Scholar Publishing Group
Water Pollution Prevention and Control Project
https://doi.org/10.38007/WPPCP.2023.040203
ISSN 2790-2730 Vol. 4, Issue 2: 18-25

Machine Learning Algorithm for Background Analysis of Remote Sensing Image Pollution Monitoring

Ilankoon Raymond[*]

Univ Adelaide, Adelaide, SA, Australia

[*]*corresponding author*

Keywords: Remote Sensing Image, Pollution Monitoring, Machine Learning Algorithm, Gray Distribution

Abstract: Traditional water quality detection methods will consume a lot of manpower and material resources. However, remote sensing data and field water quality monitoring data are used to establish water quality remote sensing inversion model to achieve water quality remote sensing monitoring, which makes up for the shortcomings of traditional methods and can monitor the water quality environment comprehensively, quickly and dynamically. However, due to the impact of objective conditions, field water quality data is difficult to obtain in large quantities, so machine learning theory is used, A large number of easily obtained remote sensing image(RSI) data can effectively solve the problem of water quality monitoring. In this experiment, 50 polluted RSIs and 50 unpolluted RSIs are used to extract the gray distribution features of RSIs as the features of the training set. The parameters of GBDT model were optimized, and the accuracy of identifying polluted and unpolluted areas reached 98.67%. In addition, by comparing the performance of three machine learning algorithms, GBDT, DT and SVM, it is found that GBDT algorithm has the highest accuracy in RSI classification. The use of GBDT algorithm enables us to automatically monitor the pollution in RSIs without any marks, which realizes the automation of pollution monitoring.

1. Introduction

Because the potential polluted water body has the characteristics of wide area, large quantity, complex spatial distribution, etc., it is difficult to effectively obtain the status quo of large-scale urban river pollution by only using ground monitoring means, which requires a considerable part of human and material resources in the process, so this task is very difficult. In order to solve this problem, remote sensing technology, which has made great progress in recent years, is the only choice. Monitoring methods based on remote sensing have the advantages of large observation area and fast data acquisition. Therefore, remote sensing technology is widely used in water pollution monitoring [1-2].

China has made some research on remote sensing technology in environmental monitoring and decision-making. For example, some scholars extract water area information based on CCD data of environmental disaster reduction satellite and monitor the flood situation of an island [3]. Some scholars have analyzed and studied flood disasters based on remote sensing images, and discussed how to use remote sensing data from various sources for collaborative flood monitoring and treatment [4]. In order to solve the serious water pollution problem, some researchers use remote sensing information to accurately predict and evaluate water quality resources, and use remote sensing data and appropriate water quality data of external water quality to establish a water quality regression model, which provides main help for analyzing remote sensing water quality data and predicting the changing trend of water pollution [5]. However, compared with the international leading level, China's application in environmental monitoring and monitoring, natural disaster assessment and efficient decision-making is still insufficient, and the monitoring and rapid response to specific scene problems appear insufficient.

This paper first introduces the machine learning algorithm GBDT algorithm and decision tree algorithm, then constructs a water quality monitoring service system based on RSIs, and analyzes the preprocessing methods of RSI data. Finally, it analyzes the application of machine learning algorithm in remote sensing water quality monitoring. By comparing the performance of three classifiers, GBDT, DT and SVM, on RSI classification, and the accuracy of classification of normal water bodies and polluted water bodies, It is verified that the accuracy, sensitivity and other indicators of GBDT algorithm are higher than those of the other two machine learning algorithms, indicating the effectiveness of GBDT algorithm in the field of water pollution monitoring.

2. Machine Learning Related Algorithms

2.1. GBDT Algorithm

GBDT algorithm is an advanced algorithm derived from machine learning which has developed to a higher level. GBDT is essentially an advanced classification algorithm, which is an algorithm based on the basic classifier [6].

GBDT algorithm is a supervised learning algorithm. The model of supervised learning algorithm is established as follows:

$$Obj(\Theta) = L(\Theta) + K(\Theta) \tag{1}$$

Where $L(\Theta)$ represents the loss function, the larger the loss, the worse the fit of the model to the training set, and $K(\Theta)$ represents the canonical term, which penalizes the training results, called the penalty term, to avoid overfitting on the training set.

