International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 14 (2018) pp. 11553-11558


# An NDVI based Spatial Pattern Analysis for Spatial Image Classification

**M.Gangappa [1]*, C. Kiran Mai, P.Sammulal**

*[1]Associate Professor, Computer Science and Engineering, VNRVJIET, Hyderabad, Telangana, India.*
*[2] Professor, Computer Science and Engineering, VNRVJIET, Hyderabad, Telangana, India.*
*[3]Assoc. Prof, Department of Computer Science and Engineering, JNTUH College of Engineering, India.*
**Corresponding author*

**Abstract**

In the recent years, the spatial data classification has been popular because its vast applications focus on classifiers with better accuracy. The current research on spatial data analysis has been using machine learning approaches and deep learning algorithms. The supervised learning methods are popular because of their robust performance in classifying the spatial data. The supervised classifier performance can be increased to analyze the classification of spatial data with the use of vegetation index. The one of the spatial index is the normalized difference vegetation index (NDVI) that uses NIR and RED electromagnetic spectral bands to identify the vegetation in the spatial image. This paper focused on the utility of NDVI to extract the pixel information relevant to the vegetation which is then useful to classify the data in the image. The various supervised classification algorithms are used to study the effect of NDVI in classification. The proposed system assessment is done with the analysis of 10 fold cross validation method. This experiment evinced the use of NDVI in increasing the accuracy of the classifier.

## INTRODUCTION

The spatial data or geo-graphic information has been an essential source of information to locate and describe the data objects on the earth surface. It has enormous applications such as environmental change detection, digital image analysis, education and science, military use and a lot more. Since last two decades, this area has been evolving as a good domain to conduct the research [22]. The spatial data is collected from remote sensing satellites and the data includes image pixels to define information. Spatial data seems to be an unstructured, non linear and complex data. Therefore, the analysis process is carried out to bring the hidden knowledge and such type of process is called spatial data analysis. The Geo-spatial data analysis plays a vital role to locate and identify the various landscape types in spatial data[2]. The landscape types in the spatial data can be identified and classified using the machine learning methods. Of late, deep learning on remote sensing data has been applied for better accuracy. The one of the better optimization mechanism in the machine learning is SVM because it takes the advantage of activation function also known as kernel function which is used to conduct the descriptive analysis on the image data. The features in the spatial data are used to identify and classify the landscape types. The landscape types include bare soil, urban land, water body, natural vegetation and forest area [2]. The spatial data also known as multi-spectral data includes many spectral bands and this image seems to be more complex. The classification task of multi spectral data is one of the most important and difficult tasks in the remote sensing data analysis because it has some critical issues such as spatial data orientation, structure of spatial data, atmospheric conditions[2][5][21][15]. For landscape type identification, the machine learning mechanisms are often used in real time scenarios. The classification techniques in machine learn are considered as supervised mechanisms. It takes the training data and learn from it and that is quite different from the AI(artificial Intelligence).Once it has been trained ,then it can be used to predict the labels in the data and classify the data into different class labels. The supervised classifier uses the features in training data set. If the data set is complex and has many features which are used to describe the data .Then the classifier may take longer amount of time and may decrease the performance of the classifier. This may lead to degrade the accuracy of the classifier. Hence, sometimes feature selection and reduction mechanisms are required in the classification process. Therefore, we used dimensionality reduction technique to bring down the low dimension space in the spatial data set. The main aim of this paper is to find out the suitable machine learning model that distinguish the label the suitable to the landscape type based on the pixel information relevant to that land cover type. Eventually, we focus on assessment of machine learning methods suitable for spatial data points.

