream CNN [9] network, in the recognition effect of a small reduction in the case of a substantial reduction in the number of parameters of the model and the amount of computation. It can be applied to devices with limited computational resources such as mobile devices and edge devices, which effectively improves the practicality of deep learning in classification, recognition and other aspects. Therefore, the vector biometric identification method based on lightweight convolutional neural network has high research value.

MobileNet ^[10] is an efficient lightweight network architecture aimed at providing higher performance to enable mobile and embedded devices to better meet the needs of various visual applications. This network architecture can fully utilize limited computing power and storage space. The key advantage of MobileNet [11] is its ability to use depthwise separable convolutions. By splitting the standard convolution operation into smaller units, deep convolution operation and point by point convolution operation are obtained. By adopting this approach, we can significantly reduce the computational cost and parameter requirements of convolution operations, enabling us to better utilize limited computing power and storage space, while also significantly reducing the complexity of the entire network. This also enables deep convolutional neural networks to adapt to the limitations of mobile and embedded devices. MobileNetv2 [12] has optimized the depthwise separable convolution module, which is more advanced and efficient in design and functionality than the original ones. MobileNetv2 proposes a linear bottleneck

reverse residual module, which has higher accuracy and faster processing speed compared to MobileNetv1.

ShuffleNet [13] is a lightweight network model with extremely high computational efficiency, which can effectively support mobile devices with weaker computing power. ShuffleNet has two important breakthroughs in its network structure. Firstly, it adopts point by point grouping convolution technology; Secondly, it adopts channel random mixing technology. By grouping convolution, feature maps can be divided equally according to channel dimensions, and each group of features is extracted by a convolution kernel, thereby achieving parallel computation of multiple convolutions, greatly improving computational efficiency. By randomly mixing feature maps from different groups, we can achieve convolutional optimization of the entire network. This method is called channel random shuffling.

In this paper, we propose a new lightweight image classification network model, the backbone network is based on a network architecture combining convolutional neural network and Transformer, a convolutional neural network model built with smaller parameters. Firstly, the fusion module is introduced to replace the depth separable convolution in shuffleNet with a common conv3×3 and the ReLU activation function [14] is replaced with h-swish to better reduce the number of parameters of the model while reducing the cost. Then this paper uses a network architecture that combines the convolution with an improved Transformer to input the image as a sequence into the Transformer to obtain global feature information, ensuring that the network model is able to extract both local and global features of the image, and at the same time improving the Transformer to reduce the number of parameters of the model.

2. Methods

In this paper, we use three sizes of network models with exactly the same network structure. the number of Transformer layers is [2, 4, 3], the dimensions are [64, 80, 96], and finally a group point convolution is performed to reduce the feature map. The overall architecture in the model is shown in Fig. 1. Only the input dimensions of Transformer are different ShuffleViT-xxs input token is [64,80,96], shufflenViT-xs input token is [96, 120, 144]. shuffleViT-S input token is [144, 192, 240]. The network structure is shown in Table 1.

2.1. Lightweight Module

ShuffleNet network is a lightweight convolutional neural network. The proposed idea lies in the fact that when the computational conditions are limited, the number of channels of the feature map in the convolutional neural network operation is also limited. Therefore ShuffleNet proposes two techniques to increase the number of channels of feature maps without significantly increasing the computational effort of the network: point-by-point group convolution and bottleneck-like structure.