2.2. Decision Tree(DT)

Building a decision tree requires an understanding of what is meant by "information gain". Information means that if the thing to be classified can be divided into multiple classifications, then the information of the symbol xi is defined as:

$$l(x_i) = -\log_2 p(x_i) \tag{2}$$

"Information gain" is a measure of information entropy. Classifying the confused data by decision tree reduces the information entropy of all data, and the information gain is a small value [7]. When a feature f is introduced, the uncertainty of the whole data, i.e., the information entropy changes to E(S|f), and the original information entropy of the data is E(S), then the information gain is :

$$D(f) = E(S) - E(S \mid f) \tag{3}$$

3. Service Driven Approach For Water Quality Monitoring Events

3.1. RSI Based Water Quality Monitoring Service System

The overall architecture of remote sensing water quality monitoring service system is shown in Figure 1. The system architecture is designed as a three-layer structure consisting of bottom-up data, service and application layers [8]. On this basis, some existing open source software and interfaces are integrated by Mashup. The traditional scientific model of centralized water quality monitoring will be used in the water quality monitoring business, realizing the online service of suspended sediment retrieval thematic map based on remote sensing data, and supporting the acquisition, analysis, publication and visualization of relevant geospatial data and services [9-10]. The system modules involved in the service layer are all based on OGC or relevant industrial standards, which makes the system services have good interoperability, reusability and extensibility [11].

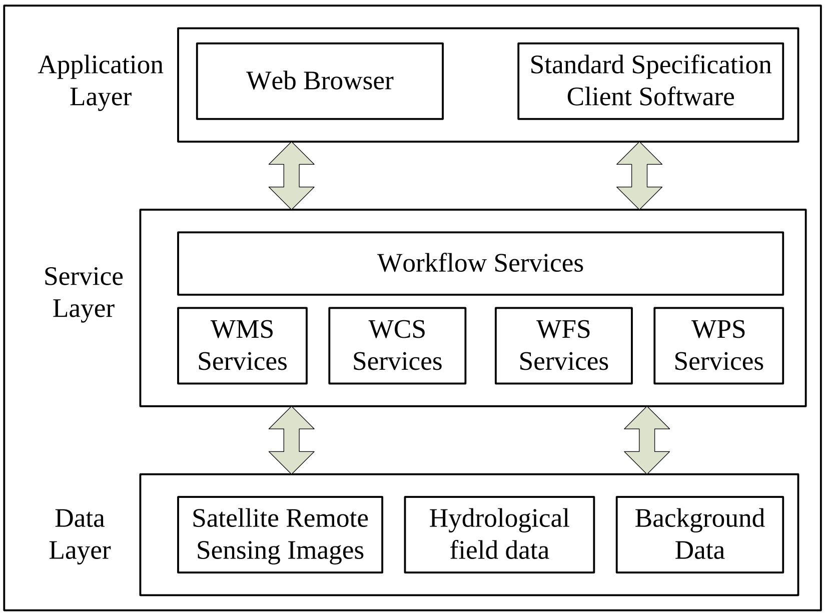

Figure 1. The general framework of the water quality monitoring service system

Figure 1. Overall framework of water quality monitoring service system

The data layer provides geospatial and remote sensing data related to water quality monitoring, including satellite remote sensing images, hydrological maps, administrative divisions and other underlying maps. According to the data type, the data layer can be subdivided into high-resolution remote sensing images, raster data, vector data and static attribute data [12].

Service layer is the core of the whole system, which mainly provides network data services, network processing services and workflow services. The network can be divided into WCS, WFS and WMS services, which are used to support the publication and collection of different data in the distribution network environment within the data layer. In the distributed network environment, WFS is mainly used to publish vector data, WCS is mainly used to publish raster data, and WMS is used to visually represent and realize network processing services, and publish data based on WPS [13-14]. It mainly includes service configuration sub-module, request/return analysis sub-module, data management sub-module and algorithm processing sub-module, which is extended by 52NorthWPS.

The top layer is the application layer that provides links to users and servers. Application types can be divided into web browsers (IE, Chrome, Firefox, etc.) and client software (QGIS, ArcGIS, etc.) that support OGC standard service specifications. Users can send requests directly through web browsers, or perform services in client programs [15-16].

