Urbanization Analysis Using Spatial Support and Improved Random forest Decision Tree Approach

P. Kalyani[†] **and P. Govindarajulu**[††],

S.V. University, Tirupati, Andhra Pradesh, India

Summary

Urbanization is the major strategy that leads to the loss of the various biodiversity and the homogenization of geological habitat. Thus, it is necessary to study and analyze the drastic changes occurred due to global urbanization periodically. The periodical assessment of urbanization gives rise to the development of various techniques and rules from several researchers. This paper is also a part of development over the urban-land cover analysis. It introduces a Budget in Random Forest Decision tree (RFDT) approach that preserves the statistical features and object boundaries and help in improving the classification accuracy. The system implements a spectral band segmentation method, which differentiates the remote sensory images as Land Cover (LC) and Land Use (LU). The proposed RF algorithm of decision trees attempts to provide an improved efficiency over the other existing methods that is obtained from the experimental verification of the earlier algorithms with the proposed. The comparison report shows the performance of the algorithm from 2007 to 2013 respectively.

Key words:

Decision Trees, Geographic Information Systems, Land Cover, Land Use, Random Forest Decision Tree.

1. Introduction

The population increase and migration of people from rural areas is the reason for drastic change in land cover areas and water surfaces that leads to urban development. The Urban development has been motivated in high rate by the growth of infrastructures, buildings, real estates and housing, sanitaries, utility grids and transportation systems [1]. Hyderabad and Secundrabad, the twin capital city of Andhra Pradesh and its surroundings are the areas facing high urbanization rate of 27% of all Indian cities [2] and for this reason, it is taken as consideration. It is necessary to study and analyze the changes occurred due to the rapid development in a country or a state, which is made possible by adopting a remote sensing method. [3], [4] Remote sensing is the art and science of making measurements of the earth using sensors on airplanes or satellites. These sensors collect data in the form of images and provide specialized capabilities for manipulating their features. [5] Geographic Information Systems (GIS) provide the most important informative contribution to the remote sensing applications and geological analysis. The collected data information from the sensor is processed, mapped and analyzed by GIS and further helps in managing location-based information. The biological invasions like species extinction and global changes caused in the ecosystem are identified by the application of remote sensing and GIS. In order to efficiently extract more reliable feature information from the satellite data, suitable classification algorithm is very essential to be selected.

Classification is defined as a function involves in mapping the input samples consisting of many features into a single class label. [6] The classification algorithms are divided into two major approaches such as supervised and un-supervised classification. Un- supervised approach provides a cluster based algorithms to partition the spectral images into number of spectral classes based on their statistical information. However, it fail to produce a prior information or definition about the classes used. To solve this problem, the supervised method is introduced, in which the land cover classes are well defined with the sufficient reference data as a training sample [7]. Though supervised method provides good classification results, it does not contain more information about spatial support data. If the data are complex in structure, then to model the data in an appropriate way can become a real problem. To overcome this issue and to collect more details about spatial existence, Decision tree (DT) algorithms are introduced [9], which is very easy to use and it offers various advantages than any other method. It is one of the data processing algorithms widely implied for both classification and regression. Each interior node corresponds to one of the input variables and is split into child nodes based on the values of the input variable.

Several methods are found in DT, which are fully non-parametric, compact and does not require any additional assumption regarding the input data processing. [10] Researchers and scientists have undergone many efforts to develop and improve classification accuracy in DT but still it is noted to be a challenging task. The proposed DT algorithm includes the improved Random forest approach, which focuses on a direct computation of statistical features for each super-pixel present in remote sensed data. It consists of a collection or ensemble of simple feature

Manuscript received April 5, 2017
Manuscript revised April 20, 2017

predictors that involve in identifying the existence of land cover areas. Super pixel and spatial support are the two concepts used in Budget based DT approach to acquire the statistical features, which preserves object boundaries as well. The spatial support consists of the sufficient information like texture, color, and shape, based on which the super pixels are generated. In addition, the spectral and texture attributes are collected to differentiate the pixels from one another. The classification algorithm is applied to the extracted spatial features to divide them according to their characteristics. Budget based Decision Tree approach used in this paper that performs a classification in remotely sensed data and improves the accuracy of urban and land cover analysis.

2. Literature Review

J. R.Otukei and T. Blaschke, [11] analyzed the potential of DTs as one technique for data mining for the analysis of the 1986 and 2001 Landsat TM and ETM+ datasets, respectively. The results were compared with those obtained using SVMs, and MLC. Overall, acceptable accuracies of over 85% were obtained in all the cases. In general, the DTs performed better than both MLC and SVMs.