## ORGANIZATION OF THE PAPER

The information in this paper has been organized as follows. The section 3 provides the relevant related work .The preliminary techniques for conducting the experiment are discussed in section 4.The input data required is studied in the section 5. The methodology is discussed in the section 6. The next section discusses the experiential study and performance analysis. Finally, The last section provides the conclusion and future directions

## RELATED WORK

The recent literature survey on remote sensing data classification using machine learning methodologies encompasses the rich information about the spatial data advantages, environmental ecology, precision agriculture, science and engineering and military use. Lately, the remote sensing data classification has been done using better approaches in machine learning and deep learning. The

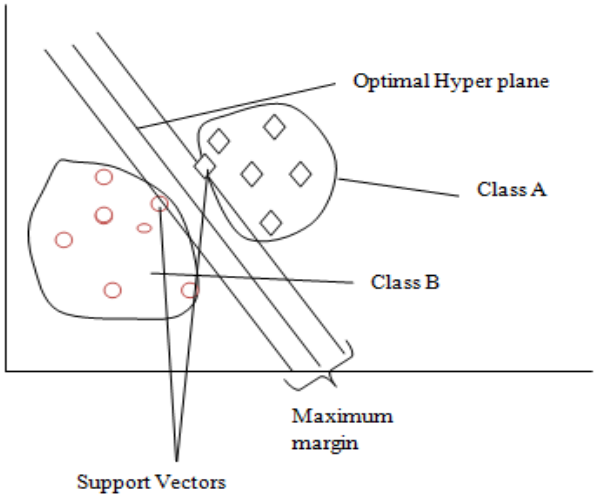
different classification methods were used on remote sensing data. The dimensionality reduction and optimal attribute selection procedure play a vital role in remote sensing data classification.

The rough set based dimensionality reduction was used to reduce the attribute size to gear up the classification tasks[22]. The SVM based classification method adopts active learning procedures to monitor the input data during optimization stage. Hence, it learns the knowledge dynamically. In high dimensional space, the limitation of dimensionality may yield good results. The high dimensional data handling is a critical task in the optimization problems [1]. Hence the optimization method like SVM may often insensible to high dimensional space[6][7].There are many classification algorithm which are often used to predict the class objects in the spatial data..The supervised learning methods for spatial data are neural networks, decision tree method, random forests classification methods ,K-means clustering and classification method, rough set based data reduction and classification method and fuzzy-rough set based classification method[16][6][9]. The fuzzy logic and neural net works are used in the spatial data classification. The sufficient amount of contribution of research work has been on fuzzy and rough based mechanisms [14]. Many other algorithms have been used in optimization problems. In the descriptive models, the training data with class labels are provided at the training of an algorithm. In order to get trained, these methods use some spatial data features such as spatial resolution, entropy, mean eleven and mean slop and other relevant features in input data. We strongly argue that prediction accuracy relies on the significant features used in that model.

## PRELIMINARIES

### A. Support vector machines

The one of the leading technique in the machine learning for descriptive data analysis is support vector machines. It is a supervised machine learning model and it has been used to recognize the patterns from the training data that is being supplied to this model and to classify complex data. The SVM is quite good for solving the problems that are relevant to binary classification which take two class labels class A and class B such as Y[+1,-1] and $X_i$ is a set of the predictor variables .To classify the data ,it generates parallel lines which have maximum margin between them and called the hyper planes. The hyper planes which have the largest gap or margin in the high dimensional data are used to separate the data into different classes. The optimal hyper plane separates data by minimizing the misclassification between data points and maximizes the distance between nearest data points.SVM finds only one optimal hyper plane from the generated many hyper planes.




.

**Figure 1.** Support Vector Machine with maximum margin

The mathematical equation for optimal hyper plane can be made with the training set feature vectors. Let X={x1,x2,x3,x4,...xN} be a set of feature vectors of X and the relevant classes are w1 and w2. From the feature vectors and class labels the equation for optimal hyper plane f(x) can be derived. The hyper plane equation separating the vectors is shown in equation (1).

$$\mathbf{w}.\mathbf{x} + \mathbf{b} = \mathbf{0} \quad \textbf{(1)}$$

Where b is a bias and weight vector is w and the maximum margin between support vectors is given by the following equation .

$$M = \frac{2}{\sqrt{\|w\|}} \quad \textbf{(2)}$$

Where $\|w\|$ is the Euclidean distance computed from the feature vectors[7]. The equation for optimal hyper plane is given as follows

$$f(X) = w_0 + w^T.x \quad \textbf{(3)}$$

Where the weight vectors $w$ and $w_0$ is bias. Using the training data, SVM finds an optimal hyper Plane. However, SVM are not confined to the binary classification problems. They can be used to classify the linear and non linear data classification problems as well.

