Artificial Intelligence Methods for Detecting Water Pollution

E P Zharikova¹, J Yu Grigoriev¹, A L Grigorieva¹

¹Faculty of Computing Technologies, Komsomolsk-na-Amure State University, 27 Lenin Prospect, Komsomolsk-na-Amure, 681013, Russian Federation

E-mail: evgenia131094@mail.ru

Abstract. In the modern world, industrial development often negatively affects the environment, including the state of water bodies. Pollution of various types, from thermal to chemical (oil spills, industrial waste dumping and thermometric disturbances), have a detrimental effect on flora and fauna. Continuous monitoring of water areas allows timely detection of pollution. One of the tasks of analyzing the state of water resources is monitoring the water surface and monitoring the coastal zone. The aim of the study is to compare classical approaches based on the application of spectral characteristics and machine learning methods to the analysis of the state of water bodies. The studies show the disadvantages of classical methods of remote sensing in solving problems of autonomous monitoring, consisting in poor resistance to noise and the need for constant expert assessment. The paper presents solutions to the problem of detecting pollution of water bodies using machine learning methods.

1. Introduction

With the active development of the directions of human activity associated with the development of fresh and sea waters, the extraction of resources, the development of the tourist network and others, it becomes necessary to control the consequences of the impact of this activity on the state of water areas. Timely assessment of the surface of the aquatic environment and the state of the coastal zone ensures the possibility of safe use of water resources.

To determine approaches to solving the problem, it is necessary to classify the types of pollution. The following types are distinguished: physical, chemical, thermal, oil.

Physical contamination is long-term decomposing debris (mainly plastic) that emits toxic substances and forms very large areas of stains. Chemical contamination is highly hazardous, often containing mercury and pesticides. Thermal pollution contributes to local changes in water temperature, which can lead to mass death of living organisms and to active reproduction of harmful flora. Oil pollution occurs as a result of tanker wrecks, man-made disasters, well drilling, as well as the discharge of oil products by sea vessels.

For timely detection of pollution and prevention of their consequences, it is necessary to take into account a sufficiently large number of parameters that determine both the nature and source of pollution and the process of its spread.

The aim of the study is to analyze the methods and select the optimal approaches to the design of intelligent systems for monitoring the state of water bodies. The considered methods operate with information from various spectral channels.

The studies involve the assessment of classical methods that allow monitoring based on a fairly small amount of data, and the use of machine learning methods, with the formation of large training samples.
2. Review of approaches to solving the problem
The classical approach for analyzing the aquatic environment will be the use of spectral indices: NDWI (normalize difference water index), NDMI (normalize difference moisture index), NDVI (normalize difference vegetation index). To calculate these indices, the near (NIR), middle (MIR) infrared channels and one of the visible spectrum channels are used. [1], [2]. The water value calculated using such indices is strictly positive or strictly negative. However, such indexes are ineffective in the tasks of searching for water pollution, and the equipment used for surveying in the MIR is expensive.

Figure 1 shows a contaminated body of water captured by Polaris Sensor Technologies, Inc. using multispectral sensors. In this article all raw pictures was created by Polaris sensor technologies, inc.

![Image 1](image1.png)

**Figure 1.** Image of a polluted reservoir in different ranges.

The results shown in Figure 2 show the inability of the water index to accurately identify areas of contamination without a lot of noise and to determine their nature. An index with parameters adjusted for one area cannot be applied to another area with different characteristics of pollution.

![Image 2](image2.png)

**Figure 2.** Water Index results. Wight is a pollution, black is the water.

The image of the ratio of different spectral ranges obtained by calculating the water index is shown in Figure 3, blue color indicates clean water, green and red - pollution.
Figure 3. A graph of the dependence of spectral ranges, with division of zones, for a contaminated surface (red, green - there is contamination, blue - pure water).

Representation of the dependence of spectral channels on a plane together with their division into zones of different states, similar to a graphical representation of the classification problem. The transition to the classification problem allows to optimize the solution, increase the accuracy of data processing due to the absence of the need to calculate classical indices and select threshold values. The paper considers the problem of classification of tabular data: n characteristics of n spectral channels, the state of the environment for a given data set. The article discusses the following methods for solving the problem: the use of fully connected neural networks, the use of bagging and boosting methods.

