Machine Learning Research
2018; 3(2): 11-17
http://www.sciencepublishinggroup.com/j/mlr
doi: 10.11648/j.mlr.20180302.11
ISSN: 2637-5672 (Print); ISSN: 2637-5680 (Online)

GIS-Based Analysis of Changing Surface Water in Rajshahi City Corporation Area Using Support Vector Machine (SVM), Decision Tree & Random Forest Technique

Mahbina Akter Mim, K. M. Shawkat Zamil

Department of Computer Science and Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh

Email address:
mahbinaaktermim@gmail.com (M. A. Mim), kmszamil@gmail.com (K. M. S. Zamil)

To cite this article:
Mahbina Akter Mim, K. M. Shawkat Zamil. GIS-Based Analysis of Changing Surface Water in Rajshahi City Corporation Area Using Support Vector Machine (SVM), Decision Tree & Random Forest Technique. *Machine Learning Research*. Vol. 3, No. 2, 2018, pp. 11-17. doi: 10.11648/j.mlr.20180302.11

Received: July 22, 2018; **Accepted**: August 1, 2018; **Published**: September 3, 2018

Abstract: Water is one of the essential natural resources of nature. All living creature depends on water. Living creatures are using water for their different purposes. Earth's large portion is covered by salt water but very less has fresh water. Freshwater can be found as groundwater and surface water. Surface water is stored as waterbodies on the surface of this world. Ponds, canals, rivers, and lakes are some of the waterbodies that provide fresh water to us. These waterbodies are fulfilling our need for fresh water. Most of the waterbodies are drying up for natural disasters or they are continuously filling by humans. These resources need some of our attention to preserve them. Rajshahi Development Authority (RDA) and United States Geological Survey (USGS) provide important data for this research. Waterbodies are detected by using Geographic Information System (GIS), GIS gives us the power of mapping and store, detect, and manipulate spatial or geographic data. Images are collected from the Landsat 4-5 Thematic Mapper (TM) and Landsat 8 Operational Land Imager (OLI). They are classified by using ArcGIS. Images are classified in maximum likelihood classification by generating signature files to extract feature. Percentage of waterbodies in each year is calculated from the attribute table. A dataset is prepared from these features and tested on different classification techniques. Support Vector Machine (SVM), Decision Tree and Random Forest Technique are implemented on this dataset. Among them, Random Forest shows 92% accuracy, which is better from other techniques. These algorithms also measure the precision, recall, and f1 scores of the classifiers. The precision, recall, and f1-score of random forest technique show 0.943, 0.920, 0.922, which indicate better accuracy than other techniques.

Keywords: Waterbodies, GIS, Remote Sensing, ArcGIS, Maximum Likelihood Classification, Support Vector Machine (SVM), Decision Tree, Random Forest

1. Introduction

Water is an important element of our nature. All living creatures like humans, animals, plants depend on water. We generally use water for drinking, cooking, washing clothes, irrigation, in industries and so on. Without lifting groundwater, surface water can be used to fill the demand of people's need. But many waterbodies are filling in an unplanned way for different reasons. Many waterbodies are continuously polluting by various harmful human wastes and industrial wastes. The earth's hydrosphere consists of a large amount of water, about, 1,386 cubic kilometers (km^3). 97.5% of this water is salt water and from that only 2.5% is stored as fresh water. The greater portion of fresh water (68.7%) is in the form of ice and permanent snow in the Antarctic, Arctic and mountain region. 29.9% fresh water is stored as groundwater and only 0.26% of fresh water is concentrated in lakes, reservoirs and river systems [1].

The main attention of this research is to find the change of waterbodies in Rajshahi City Corporation (RCC) area and by making a dataset test that dataset on different classification techniques. To find the change of waterbodies Geological Information System (GIS) is used. ArcGIS is used to classify images in maximum likelihood classification to extract features. Preparing a dataset, Support Vector Machine (SVM), Decision Tree and Random Forest technique are implemented to test the dataset.

For indiscriminate earth dumping and unplanned urbanization, almost 4,000 ponds are filled in the past few decades. Rajshahi city used to have 4,238 ponds, canals, and wetlands in 1961 which has been decreased to 2,271 in 1981 and in the year 2000 the number was 729. That means the number is decreasing rapidly. Now, this city has only 214 waterbodies [2].

