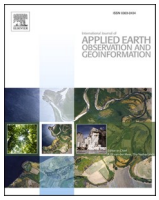
International Journal of Applied Earth Observations and Geoinformation 103 (2021) 102497

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Effectiveness of machine learning methods for water segmentation with ROI as the label: A case study of the Tuul River in Mongolia

Kai Li [a,b], Juanle Wang [a,*], Jinyi Yao [a,c]

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[b] College of Geoscience and Surveying Engineering, China University of Mining & Technology (Beijing), Beijing 100083, China
[c] School of Civil and Architectural Engineering, Shandong University of Technology, Zibo 255049, China

ARTICLE INFO

Keywords:
Water Segmentation
Pixel-based CNN
ROI
U-net
Mongolia

ABSTRACT

The carrying capacity of water resources is key to the sustainable development of arid and semi-arid regions. There are important challenges related to the detection of discontinuous and crooked water bodies in the vast Mongolian Plateau, despite the availability of remote sensing technology which has the advantage of facilitating water observations over large areas and timelines. Given the high cost and low coverage of high-resolution images and the low resolution of images with high coverage, this study proposes a pixel-based convolutional neural network (CNN) method for the application of water extracted from the region of interest (ROI) to medium-resolution Landsat images. The pixel-based CNN method combines the texture and spectral features of the ground object by connecting the center pixels of the images to the surrounding pixels. ROI is used instead of full-label datasets, reduce the difficulty of building labels in low-to-medium-resolution images. Taking the Tuul River in Mongolia as a case, the pixel-based CNN method, the normalized difference water index threshold (NDWI) method, the modified normalized difference water index (MNDWI) threshold method, U-net model in deep learning, and the pixel-based deep neural network (DNN) method were used with medium-resolution Landsat 8 images with ROI labels. The pixel-based CNN method shows better water extraction results for the cloud, cloud shadows, and building areas, compared with other methods. The method proposed in this study had the highest verification accuracy (92.07%). It also has the advantages of fewer training parameters and shorter training time. The training accuracies of the pixel-based CNN, pixel-based DNN, and U-net were 99.90%, 96.98%, and 93.70%, respectively. All training models and calling methods were uploaded to GitHub (https://github.com/CaryLee17/Pixel-based-CNN).

1. Introduction

Water is an important natural resource on which all life depends. Mongolia is a typical arid and semi-arid region with limited water resources which are unevenly distributed (Liu et al., 2015). Inadequate water in vast areas limits the development of animal husbandry. Animal husbandry, as Mongolia's traditional industry, is the foundation of the national economy and the main source of raw materials for the economy and for the fulfillment of daily necessities. Reasonable water ecology and water environments can promote the development of animal husbandry (Mubareka et al., 2013; Mekete, 2013). Meanwhile, excessive rain in a short amount of time, usually in summer, causes flooding. Through the analysis of 965 effective reports, Nara et al. (2019) found that the biggest risk to residents in Khovd Province, Mongolia is flooding (Nara and Battulga, 2019). Therefore, there is an urgent need to carry out effective water resource monitoring and management (Qu et al., 2020) and provide data services for Mongolia's flood mitigation and emergency response.

Water information extracted by remote sensing will not only provide effective basic hydrological data, but also provide guiding data for Mongolia's sustainable water resource management, flood disaster risk reduction, and animal husbandry development. There are two main

Abbreviations: CNN, convolutional neural network; ROI, region of interest; DNN, deep neural network; NDWI, normalized difference water index; MNDWI, modified normalized difference water index; FCN, Fully convolutional network; OLI, Operational Land Imager; MIoU, Mean Intersection over Union; TWR, True Water Rate; FWR, False Water Rate.

* Corresponding author.
E-mail addresses: lk@lreis.ac.cn (K. Li), wangjl@igsnrr.ac.cn (J. Wang), yaojy@lreis.ac.cn (J. Yao).

https://doi.org/10.1016/j.jag.2021.102497
Received 23 June 2021; Received in revised form 6 August 2021; Accepted 9 August 2021
Available online 20 August 2021

types of existing water extraction methods: traditional machine learning and deep learning. Traditional machine learning methods are divided into thresholds and classification methods. McFeeters (McFEETERS, 1996) used the green band and the near-infrared (NIR) band of remote sensing images to normalize and proposed the normalized difference water index (NDWI), which is believed to suppress soil and vegetation characteristics and can be used to describe the characteristics of water bodies and effectively estimate the surface water area (McFeeters, 1996). Xu Hanqiu (2006) replaced the NIR band in the NDWI index with short-wave infrared (SWIR) and constructed a new index suitable for extracting water information, resulting in a modified normalized difference water index (MNDWI) (Xu, 2006). Feyisa et al. (Feyisa et al., 2014) built an automated water extraction index to improve the classification accuracy of shadows and dark surface objects that are usually misclassified. The overall accuracy was better than that of the MNDWI and maximum likelihood classification methods (Feyisa et al., 2014). Yang et al. (Yang et al., 2017) subsequently used NDWI to enhance the SWIR band to extract Sentinel-2A urban water bodies (Yang et al., 2017). Fisher et al. (2013) created the linear discriminant analysis water index method in SPOT5 (Satellite pour l'Observation de la Terre 5) with higher accuracy than that of NDWI and MNDWI. However, distinguishing between water bodies and shadows remained difficult (Fisher and Danaher, 2013). Yang et al. extended threshold methods to obtain dynamic water area using Google Earth Engine (Yang et al., 2020). Classification methods are divided into supervised and unsupervised types. The difference between them is the presence or absence of training samples. Supervised classification requires manual labeling, whereas unsupervised classification does not. The most common supervised classification methods are the maximum likelihood classification (Deng et al., 2011), support vector machine (Jia et al., 2017; Wentao-Lv et al., 2010), minimum distance method (Japitana et al., 2019), decision tree classification (Wang et al., 2007; He et al., 2008); and random forest method (Li et al.; Wu et al., 2012). The most widely used unsupervised classifications are ISODATA (Zhao et al., 2011; Hongxia et al., 2015) and K-means (Rani and Kapinaiah, 2017; Yousefi et al., 2018). Some researchers used multi-source remote sensing data to construct different spectrum feature spaces to extract water bodies in an attempt to find feature information that could effectively distinguish water from other ground objects (Zhang et al., 2003; Zheng et al., 2019; Cao et al., 2015).

With the continuous development of artificial intelligence technology, deep learning methods have gradually penetrated the field of remote sensing. In computer vision, deep learning is mainly used in image classification (Cheng et al., 2020; Rawat and Wang, 2017), object detection (Zou et al., 2019; Jiao et al., 2019); and semantic segmentation (Garcia-Garcia et al., 2017; Asgari Taghanaki et al., 2021). Semantic segmentation is a pixel-level classification method that assigns a category to each pixel, which coincides with the classification task of remote sensing images. Therefore, it has begun to show its potential in the field of remote sensing mapping. Semantic segmentation methods such as fully convolutional network (FCN) (Long et al., 2015), SegNet (Badrinarayanan et al., 2017); and U-net (Ronneberger et al., 2015) are consistently emerging, and progress has also been made in the field of water segmentation. Li et al. (2021) used FCN to extract the water area of a high-resolution Gaofen-2 image (Li et al., 2019), and Weng et al. (Weng et al., 2020) proposed a separable residual SegNet by improving SegNet for refined water extraction (Weng et al., 2020). U-net was introduced into the attention module and the pyramid module, and the PA-U-net model was constructed and applied to the sentinel image to monitor the changing trends of Dongting Lake and Poyang Lake in China from 2017 to 2020 (Li et al., 2021).

Currently, water extraction technologies using deep learning methods are mainly applied to high-resolution satellite imagery (Zheng and Chen, 2021; Wang et al., 2020; Guo et al., 2020; Chen et al., 2020). Existing deep learning water extraction methods mostly rely on the production of fine image label data, but manual labeling is time-consuming and labor-intensive. Moreover, semantic segmentation methods have various problems, such as a large number of parameters, difficulty in fitting, and time consumption. Traditional machine learning methods make it difficult to distinguish between water and dark surface objects, and rely on the spectral information of the image and ignore the textural features. To solve these problems, we proposed a time-saving, pixel-based CNN method with few training parameters, and compared it with various existing traditional machine learning and deep learning methods in Mongolia's Tuul River region. This research introduces the region of interest (ROI) into the deep learning method to improve the accuracy of the labels for water extraction from remote sensing images, and to reduce the difficulty of labeling.