Spatial data has a complex structure and spatial data is a non linear data. Hence, SVM with linear separable can't be used to describe the data points in the spatial data. SVM with kernel trick will perform better to separate the data points in such non linear data.

### SVM with non linear kernels

The non linear data points in the classification problem are to be mapped to higher dimension space. It is clear that high dimension space provides a way to create a boundary among the separable and non linear data points and this boundary

curve can be used to identify the class labels. Let us consider the non linear data in 1D representation as shown in figure given below. The red and blue colour points have been spread in a line. These red and blue data points are separated by projecting the 1D space into 2D space as shown in figure()

**Figure 2.** non linear data with single feature

Now map the 1D feature data to high dimension, say 2D .Therefore, the data is now seen as linearly separable and is useful for classification.

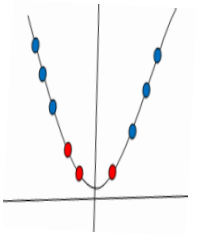


**Figure 3.** Linear separable data with two features

Therefore ,the transformation function is required to map the 1D data representation to 2D representation.In cases,this kind of transformation from low dimension to high dimension may be more difficult.To avoid such difficulties,SVM introduced a transformation function which is called a SVM kernal.The transformation function performs the inner product on the feature data points from the non linear strure and then converts them into linear and separable format for classification.Let the feature vectors x1 and x2 of each with n dimentions.The inner dot product of x1 and x2 is denoted by K(x1,x2) and this function is called kernal function [8]. The kernel function is defined as shown below.

$$K(x, x') = \emptyset(x').\emptyset(x)^T \quad (4)$$

The SVM RBF kernel function is discussed as follows: The Gaussian Kernel function is a kernel function and is represented as

$$\mathbf{K_{RBF}}\ (\mathbf{x}, \mathbf{x}') = \mathbf{exp}(-\boldsymbol{\gamma}\|\mathbf{x} - \mathbf{x}'\|^2) \quad \mathbf{(5)}$$

Where γ is a measure of similarity parameter . The kernel function in the RBF performs the projections. So the complex data points in the lower dimensions are projected to the high dimensions

When the similarity parameter is considered as 0,5 with minimizing the loss between feature data points, the new equation can be derived as follows

$$\mathbf{K_{RBF}}\ (\mathbf{x}, \mathbf{x}') = \mathbf{exp}\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2}\right) \quad \mathbf{(6)}$$

Solving the eq(5),we get the new equation which shows that RBF kernel can take the infinite sum of polynomials.

$$\mathbf{K_{RBF}}\ (\mathbf{x}, \mathbf{x}') = \mathbf{exp}\left(-\frac{\|\mathbf{x}\|^2}{2} - \frac{\|x'\|^2}{2} - \mathbf{2}(x - x')\right)$$

$$\mathbf{K_{RBF}}\ (\mathbf{x}, \mathbf{x}') = \mathbf{Ce}^{(\mathbf{x},\mathbf{x}')} \quad \mathbf{(7)}$$

Where the constant is $C = \exp\left(-\frac{\|x\|^2}{2} - \frac{\|x'\|^2}{2}\right)$ and $e^x$ is the Taylor series.

### B. Artificial Neural Networks

The ANN is a supervised learning classification method. Its functional design structure is quite the same as the structure of human brain. For each problem, its structure gets changed. The topology and network nodes are used to decide the hidden layer. It has no concept of hyper planes and multi dimensional planes as in the SVM .However it takes longer amount of processing time to carry out the data .

