Front. Struct. Civ. Eng. 2022, 16(5): 564–575
https://doi.org/10.1007/s11709-022-0829-x

RESEARCH ARTICLE

Efficient Identification of water conveyance tunnels siltation based on ensemble deep learning

Xinbin WU[a*], Junjie LI[a,b], Linlin WANG[a]

[a] *Faculty of Infrastructure Engineering, Dalian University of Technology, Dalian 116024, China*
[b] *College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China*
[*] *Corresponding author. E-mail: wxb960110@mail.dlut.edu.cn*

ABSTRACT The inspection of water conveyance tunnels plays an important role in water diversion projects. Siltation is an essential factor threatening the safety of water conveyance tunnels. Accurate and efficient identification of such siltation can reduce risks and enhance safety and reliability of these projects. The remotely operated vehicle (ROV) can detect such siltation. However, it needs to improve its intelligent recognition of image data it obtains. This paper introduces the idea of ensemble deep learning. Based on the VGG16 network, a compact convolutional neural network (CNN) is designed as a primary learner, called Silt-net, which is used to identify the siltation images. At the same time, the fully-connected network is applied as the meta-learner, and stacking ensemble learning is combined with the outputs of the primary classifiers to obtain satisfactory classification results. Finally, several evaluation metrics are used to measure the performance of the proposed method. The experimental results on the siltation dataset show that the classification accuracy of the proposed method reaches 97.2%, which is far better than the accuracy of other classifiers. Furthermore, the proposed method can weigh the accuracy and model complexity on a platform with limited computing resources.

KEYWORDS water conveyance tunnels, siltation images, remotely operated vehicles, deep learning, ensemble learning, computer vision

1 Introduction

Water conveyance tunnels play a significant role in long-distance water diversion projects. During long-term water delivery, the material carried by water is prone to deposit sediment. Further, with tunnels' service life increasing, defects and aging affect water quality and water delivery safety. If damage is not detected and repaired in time, serious accidents may occur, such as rupture, collapse and settlement of the pipeline. This threatens urban stability and development and presents a serious social hazard [1]. However, water conveyance tunnels are usually deeply buried and undulate, resulting in complicated hydraulic conditions and increasing the difficulty of route inspection.

In the past, the main detection methods for water conveyance tunnels were as follows.

1) Pre-installed monitoring facilities. However, the location of the pre-installed detection facilities underwater is fixed, and the detection range is limited. In addition, with the increase of service life, the damage to monitoring facilities also affects the detection accuracy.

2) Emptying detection. Emptying detection is restricted by many complicated conditions. It not only requires huge manpower, material and financial resources, but also may lead to sudden changes in working conditions and cause engineering problems.

3) Diver entry inspection. However, there is a high risk for divers due to long periods of time that may be required, high water pressure and fast water flow speed.

In recent years, the use of remotely operated vehicles (ROVs) has gained traction in inspection of water conveyance tunnels. Moughamian and McLeod [2] inspected and evaluated Pardee Tunnels with Falcon ROV equipped with 3D multi-beam sonar, 2D multi-beam imaging sonar and standard definition camera at a

Article history: Received Nov 22, 2021; Accepted Jan 28, 2022

water flow rate of one foot per second. Jorge et al. [3] designed a compact unmanned underwater robotic system equipped with forward-looking sonar, echosounders, and low-light cameras to effectively complete the inspection of high turbidity underwater tunnels. Lai [4] carried out research on ROV underwater inspection and adopted the method of combining real-time three-dimensional imaging sonar scanning survey and local detailed inspection by optical camera to complete a comprehensive underwater inspection with high precision and apparent full coverage of large diameter and long water tunnels. However, the optical images collected by the above methods need to be assessed by human experts. Additionally, the images may have different degrees of degradation, such as uneven illumination, low contrast, color distortion, image blur, etc., due to the complicated underwater conditions. These problems make subsequent classification tasks error-prone, inefficient and time-consuming. To address these problems, efficient data processing is needed.

With the continuing development of artificial intelligence, researchers have begun to apply machine learning methods to deal with engineering damage recognition. In the last two decades, researchers have applied many different traditional feature-based methods to structural health monitoring (SHM), including support vector machine (SVM) [5–8], k-nearest neighbor [9], decision trees [10], as well as Bayesian method [11]. However, these traditional approaches all rely on hand-designed features customized for specific tasks. In addition, the trained model may no longer be suitable for vision-based SHM applications [12].

Since deep learning methods have shown their advantages in many fields [13–19], there is a growing interest in using deep learning in civil engineering [20]. Deep learning models develop an end-to-end structure that can automatically and efficiently extract complex features from the original image, and build nonlinear mappings through stacked deep neural network layers, so that deep learning-based SHM is applicable to a wide range of problems without human intervention [12]. Inspired by the great success of deep learning in the field of computer vision, researchers have recently attempted to apply vision-based deep learning methods to civil engineering problems. Luo et al. [21] proposed a pothole detector based on deep learning that can automatically detect potholes and prevent repeated detection. Cha et al. [22] proposed a Faster R-CNN-based structural vision detection method for real-time and accurate detection of multiple types of damage. Shi et al. [23] proposed a novel underwater dam crack detection and classification approach. Liang [24] presented a three-level image-based method for post-disaster detection of reinforced concrete bridges based on deep learning and new training strategies. Savino and Tondolo [25] proposed a convolutional neural network (CNN) based classification method for concrete damage classification of bridges, tunnels and pavements.

Although deep learning methods can improve the performance of damage recognition, some challenges remain in the application of ROVs. On the one hand, ROVs are limited by volume and energy, and generally adopt embedded computer systems with limited hardware capabilities. Therefore, the large-scale deep neural network is not applicable for embedded systems of ROVs. On the other hand, the limited number of underwater training samples will lead to the overfitting problem of the deep learning model, which means the model learns specific features without abstracting general features.

To achieve the trade-off between recognition accuracy and complexity of the CNN-based method, ensemble learning combined with the deep CNN model is introduced in this study. Ensemble learning is a model integration algorithm that generates a set of weak classifiers and integrates their predicted results according to a certain strategy [26]. Several ensemble learning methods have been proposed, including bagging [27], boosting [28], Adaboost [28–30] and random forest [9]. Compared with the single training method, the ensemble method demonstrates better prediction performance in practice and has attracted attention in various fields.

In this study, an ensemble deep learning method is proposed for siltation recognition by ROVs to balance the computational cost and classification accuracy of the embedded system. In fact, we construct a deep learning model called Silt-net with good feature extraction and classification ability. During training, we adopt Bootstrap to ensure the diversity of these homogeneous classifiers. Finally, a fully-connected neural network is used as a meta-learner to combine the predictions of Silt-nets.