2. Materials

2.1. Study area

The Tuul River basin in the Selenga River Basin is in north-central Mongolia (Fig. 1). The Tuul River (46°35′33″–48°57′13″ N, 102°48′5″–108°8′40″ E) (Dorjsuren et al., 2021) originates from the Khentii Mountains and flows through Ulaanbaatar (the capital of Mongolia), Töv Province, and Bulgan Province and merges into Lake Baikal via the Selenga River. The total length of the Tuul River is approximately 704 km, with a drainage area of 49,840 km^2 (Soyol-Erdene et al., 2019; Batbayar et al., 2017). The reach in Ulaanbaatar from Gachuurt to Songino is approximately 35 km long with a catchment area of 53.2 km^2 (Munkhuu et al., 2019). The land cover types mainly include grassland, woodland, cropland, construction area, and water. Grasslands are the most dominant land cover in this region. The woodlands are mainly located in Ulaanbaatar and in the northeastern part of Töv Province. The croplands were distributed northwest of the Töv Province. The construction area is mainly located in Ulaanbaatar.

2.2. Datasets

The data were selected from Landsat 8 Operational Land Imager (OLI) images. They were obtained from the United States Geological Survey official website, (http://earthexplorer.usgs.gov/) (United States Geological Survey). The Path/Row are 133027, 132027, and 131027 (Table 1), and the corresponding dates are May 7, 2020, August 18, 2020, and July 24, 2020 (Fig. 1). Path/Row 131,027 was used as the training dataset, and all three Path/Row 131027, 132027, and 133,027 were used for verification and accuracy evaluation. The input channels were all original 7 reflectance bands of Landsat 8 images. The training datasets were built in Path/Row 131,027 image by manual annotation, while the validation data were selected by ArcGIS in the three images (Fig. 1). Although the cloud amount was constrained (<10%) when selecting the images, some clouds were still present in the obtained images because of the local weather conditions in the Mongolian Plateau.

3. Methodology

As shown in Fig. 2, this study approach is divided into three main parts: problem analysis and data preparation, new method design and implementation, and comparison of accuracy and model evaluation. First, we found that the most commonly used water segmentation methods are threshold and semantic segmentation. Considering the hard labeling in low-to-medium-resolution image problems, we prepared Landsat 8 images as data sources and performed spectral reconstruction to acquire the spectral reflection of images. We also cropped images and labeled water bodies and non-water bodies for the deep learning method in this study. In the method design and implementation part, we designed a new pixel-based CNN method and compared it with the NDWI, MNDWI, U-net, and pixel-based DNN methods. To compare the accuracy and evaluate the model, we evaluated the training time and

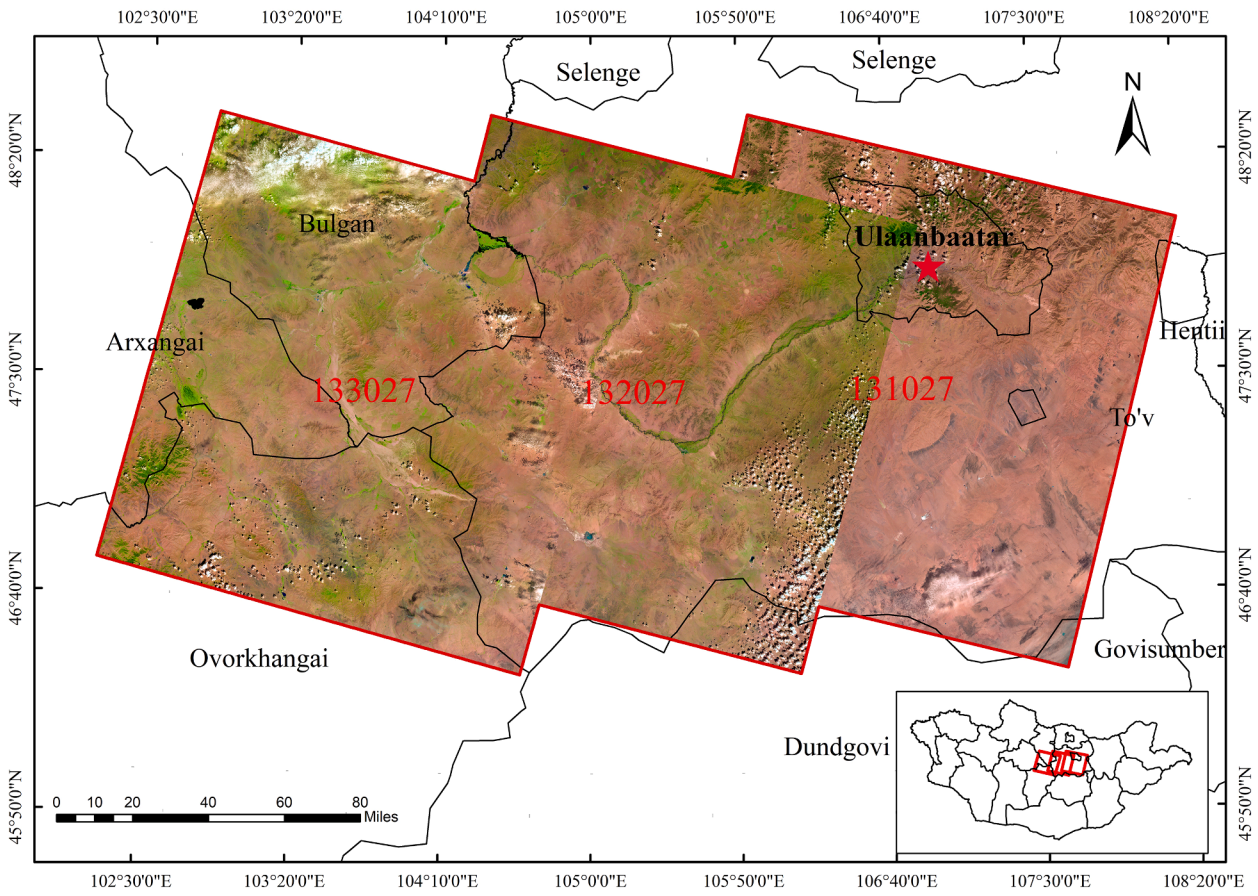


Fig. 1. Geographical location of the study area.

Table 1
Image data information.

Data Source	Resolution	Path	Row	Trainable
Landsat 8 OLI	30 m	131	027	True
		132	027	False
		132	027	False

training parameters of different deep learning models and calculated the verification accuracy of the five methods. Subsequently, we proposed some recommendations after analyzing the methods and their accuracy. The workflow of this study is illustrated in Fig. 2.

3.1. Preprocessing

3.1.1. Spectral reconstruction

The collected images were those of the top layer of the atmosphere and required spectral reconstruction (radiation correction). Spectral reconstruction is divided into radiation and atmospheric corrections. Reflectance data were obtained after correction. Where the captured image was an SR image, the image data was stored as an unsigned integer with a radiation resolution of 2^{16} bits or 65,536 bits for convenience. Thus, the image value was between 0 and 65,535. To enhance the generalization ability of the model, we first normalized the image data by dividing them by 65,535.

3.1.2. Label building

The length and width of the Landsat 8 OLI images were 7,731 and 7,841 pixels, respectively. After cropping, 200 non-full background (non-zero) images of 512 × 512 size were retained and used to outline the ROI using the annotation tool, SuperAnnotate. The final image annotation results are shown in Fig. 3. The ROI is a means of labeling important and representative objects in images. Existing deep learning labeling requires distinguishing the categories of each pixel of the entire image.

As shown in Fig. 3, the "water" label represents water bodies, and "others" represents non-water bodies. It is difficult to label all the rivers, as they are scattered and narrow. To coordinate the number of training ROIs for water bodies and non-water bodies, the typical area (including all types of features in the area as much as possible) of non-water bodies was selected in this picture. Using ROI to label the image can focus the labels on accurate water information and representative objects. It can also help improve the accuracy of the labels in low-to-medium-resolution remote sensing images where the water boundary is blurred and inaccurate.