### C. Remote Sensing Data

The spatial data is a remotely captured data which is collected using special sensors inside the satellite. The special sensors capture the reflected energy from the earth surface. There are different remote sensors such as SPOT, IRS ,AVHRR, Land sat, SAR etc. These sensors have limited spectral sensitivity which is known as spectral resolution and the range of wavelength represented on the electromagnetic spectrum is known as bands[4]. These bands are different in number for different remote sensing systems. In this study, Landsat8 data set is considered for experiments. The Landsat8 sensor images have the 11 electromagnetic spectral bands and most of the bands consists of 30 meters spectral resolution [3]

[4].The information about Land sat 8 is furnished in the table1.

**Table 1.** Information of Land sat 8 data set [3]

| Band name | Band number | Wavelength | Resolution in meters |
|---|---|---|---|
| Coastal band | 1 | 0.433–0.453 | 30 m |
| Blue band | 2 | 0.450–0.515 | 30 m |
| Green band | 3 | 0.525–0.600 | 30 m |
| Red band | 4 | 0.630–0.680 | 30 m |
| Near Infrared | 5 | 0.845–0.885 | 30 m |
| SWIR 1 band | 6 | 1.560–1.660 | 30 m |
| SWIR 2 band | 7 | 2.100–2.300 | 30 m |
| Panchromatic image | 8 | 0.500–0.680 | 15 m |
| Cirrus image | 9 | 1.360–1.390 | 30 m |
| LWIR 1 band | 10 | 10.6-11.2 | 100 m |
| LWIR 2 band | 11 | 11.5-12.5 | 100 m |

The band combination helps to visualize the certain data objects in the Landsat8 sensor. For instance, the bands 2,3and 4 provide the true colour image. The RGB colour model for true colour image uses the band combination 4-3-2. The false colour image having the bands combination 7-5-2 is useful for

analyzing crops. The vegetation index NDVI is useful for vegetation analysis on the earth surface. The NDVI is calculated by combining band5 , the near infrared(NIR) with other bands in the Landsat8 data[5]. The band combinations and useful advantages are furnished in the given table 2 below.

**Table 2.** Band combinations and their feature analysis

| Band combination | Nature of image | Feature analysis |
|---|---|---|
| 4-3-2 | True colour image | Forest, bare land, Urban area, water body |
| 7-5-2 | False colour image | Agricultural crops, Urban area, bare land |
| 5-4-3 | False colour image | Vegetation, Land cover,Water body |
| 5-6-4 | False colour image | Land and water |
| 5-6-2 | False colour image | Crops and bare land |
| 6-5-4 | False colour image | Crops and vegetation |

## STUDY AREA AND INPUT DATA

The Lanndsat8 satellite collects the earth images with the period of 16 days[4]. It consists of two instruments: OLI(operational Land Imager) sensor and TIRS(Thermal Infra Red Sensor).They collect the visible data, near infrared and panchromatic band[4]. The Landsat8 OLI/TIRS C1 level-1 data set is used for this study. The data set was collected for the study in the area ,Vijayawada, Andhra Pradesh(AP). This data set consists of different types of landscape types such as vegetation area, water body, bare land etc. The details of the data seta are furnished in the table 3.

**Table 3.** Data set attributes and their values

| Data set attribute Name | Attribute value |
|---|---|
| Image format | GeoTIFF |
| Pixel size | 30meters |
| Datum | WGS 84 |
| Orientation | North up(map) |

The data collected from the landsat8 satellite is pre-processed before using for classification. This is required because landsat8 (Level 1) is generally not corrected atmospherically.