Bangladesh has a huge population. The area of Rajshahi City Corporation is 95.56 square kilometer, and it has a total population of 3,88,811. The number of the male is 2,08,525 and female is 1,80,286. The Padma river is the main waterbody of Rajshahi City Corporation area [3]. As the population is increasing, the demand for water also increases but waterbodies are being destroyed at an alarming rate. Because of the widespread use of surface water for the increasing population, the groundwater lifting is increasing. The volume of groundwater storage is decreasing. Depletion of groundwater is causing many wells to dry up. It also causes water in streams and lakes to reduce, deterioration of water quality and an increase of arsenic contamination in drinking water [4]. Now it has become an important issue. We should take some steps to preserve these waterbodies and maintain a healthy nature not only for us but also for the future generation.

2. Literature Review

Some of the studies are on surface water like arsenic contamination area marking by using Geographic Information System (GIS), classification of Hyperspectral Remote Sensing images using Support Vector Machines (SVMs). The effects of change in waterbodies of Rajshahi City Corporation (RCC) area of past few decades is found by surveying. But GIS or other classification techniques are not used for finding or detecting the change of waterbodies. Most of the research is on the fluctuation of groundwater, groundwater pollution using GIS and the quality of surface water and groundwater [5, 6].

George, Geeja K., et al. studied the groundwater pollution in an industrial area in Chavara Taluk in Kollam district [6]. This research was on groundwater, and surface water around the study area was affected due to effluents from the industry. Water samples were analyzed, and affected areas were marked using GIS because GIS is not only useful for data capture and processing but also a powerful computational tool that facilitates multimap integrations.

Melgani, Farid, and Lorenzo Bruzzone used Support Vector Machines (SVMs) on hyperspectral remote sensing images and assessed the effectiveness of SVMs concerning the conventional feature-reduction-based approaches and their performances in hyper subspaces of various dimensionalities, applied binary SVMs to multiclass problems in hyperspectral data [7]. Different performance indicators were used in this study. The result was obtained on a real Airborne Visible/ Infrared Imaging Spectroradiometer hyperspectral dataset, and SVMs was a valid and effective alternative to conventional pattern recognition approaches for the classification of hyperspectral remote sensing dataset.

3. Methodology

For finding the change of surface water of Rajshahi City Corporation area using GIS, first, we collect images from USGS EarthExplorar. Multispectral imagery usually has 3 to 10 bands and each band is obtained by using remote sensing radiometer. In ArcGIS, the images are classified to maximum likelihood classification by generating signature files. Areas are calculated from the attribute table and values are converted to acre by using field calculator because the area in Bangladesh is generally calculated in an acre. Features are extracted from maximum likelihood classification. Finding the percentage of waterbodies, this dataset is classified into three different groups. This dataset is used to classify in Support Vector Machine (SVM), Decision Tree and Random Forest technique.

Different percentage of accuracy is measured because each of the technique is different from each other.

3.1. Image Collection

Images of Rajshahi City Corporation (RCC) area are downloaded from 1987 to 2016 for preparing a dataset. These images are collected from the United States Geological Survey or USGS EarthExplorer. The shapefile of the study area is collected from Rajshahi Development Authority (RDA). Landsat 4-5 Thematic Mapper (TM) images and Landsat 8 Operational Land Imager (OLI) images are collected for image classification. All images are not of the same type because Landsat 4 was terminated on December 14, 1993, and Landsat 5 terminated on June 5, 2013 [8]. Landsat 4-5 TM has seven different bands and Landsat 8 OLI has 11 different bands.