3.2. Automatic Pre-Processing of RSI Satellite Data

Since the original RSI is partially distorted by the sensor itself and the external atmospheric environment in the process of data acquisition, the original RSI needs to be pre-processed so as to restore the spectral characteristics of real features as much as possible [17]. The method of manually correcting RSIs is labor-intensive, time-consuming and labor-intensive, and has great limitations in the face of massive RSIs. Therefore, instead, Python is used to automate and batch the pre-processing process, thus simplifying the time and effort of pre-processing RSI data [18].

4. Remote Sensing Water Pollution Image Monitoring With Machine Learning Algorithm

The purpose of the GBDT algorithm is to classify remotely sensed images, as shown in Figure 2. In the preprocessing process, we segmented the four remotely sensed images into many different small slices and selected 50 polluted slices and 50 unpolluted slices. Then, we extracted features including gray level distribution (GLD), mean value (MV), standard deviation (STD) and new feature information entropy (IFE). Finally, we obtain the output by combining these features into the GBDT classifier.

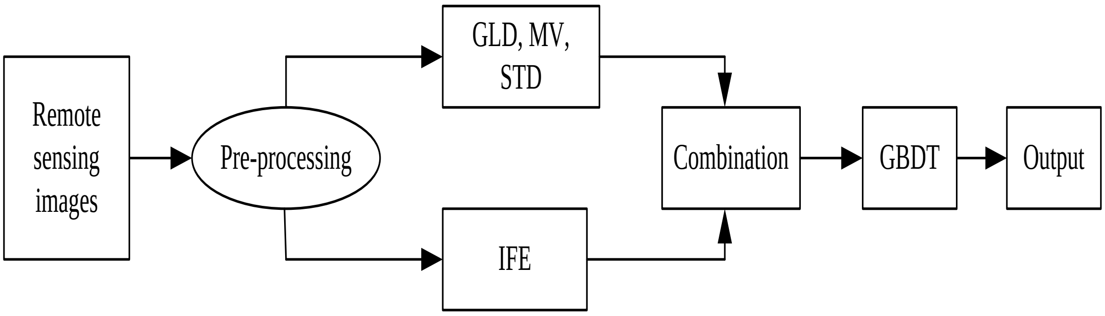

Figure 2. Block diagram of GBDT algorithm steps

The ability to distinguish contaminated and uncontaminated images depends mainly on the quality of the classifier input. In order to capture the RSI features, GLD features of water contaminated RSIs are calculated. Gray scale distribution (GLD) is the gray scale value of pixels in an image is an important feature. There are m*n pixels in each image, and the gray numbers of each pixel range from 0 to 255, with 0 indicating pure black and 255 indicating pure white. There are 256 discrete gray numbers in the interval [0, 255]. The interval is divided into 32 small intervals, each containing 8 gray digits. We calculate the frequency of each interval. Thus, for each individual image, we can get the features of GLD with 32 dimensions.

We did -some experiments on 100 RSIs to verify the validity of RSI feature characteristics and the effectiveness of GBDT method. Figure 1 shows the distribution of GLD.