The main contributions of this paper are summarized as follows.

1) To the best of our knowledge, the idea of combining ROVs with ensemble deep learning methods for underwater siltation recognition is proposed for the first time in this paper.

2) An ensemble learning method based on deep learning is proposed for siltation recognition. In the proposed method, several convolutional neural networks called Silt-net are constructed and trained as base learners. Bootstrap is used to generate the diversity of the ensemble system. A fully-connected network is used as a meta-learner to combine the outputs of the base learners to obtain the final results.

3) With the proposed ensemble deep learning method, we conduct a siltation recognition framework that is more suitable for the embedded system of ROVs. The recognition framework not only achieves 97.2% accuracy, but also has low complexity. The floating point operations

(FLOPs) are 1.2 G, and the number of parameters is 75.1 M.

The remainder of this paper is organized as follows: Section 2 describes the acquisition and preprocessing of images. Section 3 introduces the proposed ensemble deep learning siltation recognition model. Experimental design and experimental results on underwater siltation dataset are presented in Section 4. Section 5 provides concluding remarks and directions for future research.

2 Data collection

2.1 ROV system

An underwater vehicle is a complex multi-functional system, with different equipment to solve challenging tasks in specific fields [31,32]. Underwater vehicles are divided into manned underwater vehicles and unmanned underwater vehicles (UUVs). UUVs can be further divided into two categories: ROVs and Autonomous Underwater Vehicles (AUVs).

A hydraulic water conveyance tunnel is a tubular underground structure with flowing water and no light inside the pipe. To obtain images of siltation at the bottom of the tunnel, an observation class ROV equipped with various extensions and components is used in this study to obtain the underwater siltation data. Figures 1(a) and 1(b) show the underwater inspection system and the real-time ground control system, respectively.

A typical ROV system consists of an underwater system (including submersible body and a variety of expansion modules) and a ground system (such as an industrial personal computer (IPC), control system, underwater communication interface). Figure 2 shows a complete ROV system. The main body of the submersible has four cameras facing in different directions, various sensors, and optional tools. In order to solve the problem of illumination in the tunnel, we have installed several high-brightness LED lights. The submersible is equipped with six large thrusters, two thrusters facing in the horizontal direction and four thrusters in the vertical direction, ensuring precise motion control in the underwater space. In addition, the ROV main body adopts a combined navigation mode of inertial navigation and Doppler Velocimeter (DVL). The ROV body is also equipped with various sensors, which can obtain its position, speed, and attitude at any time, making it highly adaptable to the complex environment, such as high velocity water and confined spaces.

The entire system is controlled by the ground IPC unit and the switch unit. These units use the TCP/IP communication protocol to collect, organize and transmit control information through a zero-buoyancy umbilical cable, providing the connection between underwater and terrestrial systems. The zero-buoyancy umbilical cable is composed of communication unshielded twisted pair (UTP) cable and video cable. The operator controls the ROV using a control handle.

(a) (b)

Fig. 1 Image acquisition system: (a) underwater inspection system; (b) ground control system.

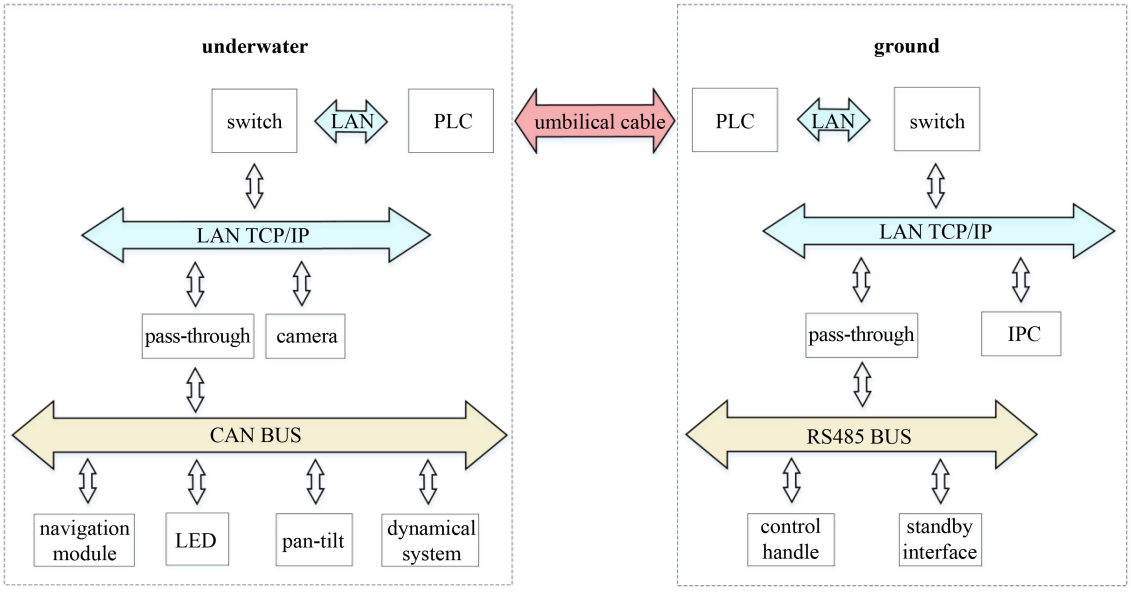

Fig. 2 An overview of ROV system.

2.2 Image acquisition of siltation

All images used in this work are collected by ROV from different water conveyance tunnels with different siltation characteristics. In order to obtain high-quality underwater siltation data, we first use ROV to obtain video data from different angles of view, altitudes and velocities of mechine, then extract images from the video frame by frame at a 60-frame interval, and finally refine the extracted images manually. In the refining process, we consider rich siltation characteristics, different image degradation characteristics, and extensive siltation content, and finally four typical types of underwater siltation are selected to train the model. Figure 3 illustrates different types of samples from different perspectives. The first category is severe siltation, in which a large proportion of the tunnel's diameter is filled with material, which then requires specialized technology and equipment to remove. The second category is general siltation. This type of siltation has a small thickness and easily becomes suspended in the water when disturbed. It can be removed by the disturbance of the ROV's propeller. No siltation is the third category, which means healthy engineering status. In underwater inspection, sometimes no useful information can be collected because of a variety of complex situations, so we set an additional category of no targets as the fourth category.

The obtained underwater siltation data set contains 6000 images, of which 5000 are used for training and 1000 are used for testing. Table 1 summarizes the number of various types of samples in the dataset.