3. Neurula network
To solve this problem, the architecture of a fully connected neural network is considered, consisting of three layers:
- input (the number of nodes corresponds to the number of input spectral channels);
- hidden layer containing 15 nodes with ReLU activation function (1);

\[
ReLU(x) = \begin{cases} 
0, & \text{if } x < 0 \\
 x, & \text{if } x \geq 0 
\end{cases}
\]  

(1)

- an output layer containing 2 nodes, each output value has a softmax activation function (2).

\[
\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}
\]  

(2)

where K is the number of classes, i - neuron number of the output layer, z - adder value.

The error is calculated using binary cross-entropy, the weights are changed using the adaptive moment estimation (Adam) optimization algorithm.

Neural network parameters are determined experimentally.

At the output of the neural network, the class to which the pixel belongs is determined: water, unknown state. The number of classes can be expanded.

The result of the neural network is shown in Figure 4.
4. Random forest

The random forest algorithm, like the next method, are representatives of ensemble machine learning methods based on “weak learners”, combined into a system to obtain a more reliable model. The decision tree is the basic unit of this algorithm. The relation of an object to a class is formed by most of the solutions obtained by each individual tree.

As parameters of the random forest algorithm, the following are considered: the number of trees, the number of variables in the subsample, the number of observations in the subsample, the minimum number of observations in a node.

The parameters of trees were determined empirically: the depth of the tree is four, the number of randomly selected parameters is two, the number of observations in the subsample is 60% of the total number of observations, while the forest should consist of two trees. These characteristics provide the best performance. The significance of the input values is shown in Figure 6.
The graph shows that, unlike methods based on the use of classical indices (NDWI, MNDWI and others), the random forest method determines the visible spectrum channels as more significant, which may allow in some situations to abandon the use of multispectral cameras in favor of more technically simple analogues. Such actions will improve the performance of the system by reducing the number of processed channels, which allows you to determine the number of sensors during the initial setup of the system.

The results of image processing by the random forest algorithm are shown in Figure 7.

![Variable Importances](image)

**Figure 6.** Input variable importance.

The random forest method, in contrast to the neural network, shows resistance to noise caused by water movement. The best performance of the method may be due to the fact that it can work with a small amount of data in the training set, as opposed to neural networks.

5. **Adaboost**

When using the random forest method, the problem of overfitting may arise; as an alternative, an adaptive acceleration algorithm is considered, which shows good results in problems of binary classification.

The basic classification algorithm and the learning parameter are considered as parameters of the adaptive acceleration algorithm. The study uses a classification based on sibling decision trees with a total of five.

The results of the algorithm are shown in Figure 8.

![Random forest results](image)

**Figure 7.** Random forest results. White is a pollution, black is the water.

The random forest method, in contrast to the neural network, shows resistance to noise caused by water movement. The best performance of the method may be due to the fact that it can work with a small amount of data in the training set, as opposed to neural networks.

5. **Adaboost**

When using the random forest method, the problem of overfitting may arise; as an alternative, an adaptive acceleration algorithm is considered, which shows good results in problems of binary classification.

The basic classification algorithm and the learning parameter are considered as parameters of the adaptive acceleration algorithm. The study uses a classification based on sibling decision trees with a total of five.

The results of the algorithm are shown in Figure 8.
Figure 8. Adaboost results. White is a pollution, black is the water.

The results obtained show not only good resistance to noise, but also the acceleration of data processing by the adaptive acceleration algorithm relative to the random forest method.

6. Conclusion
Since at this stage of the study the problem of binary classification was solved, the metric (3) is used to assess the quality of the results of the application of the above approaches.

\[
\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}
\]

where TP and TN - correct class definitions by the method, FP and FN - incorrect class definitions by the method.

Accuracy value:
- Neural network (14 nodes in each of two hidden layers, activation function - ReLu) (0.21);
- Neural network (10 nodes in a hidden layer, activation function - ReLu) (0.53);
- Neural network (15 nodes in the hidden layer, activation function - ReLu) (0.69);
- Random Forest (0.78);
- AdaBoost (0.81);

The use of classical indices for solving such a problem is not possible due to a number of factors: a separate water index can only indicate the presence or absence of water; for a more accurate characterization of the surface condition, the use of other indices is required, each of which must be adjusted depending on the terrain; the presence of severe pollution near the coastline may go unnoticed due to the absence of a visible water surface. Machine learning algorithms are able to cope with the task, provided there is a sufficiently large sample and its increase during the operation of the system.