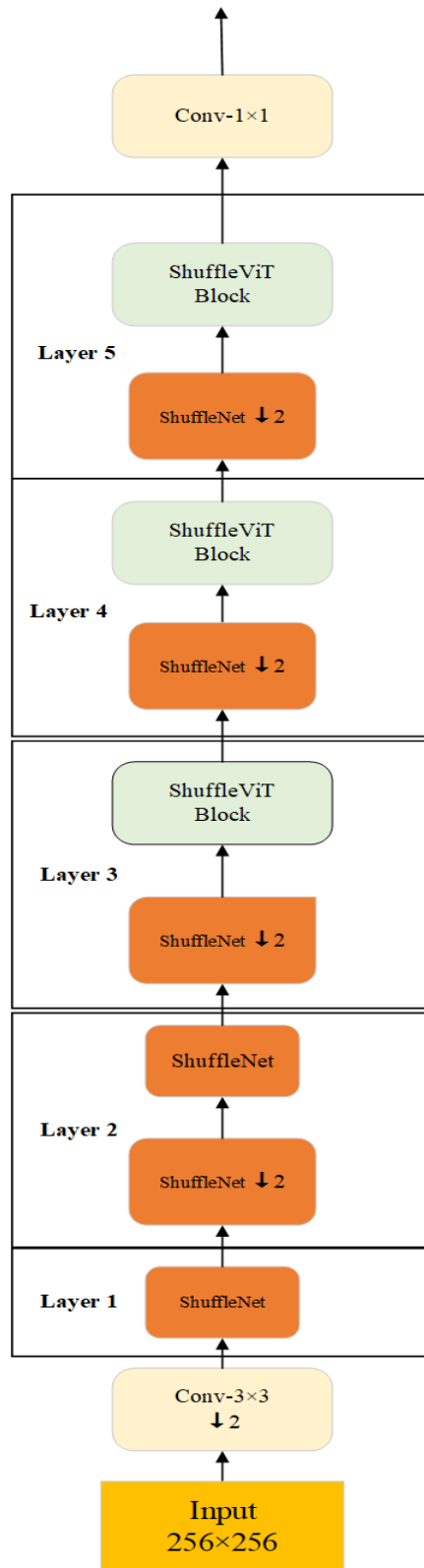


Figure 1: Lightweight image classification network model diagram

Table 1
ShuffleViT network structure table

| Layer | Output | Output stride | Repeat | Output channels | | |
|-----------------------|---------|---------------|--------|-----------------|-----------|------------|
| | | | | XXS | XS | S |
| Image | 256×256 | 1 | | | | |
| Conv-3×3, ↓2 | | | 1 | 16 | 16 | 16 |
| SV2 | 128×128 | 2 | 1 | 16 | 32 | 32 |
| SV2, ↓2 | | | 1 | 24 | 48 | 64 |
| SV2 | 64×64 | 4 | 2 | 24 | 48 | 64 |
| SV2, ↓2 | | | 1 | 48 | 64 | 96 |
| ShuffleViT block(L=2) | 32×32 | 8 | 1 | 48(d=64) | 64(d=96) | 96(d=144) |
| SV2, ↓2 | | | 1 | 64 | 80 | 128 |
| ShuffleViT block(L=4) | 16×16 | 16 | 1 | 64(d=80) | 80(d=120) | 128(d=192) |
| SV2, ↓2 | | | 1 | 80 | 96 | 160 |
| ShuffleViT block(L=3) | 8×8 | 32 | 1 | 80(d=96) | 96(d=144) | 160(d=240) |
| Conv-1×1 | | | 1 | 320 | 384 | 640 |
| Global pool | | | 1 | | | |
| Linear | 1×1 | 256 | 1 | 1000 | 1000 | 1000 |
| Parameters | | | | 1.22M | 2.21M | 5.1M |

Convolutional neural networks are divided into multiple modules for feature extraction, e.g., Xception and ResNeXt deep separable convolution with swarm convolution to form a neural network with better performance, which achieves a balance between performance and computational consumption. However, a large number of point-by-point convolutions are used in these networks, which can lead to high computational complexity overhead. The use of point-by-point convolution in lightweight convolutional neural networks fills the feature channels with constraints, which can drastically reduce the performance of the network model. Point-by-point group convolution uses group convolution to make the convolution operation on the corresponding channel, and group convolution reduces the computational loss of the model. ShuffleNet is an improvement on the residual structure: it first does a 1×1 group convolution, followed by a channel mixing operation, then a 3×3 depth-separable convolution, and finally a group convolutional transformation of the output size to be directly summed with the input. When the step size is not 1, it is necessary to do a convolution operation to transform the size of the input and then splice it with the output. Subsequently, there was a V2 version, ShuffleNetV2, which proposed that an efficient network structure should have the same convolution width, smaller cost of convolution, reduced minutiae operations and as lean a network model as possible. Since the point-by-point group convolution and bottleneck structure proposed in ShuffleNetV1 increased the operation cost too much for a lightweight network model, ShuffleNetV2 replaced the group convolution with ordinary convolution, when the group convolution is replaced with ordinary convolution, the channel shuffling effect behind the model will disappear, so the channel shuffling is put into the input and output splicing after the shuffling. Improved ShuffleNet structure schematic shown in Figure 2. In contrast