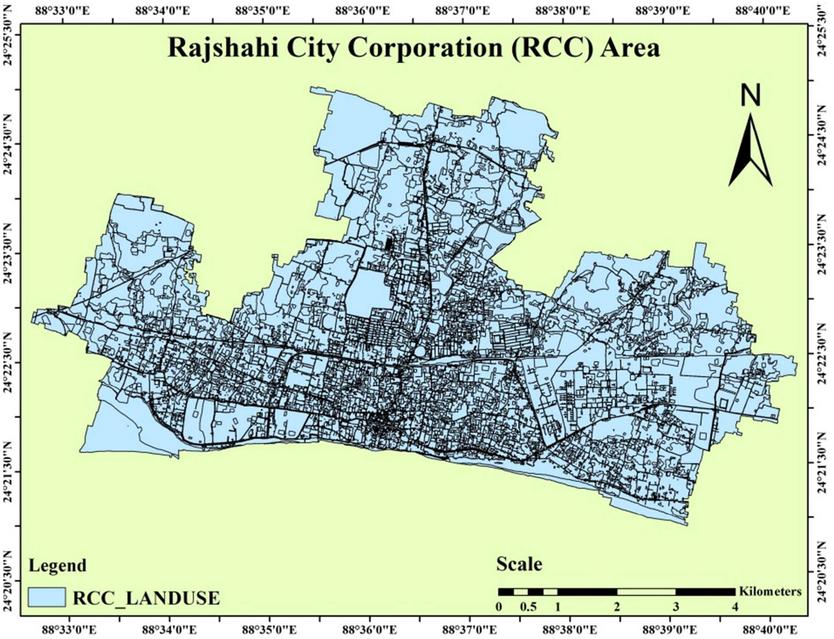

Figure 1. *The study area of Rajshahi Division.*

3.2. Image Classification

To classifying the images, ArcGIS tool is used. Data classification can be done using three different techniques; these are supervised, unsupervised classification and object-based analysis. Generating a signature file the images are classified in maximum likelihood classification. Maximum likelihood is a very popular method for remote sensing. It has some advantages such as it can be developed with a variety of estimation situations and the method is generally used for mathematical and optimality properties [9].

Maximum likelihood classification is a supervised classification technique. When instances are given with known labels, they are called supervised learning [16]. This can be done in ArcGIS. This can be done is four basic steps [17].

1. Firstly, enable image analysis toolbar from ArcMap.
2. Secondly, select training areas by drawing polygons which denote the specific areas. For training areas, we have to multi-select polygons and merge into a single class.
3. Thirdly, generate a signature file by merging and renaming accordingly.
4. Finally classify using maximum likelihood classification, iso cluster, class probability or principal component. Each one of them has its own advantages. From these techniques, maximum likelihood classification is used.

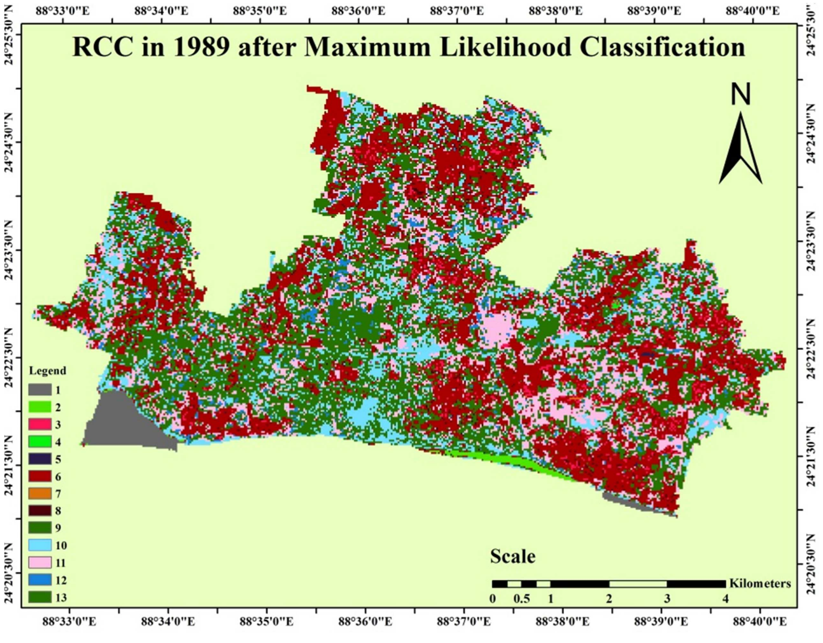

Figure 2. *Study area in 1989 after using maximum likelihood classification.*

For many advantages, this technique has been used to get the values of waterbodies, vegetation, urban area and open space.

3.3. Feature Extraction

For feature extraction, the classified images are used in reclassification. Each image has four features, and they are waterbodies, vegetation, urban area and open space. In Figure 3 red color is an urban area, the blue color is waterbodies, the green color is a vegetation area, and the yellow color is for open space. These have been detected by changing the color bands of the images. After reclassifying each image the values of the attribute table are calculated in square kilometer and then in the acre. Values are converted to the acre as the area of Bangladesh in generally measured in an acre.