3.1.3. Data augmentation

Because of the small amount of water, the number of "others" pixels is much greater than the number of "water" labels. In this case (the two categories have an obvious difference in the number of levels), in the process of gradient descent of loss, the weight of training "others" will be greatly increased and the loss of "water" labels will be ignored. Therefore, to coordinate the data volume of the two categories, we augmented the data of the pixels with the ROI label of "water. " by repeatedly copying the water samples to increase their numbers, until they balanced up with the "others" samples. The data volume of "water" and "others" before data augmentation were 5,969 and 697,453, respectively. After augmentation, the data volumes were 65,659 and 697,453, respectively.

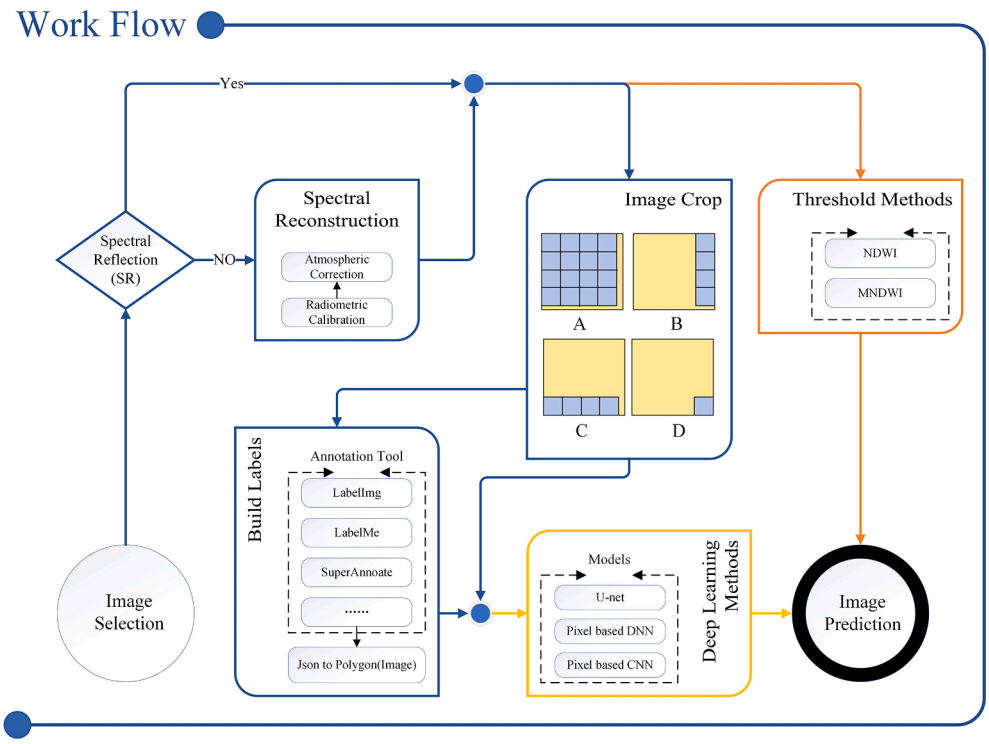


Fig. 2. Flowchart of Water Segmentation.

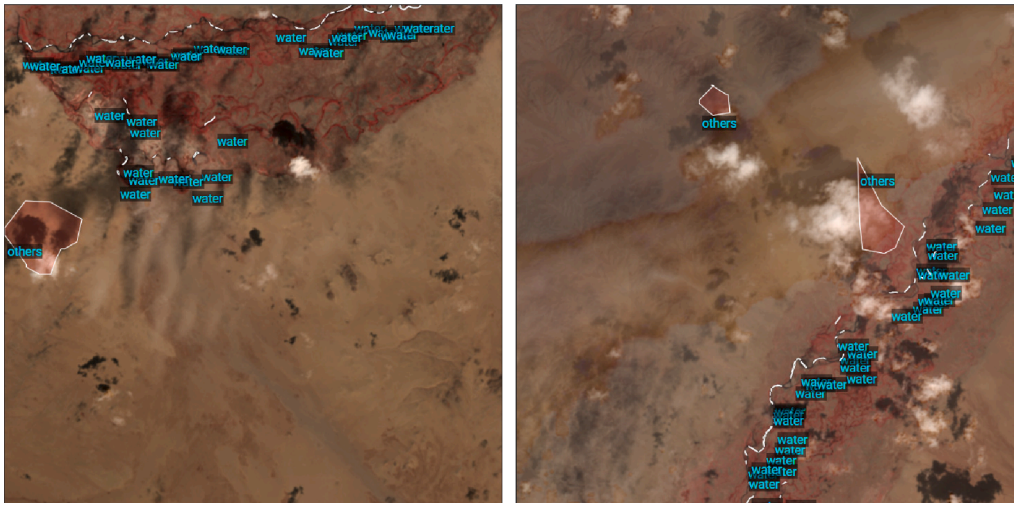


Fig. 3. Annotation examples.

3.2. Threshold classification

In traditional machine learning classification methods, NDWI and MNDWI threshold methods are used to extract water bodies.

The NDWI formula is as follows (McFEETERS, 1996):

$$NDWI = \frac{Grenn - NIR}{Green + Nir} \quad (1)$$

where *Green* is Band 3 in the OLI sensor, and *NIR* is Band 5 in the OLI sensor.

The formula of MNDWI is as follows (Xu, 2006):

$$MNDWI = \frac{Green - MIR}{Green + Mir} \quad (2)$$

where *Green* is Band 3 in the OLI sensor, and mid-infrared (*MIR*) is Band 6 in the OLI sensor. The band paprmaters can be found in Table 2.

Table 2
Landsat 8 OLI band parameters.

Sensor	Band	Wavelength(μm)	Resolution(m)
OLI	Band 1 Coastal	0.433–0.453	30
	Band 2 Blue	0.450–0.515	30
	Band 3 Green	0.525–0.600	30
	Band 4 Red	0.630–0.680	30
	Band 5 NIR	0.845–0.885	30
	Band 6 SWIR 1	1.560–1.660	30
	Band 7 SWIR 2	2.100–2.300	30

3.3. Deep learning classification

Among the deep learning classification methods, this study used the semantic segmentation model U-net, pixel-based DNN model, and pixel-based CNN method to extract water bodies. The sample involves 3 pixel types, "water," "others," and unlabeled pixels (the label values are 01, 10, 00). Therefore, in the training process, the water bodies and others in the label need to be processed in the form of multi-classification (ignoring the ground objects to be classified), and Softmax (Gao and Pavel, 1704) was selected as the activation function to map the output of multiple neurons to 0–1, which can be understood as a probability of each category. Cross-entropy (Murphy, 2012) was used as the loss function to measure the difference between the true label and the predicted value.

Softmax activation function:

$$\widehat{y}_1 = \frac{exp(o_1)}{\sum_{i=1}^{2} exp(o_i)}, \widehat{y}_2 = \frac{exp(o_2)}{\sum_{i=1}^{2} exp(o_i)} \tag{3}$$

where $\widehat{y}_1$ and $\widehat{y}_2$ represent the predicted values of water bodies and non-water bodies, respectively, and $\widehat{y}_1 + \widehat{y}_2 = 1$. o_i is the value before the activation of Softmax, the linear function is non-linearized after the Softmax function is passed, and the predicted value is also normalized.

Cross entropy loss function:

$$L = -\frac{1}{N}\sum_{i}[y_i ln\widehat{y}_i + (1 - y_i)ln(1 - \widehat{y}_i)] \tag{4}$$

where L is the loss value, N is the number of samples, i is the i-th sample, y_i is the actual label value (0 or 1), and $\widehat{y}_i$ is the prediction result. When the labels are all zero, the loss function is zero, and it does not contribute to the loss. Therefore, the unlabeled pixels (values of 00) did not affect the loss.