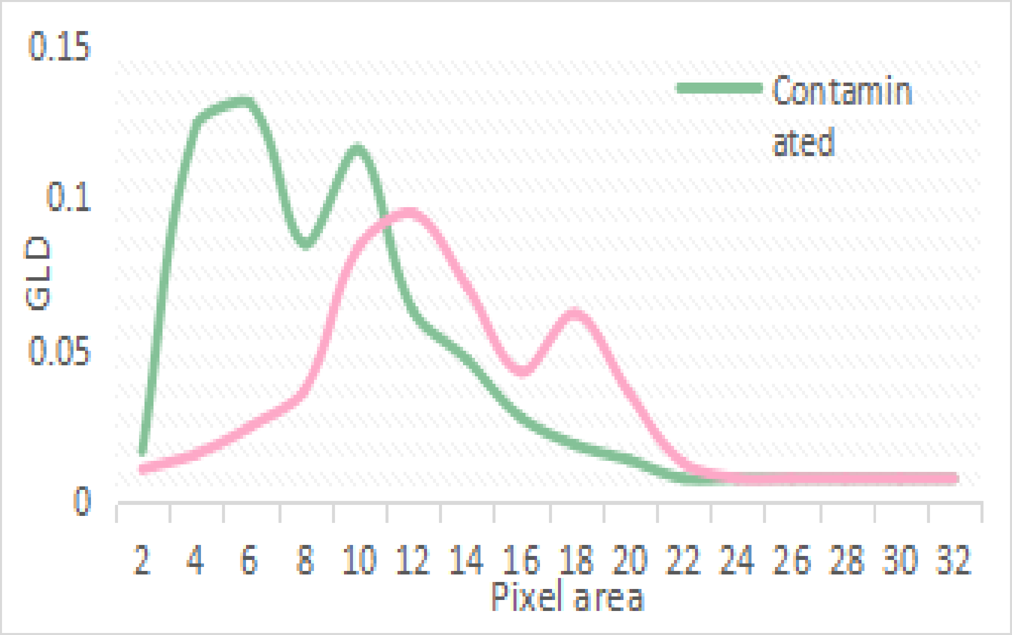

Figure 3. Grayscale distribution of contaminated and uncontaminated samples

Figure 3 shows the GLD of contaminated and uncontaminated samples. we divide the pixels of grayscale values into 32 intervals. the x-axis represents 32 small intervals. the grayscale of 50 contaminated and 50 uncontaminated samples is calculated in 32 intervals. the y-axis represents the distribution of these 32 intervals. The peaks of these two lines are different, and this phenomenon implies that GLD is an important feature to classify contaminated and uncontaminated samples.

An important issue in the automatic classification of RSIs is the development of a high accuracy classifier. For this purpose, we compared the performance of GBDT with other commonly used classifiers (including SVM and DT) under the same conditions (using the same data and feature set).

To reduce the computational time, we also used 5-fold cross-validation for the GBDT model, and obtained the final results by averaging the results generated from the corresponding rounds.

To evaluate the effectiveness of the GBDT model in RSI classification, the classification metrics including precision, sensitivity, specificity, accuracy, recall and F1_score results of the three classifiers were calculated as shown in Table 1. GBDT and the other two methods were implemented by using the python package sklearn which contains many machine learning methods and has been used in every python version has been well established.

Table 1. Performance of different classifiers

	Gbdt	Dt	Svm
Accuracy	98.67%	96.43%	87.92%
Sensitivity	95.33%	91.78%	76.35%
Specificity	100%	98.81%	63.48%
Precision	99.86%	98.42%	90.55%
Recall	96.25%	93.06%	75.79%
F1_score	98.12%	96.57%	83.61%
Time (s)	0.051	0.284	0.224