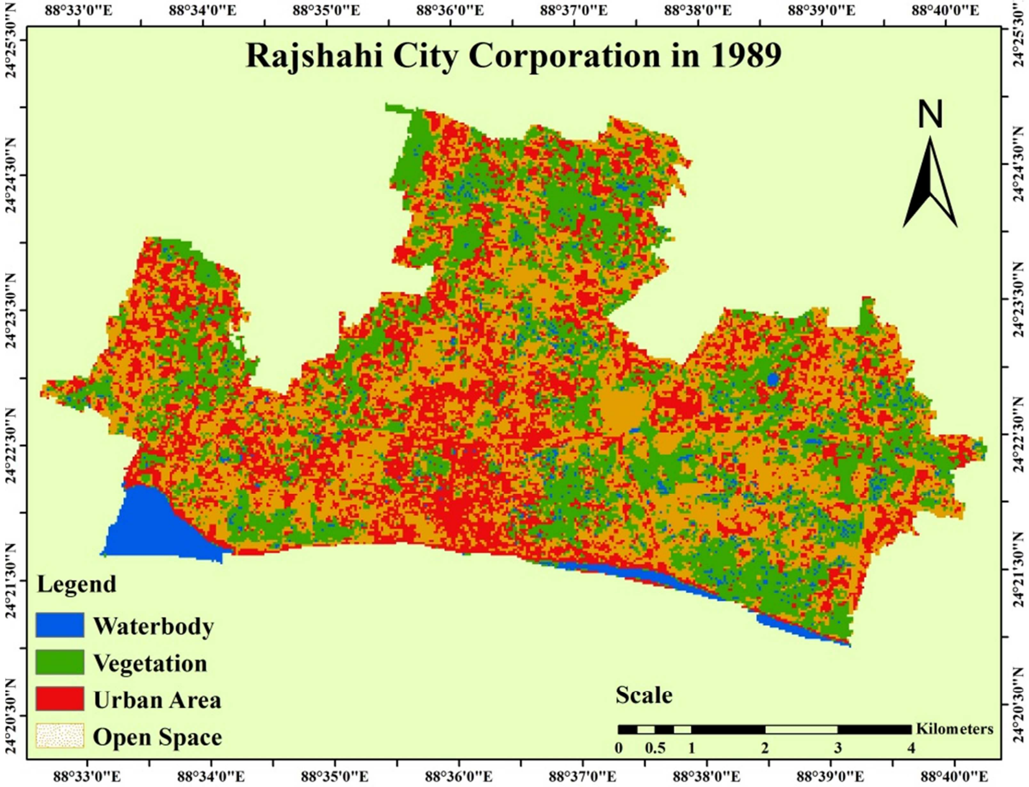

Figure 3. *Study area in 1989 after reclassify.*

3.4. Dataset Preparation

After reclassifying all the images from 1987 to 2016, a dataset is generated. This dataset has 25 values. This dataset has 25 rows and 6 columns.

3.5. Classifier

For classifying this type of dataset Support Vector Machine (SVM), Decision Tree and Random Forest technique has been used.

3.5.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a biased classifier that is defined by a separable hyperplane. In a two dimensional space, the hyperplane is a line dividing a plane into two parts. These two parts lay on either side of the hyperplane [11]. Generally, SVM and neural networks give better performance in dealing with multiclass and continuous features, and logic-system performs better regarding discrete values [16].

It is a supervised machine learning algorithm that can be used in classification or regression but mostly used in classification. Standard Support Vector Machines are designed for dichotomic classification problem, but multi-class classification problem is solved by decomposition to several binary problems for which standard SVM can be used [10]. For instance, one-against-all decomposition is normally applied. Different types of kernel tricks are used in SVM.

A training set of instance-label pairs (x_i, y_i) , $i = 1,2,3, \dots \dots \dots l$ where $x_i \in R^n$ and $y \in \{1, -1\}^l$, the SVM require the solution of the following optimization problem:

$$\min_{d,\, b,\, \xi,} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

$$\text{subject to } y_i(w^T \emptyset(x_i + b) \geq 1 - \xi_i \qquad (1)$$

$$\xi_i > 0$$

By the function $\emptyset$ training vectors x_i are mapped into a higher dimensional space. In the higher dimensional space SVM finds a linear separating hyperplane with the maximal margin. For the error term $C > 0$ the penalty parameter. The kernel function is formed as $K(x_i, x_j) \equiv \emptyset(x_i)^T \emptyset(x_j)$. There are four basic kernels:

1. Linear:

$$K(x_i, x_j) = x_i{}^T x_j \tag{2}$$

2. Polynomial:

$$K(x_i, x_j) = (\gamma x_i{}^T x_j + r)^d, \gamma > 0 \tag{3}$$

3. Radial Basis Function (RBF):

$$K(x_i, x_j) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \gamma > 0 \tag{4}$$

4. Sigmoid:

$$K(x_i, x_j) = \tanh(\gamma x_i{}^T x_j + r) \tag{5}$$

These are the kernels of SVM and here γ, r and d are kernel parameters [12].

3.5.2. Decision Tree [13]

To determine suitable property for each node of a generated decision tree, information gain approach is used. The attribute that has the highest information gain is selected as the test attribute of the current node. Use of the property to partition the sample contained in the current node makes the mixture degree of different types for all the generated sample subsets reduce to a minimum.

S is a set that includes s a number of the data sample. These types of attributes can take m potential different values corresponding to m different types of C_i where $i = 1,2,3, \dots\dots\dots\dots\dots, m. S_i$ the sample number of C_i. The amount of information to classify a given data is,

$$I(s_1, s_2, \dots..s_m) = -\sum_{i=1}^{m} p_i \log(p_i) \tag{6}$$

Where the probability is $P_i = S_i/|S_j|$ which is any subset of data samples belonging to categories C_i. Where S_j contains the data sample whose attribute A are equal a_j in S set.

Consider that A is the property which has v different values $\{a_1, a_2, \dots\dots\dots, a_v\}$. By using the property of A, S can be divided into v different number of subsets $\{S_1, S_2, \dots\dots, S_v\}$. If property A is selected for test that is used to make a partition for current samples, suppose that S_{ij} is the sample set of type C_i in the subset S_i, the information entropy is,

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}, s_{2j}, \dots\dots, s_{mj}}{s} I(s_{1j} s_{2j} \dots\dots s_{mj}) \tag{7}$$

The obtained information gain is,

$$Gain(A) = I(s_1, s_2, \dots\dots, s_m) - E(A) \tag{8}$$

3.5.3. Random Forest

Random Forest is a supervised classification algorithm. It is an effective tool in prediction because of the Law of Large numbers they do not overfit. By injecting the right kind of randomness, it can be made accurate classifiers and regressors. Random features and random inputs produce excellent results in classification but less in regression [14].

Instead of one decision tree, random forest uses a collection of decision trees [15]. Θ denotes the set of possible attributes, and h (x, Θ) denotes a tree grown using Θ to classify a vector x. By using the above notations the random forest f can be defined as,

$$f = \{h(x, \Theta_k)\} \tag{9}$$

Where $k = 1,2,3, \dots\dots, K$ and $\Theta_k \subseteq \Theta$. It means that the random forest is a collection of trees where a tree is grown with a subset of possible attributes. For k_{th}tree Θ_k is randomly selected and it is independent from the past random vectors $\Theta_1, \Theta_2, \dots\dots, \Theta_{k-1}$.

Random Forest performs better than a single decision tree. This can utilize unnecessary features and the independence of the different classifiers (trees) use.

4. Experiments

4.1. Dataset Description

In this work, there was no prepared dataset. So we have to prepare a dataset from all the classified images in ArcGIS. This dataset has mainly six attributes. It has been possible to collect 25 images, so the dataset has 25 values.

Table 1. *Description of the dataset.*

Attribute	Description	Values
Waterbodies	Amount of area that is covered by water	Continuous value in Acre
Vegetation	Amount of area that is covered by greenery	Continuous value in Acre
Urban Area	Amount of area that is covered by buildings	Continuous value in Acre
Open Space	Amount of area that is open or empty	Continuous value in Acre
Percentage of waterbodies	Percentage of waterbodies in each year	Values in percentage
Class	For different amount range values differ	Multiclass value 0-5 = 0 value 5.1-10 = 1 value 10.1 - above = 2

This dataset is built by classifying all the collected images from USGS in ArcGIS. This paper aims to implement this dataset in different classification techniques.

4.2. Experimental Setup

To conduct this experiment, we have three different classifiers namely Support Vector Machine (SVM), Decision Tree and Random Forest. The experiment is performed based on five-fold cross-validation as the dataset consists of only 25 values. For measuring the performance, we have selected the split ratio = 0.6 that means 60% data is used for training and 40% data is used for testing.