3.3.1. U-net method

The semantic segmentation model used the U-net neural network model (Ronneberger et al., 2015) as an example. U-Net (Fig. 4) is an end-to-end image-segmentation technology. First, we imported the cropped 512 × 512 images and labeled the images with the ROI in batches. Then, the image features (encoding) are extracted after convolution and pooling (decoding). Subsequently, the images were restored through up-sampling, concatenation, and pooling. Finally, the probability of each category of the restored images was determined using Softmax, and the cross-entropy loss was calculated.

3.3.2. Pixel-based DNN

Fig. 5 shows the structure of the Pixel-based DNN model. The pixel-based DNN model is a deep neural network model. Taking the 7-band-labeled reflectance pixels as input and then passing them through three hidden layers (the neuro number of the three hidden layers are 16, 64, and 128, respectively), the output layer has two neurons. Finally, Softmax calculates the probability of each category of the center pixel, and the cross-entropy function calculates the loss between the prediction and labels.

3.3.3. Pixel-based CNN

Fig. 6 shows the model structure of the pixel-based CNN model. In Fig. 6, @ is used as a special separator to separate the image size and the number of channels. 512 × 512@7 means that the size of the input image is 512 × 512, and 7 means seven channels or bands, the same below. The model traverses the entire image (512 × 512@7) for iterative training. First, the pixels in the 7 × 7 neighborhood around the labeled pixels in the ROI label are used as input (because there are seven bands, the input shape is 7 × 7@7). After two 3 × 3 convolutions, the shape is 3 × 3@32, which becomes 3 × 3@39 after concatenating it with the input center 3 × 3 area. After another convolution and concatenation, the shape was 1 × 1@71. The 1 × 1 convolution and Softmax function are used to determine the category probability of each center pixel. Finally, the label pixel and category probabilities are used to calculate the loss. Because images undergo three times 2 × 2 convolution, traversing the entire cropped images (512 × 512), the shape will become 506 × 506. Therefore, it is necessary to add three zero-padding layers to the images before prediction to ensure that the shape of the masks is the same as that of the images.

4. Results

Fig. 7 shows the results of water extraction from the NDWI, MNDWI, U-net, pixel-based DNN, and pixel-based CNN methods. When there are clouds and small rivers (Fig. 7 lines 1–2), the U-net water mask is the worst, and the rivers are not well extracted. The pixel-based DNN method is slightly better than U-net, but it cannot extract the whole water body. Although NDWI and MNDWI can retain the water body adequately, the cloud body is preserved. The water masks of the pixel-based CNN are the best. It not only eliminates the interference of the cloud, but also extracts the water body completely. In lines 3–4, there is no water in the original satellite images, and NDWI and MNDWI show the worst performance, with many clouds misclassified as water. The pixel-based DNN and U-Net methods showed moderate results. The pixel-based CNN method works best because it does not extract clouds from images. Lines 5–6 select the urban area of Ulaanbaatar, but because the cloud covers the main body of the river, the mask of the river is discontinuous. As mentioned earlier, the pixel-based CNN masks also show the best results and the strongest anti-interference in the construction area. The lake images are selected in lines 7–8, and the threshold methods show poor results (b7, c7), mainly because of the threshold setting problem (if the threshold is 0, the water mask will be

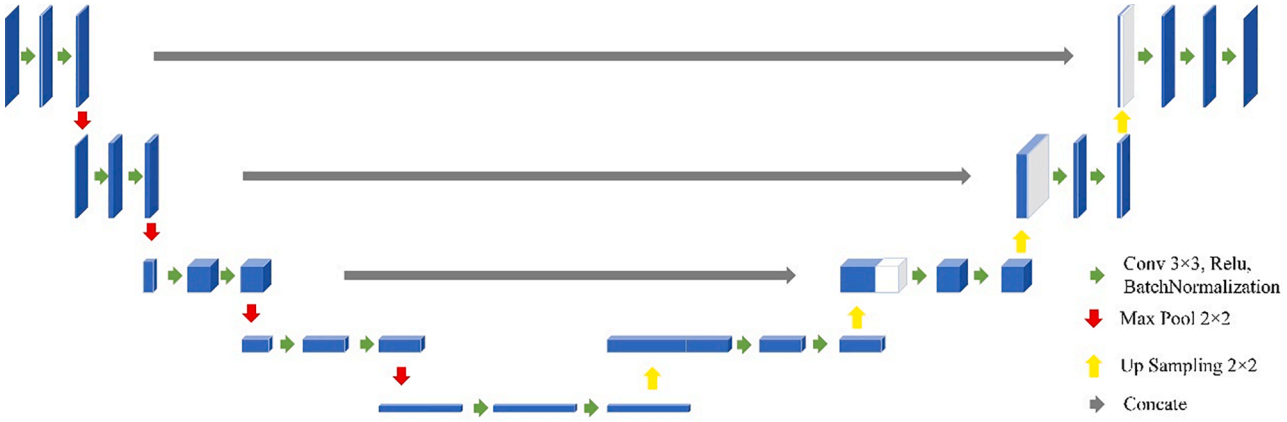


Fig. 4. U-net model structure.

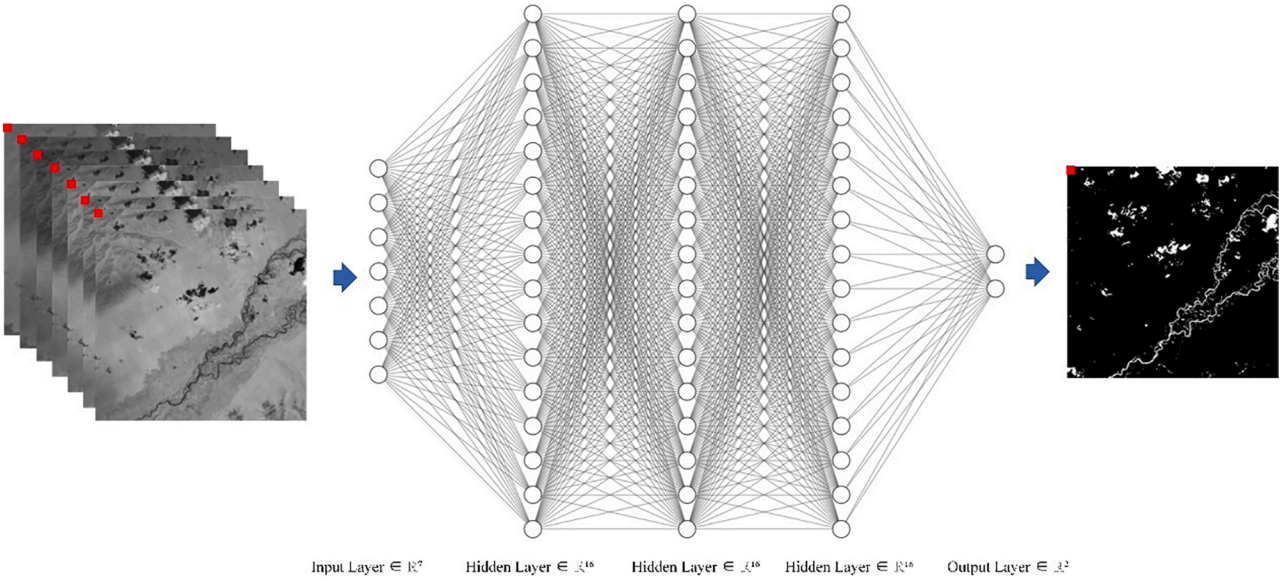


Fig. 5. Pixel-based DNN model structure.