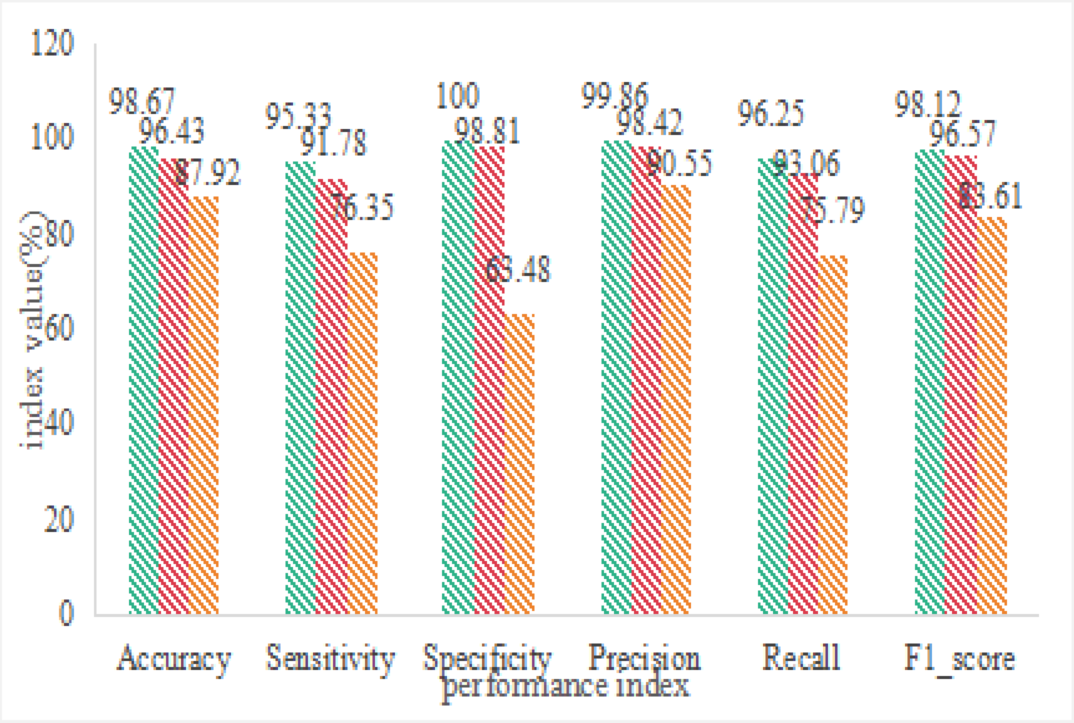

Figure 4. Various performance values of the classifier

Table 1 and Figure 4 show that the overall performance of GBDT is the best of the three classifiers. The precision of GBDT can reach 98.67%, 2.24 percentage points higher than DT and 11 percentage points higher than SVM. The last row in the table shows the running time of each classifier. The GBDT method reduces the running time to 0.051 seconds, which is 5.57 times faster than DT and 4.39 times faster than SVM. We can conclude that the GBDT classifier is superior to the other two methods. In general, the performance of SVM and DT is poor, because they do not use aggregate strategy. We noticed that the specificity and accuracy of GBDT were very high, reaching 100.00% and 99.86%. This may be caused by insufficient input data (only 100 input data sets). Among the three classifiers, SVM has the worst performance, with each index only between 75% and 90%.

Table 2. Water body classification accuracy of machine learning algorithm

	Precision		Recall		F1_score	
	Normal water	Polluted water	Normal water	Polluted water	Normal water	Polluted water
GBDT	1.00	0.018	0.98	0.95	0.98	0.037
DT	1.00	0.007	0.91	0.96	0.95	0.014
SVM	1.00	0.004	0.87	0.90	0.91	0.006

In Table 2, the three precision indicators of normal water body are very high. However, on the contrary, due to the misjudgment of normal water bodies, the accuracy of polluted water bodies is surprisingly low, such as the accuracy of GBDT, DT and SVM algorithms are 0.018, 0.007 and 0.004 respectively; F1_ The scores are 0.037, 0.014 and 0.006 respectively. The three machine learning algorithms misjudge normal water bodies in areas where water quality is obviously poor, such as ponds and irrigated farmland. Although the water body in these areas is not black and smelly, poor water circulation and organic pollutants in these areas all lead to poor water quality in these areas.

5. Conclusion

The problem of water pollution is widely concerned all over the world. In order to protect the ecological stability of the water environment, this paper uses remote sensing technology to obtain RSIs of water bodies, and then analyzes the image features through machine learning algorithms to achieve water quality safety monitoring. After comparing the performance indicators of GBDT, DT and SVM in water quality RSI classification, it is found that the GBDT classifier has the shortest running time and the highest accuracy in water quality monitoring, and the GBDT algorithm has the highest classification accuracy for normal water bodies and polluted water bodies, which reflects the efficiency and accuracy of GBDT algorithm in RSI water quality monitoring.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Priyanka, N. Sravya, Shyam Lal , J. Nalini, Chintala Sudhakar Reddy, Fabio DellAcqua: DIResUNet. Architecture for multiclass semantic segmentation of high resolution RSIry data. Appl. Intell. (2022) 52(13): 15462-1 5482. https://doi.org/10.1007/s10489-022-03310-z

[2] Tran Manh Tuan, Tran Thi Ngan, Nguyen Tu Trung. Object Detection in Remote Sensing lmages Using Picture Fuzzy Clustering and MapReduce. Comput. Syst. Sci. Eng. (2022) 43(3): 1241-1253. https://doi.org/10.32604/csse.2022.024265

[3] K. Kala, N. Padmasini, B. Suresh Chander Kapali, P. G. Kuppusamy. A new framework for object detection using fastcnn- Naive Bayes classifier for RSI extraction. Earth Sci. Informatics. (2022) 15(3): 1779-1787. https://doi.org/10.1007/s12145-022-00834-3

[4] A. Azhagu Jaisudhan Pazhani, S. Periyanayagi. A novel haze removal computing architecture for RSIs using multi-scale Retinex technique. Earth Sci. Informatics. (2022) 15(2): 1147-1154. https://doi.org/10.1007/s12145-022-00798-4

[5] Samuel A. Ajila, Chung-Horng Lung, Anurag Das. Analysis of error-based machine learning algorithms in network anomaly detection and categorization. Ann. des Telecommunications. (2022) 77(5-6): 359-370. https://doi.org/10.1007/s12243-021-00836-0