4.3. Result and Discussion

Precision, recall, and f1-score are calculated because accuracy measurement is not enough for evaluating the performance of any classification algorithm. The confusion matrix is a performance indicator that gives information about the correctly and incorrectly classified instances number for each classifier. As we have three outcomes, each classifier generates a 3X3 confusion matrix. There are four elements in the confusion matrix and they are:

1. True Positive (TP): The total number of positive predictive instances that are correctly and truly positive.
2. True Negative (TN): The total number of negative predictive instances that are classified correctly and truly negative.
3. False Positive (FP): The total number of predictive instances that are classified incorrectly and truly negative.
4. False Negative (FN): The total number of negative predictive instances that are classified incorrectly and truly positive.

For three different classes the assumptions are:

1. The sum of the corresponding row would be the total number of test examples of any class that means, the TP+FN for that class.
2. The sum of values in the corresponding row (excluding the TP) is the total number of FN's for a class.
3. The sum of values in the corresponding column (excluding the TP) is the total number of FP's for a class.
4. The sum of all columns and rows (excluding that class's column and row) will be the total number of TN's for a certain class.

The following table shows the confusion matrix for 3 classes.

Table 2. Confusion matrix for three class.

	Predicted			
		A	B	C
Actual	A	TP_A	E_{AB}	E_{AC}
	B	E_{BA}	TP_B	E_{BC}
	C	E_{CA}	E_{CB}	TP_C

The following table shows the confusion matrix of different classifiers for 3 different classes.

Table 3. Confusion matrix of different classifiers.

Classification	Confusion matrix			Accuracy
Support Vector Machine (SVM)	TP_A (0)	E_{AB} (5)	E_{AC} (0)	72%
	E_{BA} (0)	TP_B (17)	E_{BC} (0)	
	E_{CA} (2)	E_{CB} (0)	TP_C (1)	
Decision Tree (ID3)	TP_A (4)	E_{AB} (0)	E_{AC} (1)	88%
	E_{BA} (1)	TP_B (16)	E_{BC} (0)	
	E_{CA} (1)	E_{CB} (0)	TP_C (2)	
Random Forest	TP_A (5)	E_{AB} (0)	E_{AC} (0)	92%
	E_{BA} (1)	TP_B (16)	E_{BC} (0)	
	E_{CA} (1)	E_{CB} (0)	TP_C (2)	

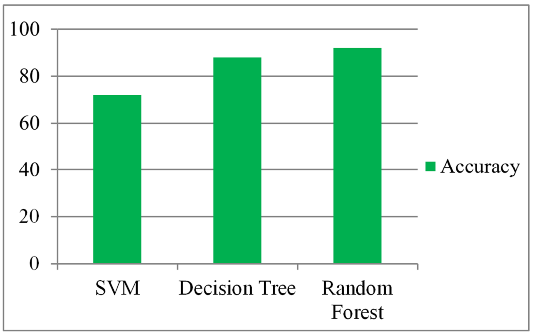

Figure 4. Graphical representation of accuracy for each classifier.

Figure 4 shows the graphical view of the accuracy of Support Vector Machine (SVM), Decision Tree and Random Forest techniques.

In three class confusion matrix:

Precision A = TP_A / (TP_A+E_{BA}+E_{CA})
Precision B = TP_B / (TP_B+E_{AB}+E_{CB})
Precision C = TP_C / (TP_C+E_{AC}+E_{BC})
Recall A = TP_A / (TP_A+E_{AB}+E_{AC})
Recall B = TP_B / (TP_B+E_{BA}+E_{BC})
Recall C = TP_C / (TP_C+E_{CA}+E_{CB})

Classification report for each classifier is given below:

Table 4. Classification report of classifiers of three classes.

Classification	Precision	Recall	F1-score
Support Vector Machine (SVM)	0.645	0.720	0.653
Decision Tree (ID3)	0.893	0.880	0.885
Random Forest	0.943	0.920	0.922

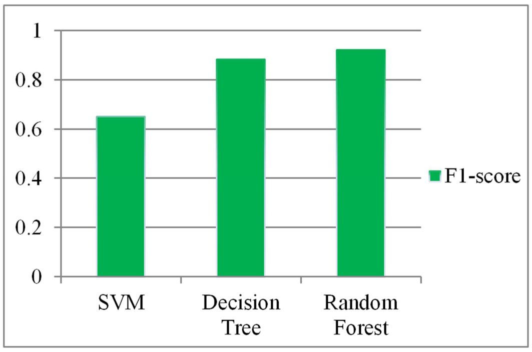

Figure 5. Graphical representation of F1-score for each classifier.