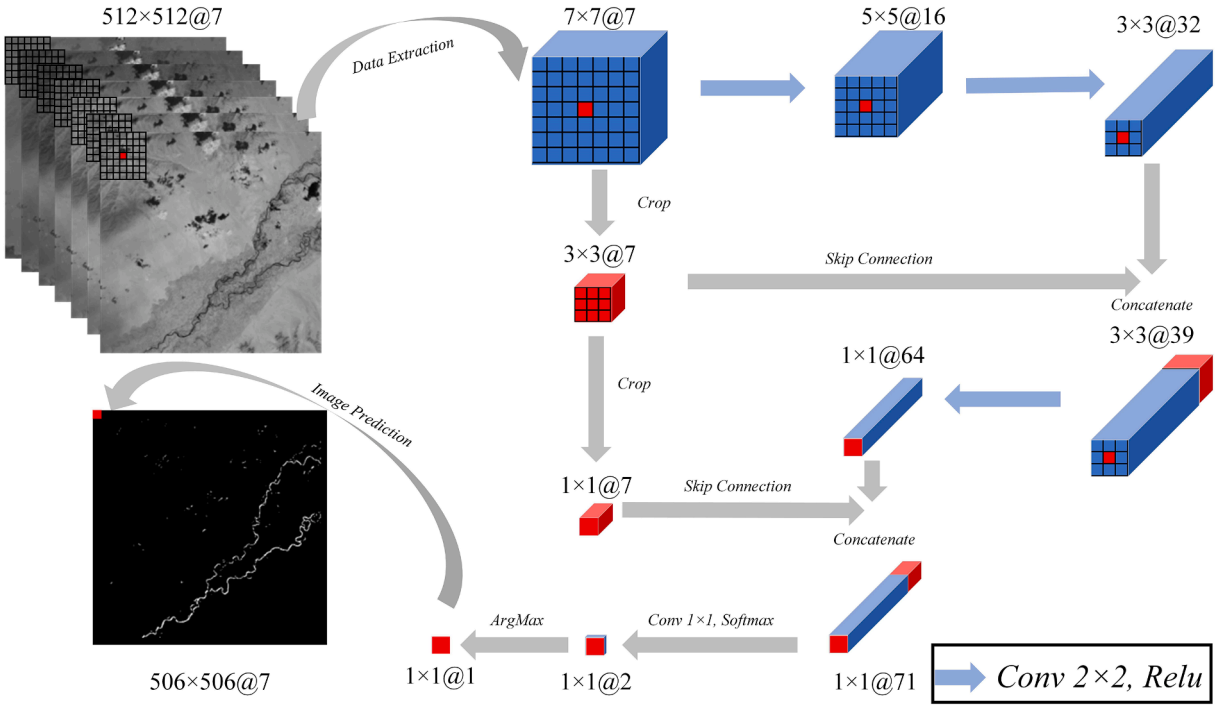


Fig. 6. Pixel-based CNN model structure.

better). In terms of the results of the lakes, the water masks of all the deep learning methods were good. However, the pixel-based CNN method can better distinguish between mountain shadows and water bodies.

In terms of model parameters (Table 3), there are a total of 37,666 training parameters for pixel-based CNN, 9,794 training parameters for pixel-based DNN, and 492,560 parameters, including 491,568 training parameters, for U-net. U-net has the most training parameters, the pixel-based CNN is centered, and the pixel-based DNN is the least. In terms of training time, each iteration of the pixel-based CNN requires the 6 s ranked center. The pixel-based DNN takes the shortest time of 1 s, and the training time for U-net is the longest at 35 s. In terms of accuracy, after 50 iterations of training, pixel-based CNN is the highest at 99.90%, pixel-based DNN is the second-highest at 96.98%, and U-net is the lowest at 93.70%.

The formula for calculating accuracy is as follows:

$$Accuracy = \frac{TP + TN}{TP + TP + FP + FN} \tag{5}$$

The water segmentation accuracy was determined using formulae 5–12. True Positive (TP) is the number of pixels whose label and prediction are both “water,” whereas True Negative (TN) is the pixel whose label and prediction are both “others.” False Positive (FP) represents the number of pixels with the label ”water“ and predicted as ”others,“ and False Negative (FN) represents the number of pixels with the label ”others“ and predicted as ”water.“

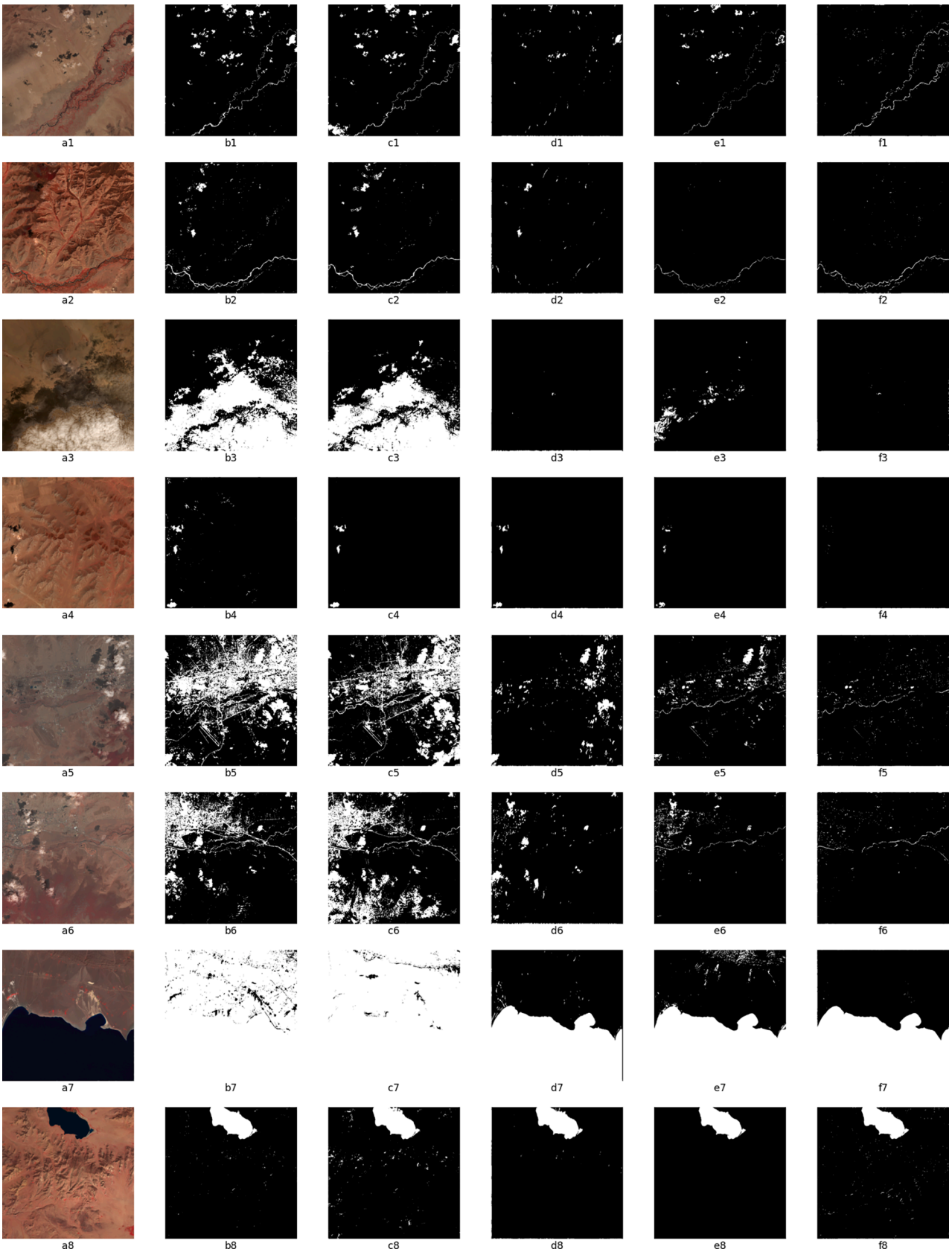


Fig. 7. Water masks in five methods. (a1-a8) Cropped image false color synthesis (RGB: NIR, R, G) linear stretch visualization image, (b1-b8) NDWI water masks (threshold −0.11), (c1-c8) MNDWI water masks (Threshold −0.15), (d1-d8) U-net water masks, (e1-e8) pixel-based DNN water masks, (f1-f8) pixel-based CNN water masks.

Table 3
Comparison of model parameters and accuracy.

	Pixel-based CNN	Pixel-based DNN	U-net
Total params	37,666	9,794	492,560
Trainable params	37,666	9,794	491,568
Non-trainable params	0	0	992
Training Time (/epoch)	6 s 8 ms	1 s 2 ms	35 s 571 ms
Accuracy	99.90%	96.98%	93.70%

The parameters, accuracy, recall, MIoU (Takikawa et al., 2019), TWR (Guo et al., 2020), FWR (Guo et al., 2020) and Kappa (McHugh, 2012) were used for validation. The higher the accuracy, recall, MIoU, TWR and Kappa, means the better the water extraction results. The lower the FWR, the worse the water body extractions. ArcGIS software was used to uniformly select the verification points for the three scene images. A total of 328 verification points were selected, and the number of verification points for water bodies and non-water bodies was 164. Spatial Joining was then used in ArcGIS to process the vectorized water mask file, and the verification accuracy is shown in Table 4. The pixel-based CNN model was superior to other methods in the verification of accuracy, MIoU, TWR, FWR, and Kappa.