the ReLU6 activation function is limiting the ReLU function to a maximum output value of 6, which allows for excellent numerical resolution on low precision devices. When the ReLU function is not limited to a maximum value, it will not be able to accurately characterize a large range of values on low precision devices, resulting in a loss of accuracy. The use of the h-swish activation function reduces the number of parameters compared to ReLU6 and h-swish.

In this paper, the convolution part is further optimized on the basis of ShuffleNetV2 by replacing the beginning 1×1 convolution with 3×3 depth-separable convolution structure in the branch to 3×3 ordinary convolution. Then the activation function ReLU is replaced with h-swish activation function:

$$h - swish(x) = x \cdot \frac{ReLU6(x+3)}{6}, \quad (1)$$

$$ReLU6(x) = \min(\max(x, 0), 6), \quad (2)$$

In contrast the ReLU6 activation function is limiting the ReLU function to a maximum output value of 6, which allows for excellent numerical resolution on low precision devices. When the ReLU function is not limited to a maximum value, it will not be able to accurately characterize a large range of values on low precision devices, resulting in a loss of accuracy. The use of the h-swish activation function reduces the number of parameters compared to ReLU6 and h-swish.

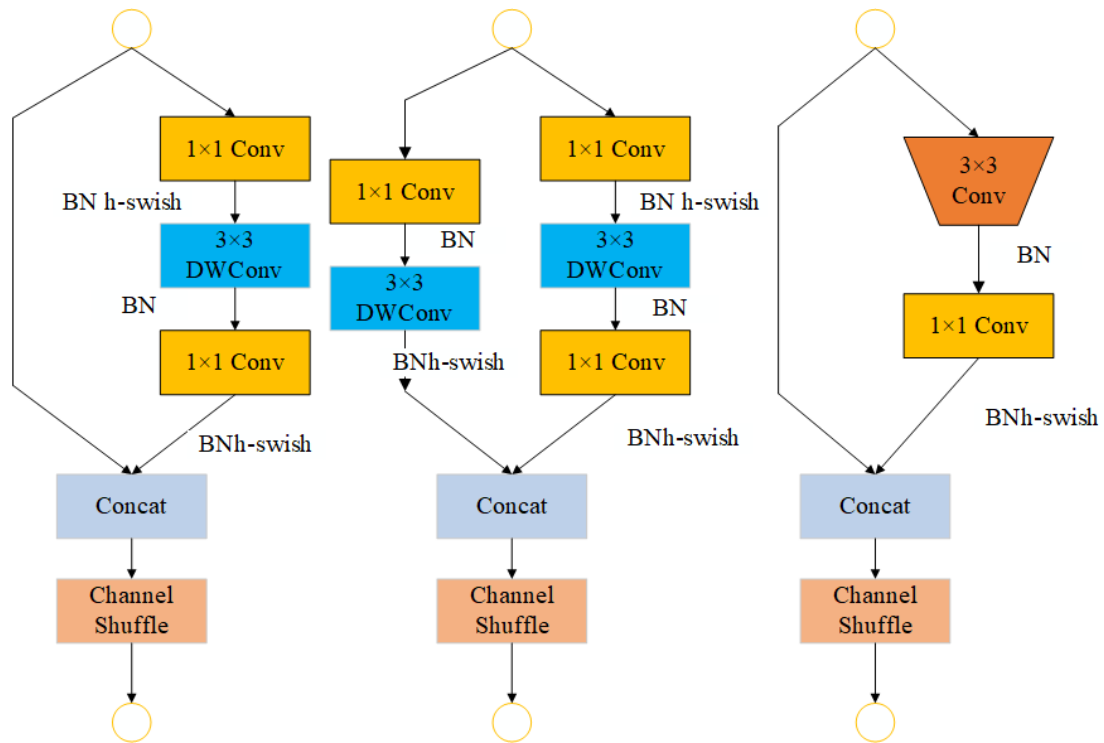


Figure 2: Schematic diagram of the improved ShuffleNet structure

2.2. Attention Module

To solve the problem of insufficient spatial generalization bias, this paper uses a convolutional neural network combined with Transformer. Convolutional neural network has excellent spatial imputation bias ability, because the picture has a strong two-dimensional local structure, spatial neighboring pixels are often highly correlated. Convolutional neural network can forcefully capture the spatial inductive bias by using local sensory fields, shared weights and spatial sampling, which to some extent realizes the non-deformation of the transformation in each of the pictures. Whereas Transformer is more capable of capturing global features using a multi-head attention mechanism.