5. Conclusion

The objective of this research is to classify the images to find out the changes of the waterbodies in Rajshahi City Corporation (RCC) area and implement the dataset on different classification techniques. Because waterbodies are decreasing drastically but some years, this area faces flood and some year faces drought. So the amount of waterbodies is not constant each year. In this research paper, the dataset has only six attributes with three different classes. Support Vector Machine (SVM), Decision Tree and Random Forest are implemented on this dataset. Most of the research papers have shown the quality of the surface water and groundwater, amount of arsenic in tube-well water, fluctuation of groundwater, applied GIS on the arsenic contaminated area but generating a dataset from images and implement that dataset on any data mining algorithm has rarely done. In this research, used images are multispectral images as they have 3-10 bands and it has more than two classes, so it is a multiclass problem. Precision, recall, and f1-score are calculated for three class problems. We found 92% accuracy in Random Forest Techniques. Which indicate it performs better in this kind of datasets. This type of dataset can be used in different classification techniques. By doing efficient coding the accuracy can be increased.

Acknowledgements

The authors would like to give special thanks to their honorable Prof. Dr. Md. Shahid Uz Zaman, Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Bangladesh for his valuable suggestion, encouragement, guiding and his constant support.

References

[1] Shiklomanov, Igor A. "Appraisal and assessment of world water resources." *Water international* 25.1 (2000): 11-32.

[2] Md. Habibur Rahman. "Pond filling plagues Rajshahi city." DhakaTribune, 2 Sept. 2014, archive.dhakatribune.com/environment/2014/sep/02/pond-filling-plagues-rajshahi-city.

[3] "Rajshahi City Corporation." *BANGLAPEDIA,* 9Mar. 2015, en.banglapedia.org/index.php?title=Rajshahi_City_Corporation.

[4] Perlman, USGS Howard. "Groundwater depletion." Groundwater depletion, USGS water science, 9 Dec. 2016, water.usgs.gov/edu/gwdepletion.html.

[5] Ahmeduzzaman, Mohammad, Shantanu Kar, and Abdullah Asad. "A Study on Ground Water Fluctuation at Barind Area, Rajshahi." *International Journal of Engineering Research and Applications (IJERA) ISSN* (2012): 2248-9622.

[6] George, Geeja K., et al. "Study of Ground Water Pollution around an Industry Using GIS."

[7] Melgani, Farid, and Lorenzo Bruzzone. "Classification of hyperspectral remote sensing images with support vector machines." *IEEE Transactions on geoscience and remote sensing* 42.8 (2004): 1778-1790.

[8] "Landsat Program." *Wikipidea,* Wikimedia Function, 29 Nov. 2017, en.wikipidea.org/wiki/Landsat_program.

[9] Natrella, Mary. "NIST/SEMATECH e-handbook of statistical methods." (2010).

[10] Franc, Vojtech, and Václav Hlavác. "Multi-class support vector machine." *Pattern Recognition, 2002. Proceedings. 16th International Conference on*. Vol. 2. IEEE, 2002.

[11] Patel, Savan. "Chapter 2: SVM (Support Vector Machine) - Theory – Machine Learning 101 – Medium." *Medium,* Machine Learning 101, 3 May 2017, medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72.

[12] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003): 1-16.

[13] Jin, Chen, Luo De-Lin, and Mu Fen-Xiang. "An improved ID3 decision tree algorithm." *Computer Science & Education, 2009. ICCSE'09. 4th International Conference on*. IEEE, 2009.

[14] Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

[15] Qi, Yanjun, Judith Klein-Seetharaman, and Ziv Bar-Joseph. "Random forest similarity for protein-protein interaction prediction from multiple sources." *Biocomputing 2005*. 2005. 531-542.

[16] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." (2007): 3-24.

[17] "Supervised and Unsupervised Classification in ArcGIS."GISGeography, 22 Jan. 2017, gisgeography.com/supervised-unsupervised-classification-arcgis/.