2.3 Image preprocessing

5000 samples are randomly selected from the dataset as a set in the subspace of the dataset by the bootstrap method. Multiple different training sets are generated using the bootstrap method multiple times. Due to limited memory, we adjust the pixel size of training data from 1920 × 1080 to 224 × 224. In order to improve the generalization performance of the model and make the model more robust to complex and different underwater siltation images, this paper applies real-time data augmentation processing to maximize the effect of a small sample

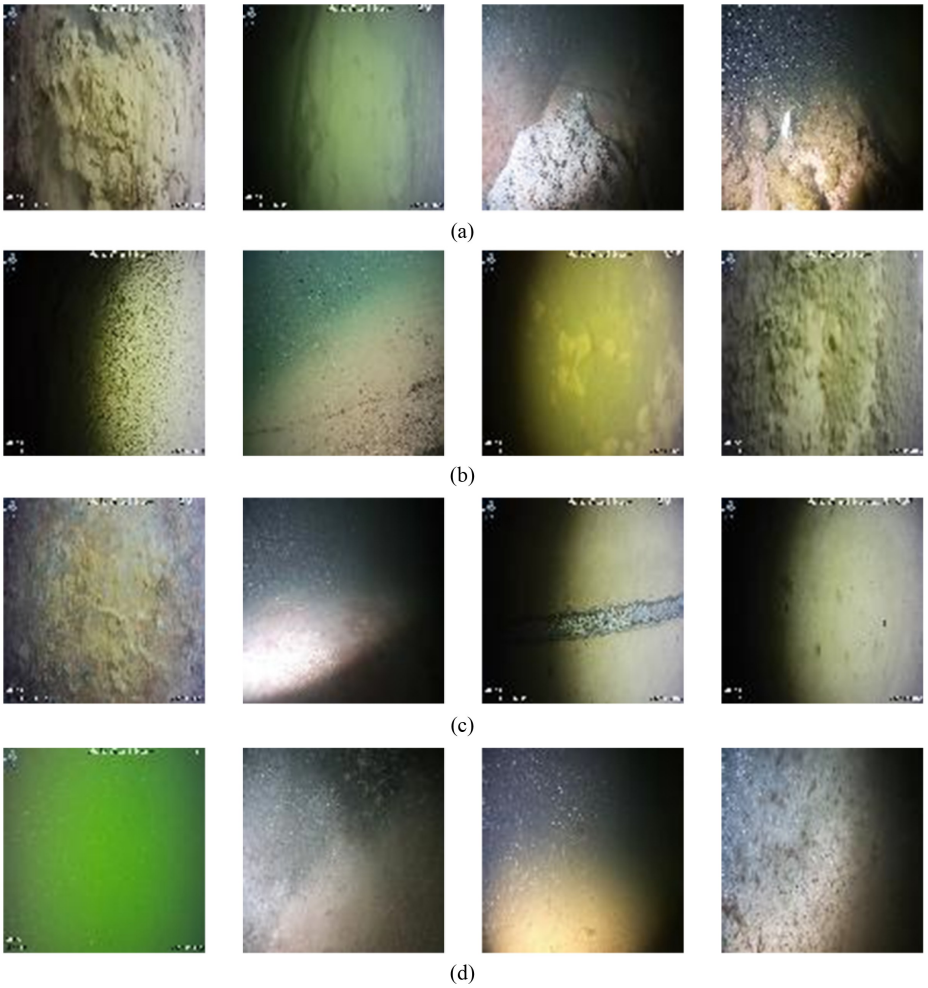

Fig. 3 Different types of samples from different perspectives: (a) severe siltation; (b) general siltation; (c) no siltation and (d) no targets.

Table 1 Dataset for training and testing

dataset	category	number
train 5000	severe siltation	1250
	general siltation	1250
	no siltation	1250
	no targets	1250
test 1000	severe siltation	250
	general siltation	250
	no siltation	250
	no targets	250

training set. The data augmentation methods used are as follows. 1) The image is randomly rotated by 90°; 2) the image is randomly shifted horizontally and vertically, with a translation ratio of 0.1; 3) the image is cropped randomly with a cropping ratio of 0.2; 4) the image is randomly scaled with a scaling ratio of 0.2; 5) the image flips horizontally at random; 6) image brightness, contrast and saturation are adjusted randomly.

3 Deep CNN-based ensemble method for siltation recognition

In this section, the deep learning ensemble framework is discussed. The workflow of the proposed method is shown in Fig. 4. The proposed method utilizes Silt-nets as base classifiers and then integrates the predictions from Silt-nets using a meta classifier. Furthermore, the use of Bootstrap ensures the diversity of homogeneous learners. In the following, we will explain the stacking method in more detail as well as the base and meta learners in the stacking method.

3.1 Deep learning and convolutional neural network

Deep learning methods are representation-learning methods composed of several simple representations, which are converted into multi-level representations through simple but nonlinear modules, and the level of representations is gradually higher and slightly more abstract relative to the original input [33]. Several effective deep learning architectures have been proposed for various tasks including CNN [34], RNN [35], Autoencoder [36,37], GAN [38] and deep belief network [39]. In recent years, CNN has achieved promising success in classification problems [40]. In this study, well-designed CNNs are used as base learners.

There are four key ideas in CNN's architecture design: local receptive fields, shared weights, pooling and the use of stacked layers [33]. In a complete CNN, feature extraction and aggregation operations are performed by stacking multiple convolutional layers, non-linear and pooling layers. At this time, the mapping from original data to feature space is realized. In order to map the learned feature representation to the sample label space, the fully-connected layers are used. Through the mini-batch-based back propagation algorithm, all the learnable parameters in the CNN can be trained.

3.2 Stacking ensemble method

In order to obtain an accurate classifier for siltation, stacking ensemble learning with deep CNNs is investigated in this paper. Stacking is a strategy that combines the results of individual base classifiers using another machine learning algorithm. It can be regarded as a meta learning approach in which the base classifiers are called first-level classifiers and a second-level classifier learns to combine the first-level classifiers [41]. The steps of the stacking ensemble algorithm are as follows.

