Integrating various satellite images for identification of the water bodies through using machine learning: A case study of Salah Adin, Iraq

Ahmed Fouad Nashait¹, O Z Jasim², M M Ismail³ and F H Saad⁴

¹Lecturer Ahmed Fouad Nashait, Civil Engineering Department, University of Technology.
²Prof. Dr. Oday Zakariya Jasim, Civil Engineering Department, University of Technology.
³Lecturer Muayad Mohammed Ismail, Civil Engineering Department, University of Technology.
⁴Assist chief of engineers Fouad Hamdan Saad, Ministry of Labors and Social Affairs.

Email: 40122@uotechnology.edu.iq

Abstract. This research aims to extract water bodies from several types of satellite images by using machine learning. There are several methods to apply the extraction of information about water bodies such as unsupervised classification, supervised classification. This project applied the supervised classification method to extract water bodies and building geodatabase for the water bodies in Salah Adin, Iraq by applying remote sensing and GIS technique. The satellite images which have been used in this research include satellite images from RapidEye satellite with spatial resolution (5x5) m at 2011, where these images used to extract canals of water, also used images from Sentinel-2 satellite with spatial resolution (10 x 10) m at 2017 to extract lakes, and broad rivers. In addition, DEM raster (90) m at 2011 from ASTER satellite was used to extract the water streams, which are expected to be channels of water discharge in flood situations. During implementation of the methodology of the research the most important issue which appear is using multi data to extract all types of water bodies in study area also to avoid the pixel mixed problem, this case was evident when using RapidEye satellite images in the confluence of the river in the surrounding wetlands, which led to inaccurate results in the geometric dimensions of the river, although the high spatial resolution but the influenced element on the accuracy of results is radiometric resolution. The result shows the random forest of machine learning algorithm is overcome on the other algorithms such as decision tree machine learning, maximum likelihood and support vector machine. The high accuracy of image classification to extract the water bodies depend on integrating the three satellite images.

Keywords: Remote sensing, geographic information system, supervised classification method, unsupervised classification method, random forest, decision tree, maximum likelihood and support vector machine.

1. Introduction and Literature Review

Water is considered a vital component of the life cycle in the nature (Jing et al., 2018; Ghimire et al., 2017; Vaghefi and Yu, 2017; Shi et al., 2015; Zhao et al., 2012). Since the issue of water is one amongst the necessary topics that have great impact on human life (Seligman et al., 2014; Vörösmarty et al., 2010; Vitousek et al., 1997), so this has acquired great attention by researchers over the last years as the problem of water has become an international problem, which needs study and preparation of spatial databases and accurate description of water bodies. Water body is generally an accumulation of water on the Earth (Ameliet al., 2018; Lu et al., 2011; Alsdorfer et al., 2007; Hem, 1985). They are ocean, seas, lakes, rivers, ponds and others. It is very intractable to extract water body.
by traditional survey methods. Therefore, remote sensing (RS) records ground objects in a comprehensive way. Water observed by RS images include not only lakes, rivers, ponds, but also marshes and canals. The appropriate methodology to apply this type of study by employing the remote sensing science and GIS.

Nowadays, satellite images are the essential data for producing and modernizing cartographic and geographic data. The geomatics engineering stands out appropriate methods to acquire accurate results by utilizing GIS techniques to employ satellite images to extract the features of water bodies. The classification of satellite images technique by using GIS is considered as a base for other hydrology studies, which is a widespread topic in the image processing dealing with remote sensing data which include land cover, land use and change detection studies.

Several studies were used to extract the water bodies by using the satellite images such as Mishra et al, 2015 was used perceptron model to extract the water bodies automatic from Landsat ETM imagery in Hyderabad city, India. The perceptron model includes supervised classification based on the linear function predators through used some characteristic properties of the object in the Landsat image. All these features were combined to gather to extract the water bodies. The proposed method accurately and quickly discriminated water from no water features.

In 2015 Yang et al, also was used Landsat ETM imagery to extract water body through tacked sparse auto encoder (SSAE) one deep learning method. This study used SSAE with NDWI, NDVI, NDBI to extract the water body by collecting the construct unique feature matrix for each pixel of study area. After applying these methods, other methods such as Support Vector Machine (SVM) and traditional artificial neural network (ANN) were also applied. The results showed the proposed method was outperformed the SVM) and ANN.

While Acharya et al., 2016 was used J48 decision tree (JDT) with Landsat 8 OLI image to identify the water bodies in the Northern Han River Basin, Korea. Pixels of image were separated to 7070 (70%) and 3065 (30%) pixels for training and testing. This model was used as input such as density slicing, NDWI, MNDWI. The accuracy of JDT model was outperformed SVM.