[6] Koushiki Dasgupta Chaudhuri, Bugra Alkan. A hybrid extreme learning machine model with harris hawks optimisation algorithm: an optimised model for product demand forecasting applications. Appl. Intell. (2022) 52(10): 11489-11505. https://doi.org/10.1007/s10489-022-03251-7

[7] Paul D. Rosero-Montalvo, Vivian F. Lopez Batista, Ricardo P. Arciniega-Rocha, Diego Hernan Peluffo-Ordonez. Air Pollution Monitoring Using WSN Nodes with Machine Learning Techniques: A Case Study. Log. J. IGPL. (2022) 30(4): 599-610. https://doi.org/10.1093/jigpal/jzab005

[8] Davut Ari, Baris Baykant Alagoz. An effective integrated genetic programming and neural network model for electronic nose calibration of air pollution monitoring application. Neural Comput. Appl. (2022) 34(15): 12633-12652. https://doi.org/10.1007/s00521-022-07129-0

[9] Ekta Dixit, Vandana Jindal. IEESEP: an intelligent energy efficient stable election routing protocol in air pollution monitoring WSNs. Neural Comput. Appl. (2022) 34(13): 10989-11013. https://doi.org/10.1007/s00521-022-07027-5

[10] Yiannis N. Kontos, Theodosios Kassandros, Konstantinos Perifanos, Marios Karampasis, Konstantinos L. Katsifarakis, Kostas D. Karatzas. Machine learning for groundwater pollution source identification and monitoring network optimization. Neural Comput. Appl. (2022) 34(22): 19515-19545. https://doi.org/10.1007/s00521-022-07507-8

[11] Juan Jesus Roldan-Gomez, Pablo Garcia Aunon, Pablo Mazariegos, Antonio Barrientos. SwarmCity project: monitoring traffic, pedestrians, climate, and pollution with an aerial robotic swarm. Pers. Ubiquitous Comput. (2022) 26(4): 1151-1167. https://doi.org/10.1007/s00779-020-01379-2

[12] Mowva Pavani, K. Kishore Kumar. Large scale air pollution monitoring using static multi-hop wireless sensor networks. Int. J. Comput. Aided Eng. Technol. (2021) 15(2/3): 294-305 . https://doi.org/10.1504/IJCAET.2021.117139

[13] Pau Ferrer-Cid, Jose M. Barcel6-Ordinas, Jorge Garcia-Vidal. Graph Learning Techniques Using Structured Data for IoT Air Pollution Monitoring Platforms. IEEE Internet Things J. (2021) 8(17): 13652-13663. https://doi.org/10.1109/JIOT.2021.3067717

[14] Huber Flores, Naser Hossein Motlagh, Agustin Zuniga, Mohan Liyanage, Monica Passananti, Sasu Tarkoma, Moustafa Youssef, Petteri Nurmi. Toward Large-Scale Autonomous Marine Pollution Monitoring. IEEE Internet Things Mag. (2021) 4(1): 40-45. https://doi.org/10.1109/IOTM.0011.2000057

[15] Swati Chopade, Hari Prabhat Gupta, Rahul Mishra, Preti Kumari, Tanima Dutta. An Energy-Efficient River Water Pollution Monitoring System in Internet of Things. IEEE Trans. Green Commun. Netw. (2021) 5(2): 693- 702. https://doi.org/10.1109/TGCN.2021.3062470

[16] Aayushi Gautam, Gaurav Verma, Shamimul Qamar, Sushant Shekhar. Vehicle Pollution Monitoring, Control and Challan System Using MQ2 Sensor Based on Internet of Things. Wirel. Pers. Commun 11. (2021) 6(2): 1071-1085. https://doi.org/10.1007/s11277-019-06936-4

[17] Vahid Sadeghi, Hossein Etemadfard. Optimal cluster number determination of FCM for unsupervised change detection in RSIs. Earth Sci. Informatics. (2022) 15(2): 1045-1057. https://doi.org/10.1007/s12145-021-00757-5

[18] Rajni Sharma, M. Ravinder, Nitin Sharma, Kanchan Sharma. An optimal RSI enhancement with weak detail preservation in wavelet domain. J. Ambient Intell. Humaniz. Comput. (2022) 13(4): 1941-1 952. https://doi.org/10.1007/s12652-021-02957-9