1) Learn base classifier based on the subspace of the original training set. For a given dataset $D = \{(\boldsymbol{x}_i, y_i) | 1 \leqslant i \leqslant M\}$, which is composed of M training samples, where $(\boldsymbol{x}_i, y_i)$ is the ith training sample with the corresponding category label $y_i \in Y = \{1, 2, ..., C\}$, and Y is

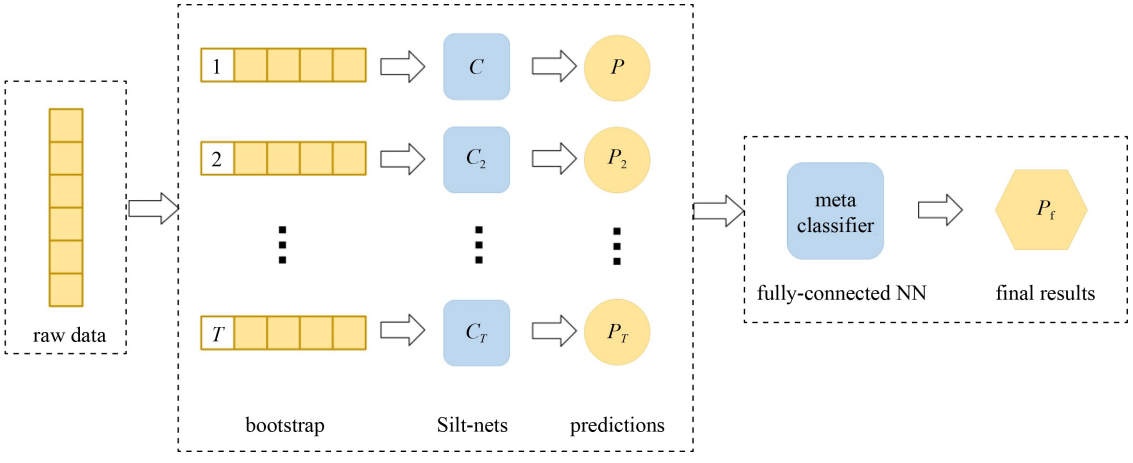

Fig. 4 An overview of proposed Silt-nets stacking framework.

the set of all the labels, C is the total number of categories. During the learning process, each base classifier is trained in a random data subspace sampled by bootstrap method individually, in which the number of samples m is equal to M. After repeated T times, random data subspace $\widetilde{D} = \left\{\widetilde{D_1}, \widetilde{D_2}, ..., \widetilde{D_T}\right\}$ and corresponding homogeneous base classifiers $H = \{h_1, h_2, ..., h_T\}$ are obtained, where T is the number of the base classifiers.

2) Construct new data space based on the predictions of base classifiers. The prediction results of first-level classifier $P = \{P_1, P_2, ..., P_T\}$ are used as the new input, and the original category labels remain as the labels of the new data space. The newly created data space is as follows: $D' = \{(\boldsymbol{x}'_i, y_i) | 1 \leqslant i \leqslant M\}$, where $\boldsymbol{x}_i' = \{h_1(\boldsymbol{x}_i), h_2(\boldsymbol{x}_i), ..., h_T(\boldsymbol{x}_i)\}$.

3) The second-level classifier is learned using the newly constructed training set. Any machine learning method, such as decision tree, SVM, Bayesian classifier, neural network, can be used for second-level classifier learning.

Once we have generated the first-level and second-level classifiers, we can use stacking for high-precision classification tasks. For a test sample $\boldsymbol{x}$ (not seen by the classifiers), the predicted category result for stacking is $h'(h_1(\boldsymbol{x}_i), h_2(\boldsymbol{x}_i), ..., h_T(\boldsymbol{x}_i))$, where $\{h_1, h_2, ..., h_T\}$ are first-level classifiers and h' is the second-level classifier.

3.3 Silt-net stacking framework

Under the influence of many factors such as reservoir scheduling, rainfall and temperature, the water and sand transfer process in long distance water conveyance tunnels is very complicated. When the sand-bearing water flows through the tunnel, the water flow conditions change due to various uncertainties such as bottom friction, the water flow path and the shape of the tunnel cross-section. If the water flow velocity is less than the sediment-moving incipient velocity, siltation is generated. In particular, cumulative siltation may occur in local sections. Using ROVs for long-distance water conveyance tunnel inspection is currently an effective non-destructive inspection method, but the large amount of data collected by ROVs cannot be processed automatically. In addition, the embedded platform with extremely limited computational resources cannot support the deployment of large-scale neural networks. We propose the Silt-net stacking framework to provide a trade-off between the accuracy of automated silt identification and model complexity.

We design an extremely compact and effective CNN based on VGG16 as the base learner called Silt-net. The Silt-net is designed to reduce the complexity of the model. Since there are fewer types of siltation in long water conveyance tunnels and the features of each category are relatively simple and spread over the whole image, the redundancy of the network can be reduced by simplifying VGG16. We reduce the number of network layers as well as the number of channels per layer. Tables 2 and 3 show the network structure and some parameter settings of VGG16 and Silt-net respectively. In Silt-net, we used stacked convolutional layers, where we use filters with a spatial resolution of 3 × 3. The convolution stride and padding are fixed at 1 pixel. To reduce the dimensionality of each feature map but to retain the most important information, three max-pooling layers are applied, which follow some of the convolutional layers. Max-pooling is performed over a 2 × 2 pixel window, with stride 2. There are three stages of convolution layers and pooling layers. Each convolutional layer uses a ReLU activation function, $\sigma(x) = \max(0, x)$.

After the stacked convolutional layers, we add the flatten operation to convert the multi-dimensional feature map matrixes to vectors. This is followed by two fully-connected layers: the first has 512 channels, using ReLU as the activation function, and the second performs the classification of siltation and thus contains four channels with softmax as the classifier.

Table 2 VGG16 architecture

layer size	stride	operator	output shape	parameters
$224^2 \times 3$	s1	2*(conv2d, 3 × 3)	$224^2 \times 64$	38720
$224^2 \times 64$	s2	max pooling2d	$112^2 \times 64$	–
$112^2 \times 64$	s1	2*(conv2d, 3 × 3)	$112^2 \times 128$	221440
$112^2 \times 128$	s2	max pooling2d	$56^2 \times 128$	–
$56^2 \times 128$	s1	3*(conv2d, 3 × 3)	$56^2 \times 256$	1475328
$56^2 \times 256$	s2	max pooling2d	$28^2 \times 256$	–
$28^2 \times 256$	s1	3*(conv2d, 3 × 3)	$28^2 \times 512$	5899776
$28^2 \times 512$	s2	max pooling2d	$14^2 \times 512$	–
$14^2 \times 512$	s1	3*(conv2d, 3 × 3)	$14^2 \times 512$	7079424
$14^2 \times 512$	s2	max pooling2d	$7^2 \times 512$	–
$7^2 \times 512$	–	flatten	25088	–
25088	–	2*dense	4096	119545856
4096	–	dense	4	16388

Table 3 Silt-net architecture

layer size	stride	operator	output shape	parameters
$224^2 \times 3$	s1	conv2d, 3 × 3	$224^2 \times 16$	512
$224^2 \times 16$	s2	max pooling2d	$112^2 \times 16$	–
$112^2 \times 16$	s1	2*(conv2d, 3 × 3)	$112^2 \times 32$	14144
$112^2 \times 32$	s2	max pooling2d	$56^2 \times 32$	–
$56^2 \times 32$	s1	2*(conv2d, 3 × 3)	$56^2 \times 64$	55936
$56^2 \times 64$	s2	max pooling2d	$28^2 \times 64$	–
$28^2 \times 64$	–	flatten	43264	–
43264	–	dense	512	22151680
512	–	dense	4	2052

In order to improve the performance of the Silt-net, Batch Normalization (BN) is adopted in the convolutional layer before the ReLU activation function. BN has the following advantages. First, it can adjust the data distribution of output activation values in the network and accelerate the learning speed of the model. Second, BN makes the model less sensitive to the parameters in the network, simplifies the parameter adjustment process, and makes the network learning more stable. Third, BN can effectively suppress overfitting. Simultaneously, we use a dropout layer with 50% drop rate to prevent overfitting. The input given to the CNN is an image with a pixel size of 224 × 224 and 3 channels. The output of the last layer of the network is the probability that an image belongs to each class.