## METHODOLOGY

In this section model frame work is designed to deal with the spatial data .The major essential part of handling the spatial data is to prepare the data for conducting the experiment. In this step the instructed data is made to be structured and this step in the model frame is generally called as pre-processing stage. At First, the collected spatial data is considered as input. This data generally depends on the source of electromagnetic radiation on the earth surface and reflection received by sensor. The reflected data which is received by the sensor is assigned by a digital number (DN). The digital number (DN) represents the intensity at each pixel in the spectral band. The DN values are then converted to top of atmospheric reflectance. The data preparation uses some techniques on the raw data to correct errors and to avoid noise in the spatial data. The techniques are considers as

- Radiometric Correction
- Geometric Correction
- Noise removal

After this stage, The geo referenced data is to be aligned to the coordinates of real word .This is a Geometric correction and align the data to the real word coordinates.

In the next step, the spatial bands are stacked together to form a stack of bands. Here, the vegetation indices are calculated which are used in many application such as climate change detection and monitoring, vegetation system modelling, agriculture studies etc. It has special spectral signature and combines information from different bands. The following equation helps to find the NDVI (Normalized Difference Vegetation Index)[5].

$$NDVI = \left(\frac{NIR-RED}{NIR+RED}\right) \quad (8)$$

Then the shape file is created for region of interests in the image. The training and testing samples are extracted from the image using the shape file which consists of region of interests. The detailed procedure is given in below.

Procedure begin

1. Convert the digital number (DN) to the reflectance for each pixel.
2. Construct the stack of raster image by combining the bands.
3. Crop the raster image to the specific interested area.
4. Calculate NDVI using required bands as given below
   $$NDVI = \left(\frac{NIR - RED}{NIR + RED}\right)$$
5. Find the shape file using Region of Interest(ROI) to collect the required pixel information for different land cover types.
6. Prepare the training set and test set for classifying image.
7. Build the training model using machine learning approach.
8. Find the performance analysis of the model and compare the results.

Procedure ends.

The model frame work in figure illustrates the step by step process of conducting the image classification task for spatial data set.

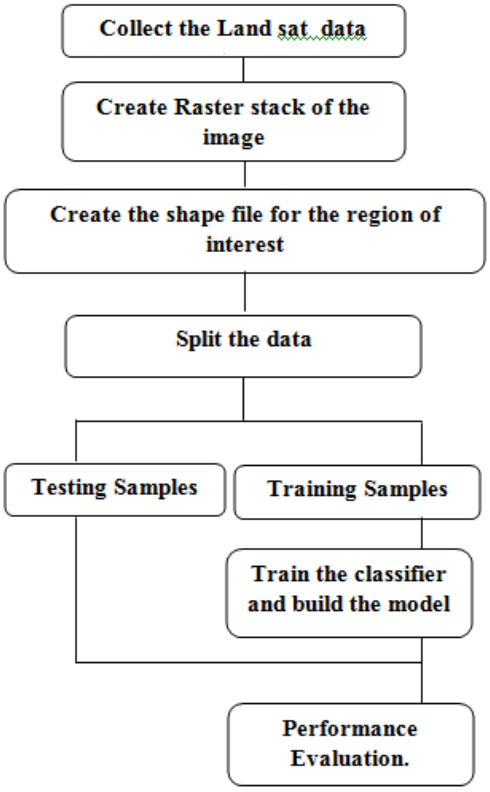
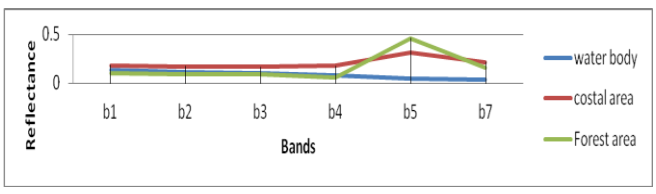



**Figure 4.** Model spatial data classification frame work

## EXPERIMENTS AND RESULTS

The Land sat data is pre-processed using the QGIS3.0 tool[12]. Then the geometrically corrected data is used to create various class labels such as water body. Forest Area, dry land etc., for the training samples using the polygon structures and then it produces a shape file. With the help of shape file ,the tainting pixels are collected for the classes used in the shape file.