S. Moustakidis et al. [12] proposed a novel fuzzy decision tree (the FDT-support vector machine (SVM) classifier), where the node discriminations were implemented via binary SVMs. The tree structure was determined via a class-grouping algorithm, which formed the groups of classes to be separated at each internal node, based on the degree of fuzzy confusion between the classes. In addition, effective feature selection was incorporated within the tree building process, selecting suitable feature subsets required for the node discriminations individually. FDT-SVM exhibited a number of attractive merits, for example, enhanced classification accuracy, interpretable hierarchy, and low model complexity. Furthermore, it provided hierarchical image segmentation and had reasonably low computational and data storage demands. Their approach was tested on two different tasks: natural forest classification using a QuickBird multispectral image and urban classification using hyperspectral data. Exhaustive experimental investigation demonstrated that FDT-SVM was favorably compared with six existing methods, including traditional multiclass SVMs and SVM-based binary hierarchical trees. Comparative analysis was carried out in terms of testing rates, architecture complexity, and computational times required for the operative phase.

K. L. Bakos and P. Gamba [13] introduced a novel methodology to build a multistage hierarchical data processing approach that was able to combine the advantages of different processing chains, which may be best suited for specific classes, or simply already available to the data interpreters. The combination process was carried out using a hierarchical hybrid decision tree architecture where, at each node, the most useful input information source, i.e., the processing chain was used. The structure of the tree was created by using the predicted accuracy level of the whole structure estimated on a validation set. The final maps were achieved by applying the designed framework to the whole data set. The usefulness of the procedure was proved by two instances of a specific application, i.e., vegetation mapping, in mountainous and plain areas.

A. Baraldi et al. [14] provided a quantitative assessment of ISRC accuracy and robustness to changes in the input data set consisting of 14 multisource space borne images of agricultural landscapes selected across the European Union. The collected experimental results showed that, first, in a dichotomous vegetation/non-vegetation classification of four synthesized VHR images at regional scale, ISRC, in comparison with LSRC, provided a vegetation detection accuracy ranging from 76% to 97%, rising to about 99% if pixels featuring a low leaf area index were not considered in the comparison. Second, in the generation of a binary vegetation mask from ten panchromatic-sharpened QuickBird-2 and IKONOS-2 images, the operational performance measurement of ISRC was superior to that of an ordinary normalized difference vegetation index thresholding technique. Finally, the second-stage automatic stratified texture-based separation of low-texture annual cropland or herbaceous range land (land cover class AC/HR) from high-texture forest or woodland (land cover class F/W) was performed in the discrete, finite, and symbolic ISRC map domain in place of the ordinary continuous varying, sub-symbolic, and multichannel texture feature domain. In addition, they demonstrated that the automatic ISRC was eligible for use in operational VHR satellite-based measurement systems such as those envisaged under the ongoing Global Earth Observation System of Systems (GEOSS) and Global Monitoring for the Environment and Security (GMES) international programs.

V. E. G. Millán et al. [15] investigated the best season (wet or dry season) and angle of observation to map tropical dry forest succession in Brazil. Nonparametric decision trees were used to build up classification maps based on principal component analysis (PCA) inputs. The results indicated that the use of off-nadir data improved the map accuracy of successional stages of tropical dry forests and riparian forests. Particularly, extreme and negative angles of observation generated higher map accuracies, suggesting that tree shadows were enhancing spectral differences between the studied vegetation classes. Images from the dry season provided better total and classes map accuracies for late and intermediate stages of tropical dry forests. On the other hand, some classes, such as riparian forests and early stage of tropical forests needed the use of off-nadir angles of observation to reach a minimum accuracy and best scores were reached using wet season's images.

R. B. Kheir et al. [16] discussed about the use of Geographic Information Systems (GIS), remote sensing, and, more specifically, structural classification techniques and decision-tree modeling to map erosion risks and design priority management planning over a representative region of Lebanon. The structural classification organization and analysis of spatial structures (OASIS) of Landsat TM satellite imagery (30 m) was used to define landscapes that prevail in this area and their boundaries, depending on their spectral appearance. The landscape map produced was overlaid sequentially with thematic erosion factorial maps (i.e., slope gradient, drainage density, rainfall quantity, vegetal cover, soil infiltration, soil erodibility, rock infiltration and rock movement). The overlay was visual and conditional using three visual interpretation rules (dominance, unimodality and scarcity conservation), and landscape properties were produced. Rills and gullies were measured in the field, and a decision-tree regression model was developed on the landscapes to statistically explain gully occurrence. This model explained 88% of the variability in field gully measurements. The erosion risk map produced corresponds well to field observations (accuracy of 82%). The landscapes were prioritized according to anti-erosive remedial measures: preventive (Pre), protective (Pro), and restorative (Res). This approach was useful in Lebanon.