Another key factor that restricts the application of ROV in water conveyance tunnel inspection is the impact of the complex underwater environment on the image. On the one hand, due to the high velocity of water flow, the water body carries a large amount of sand, debris-like material and other fine particles. And disturbed by the ROV thrusters, the ground silt will be raised rapidly, resulting in high turbidity of the water body, which seriously affects the quality of optical vision imaging. In addition, there is a lack of natural light in the hydrographic tunnel, and the illumination depends entirely on the LED lighting system equipped with the ROV. The LED light intensity is high, and the color distortion is less compared to the situation in the ocean. However, there is a phenomenon of uneven illumination, with directly lit foreground areas appearing completely white, while the area away from the source of light appears darker.

Combining the above problems, we hope to mitigate the impact of the complex underwater environment on the image quality by joint decision making of multiple classifiers, thus improving the accuracy of the whole model. Specifically, we use a fully-connected network with 3 hidden layers as a meta-learner. The fully connected neural network takes the output of our multiple base learners as input and returns the final prediction. Similarly, the output of the network is the probability that an image belongs to each class. Moreover, compared with a single deep learning model, although the proposed classification method requires several separate training of the base classifier as well as the meta-classifier, the introduction of the ensemble strategy allows several sufficiently simple classifiers with limited classification power to maximize their potential. Especially in the inference stage, the model complexity of the method using the integration of multiple simple classifiers is much less than that of large deep neural network models. In summary, our model can obtain comparable or even better accuracy than a single powerful classifier with reduced model parameters and computational effort.

4 Experimental results

In this section, we introduce the datasets used in this work, and the experimental results are analyzed according to the evaluation criteria. All the experiments are carried out on a PC with NVIDIA RTX 2080Ti GPU, using Keras with TensorFlow (version 1.14.0) as the backend. The details of the experimental process are introduced as follows.

4.1 Model training and validation

The proposed model is trained in two stages. Table 4 details the hyperparameters used at each stage. During first-level classifier training, the training is carried out by optimizing the cross entropy loss function using RMSprop with rho = 0.9 and decay = 10^{-6}. The batch size is set to 32. The learning rate is initially set to 10^{-3}, and then adjusted by the cosine annealing method. The filter weights of each layer are initialized by He initialization. Bias is initialized as a constant. Finally, we use 100 epochs to train a base classifier. During second-level classifier training, we use the Adam to optimize the loss function. To prevent overfitting, L2 regularization is adopted for the weight of the fully-connected layer, and the regularization parameter is set to 0.0001. The dropout layers with a dropout ratio of 0.5 are used after each hidden layer. In addition, the early stopping strategy is also introduced, and the training is stopped after 40 epochs.

We use the total loss to evaluate the convergence and fit level of the training and validation processes to follow up whether the model is overfitting. Figure 5 shows the loss curves of the three different base classifiers as well as the meta-classifier. As can be seen from Figs. 5(a)–5(c), the training and validation processes of all three base classifiers converge. Besides, Fig. 5(d) also gives the loss curve of the meta-classifier at 100 epochs. Obviously, the loss curve on the validation dataset increases in the 20th epoch, which further proves that it is very necessary to use the early stopping strategy.

In order to evaluate the performance of the proposed ensemble deep learning method, three individual

Table 4 Experimental hyper-parameter setting

stage	variable	hyper-parameters
stage 1	RMSprop	$Lr = 0.001$
	epoch	100
	batch size	32
stage 2	Adam	$Lr = 0.001$
	L2 regularization	0.0001
	early stopping	*patience* = 20
	dropout	*ratio* = 0.5