Although the researchers tried to improve the extraction of water bodies based on models only, they success to extract water body from different satellite images was used. In this paper the authors focused on the satellite images by integrating three satellite images to improve the extraction of the water bodies. On the other hand, the methodology based on the machine learning such as decision tree and random forest, and compare the results with traditional methods such as maximum likelihood and support vector machine.

2. Study area and datasets
2.1 Study area
The area of study (Salah Adin province) is in the north of Baghdad province also in the south of Ninava province, Erbil, Kirkuk, Diyala, Anbar provinces east and west, respectively between (42°-45°E) and (33°-35°N), with about (26220) km² in area. As shown in figure(1).
2.2 Data sets

In this study three satellite images were used and two images are available on USG website. They are sentinel satellite with (10*10) m (2014) as spatial resolution, and also RapidEye satellite with (5*5) m (2014) as spatial resolution and the DEM raster (90) m from ASTER satellite at 2014. These images need reprocessing to enhance, correct the geometry image and correct the reflectance and radiometric of images in order to solve all problems before starting with integration of images.

3. Method

The extraction of information is considered as the most important step in digital image processing, which involves data acquisition, pre-processing and image display and enhancement as shown in 'figure 2, 3, 4 and 5'. After finishing all the steps to treat the images, the current step is from digitizing the samples on the proposed image which has high spatial resolution 5 m with four bands (red, gray, blue and nearinfrared). This study used two extra raster as input to the classification method, the normalized difference water index (NDWI) to remove portion of background noise but fault built-up land for water bodies and the modified normalized difference water index (MNDWI) to differentiate between buildings and water. Commonly the NDWI raster and MNDWI raster were used to extract the water bodies and called about them raster is “Water index”. On the other hand, water index raster is useful for extracting water bodies through the vital element, which influence the accuracy of the water bodies extraction, whereas some small water bodies not fully extracted such as small canals so the water index raster can help about that. Now is ready to apply classification algorithms for this proposed image by using several algorithms such as decision tree, random forest, maximum likelihood and support vector machine. The reason for using this image, to extract canals of
water and lakes, broad rivers and the water streams through the satellite images RapideEye, sentinel-2 and DEM of ASTER, respectively.

**Figure 2.** Add RapidEye data to ArcMap program.  **Figure 3.** RapidEye Raster after mosaic process.

**Figure 4.** RapidEye Raster after Copy Raster.  **Figure 5.** Sentinel Raster after clip process.

In this study, the classification method based on supervised classification and several software such as ArcGIS were used to produce the mapping and all the raster needed to use on proposed model, ENVI software was utilized to reprocess the images. Lastly, the Weka software for applying classification algorithms. On the other hands, the accuracy assessment based on the overall accuracy and kappa accuracy to examine and check which is the best algorithm to get high accuracy assessment of classification.
4. Results

4.1 Classification image:
After applying the four algorithms to classify the water bodies, we found the random forest algorithm is the best and surpasses the other algorithms such as diction tree, support vector machine and maximum likelihood. Whereas, the overall accuracy and the kappa accuracy of random forest algorithm is 0.982, 0.973 and 0.969, 0.946 for training and testing data, respectively. While, the overall accuracy of the other algorithms is (0.978, 0.967), (0.967, 0.953) and (0.918, 0.904) for training and testing of diction tree, support vector machine and maximum likelihood, respectively as given in table 1.

Table 1. The accuracy of diction tree, support vector machine and maximum likelihood classifiers.

| Classifier                  | Overall accuracy | kappa accuracy |
|-----------------------------|------------------|----------------|
|                             | Training         | Testing        | Training | Testing |
| Random Forest Model         | 0.982            | 0.973          | 0.969    | 0.946   |
| Diction Tree Model          | 0.978            | 0.967          | 0.972    | 0.957   |
| Support Vector Machine Model| 0.967            | 0.953          | 0.954    | 0.941   |
| Maximum Likelihood Model    | 0.918            | 0.904          | 0.909    | 0.883   |

4.2 Extraction of the features of water bodies:
In section we describe how to extract the water bodies as shape file format, after getting the classification image from the best algorithms which was the random forest algorithm. Also in this study, the DEM was used to apply flow direction of the river network (streams) through the Flow Direction algorithm, which was available in the spatial analysis tool. After that, the raster is converted to polygon in order to extract the shape of water bodies by applying Raster To Polygon algorithm, which was available in the ArcGIS software as shown in 'figures 6, 7, 8, 9 and 10'.