A shape file contains the features that are used to prepare the data set for classification problem. The features such as water body, forest area etc ., of earth surface is represented by pixel values with certain intensity. A spectral profile is a spectrum which is made of all bands contains pixels representing the features of earth surface .The figure shows the various spectral properties which are extracted from spatial data set.




The NDVI was calculated .The training spatial image was formed with NDVI feature and was used as image for classification .From that image, the training samples and testing samples were collected with 70% and 30% basis. Then the SVM RBF l and NN models were used to classify the spatial image . The confusion matrices for both models are given in the table 4 and table 5. The model accuracy and kappa statistics were also recorded in the table 6. This shows that SVM RBF model is quite superior to NN for spatial data classification.

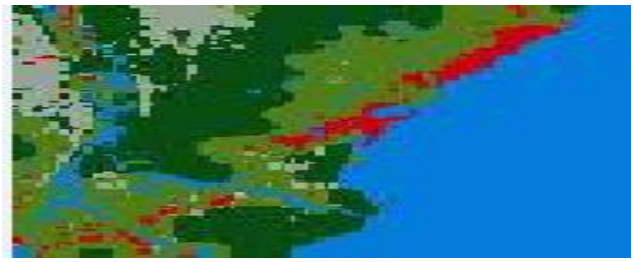


**Figure 5.** The classified image generated by SVM RBF

**Figure 6.** The classified image generated by NN

**Table 4.** Confusion Matrix for SVM

| Prediction/ Observed | Water body | Forest area | Dry land |
|---|---|---|---|
| Water body | 1586 | 220 | 0 |
| Forest Area | 342 | 828 | 0 |
| Costal Area | 0 | 0 | 510 |

**Table 5.** Confusion Matrix for NN

| Prediction/ Observed | Water body | Forest area | Dry land |
|---|---|---|---|
| Water body | 1529 | 223 | 0 |
| Forest Area | 399 | 825 | 0 |
| Costal area | 0 | 0 | 510 |

**Table 6.** Model accuracy and kappa statistics

| Model Name | Accuracy | Kappa statistics |
|---|---|---|
| SVM RBF | 86.9428% | 79.2901% |
| NN | 85.626% | 76.302% |

## CONCLUSION AND FUTURE DIRECTION

This paper discussed the NDVI based spatial data classification methods. The classification accuracy of various methods gets increased with the new band formed with NDVI in the stack of bands of spatial image. The SVM RBF model

shows the better prediction accuracy over the NN(Neural Network) model. The spatial classification can be performed with the spatial vegetation index like NDVI. In future, the prediction performance and accuracy of remote sensing image classification models can be further improved with the other vegetation indices.

## REFERENCES

[1] Pozdnoukhov A., &Kanevski M. , "Monitoring network optimisation for spatial data classification using support vector machines", International journal of environment and pollution, 28(3-4), 465-484, 2006.

[2] M.Gangappa, Dr C.Kiran mai, Dr P.Sammulal ," ESDAS: Explorative Spatial Data Analysis Scale to predict spatial structure of landscape", International Journal of Scientific & Engineering Research 7(12), pp. 466-472,Dec 2016.

[3] Tri Dev Acha,Intae Yang (April 2015) 'Exploring Landsat 8', *International Journal of IT, Engineering and Applied Sciences Research,* 4(No. 4), pp. 4-10.

[4] USGS (2014), United States Geological Survey, Landsat 8 Quality Assessment Band, https://landsat.usgs.gov/L8QualityAssessmentBand.

[5] Ravi Prakash Singh, Neha Singh, Saumya Singh, and Saumitra Mukherjee (2016) 'Normalized Difference Vegetation Index (NDVI) Based Classification to Assess the Change in Land Use/Land Cover (LULC) in Lower Assam, India', International Journal of Advanced Remote Sensing and GIS, 5(10), pp. 963-1970.