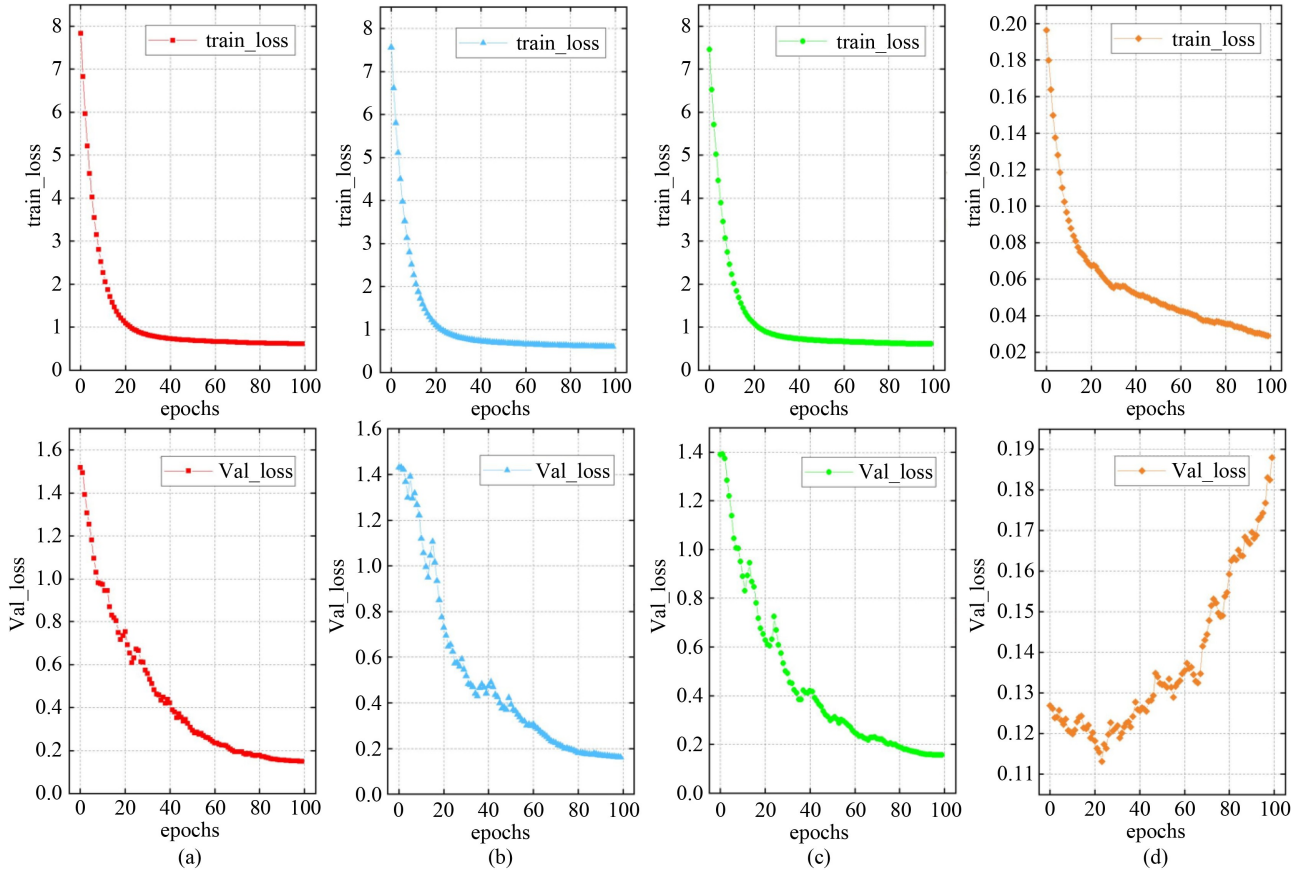

Fig. 5 Loss curves of base classifiers and meta-classifier in training and validation process: (a) base classifier 1; (b) base classifier 2; (c) base classifier 3; (d) meta-classifier.

classifiers include SVM, Silt-net and VGG16, and three ensemble learning methods include bagging (using major voting), bagging (using weighted average) and Adaboost are used for comparison.

4.2 Compare with individual classifiers

In the quantitative comparison with single classifiers, overall accuracy (OA), the number of parameters (Params), and FLOPs are used as the evaluation criteria. Specifically, OA is used to evaluate the model accuracy, and Params and FLOPs are used to evaluate the complexity of the model in terms of model size and computational cost, respectively.

OA is the most common metric used to evaluate classification problems. It refers to the ratio of the samples with correct prediction to the total samples. OA is defined as follows:

$$Overall\ accuracy = \frac{TS + TG + TNS + TNT}{N}, \quad (1)$$

where TS, TG, TNS, TNT are number of true severe siltation, true general siltation, true no siltation and true no targets, respectively. N is the total number of test images.

In the embedded system of ROV, Params and FLOPS are the key evaluation metrics to be considered emphatically. Params determines the memory space occupation of the model, and FLOPS is used to measure the computational consumption of the model. It is worth noting that this paper considers the impact of convolutional layers as well as fully-connected layers on the model complexity.

The parameters of the convolutional layer and fully-connected layer are calculated respectively as follows:

$$Param_{\text{conv}} = (k_{\text{w}} * k_{\text{h}} * c_{\text{in}}) * c_{\text{out}} + c_{\text{out}}, \quad (2)$$

$$Param_{\text{fc}} = (n_{\text{in}} * n_{\text{out}}) + n_{\text{out}}. \quad (3)$$

To calculate the number of FLOPs, it is assumed that the convolution is implemented as a sliding window, and the non-linear function is calculated for free. Different layers require different computational consumption due to the size and number of input-output feature maps and the number of convolution kernels [42]. The FLOPs of the convolutional layer and fully-connected layer are calculated as follows:

$$FLOPs_{\text{conv}} = 2HW(k_{\text{w}} * k_{\text{h}} * c_{\text{in}} + 1) * c_{\text{out}}, \quad (4)$$

$$FLOPs_{\text{fc}} = (2n_{\text{in}} - 1) * n_{\text{out}}, \quad (5)$$

where k_w, k_h are the width and height of the kernel, respectively. W, H are width and height of input feature map, c_{in}, c_{out} are the number of channels of the input feature map and output feature map, respectively, and n_{in} and n_{out} are the input and output dimensions of the fully-connected layer, respectively.

Since it is not reliable to perform the training and testing phases on the model only once, we separately train and test each model five times, and the results are shown in Fig. 6. The results show that the classification accuracy of the deep learning method is much higher than that of traditional RBF-SVM method. It can be seen that the classification accuracy of the deep learning method is at least 4% higher than that of the SVM method. Among the deep learning methods, the classification accuracy of Silt-net with a simple network structure is lower than that of VGG16. Moreover, the feature extraction ability of Silt-net is relatively poor due to the lack of network depth, which leads to its large fluctuation of accuracy in the testset and poor robustness to data. But with an effective ensemble method, the classification accuracy of the model exceeds that of VGG16, and is also more robust in complex siltation situations.

Then, we show the classification performance of the classifier using a model with median classification accuracy obtained in five training sessions for each algorithm. Table 5 shows the classification performance of each classifier in various categories. The best accuracy is highlighted in bold. It can be seen that the accuracy of our proposed method is as high as 98% and 99.2% for severe siltation and general siltation, respectively.

In addition, the Params and FLOPs of the model are shown in Table 6. As shown in Table 6, the proposed method is significantly superior to VGG16 in terms of the number of parameters and FLOPs, which are reduced by 44.1% and 92.3%, respectively. So ensemble deep learning has a better balance between model complexity and model accuracy. Under the premise of ensuring classification accuracy, ensemble deep learning significanty reduces the complexity of the model, making it more conducive to its application in ROVs systems.

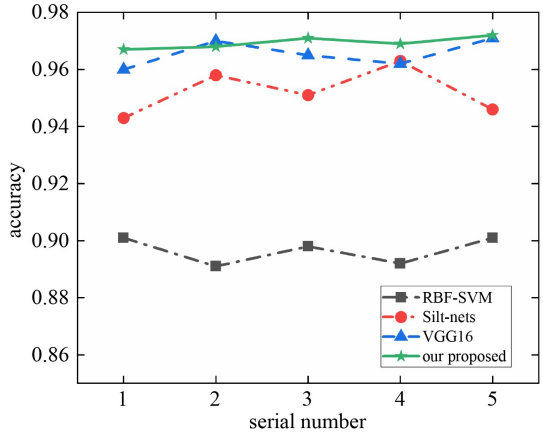

Fig. 6 Classification results of different classifiers.