Calculated as follows:

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{6}$$

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + TP + FP} \times 100\% \tag{7}$$

$$TWR = \frac{TP}{FP + TP} \times 100\% \tag{8}$$

$$FWR = \frac{FP}{FP + TP} \times 100\% \tag{9}$$

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{10}$$

$$p_o = Accuracy = \frac{TP + TN}{TP + TP + FP + FN} \tag{11}$$

$$p_e = \frac{(TP + TN) \times (TP + FP) + (FP + FN) \times (TN + FN)}{N^2} \tag{12}$$

5. Discussion

Although both NDWI and MNDWI can be used for water body extraction in traditional methods, clouds, cloud shadows, and building areas cannot be effectively distinguished from water bodies. Second, the threshold setting for different scenes is a big challenge (the threshold of a large-area lake scene is different from that of a small river). In the deep learning methods, after training with the same ROI sample, it can be concluded that the representative U-net model of semantic segmentation has the worst performance in the extraction of small rivers, but the extraction of large lakes is good (Figure 7, d7 and d8). Although the pixel-based DNN model performs better than the U-net model in the extraction of small rivers, the rivers are scattered and cannot be extracted completely. While clouds can be avoided by the DNN model, cloud shadows and buildings are still difficult to distinguish from rivers and other water bodies. However, the pixel-based CNN model has the best extraction effect, ensuring the integrity and continuity of small rivers. Clouds, cloud shadows, and building areas will not be misclassified as water bodies, and the extraction results of large-area rivers are better.

Table 4
Validation results.

Methods	NDWI	MNDWI	U-net	Pixel-based DNN	Pixel-based CNN
Accuracy	80.79%	85.97%	75.30%	89.33%	**92.07%**
Recall	**100.0%**	98.36%	94.62%	99.23%	98.59%
MIoU	66.91%	75.09%	59.16%	80.56%	**85.25%**
TWR	61.59%	73.17%	53.67%	79.26%	**85.37%**
FWR	38.41%	26.83%	46.34%	20.73%	**14.63%**
Kappa	0.62	0.72	0.51	0.79	**0.84**

The following reasons can explain why the pixel-based CNN method is the best for water extraction. Traditional methods often ignore texture information and focus mainly on spectral characteristics. However, the main interpreted characteristics are shape, size, color and tone, shadow, position, structure, texture, and resolution. In traditional machine learning methods, because of the different spectral characteristics of different features in most cases, color and tone are usually the most important signs that determine the attribution of features (Wen et al., 2020; PEREIRA and SETZER, 1993; Herold and Roberts, 2005). There is also the phenomenon of 'same thing with different spectra, different things with the same spectrum' (Zhang et al., 2005). Therefore, it is not sufficient to consider only spectral characteristics. In this study, the spectral characteristics of water bodies, clouds, cloud shadows, and building area shadows are similar, so pixel-based DNN (Figure 7 e1, e3, and e5) and threshold methods (Figure 7 b1, c1, b3, c3, b5 and c5) perform not so good in the differentiation of water from dark pixels. In the pixel-based CNN method, the entire (7 × 7) area assists the center pixel to complete the classification, and the concatenation step (Fig. 6) takes into account the regional texture and color characteristics in the center.

Moreover, the semantic segmentation model is not suitable for remote sensing image classification using ROI labels. It is mostly achieved by down-sampling and extracting features, followed by up-sampling to restore images (Long et al., 2015; Badrinarayanan et al., 2017; Ronneberger et al., 2015). Alternatively, it is widely used in medical images (Ronneberger et al., 2015; Zhu et al., 2020; Guo et al., 2020) because the positions of human organs and structures are relatively fixed, while remote sensing images do not have this feature. However, the semantic segmentation model has the problems of large number of parameters, long training time, and difficulty in fitting the training process. Compared with U-net, which represents semantic segmentation method, the pixel-based CNN model is more suitable for remote sensing images, and it has fewer parameters and is easier to train.

Fig. 8 shows the iterative curve of water extraction training based on the pixel-based CNN technology applied to the ROI label. The evolution of the loss is indicated with the red line, and the accuracy (Acc) is indicated with the blue line. In the 50 iterations, the loss decreased from 0.1716 to 0.0031, and was the fastest in the first three iterations. After 10 iterations, the speed of loss slowed, but it steadily decreased. The acceleration of Acc in the first six iterations was fast, and then gradually slows down and stabilized. Acc increased from 94.30% to 99.90%. It can

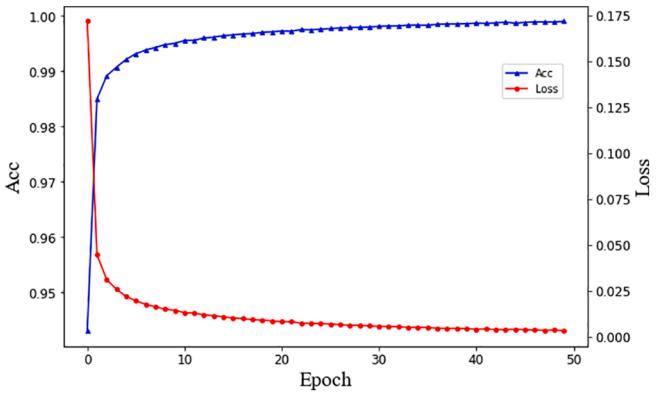


Fig. 8. Pixel-based CNN training iteration curve.

be concluded that the pixel-based CNN water extraction method is efficient and robust, with the ROI as the label.

Fig. 9 shows the water extraction results of the pixel-based CNN method after visualization using ArcGIS. The Tuul River is extracted relatively completely, which twists and turns, passing through the capital Ulaanbaatar to the southwest, and then turning to the northwest to coincide with the provincial borders of Bulgan and Töv.

The pixel-based CNN method can be used for water segmentation in small rivers, which can reduce the interference of dark pixels to a certain extent. However, there are still unresolved problems. First, there are noise points after water segmentation, which requires post-filtering and other operations. Second, Mongolia's land cover type is quite simple, and the effect of water extraction on complex land surfaces needs to be verified. Additionally, directly selecting the seven bands of Landsat 8 as input for training may not be the best band choice. There is a large amount of redundant information for the seven bands. Feature engineering technology can be used for follow-up research to remove redundant bands and improve the accuracy of water body segmentation.

In terms of flood monitoring, with the support of high-quality remote sensing data sources, this method can theoretically achieve high-frequency real-time water (flooding area) calculations. However, in real situations, the key problem lies in the availability of high-quality cloud-free remote sensing data in different flooding situations. For flash floods on the plains, or upstream floods in the downstream regions, this method can provide flood monitoring, because the flooded plain or downstream area is without heavy rain and dense cloud covering. For areas where heavy rains cause floods, this method is limited because it is not easy to provide real-time high-quality optical images in bad weather.

In future work, we hope to find a multi-scale, visible light band (most satellites cover the visible light bands) water extraction method suitable for most satellite images. In addition, this method can be extended to other ground-object classifications.

The models and calling methods have been uploaded to GitHub (https://github.com/CaryLee17/Pixel-based-CNN) in private mode. They can be trained and called in Python.

6. Conclusion

Through a case study of the Tuul River in Mongolia, the effectiveness of machine learning methods for water segmentation with ROI as the label was studied. The following conclusions were drawn: (1) The pixel-based CNN method showed the best masks for water segmentation. NDWI, MNDWI threshold methods, and pixel-based DNN methods have difficulty differentiating water bodies and dark pixels, such as city and cloud shadows. The U-net semantic segmentation also has a poor segmentation result for water bodies. (2) The use of the ROI simplifies label building and ensures the accuracy of the artificial labels. It also makes it possible for low-to-medium-resolution images to be used in deep learning. (3) The pixel-based CNN method processed both spectral and texture information, the training accuracy was 99.90%, and the overall verification accuracy of the three scene images was as high as 92.07%. It had the highest accuracy compared to the other methods. (4) The pixel-based CNN model is superior in terms of training speed and accuracy, in addition to the advantages of few parameters and easy training. In summary, the pixel-based CNN method is an efficient water extraction technology suitable for medium-resolution remote sensing images with ROI labels. The development of these methods will enable large-scale water detection and mapping in the Mongolian Plateau and other remote regions worldwide.