For the part of the convolutional neural network combined with the Transformer we first pass the feature map through a convolution operation with a convolution kernel size of $n \times n$ to obtain a local feature modeling of the model, and then adjust the number of channels of the feature map through a convolution layer with a convolution kernel size of 1×1 . The feature map is then tiled by performing a tiling operation on the already tuned feature map, similar to the operation of Vision Transformer, the tiling operation in this paper introduces channel blending to reduce the computation. The block size of the feature map is first divided, e.g., when the block size is 2×2 , i.e., each block consists of 2×2 pixels. Since a convolution operation with a convolution kernel size of $n \times n$ has already been performed on the input feature map to obtain local feature modeling, operating each token with other tokens while doing attention operation for global feature modeling will result in wastage of arithmetic, and the obtained local feature modeling has already been obtained by the previous convolution operation. In the operation on token. In this paper, the data is first transformed to make it into the data format needed for self-attention. For the common attention mechanism is to directly spread the two dimensions of height and width to make it into a sequence of tokens, whereas in this paper, we choose to generalize the corresponding pixels between each block to perform the attention operation on the tokens in the same position. It is divided into different sequences by the pixel position of each block. As shown in Fig. 3.

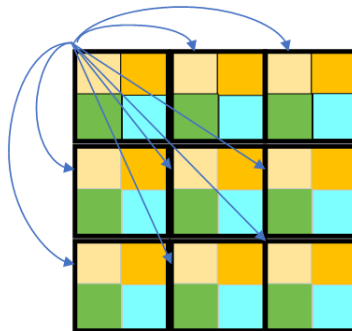


Figure 3: Schematic diagram of ShuffleViT image slice structure

Each pixel operates attentively only with other pixels in the same position. The number of sequences formed is the pixel value of each block. Assuming that the width, height and number of channels of the feature map are H , W and C respectively, then the number of

sequences in each sequence is $W \times H/P$ and P is the size of the block. Transformer operation is performed on the model after obtaining the sequences.

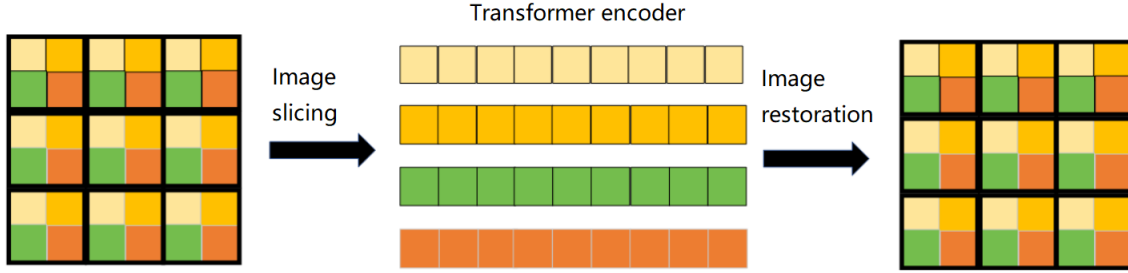


Figure 4: Schematic diagram of ShuffleViT Block picture flow

The amount of long attention is calculated as:

$$[q, k, v] = zU_{qkv}, \quad (3)$$

where q , k , and v denote the queue, key, and key value in the attention mechanism, respectively, and denote the amount of computation used to compute q , k , and v .