Finally, layout process is applied to produce the map of water bodies as shown in 'figure 11'.
**Figure 6.** Reclassify process results on RapidEye data.

**Figure 7.** Conversion Raster to Polygon process results from RapidEye data.
Figure 8. New Feature Class in ArcCatalog.

Figure 9. Copy and Paste processes results.

Figure 10. Contour Line process results
5. Conclusions

While applying the classification technique on satellite images to extract the water bodies class a problem of mixing pixel occurred, wherein RapidEye satellite images due to the low value of radiometric resolution being equal to (8 bit) although the high spatial resolution being equal to (5x5)m but that led to extract only canals in study area while the sentinel-2 satellite images led to extract lakes and broad rivers due to high radiometric resolution being equal to (12 bit) with low spatial resolution being equal to (10x10) m. Therefore, some objects may not appear in the classification process and sometimes there are overlaps between topographic classes, for example wet areas near the water bodies impact the shape accuracy of water bodies, that need appropriate selection of the type of satellite images and compatibility with the nature of the project to be achieved, also multi-source satellite images need to be used to extract maximum possible types of water bodies for the study area.

The water streams extracted from DEM data practically represented the flow path of water in case of the flood, where that is considered as an assistant element in the management of water by preparing an overview for specialists about the technique of water drainage, collection, and utilization. Finally, it is very important to take into consideration the radiometric resolution for satellite images during apply automatic classification methods, which influence the accuracy of results. Therefore, that requires adequate data selection.

Reference

[1] Acharya T, Lee D, Yang I & Lee J 2016 Identification of water bodies in a Landsat 8 OLI image using a J48 decision tree. Sensors, 16(7), p 1075
[2] Alsdorf D E, Rodriguez E & Lettenmaier D P 2007 Measuring surface water from space. Reviews of Geophysics, 45(2)
[3] Ameli A A, Gabrielli C, Morgenstern U & McDonnell J J 2018 Groundwater subsidy from headwaters to their parent water watershed: A combined field-modeling approach. Water Resources Research, 54(7), pp 5110-5125
[4] Ghimire S R, Johnston J M, Ingwersen W W & Sojka S 2017 Life cycle assessment of a commercial rainwater harvesting system compared with a municipal water supply system. *Journal of cleaner production*, 151, pp 74-86

[5] Hem J D 1985 Study and interpretation of the chemical characteristics of natural water Vol. 2254. Department of the Interior, US Geological Survey.

[6] Jing X, Zhang S, Zhang J, Wang Y, Wang Y & Yue T 2018 Analysis and modelling of stormwater volume control performance of rainwater harvesting systems in four climatic zones of China. *Water resources management*, pp 1-16

[7] Lu S, Wu B, Yan N, & Wang H (2011). Water body mapping method with HJ-1A/B satellite imagery. *International Journal of Applied Earth Observation and Geoinformation*, 13(3), pp 428-434

[8] Mishra K & Prasad P 2015 Automatic extraction of water bodies from Landsat imagery using perceptron model. *Journal of Computational Environmental Sciences*, 2015

[9] Seligman M E & Csikszentmihalyi M 2014 Positive psychology: An introduction. In Flow and the foundations of positive psychology pp. 279-298 Springer, Dordrecht

[10] Shi, Tiange, Xiaolei Zhang, Hongru Du, and Hui Shi. "Urban water resource utilization efficiency in China." *Chinese geographical science* 25, no. 6 (2015): pp 684-697

[11] Vaghefi P & Yu B 2017 Validation of CLIGEN Parameter Adjustment Methods for Southeastern Australia and Southwestern Western Australia. *Journal of Hydrometeorology*, 18(7), pp 2011-2028

[12] Vitousek, Peter M, Harold A Mooney, Jane Lubchenco and Jerry M. Melillo 1997 Human domination of Earth's ecosystems *Science* 277, no. 5325 pp 494-499

[13] Vörösmarty C J, McIntyre P B, Gessner M O, Dudgeon D, Prusevich A, Green P, ... & Davies P M 2010. Global threats to human water security and river biodiversity. *Nature*, 467(7315), 555.

[14] Yang L, Tian S, Yu L, Ye F, Qian J & Qian Y 2015. Deep learning for extracting water body from Landsat imagery. *Int. J. Innov. Comput. Inf. Control*, 11, pp 1913-1929

[15] Zhao Z & Xu H 2012 Study on the Supplying System of Cooling Water of Air Conditioner based on the Urban Street Rainwater. *Energy Procedia*, 16, pp 8-13