[6] Mingmin Chi,Rui Feng,Lorenzo Bruzzone (accepted 6 February 2008) 'Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem', Advances in Space Research, 41(2008), pp. 1793–1799.

[7] Pal M, "Random forest classifier for remote sensing classification", International Journal of Remote Sensing, 26(1), 217-222, 2005.

[8] D.Shanthini, M.Shanthi, Dr.M.C.Bhuvaneswari (2017) 'A Comparative Study of SVM Kernel Functions Based on Polynomial Coefficients and V-Transform Coefficients', International Journal Of Engineering And Computer Science, 6(3), pp. 20765-20769.

[9] Foody G. M., &Mathur A, "A relative evaluation of multiclass image classification by support vector machines". IEEE Transactions on geoscience and remote sensing, 42(6), 1335-1343, 2004.

[10] Feras Al-Obeidat, Ahmad T. Al-Taani, Nabil Belacel, Leo Feltrin, Neil Banerjee (2015) 'A Fuzzy Decision Tree for Processing Satellite Images and Landsat Data', Procedia Computer Science, 52(2015), pp. 1192 – 1197.

[11] Ham, J., Chen, Y., Crawford, M. M., &Ghosh, J," Investigation of the random forest framework for classification of hyperspectral data", IEEE Transactions on Geoscience and Remote Sensing, 43(3), 492-501, 2005.

[12] https://www.qgis.org/en/site/

[13] J.S. Rawat a, Manish Kumar (2015) ", Monitoring land use/cover change using remote sensing and GIS techniques: A case study of Hawalbagh block, district Almora, Uttarakhand, India, 18, pp. 77-8.

[14] Mahmon, Nur Anis, and Norsuzila Ya'acob. "A review on classification of satellite image using Artificial Neural Network (ANN)." Control and System Graduate Research Colloquium (ICSGRC), 2014 IEEE 5th.IEEE, 2014.

[15] M.Gangappa,Dr C.Kiran Mai, Dr P.Sammulal (2017) 'Analysis of Advanced Data Mining prototypes in Spatial data Analysis', Journal of Engineering and Applied Sciences, 12(12), pp. 3213-3219.

[16] Abhishek Maity,'Supervised Classification of RADARSAT-2 Polarimetric Data for Different Land Features', (arXiv:1608.00501v1), (1 Aug 2016).

[17] KUN-CHE LU AND DON-LIN YANG (2009) 'Image Processing and Image Mining using Decision Trees', JOURNAL OF INFORMATION SCIENCE AND ENGINEERING, 25(), pp. 989-1003.

[18] Hexiang Bai, Yong Ge, Jinfeng Wang, Deyu Li, Yilan Liao, Xiaoying Zheng (2014) 'A method for extracting rules from spatial data based on rough fuzzy sets', Knowledge-Based Systems, vol.57, pp. 28-40.

[19] Mahmon, Nur Anis, and Norsuzila Ya'acob. "A review on classification of satellite image using Artificial Neural Network (ANN)." Control and System Graduate Research Colloquium (ICSGRC), 2014 IEEE 5th.IEEE, 2014.

[20] Foody G. M., &Mathur A, "A relative evaluation of multiclass image classification by support vector machines". IEEE Transactions on geoscience and remote sensing, 42(6), 1335-1343, 2004.

[21] Gahegan M, "On the application of inductive machine learning tools to geographical analysis", Geographical Analysis, 32(2), 113-139, 2000.

[22] B Rokni Deilmai, B Bin Ahmad, H Zabihi ," Comparison of two classification methods (MLC and SVM) to extract land use and land cover in Johor Malaysia", Earth and Environmental Science 20 (2014) 012052.

[23] Gangappa M,Kiranmai C, Sammulal P (2017). New machine learning based approach for predictive modeling on spatial data. *In International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Udupi, India, 13-16 Sept. 2017.(*17414680) : IEEE. 79-84.*