$$A = softmax\left(\frac{qk^T}{\sqrt{D_h}}\right), \quad (4)$$

A is the amount of matrix computation for the attention single key value and queue, where $A \in R^{N \times N}$.

$$SA(z) = Av\phi, \quad (5)$$

$SA(z)$ denotes the amount of matrix computation computed by the attention mechanism at one time.

$$MSA(z) = [SA_1(z); SA_2(z); \dots; SA_k(z)]U_{msa}, \quad (6)$$

$MSA(z)$ denotes the amount of multi-attention mechanism computation. included among these $U_{msa} \in R^{k \cdot D_h \times D}$. It can be concluded that the arithmetic power used by the multi-head attention in performing the operation occupies the majority of the entire coding module, which is calculated to be one-fourth of the computational effort of the multi-head attention mechanism of the Improved Transformer as compared to the computational effort of the ViT. Therefore dividing the input feature map sequence will lead to a significant reduction in the computation of the model. After completing the attention computation a stacking operation is performed on the output results, where the output feature values are recovered from the one-dimensional feature map by feature changes according to their original positions. The number of channels is then adjusted to the original size by a convolutional layer with a convolutional kernel size of 1×1 . Then the original input feature map is spliced with the original input feature map using a shortcut branch.

3. Experiments

In this paper, we use the dataset which contains 3 species of cockroaches, 6 species of mosquitoes and 20 species of flies, with a total of 338 image datasets, and the background of the images is a single, unobstructed environment in the laboratory, where the various categories of vector organisms are displayed in the form of pin-inserted specimens. The dataset suffers from uneven distribution of the number of images in each category and differences in the angle of vector organism display, with the number of sample images in each category ranging from 9 to 53. Fig 5 shows the main categories of the dataset displayed.



Figure 5: Presentation of the main categories of the dataset

Table 2

Accuracy on the recognition task using ResNet-18, ResNet-34, ResNet-50, MobileNetV1, MobileNetV2, MobileNetV3, ShuffleNetV2 models vs. our model

| Model | MaxAcc(%) | Acc(%) | #Params |
|-------------------|-----------|--------|---------|
| ResNet-18 | 81.27% | 80.27% | 11.18M |
| ResNet-34 | 81.57% | 81.93% | 21.29M |
| ResNet-50 | 81.60% | 81.20% | 23.53M |
| MobileNetV1 | 78.57% | 78.36% | 4.22M |
| MobileNetV2/1.0 | 79.36% | 79.07% | 2.24M |
| MobileNetV2/2.0 | 81.02% | 80.89% | 8.72M |
| MobileNetV3-Large | 80.28% | 80.06% | 3.88M |
| MobileNetV3-small | 77.15% | 77.12% | 1.84M |

| | | | |
|-------------------|---------------|--------|-------|
| ShuffleNet_V2_0.5 | 76.71% | 76.32% | 1.42M |
| ShuffleNet_V2_1.0 | 79.24% | 79.22% | 2.28M |
| ShuffleViT | 82.60% | 82.12% | 2.21M |

The results in Table 2 show that ShuffleViT outperforms other convolutional neural networks. For the ResNet series of convolutional neural networks, ShuffleViT outperforms ResNet-18 by 1.33%, ResNet-34 by 1.03%, and ResNet-50 by 1.00% in the accuracy of the image classification task. For the mobileNet family of convolutional neural networks, ShuffleViT outperforms MobileNetV1 by 4.03%, MobileNetV2/1.0 by 3.24%, MobileNetV2/2.0 by 1.58%, MobileNetV3-Large by 2.37% higher than MobileNetV3-Smalle, and 5.45% higher than MobileNetV3-Smalle. For the shuffleNet family of convolutional neural networks, ShuffleViT is 5.89% more accurate than ShuffleNet_V2_0.5 and 3.36% more accurate than ShuffleNet_V2_1.0 for image classification tasks.