CRediT authorship contribution statement

Li Kai: Conceptualization, Methodology, Formal analysis. **Wang Juanle:** Supervision. **Yao Jinyi:** Validation, Resources, Data curation.

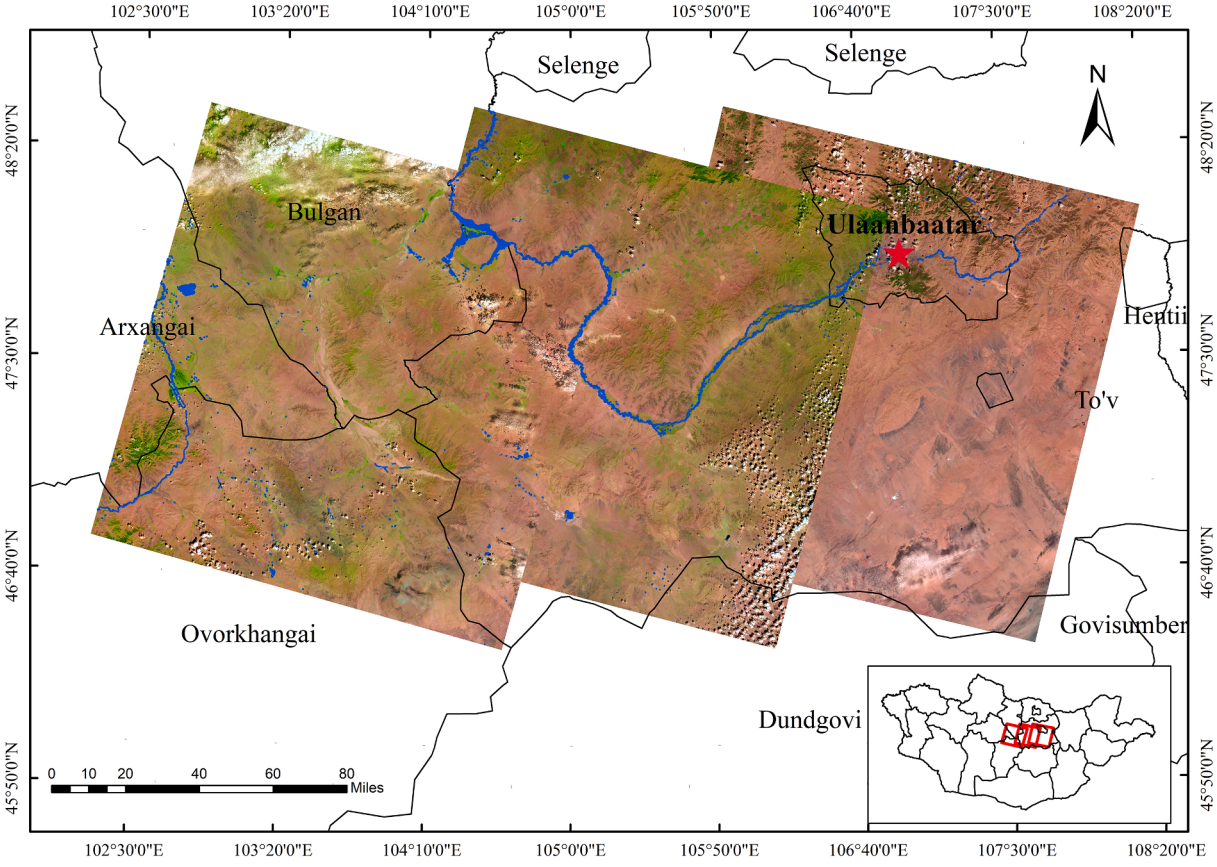


Fig. 9. Water extraction results from three scenes images.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was funded by the National Natural Science Foundation of China (grant number 41971385), the Strategic Priority Research Program (Class A) of the Chinese Academy of Sciences (grant number XDA2003020302), and the Construction Project of the China Knowledge Center for Engineering Sciences and Technology (grant number CKCEST-2021-2-18).