4.3 Compare with ensemble learning methods

In the comparison of ensemble learning methods, in addition to using accuracy as an evaluation criterion, diversity is also introduced to evaluate the performance of ensemble learning [43]. In this paper, the disagreement measure is used as the diversity criterion. The diversity of the two classifiers (h_i, h_j) is defined as follows:

$$DIV(h_i, h_j) = \frac{N_{\text{disagreement}}}{N}, \tag{6}$$

where $N_{\text{disagreement}}$ is the number of samples in which the two classifiers classify the same sample in the test samples with different results, and the denominator N is the total number of test samples. The diversity of the entire ensemble learning method is the average of the diversity of all pairwise base learners, and the calculation formula is as follows:

$$DIV(Ensemble) = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} DIV(h_i, h_j)}{2K},\ i \neq j, \tag{7}$$

where K is the number of base classifiers.

Table 7 shows the diversity and classification results of ensemble deep learning with different base learners. Since the base learner has good generalization performance and the diversity fluctuates in a small range, the

Table 5 Classification performance of each classifier in various categories

method	RBF-SVM	Silt-net	VGG16	our proposed
heavy silt	94.8%	97.2%	94.0%	98.0%
general silt	96.0%	96.8%	98.8%	99.2%
no silt	84.4%	96.0%	96.4%	95.6%
no targets	84.0%	90.4%	96.8%	94.8%
OA	89.8%	95.1%	96.5%	96.9%

Table 6 Comparison of model complexity between different models

method	Silt-net	VGG16	our proposed
params (M)	22.22	134.28	75.1
FLOPs (G)	0.4	15.5	1.2

Table 7 Ensemble results of different base learner combinations

item	base learner 1	base learner 2	base learner 3	DIV	ensemble results
1	94.3%	94.6%	95.1%	0.042	96.4%
2	95.1%	95.8%	96.0%	0.030	96.8%
3	94.3%	94.6%	96.3%	0.043	96.9%
4	94.3%	95.1%	96.3%	0.040	96.9%
5	95.8%	96.0%	96.3%	0.032	97.2%

effect of diversity on the total accuracy in this experiment is less than the effect of the accuracy of the base classifier on the total accuracy. Through the experimental results, we can see that the combination of higher classification accuracy and greater diversity base classifiers will obtain more accurate classification results.

The ensemble learning method based on deep learning has been proved to be an effective classification method. To verify the superiority of the proposed model in classification, we compare the classification accuracy of the proposed model with that of the state-of-the-art ensemble learning method using different numbers of base learners. The comparison results are shown in Fig. 7. As a result of relearning what was learned at an earlier stage, the proposed model is superior to other comparative ensemble learning methods based on overall classification accuracy.

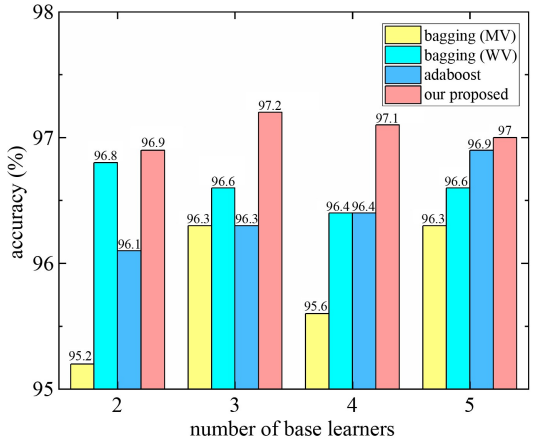

Fig. 7 Results of different ensemble methods under different numbers of base learners.

5 Conclusions and future work

In recent years, the use of ROVs for long-distance water conveyance tunnels has received extensive attention from researchers, and they have gradually become an important means for detection of underwater damage in engineering.

Correctly identifying damage can effectively reduce engineering risks and ensure engineering safety and reliability. Emerging ROVs need more effective damage detection techniques, and the development of artificial intelligence provides favorable conditions for improving the performance of damage detection models.

In this paper, for the first time, the combination of ROVs and ensemble deep learning method is applied to the siltation recognition problem of water conveyance tunnels. A deep learning-based stacking ensemble learning framework for siltation images acquired by ROVs is proposed. In the proposed ensemble deep learning, Siltnets are used as individual base classifiers, and bootstrap is used to ensure the diversity of ensemble methods. Finally, a fully-connected network is used to combine the predictions of the base classifiers to obtain the final result. The experimental results demonstrate that under the simple and efficient framework, the stacking ensemble deep learning method achieves better classification performance than traditional SVM, single CNN and some other state-of-the-art ensemble learning methods. Furthermore, the ensemble approach significanty reduces the complexity of the model, which makes it more suitable for the embedded system of ROVs.

The aforementioned ensemble deep learning method presents a promising prospect for the intelligent identification of siltation by ROVs. Although this study involves some new aspects, there are still some explorations to be done in future work. 1) Some advanced lightweight deep learning models such as MobileNet and ShuffleNet have been developed. These methods will be used to attempt to further compress the size of the model. 2) Transfer learning can accelerate the model training process and improve the classification accuracy at the same time. However, transfer learning is not used in the model training process in this study. Therefore, transfer learning of underwater data sets will be introduced. 3) To further improve the performance of ensemble deep learning, the ensemble between high-performance heterogeneous base learners can be explored.

Acknowledgements Thanks to South to North Water Diversion Central Route Information Technology Co., Ltd. for providing the underwater video of the water conveyance tunnels for research purposes. This work was supported by the National Key R&D Program of China (No. 2016YFC0401600), the National Natural Science Foundation of China (Grant Nos. 51979027, 52079022, 51769033, and 51779035). It should be understood that none of the authors have any financial or scientific conflicts of interest with regard to the research described in this manuscript.

References

1. Zhu X M, Wang T H, Liu Y B, Huang T. A new defect detection technology for long-distance water conveyance tunnel. Water Resources and Hydropower Engineering, 2010, 41(12): 78−81
2. Moughamian R, McLeod M. Pardee tunnel inspection and condition assessment. In: Conference on Pipeline Engineering—Concepts in Harmony (PIPELINES). Nashville: American Society of Civil Engineers, 2019
3. Jorge V A M, Gava P D D, Silva J R B F, Mancilha T M, Vieira W, Adabo G J, Nascimento C L. VITA1: An unmanned underwater vehicle prototype for operation in underwater tunnels. In: the 15th Annual IEEE International Systems Conference (Syscon 2021). IEEE, 2021
4. Lai J T. Research and application of underwater full coverage unmanned detection technology for large diameter and long diversion tunnel. Yangtze River, 2020, 51(05): 228−232 (in Chinese)