The AdaBoost and Random Forest methods show better results compared to the solutions obtained using the water index. Neural networks, in turn, have great potential despite the lower accuracy, which could be due to a number of factors: uniformity of the sample and its small size, the use of a neural network of simple architecture, frame-by-frame data analysis. One of the possible architectural improvements is the use of time series, recurrent neural networks (RNN), the use of dropout in training, and the use of ensembles of neural networks.

7. References
[1] Zharikova E P, Grigoryev J U, Grigoryeva A L 2021 Applications of computer vision in cross-sectoral tasks Lecture Notes in Networks and Systems 200 pp 415–426
[2] Popova O V, Grigoryev J U, Zharikova E P 2020 Lighting Control System Modeling International Multi-Conference on Industrial Engineering and Modern Technologies FarEastCon 2020 9271577
[3] Zharikova E P, Grigoryev J U, Grigoryeva A L 2019 Surface State Detection. International Multi-Conference on Industrial Engineering and Modern Technologies FarEastCon 2019 8934205
[4] Zharikova E P, Grigoryev J U, Grigoryeva A L 2019 Methods of remote sensing in forest fund assessment problems 2019 International Science and Technology Conference "EastConf" EastConf 2019 8725343

[5] Audebert N, Le Saux B, Lefèvre S 2018 Beyond RGB: very high resolution urban remote sensing with multimodal deep networks ISPRS J. Photogramm. Remote Sens 140 20–32

[6] Ball J E, Anderson D T, Chan C S 2017 Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community J. Appl. Remote Sens 11(4) 42609

[7] Belgiu M, Drăguţ L 2016 Random forest in remote sensing: a review of applications and future directions ISPRS J. Photogramm. Remote Sens 114 24–31

[8] Bengio Y, Simard P, Frasconi P 1994 Learning long-term dependencies with gradient descent is difficult IEEE Trans Neural Networks 5 157–166

[9] Bengio Y, Lamblin P, Popovici D, Larochelle H 2007 Greedy layer-wise training of deep networks Proc. Adv. Neural Inf. Process. Syst. 19 153–160

[10] Hamida B A, Benoit A, Lambert P 2018 3-D deep learning approach for remote sensing image classification IEEE Trans. Geosci. Remote Sens 56(8) 4420–4434

[11] Cheng G, Han J 2016 A survey on object detection in optical remote sensing images ISPRS J. Photogramm. Remote Sens 117 11–28

[12] Guo R, Liu J, Li N, Liu S, Chen F, Cheng B, Duan J, Li X, Ma C 2018 Pixel-wise classification method for high resolution remote sensing imagery using deep neural networks ISPRS Int. J. Geo-Inf. 7(3) 110

[13] Zhu L, Chen Y, Ghamisi P, Benediktsson J A 2018 Generative adversarial networks for hyperspectral image classification IEEE Trans Geosci. Remote Sens. 56(9) 5046–5063

[14] Wei Y, Yuan Q, Shen H, Zhang L 2017 Boosting the accuracy of multispectral image pansharpening by learning a deep residual network IEEE Geosci Remote Sens. Lett. 14(10) 1795–1799

[15] Zou Q, Ni L, Zhang T, Wang Q 2015 Deep learning based feature selection for remote sensing scene classification IEEE Geosci Remote Sens. Lett. 12(11) 2321–2325

[16] Bosilj P, Aiptoula E, Lefèvre S, Kijak E 2016 Retrieval of remote sensing images with pattern spectra descriptors ISPRS Int. J. Geo-Inf. 5 228

[17] Cao Y, Wang C, Li Z, Zhang L, Zhang H 2010 Spatial-bag-of-features In: Comput. Vis. Pattern Recognit (CVPR) 2010 IEEE Conf. IEEE pp 3352–3359

[18] Chandrasekhar V, Lin J, Morère O, Goh H, Veillard A 2016 A practical guide to CNNs and Fisher Vectors for instance retrieval Signal Process 128 426–439

[19] Oliva A, Torralba A 2001 Modeling the shape of the scene: a holistic representation of the spatial envelope Int. J. Comput. Vis. 42 145–175

[20] Özkhan S, Ate’s T, Tola E, Soysal M, Esen E 2014 Performance analysis of state-of-the-art representation methods for geographical image retrieval and categorization IEEE Geosci Remote Sens. Lett. 11 1996–2000