References

Liu, F.; Li, G.; Wang, S.; Fu, H., Lake Water Environment Protection and Management in Mongolia. Journal of Environmental Management College of China 2015, 25 (06), 44-47+74.
Mubareka, S., Maes, J., Lavalle, C., de Roo, A., 2013. Estimation of water requirements by livestock in Europe. Ecosystem Services 4, 139–145.
Mekete, B., 2013. The Livestock-Water Nexus Under Mixed Crop-Livestock Production System. The Livestock-Water Nexus Under Mixed Crop-Livestock Production System.
Nara, Y., Battulga, S., 2019. Observations on Residents' Risk Awareness and Practice of Countermeasures against Natural Disasters in Mongolia: Questionnaire Survey Data of Khovd Citizens. Procedia Comput Sci 159, 2345–2354.
Qu, S., Lv, J.J., Liu, J.X., 2020. Visualization Analysis for Global Water Resources Based on Digital Earth. J Coastal Res 47–50.
McFeeters, S.K., 1996. The use of the normalized difference water index (NDWI) in the delineation of open water features. Int J Remote Sens 17 (7), 1425–1432.
Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. Int J Remote Sens 27 (14), 3025–3033.
Feyisa, G.L., Meilby, H., Fensholt, R., Proud, S.R., 2014. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. Remote Sens Environ 140, 23–35.
Yang, X., Zhao, S., Qin, X., Zhao, N.a., Liang, L., 2017. Mapping of Urban Surface Water Bodies from Sentinel-2 MSI Imagery at 10 m Resolution via NDWI-Based Image Sharpening. Remote Sens-Basel 9 (6), 596. https://doi.org/10.3390/rs9060596.
Fisher, A., Danaher, T., 2013. A Water Index for SPOT5 HRG Satellite Imagery, New South Wales, Australia, Determined by Linear Discriminant Analysis. Remote Sens-Basel 5 (11), 5907–5925.
Yang, X., Chen, Y., Wang, J., 2020. Combined use of Sentinel-2 and Landsat 8 to monitor water surface area dynamics using Google Earth Engine. Remote Sensing Letters 11 (7), 687–696.
Deng, R., Huang, J.F., Wang, F.M., 2011. Research on Extraction Method of Water Body with DS Spectral Enhancement Based on HJ-1 Images. Spectrosc Spect Anal 31 (11), 3064–3068.
Jia, L. J.; Shang, K.; Liu, J.; Sun, Z. Q., Comparison of Water Extraction Methods in Tibet Based on GF-1 Data. Mippr 2017: Remote Sensing Image Processing, Geographic Information Systems, and Other Applications 2018, 10611.
Wentao-Lv; Qiuze-Yu; Wenxian-Yu, Water Extraction in SAR Images Using GLCM and Support Vector Machine. 2010 Ieee 10th International Conference on Signal Processing Proceedings (Icsp2010), Vols I-Iii 2010, 740-743.
Japitana, M.V., Ye, C.-S., Burce, M.E.C., 2019. Combining Water Indices to Detect Water Bodies using Landsat 8 OLI. Journal of Institute of Control, Robotics and Systems 25 (5), 470–475.
Wang, C. H.; Zhao, S.; Ma, R. H.; Tang, W.; Zhang, S. X., Hydrophytes extraction from Landsat TM multi-spectral image in Taihu Lake, China: an approach of decision tree. Geoinformatics 2007: Remotely Sensed Data and Information, Pts 1 and 2 2007, 6752.
He, H.; Zhang, X. N.; Xue, X. W., Water body extraction using MODIS data in the Yangtze River. 2008 Proceedings of Information Technology and Environmental System Sciences: Itess 2008, 2008, 2, 1232-1237.
Li, L.; Yao, Y.; Meng, L. Method for extracting river and lake water edges from remote sensing images based on random forest, involves performing non-target removing process to water-land binary image to obtain target image, and obtaining land binarization image. CN112069938-A.
Wu, C., Du, P., Xia, J., 2012. A Method of Water Extraction Based on Voting Method Fusion for ASTER Remote Sensing Image. Remote Sensing Information 27 (2), 51–56.
Zhao, M.; Shang, H. Z.; Huang, W. C.; Zou, L. Z.; Zhang, Y. J., Water Area Extraction from RGB Aerophotograph Based on Chromatic and Textural Analysis. Proceedings of the Third International Conference on Advanced Geographic Information Systems, Applications, and Services (Geoprocessing 2011) 2011, 46-52.
Hongxia, J.I., Xingwang, F.A.N., Guiping, W.U., Yuanbo, L.I.U., 2015. Accuracy comparison and analysis of methods for water area extraction of discrete lakes. Journal of Lake Sciences 27 (2), 327–334.
Rani, G.M.D., Kapinaiah, V., 2017. In: Extraction of River from Satellite Images. Information & Communication Technology (Rteict), pp. 226–230.
Yousefi, P., Jalab, H.A., W. Ibrahim, R., Mohd Noor, N.F., Ayub, M.N., Gani, A., 2018. Water-Body Segmentation in Satellite Imagery Applying Modified Kernel K-Means. Malays J Comput Sci 31 (2), 143–154.
Zhang, Z.; Prinet, V.; Ma, S., Water body extraction from multi-source satellite images. In IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477), 2003; pp 3970-3972.
Zheng, W., Shao, J.L., Gao, H., 2019. Songhua River Basin Flood Monitoring Using Multi-Source Satellite Remote Sensing Data. Int Geosci Remote Se 9760–9763.
Cao, B.; Kang, L.; Yang, S.; Tan, D.; Wen, X. In Monitoring the Dynamic Changes in Urban Lakes Based on Multi-source Remote Sensing Images, Berlin, Heidelberg, Springer Berlin Heidelberg: Berlin, Heidelberg, 2015; pp 68-78.
Cheng, G., Xie, X.X., Han, J.W., Guo, L., Xia, G.S., 2020. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13, 3735–3756.
Rawat, W., Wang, Z., 2017. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Comput 29 (9), 2352–2449.
Zou, Z., Shi, Z., Guo, Y., Ye, J., 2019. Object Detection in 20 Years. A Survey.
Jiao, L.C., Zhang, F., Liu, F., Yang, S.Y., Li, L.L., Feng, Z.X., Qu, R., 2019. A Survey of Deep Learning-Based Object Detection. Ieee Access 7, 128837–128868.
31. Garcia-Garcia, A.; Orts, S.; Oprea, S.; Villena Martinez, V.; Rodríguez, J., A Review on Deep Learning Techniques Applied to Semantic Segmentation. 2017.
Taghanaki, S.A., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G., 2021. Deep semantic segmentation of natural and medical images: a review. Artif Intell Rev 54 (1), 137–178.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. Proc Cvpr Ieee 3431–3440.
Badrinarayanan, Vijay, Kendall, Alex, Cipolla, Roberto, 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. Ieee T Pattern Anal 39 (12), 2481–2495.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Lect Notes Comput Sc 9351, 234–241.
Li, Liwei, Yan, Zhi, Shen, Qian, Cheng, Gang, Gao, Lianru, Zhang, Bing, 2019. Water Body Extraction from Very High Spatial Resolution Remote Sensing Data Based on Fully Convolutional Networks. Remote Sens-Basel 11 (10), 1162. https://doi.org/10.3390/rs11101162.
Weng, L.G., Xu, Y.M., Xia, M., Zhang, Y.H., Liu, J., Xu, Y.Q., 2020. Water Areas Segmentation from Remote Sensing Images Using a Separable Residual SegNet Network. Isprs Int J Geo-Inf 9 (4).
Li, J.J., Wang, C., Xu, L., Wu, F., Zhang, H., Zhang, B., 2021. Multitemporal Water Extraction of Dongting Lake and Poyang Lake Based on an Automatic Water Extraction and Dynamic Monitoring Framework. Remote Sens-Basel 13 (5).
Zheng, X.X., Chen, T., 2021. High spatial resolution remote sensing image segmentation based on the multiclassification model and the binary classification model. Neural Comput Appl.
Wang, Z.B., Gao, X., Zhang, Y.N., Zhao, G.H., 2020. MSLWENet: A Novel Deep Learning Network for Lake Water Body Extraction of Google Remote Sensing Images. Remote Sens-Basel 12 (24).
Guo, H.X., He, G.J., Jiang, W., Yin, R.Y., Yan, L., Leng, W.C., 2020. A Multi-Scale Water Extraction Convolutional Neural Network (MWEN) Method for GaoFen-1 Remote Sensing Images. Isprs Int J Geo-Inf 9 (4).
Chen, Y., Tang, L.L., Kan, Z.H., Bilal, M., Li, Q.Q., 2020. A novel water body extraction neural network (WBE-NN) for optical high-resolution multispectral imagery. J Hydrol 588.
Dorjsuren, B., Batsaikhan, N., Yan, D.H., Yadamjav, O., Chonokhuu, S., Enkhbold, A., Qin, T.L., Weng, B.S., Bi, W.X., Demberel, O., Boldsaikhan, T., Gombo, O., Gedefaw, M., Girma, A., Abiyu, A., 2021. Study on Relationship of Land Cover Changes and Ecohydrological Processes of the Tuul River Basin. Sustainability-Basel 13 (3).
Soyol-Erdene, T.O., Lin, S., Tuuguu, E., Daichaa, D., Huang, K.M., Bilguun, U., Tseveendorj, E.A., 2019. Spatial and temporal variations of sediment metals in the Tuul River. Mongolia. Environ Sci Pollut R 26 (31), 32420–32431.
Batbayar, G., Pfeiffer, M., von Tumpling, W., Kappas, M., Karthe, D., 2017. Chemical water quality gradients in the Mongolian sub-catchments of the Selenga River basin. Environmental Monitoring and Assessment 189 (8).
Munkhuu, A., Rybkina, I.D., Kurepina, N.Y., 2019. Assessing the Geoecological Status of the Floodplain-Terrace Complex of the Tuul River Within Ulaanbaatar (Mongolia). Geography and Natural Resources 40 (4), 404–412.
United States Geological Survey. https://earthexplorer.usgs.gov/.
Gao, B.; Pavel, L., On the properties of the softmax function with application in game theory and reinforcement learning. arXiv preprint arXiv:1704.00805 2017.
Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.
Takikawa, T., Acuna, D., Jampani, V., Fidler, S., 2019. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. IEEE/CVF International Conference on Computer Vision (ICCV) 2019, 5228–5237.
McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 22 (3), 276–282.
Wen, J., Liu, G., Huang, Y., Pang, Y., Hu, Y., Xua, J., 2020. Canopy Spectral Characteristics Under Different Backgrounds of Wetland Aquatic Vegetation. J Appl Spectrosc+ 87 (1), 62–66.
Pereira, M.C., Setzer, A.W., 1993. Spectral characteristics of fire scars in Landsat-5 TM images of Amazonia. Int J Remote Sens 14 (11), 2061–2078.
Herold, M., Roberts, D., 2005. Spectral characteristics of asphalt road aging and deterioration: implications for remote-sensing applications. Appl Optics 44 (20), 4327–4334.

Bao-Lei, Z., Meng-Qiang, S., Wan-Cun, Z., 2005. Exploration on Method of Auto-Classification for Main Ground Objects of Three Gorges Reservoir Area. Chinese Geogr Sci 15 (2), 157–161.

Zhu, Q.K., Du, B., Yan, P.K., 2020. Boundary-Weighted Domain Adaptive Neural Network for Prostate MR Image Segmentation. Ieee T Med Imaging 39 (3), 753–763.

Guo, D.Z., Jin, D.K., Zhu, Z.T., Ho, T.Y., Harrison, A.P., Chao, C.H., Xiao, J., Lu, L., 2020. Organ at Risk Segmentation for Head and Neck Cancer using Stratified Learning and Neural Architecture Search. Proc Cvpr Ieee 4222–4231.