5. Pan Y, Zhang L M, Wu X G, Skibniewski M J. Multi-classifier information fusion in risk analysis. Information Fusion, 2020, 60: 121−136
6. Pan H, Azimi M, Yan F, Lin Z. Time-frequency-based data-driven structural diagnosis and damage detection for cable-stayed bridges. Journal of Bridge Engineering, 2018, 23(6): 04018033
7. Wirtz S F, Beganovic N, Soffker D. Investigation of damage detectability in composites using frequency-based classification of acoustic emission measurements. Structural Health Monitoring, 2019, 18(4): 1207−1218
8. Babajanian Bisheh H, Ghodrati Amiri G, Nekooei M, Darvishan E. Damage detection of a cable-stayed bridge using feature extraction and selection methods. Structure and Infrastructure Engineering, 2019, 15(9): 1165−1177
9. Kurian B, Liyanapathirana R. Machine learning techniques for structural health monitoring. In: the 13th International Conference on Damage Assessment of Structures. Porto: Springer, 2020
10. Mechbal N, Uribe J S, Rebillat M. A probabilistic multi-class classifier for structural health monitoring. Mechanical Systems and Signal Processing, 2015, 60-61: 106−123
11. Huang Y, Beck J L, Li H. Hierarchical sparse Bayesian learning for structural damage detection: Theory, computation and application. Structural Safety, 2017, 64: 37−53
12. Azimi M, Eslamlou A D, Pekcan G. Data-driven structural health monitoring and damage detection through deep learning: State-of -the-art review. Sensors (Basel), 2020, 20(10): 2778
13. Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv:1810.04805
14. Bochkovskiy A, Wang C Y, Liao H Y M J. YOLOv4: Optimal speed and accuracy of object detection. 2020, arXiv:2004.10934
15. Chen L C E, Zhu Y K, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: the 15th european conference on computer vision (ECCV). Munich: Springer, 2018: 833−851
16. Moritz N, Hori T, Le Roux J. Triggered attention for end-to-end speech recognition, In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019
17. Anitescu C, Atroshchenko E, Alajlan N, Rabczuk T. Artificial neural network methods for the solution of second order boundary value problems. Computers, Materials and Continua, 2019, 59(1): 345−359
18. Guo H W, Zhuang X Y, Rabczuk T. A deep collocation method for the bending analysis of kirchhoff plate. Computers, Materials and Continua, 2019, 59(2): 433−456
19. Samaniego E, Anitescu C, Goswami S, Nguyen-Thanh V M, Guo H, Hamdia K, Zhuang X, Rabczuk T. An energy approach to the solution of partial differential equations in computational mechanics via machine learning: Concepts, implementation and applications. Computer Methods in Applied Mechanics and Engineering, 2020, 362: 112790
20. Ye X W, Jin T, Yun C B. A review on deep learning-based structural health monitoring of civil infrastructures. Smart Structures and Systems, 2019, 24(5): 567−585
21. Luo L X, Feng M Q, Wu J P, Leung R Y. Autonomous pothole detection using deep region-based convolutional neural network with cloud computing. Smart Structures and Systems, 2019, 24(6): 745−757
22. Cha Y J, Choi W, Suh G, Mahmoudkhani S, Büyüköztürk O. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. Computer-Aided Civil and Infrastructure Engineering, 2018, 33(9): 731−747
23. Shi P F, Fan X N, Ni J J, Wang G. A detection and classification approach for underwater dam cracks. Structural Health Monitoring, 2016, 15(5): 541−554
24. Liang X. Image-based post-disaster inspection of reinforced concrete bridge systems using deep learning with Bayesian optimization. Computer-Aided Civil and Infrastructure Engineering, 2019, 34(5): 415−430
25. Savino P, Tondolo F. Automated classification of civil structure defects based on convolutional neural network. Frontiers of Structural and Civil Engineering, 2021, 15(2): 305−317
26. Dietterich T G. Ensemble methods in machine learning. In: 1st International Workshop on Multiple Classifier Systems. Cagliari: Springer Science & Business Media, 2000
27. Bui D T, Ho T C, Pradhan B, Pham B T, Nhu V H, Revhaug I. GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. Environmental Earth Sciences, 2016, 75(14): 1−22
28. Tsiapoki S, Bahrami O, Hackell M W, Lynch J P, Rolfes R. Combination of damage feature decisions with adaptive boosting for improving the detection performance of a structural health monitoring framework: Validation on an operating wind turbine. Structural Health Monitoring, 2021, 20(2): 637−660
29. Kadavi P R, Lee C W, Lee S. Application of ensemble-based machine learning models to landslide susceptibility mapping. Remote Sensing, 2018, 10(8): 1252
30. Li Z R, Guo J Q, Liang W S, Xie X, Zhang G, Wang S. Structural health monitoring based on realadaboost algorithm in wireless sensor networks. In: the 9th International Conference on Wireless Algorithms, Systems, and Applications (WASA). Harbin: Springer, 2014
31. Christ R D, Wernli R. The ROV Manual: A User Guide for Observation Class Remotely Operated Vehicles. Oxford: Elsevier, 2013
32. Vukšić M, Josipović S, Čorić A, Kraljević A. Underwater ROV as inspection and development platform. Transactions on Maritime Science, 2017, 6(1): 48−54
33. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436−444
34. Dorafshan S, Thomas R J, Maguire M. Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete. Construction & Building Materials, 2018, 186: 1031−1045
35. Perez-Ramirez C A, Amezquita-Sanchez J P, Valtierra-Rodriguez M, Adeli H, Dominguez-Gonzalez A, Romero-Troncoso R J. Recurrent neural network model with Bayesian training and mutual information for response prediction of large buildings. Engineering Structures, 2019, 178: 603−615
36. Pathirage C S N, Li J, Li L, Hao H, Liu W, Ni P. Structural

damage identification based on autoencoder neural networks and deep learning. Engineering Structures, 2018, 172: 13−28

37. Zhuang X, Guo H, Alajlan N, Zhu H, Rabczuk T. Deep autoencoder based energy method for the bending, vibration, and buckling analysis of Kirchhoff plates with transfer learning. European Journal of Mechanics—A/Solids, 2021, 87: 104225
38. Zhang W, Li X, Jia X D, Ma H, Luo Z, Li X. Machinery fault diagnosis with imbalanced data using deep generative adversarial networks. Measurement, 2020, 152: 107377
39. Tamilselvan P, Wang Y B, Wang P F. Deep belief network based state classification for structural health diagnosis. In: IEEE Aerospace Conference, Big Sky. Montana: IEEE, 2012
40. Simonyan K, Zisserman A J C S. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. San Diego: OpenReview, 2015
41. Zhou Z H. Ensemble Methods: Foundations and Algorithms. Boca Raton: Taylor & Francis, 2012
42. Molchanov P, Tyree S, Karras T, Aila T, Kautz J. Pruning convolutional neural networks for resource efficient inference. 2016, arXiv:1611.06440
43. Bi Y X. The impact of diversity on the accuracy of evidential classifier ensembles. International Journal of Approximate Reasoning, 2012, 53(4): 584−607