4. Conclusion

The improvement of image classification task performance requires a very large dataset to support. When the dataset is too small, overfitting may occur, and a larger network structure will bring better classification accuracy. With the deepening of neural networks, image classification tasks have achieved astonishing accuracy. However, improving performance requires state-of-the-art network structures and larger computing resources. In the current era of widespread use of intelligent mobile devices, the enormous amount of computing has far exceeded the capabilities of many mobile devices and embedded application devices. How to successfully deploy image classification tasks on mobile devices has become a key research issue. This article proposes a new lightweight image classification network model for vector recognition. Firstly, a fusion module is introduced to replace the depthwise separable convolution in shuffleNet with a regular conv3 × 3, and h-swish is used to replace the ReLU activation function, which reduces costs while better reducing the number of model parameters. Then, this article uses a network architecture that combines convolution and improved Transformer to input images in sequence into the Transformer to obtain global feature information, ensuring that the network model can extract local and global features of the image. At the same time, the Tranformer is improved to reduce the number of model parameters.

Reference

- [1] Tاتفeng Y M, Usuanlele M U, Orukpe A, et al. Mechanical transmission of pathogenic organisms: the role of cockroaches[J]. Journal of vector borne diseases, 2005, 42(4): 129.
- [2] Nicholson W L, Allen K E, McQuiston J H, et al. The increasing recognition of rickettsial pathogens in dogs and people[J]. Trends in parasitology, 2010, 26(4): 205-212.
- [3] Gourisaria M K, Das S, Sharma R, et al. A deep learning model for malaria disease detection and analysis using deep convolutional neural networks[J]. International Journal of Emerging Technologies, 2020, 11(2): 699-704.

- [4] Nouredine S, Zineeddine B, Toumi A, et al. A new predictive medical approach based on data mining and Symbiotic Organisms Search algorithm[J]. *International Journal of Computers and Applications*, 2022, 44(5): 465-479.
- [5] Rani P, Kotwal S, Manhas J, et al. Machine learning and deep learning based computational approaches in automatic microorganisms image recognition: methodologies, challenges, and developments[J]. *Archives of Computational Methods in Engineering*, 2022, 29(3): 1801-1837.
- [6] Liu F, Xu H, Qi M, et al. Depth-wise separable convolution attention module for garbage image classification[J]. *Sustainability*, 2022, 14(5): 3099.
- [7] Huang T, Chen J, Jiang L. DS-UNeXt: depthwise separable convolution network with large convolutional kernel for medical image segmentation[J]. *Signal, Image and Video Processing*, 2023, 17(5): 1775-1783.
- [8] Xia Q, Dong S, Peng T. An Abnormal Traffic Detection Method for IoT Devices Based on Federated Learning and Depthwise Separable Convolutional Neural Networks[C]//2022 IEEE International Performance, Computing, and Communications Conference (IPCCC). IEEE, 2022: 352-359.
- [9] Jiang K, Zhang C, Wei B, Li Z, Kochan O. Fault diagnosis of RV reducer based on denoising time–frequency attention neural network [J]. *Expert Systems with Applications*, 2024, 238: 121762.
- [10] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.
- [11] Liu X, Qi P, Siarry P, et al. Mining security assessment in an underground environment using a novel face recognition method with improved multiscale neural network [J]. *Alexandria Engineering Journal*, 2023, 80: 217-228.
- [12] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.
- [13] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848-6856.
- [14] Xu X, Przystupa K, Kochan O. Social Recommendation Algorithm Based on Self-Supervised Hypergraph Attention [J]. *Electronics*. 2023, 12(4):906.