3. Proposed System for Spectral Image Classification

3.1 Study Area and Site Description

Urban areas have been recognized as “engines of inclusive economic growth”. Out of the 121 crores Indians, 83.3 crores live in rural areas while 37.7 crores stay in urban areas i.e. approximately 32% of the population due to globalization. It is because the better opportunities that cities provide for individual development and that there are very few occupational choices available in the villages. Cities are the better equipped habitat to provide livelihood opportunities. Satisfying people’s day today needs is still a challenging task in metro cities, that too in and around South India. We can say that, people will have a tendency of migrating to cities as long as cities provide better lifestyle and employment opportunities. Hyderabad is the one of the fastest growing city in southern India. Hyderabad is one of the major metropolitan cities visibly developed a lot. In 1869, Hyderabad Municipal Corporation was introduced by British Government so as monitor population growth in Hyderabad due to the urban development. Hyderabad is the capital city of Telangana and its geographical coordinates are from 17.3850° north to 78.4867° east. The city is divided into 4 major sub-urban regions and each sub-urban regions are divided into five division for official use. Many investigations were performed as a case study on Hyderabad region.

3.2 Data collection and Class selection

The rapid inflows of rural population to urban places give rise to housing problem and the development of slums in these places. The increase in population of urban places pressurize the demand of water and sanitation facilities, which results in environmental pollution, health hazards etc. So the urbanization in metro cities should be regularised. This Research mainly concentrate on classes such as vegetation, water, urban area and barren land cover in Hyderabad. Water class includes rivers, lakes, pools and streams etc., in which the rivers and streams may sometimes referred as barren areas because it may not have flowing water throughout the year. In addition, the classes such as Forest regions, Small and Medium Parks are consider as vegetative lands. However during summer and spring season, forest areas may also considered as barren lands abruptly. Cloud free Landsat satellite data of two different years and month taken from USGS Earth Explorer website.

Based on U.S Geological Survey of Earth Resources observation and Science (USGS EROS), Month of July has minimum cloud cover of 4% and the vegetation peak for pasture, which period is suitable to get high spectral signals in vegetation. Different size of samples are require to find urbanization in metro cities as 30m, 40m, and 50m respectively, where 30 m resolution is suitable for mapping in regional scales. High resolution sensors are impractical to apply to the total study area due to their high cost and this process requires a longer period to analysis than the medium spectral images.

3.3 Data pre-processing and Class Definition

Satellite spectral data are determined with the help of pre-processing, which enhance image quality on the basis of medium spectral data usage. Improvisation of band quality can be made possible in terms of histogram and normalization procedure. The pre-processed images are then classified by both existing and proposed classification methods. The classification results predicts that the Land Use (LU) / Land Cover (LC) classes are named as vegetation, water, urban area and barren land. Land use is a series of operations on land, carried out by humans, with the intention to obtain products and benefits through using land resources. Land cover is defined as the vegetation (natural or planted) or man-made constructions (buildings, etc.) which occur on the earth surface. It also includes water, ice, rock, sand and bare land. Thus, the overall research classifies the spectral data as LU and LC.

3.4 Segmentation

The proposed Urbanization growth finding algorithm is an automatic region based approach and is implemented in four steps. In step 1, the color from the input image is segmented. In step 2, vegetation regions and water bodies are identified. In step 3, urban areas and barren land area are identified and at last in step 4, the original color image is segmented again using a band based segmentation algorithm. In our research, the classes like building and barren land are segmented as LU regions and vegetation, water bodies are segmented as LC regions. The segmented LC, LU regions are further analyzed to extract features from it.

3.5 Feature Space Extraction

Feature space extraction and selection plays a major role in spectral data classification. Spectral attributes are necessary to be computed on each band of the input image to further processing it. The attribute value for a particular pixel cluster is detected from input data band, where the segmentation label image has the same value of attributes (i.e., all pixels in the same pixel cluster contribute to the attribute calculation).

Spectral_Mean	Mean value of the pixels comprising the region in band x
Spectral_Max	Maximum value of the pixels comprising the region in band x
Spectral_Min	Minimum value of the pixels comprising the region in band x
Spectral_STD	Standard deviation value of the pixels comprising the region in band x

In addition, Texture attributes are also derived from each band, in which its computation has a two-step process, where the first pass applies a square kernel of pre-defined size to the input image band. The spectral and texture attributes are calculated for all pixels in the kernel window as said above and the result is referenced as centre kernel pixel, which are averaged across each pixel in the pixel cluster to create the attribute value to provide band segmentation label.

Texture_Range	Average data range of the pixels comprising the region inside the kernel (whose size you specify with the Texture Kernel Size parameter in segmentation)
Texture_Mean	Average value of the pixels comprising the region inside the kernel
Texture_Variance	Average variance of the pixels comprising the region inside the kernel
Texture_Entropy	Average entropy value of the pixels comprising the region inside the kernel

Spatial attributes are computed from the polygon defining the boundary of the pixel cluster and it does not require any band information.

Attribute	***Description***
Area	Total area of the polygon, minus the area of the holes. If the input image is pixel-based, the area is the number of pixels in the segmented object. For a segmented object with 20 x 20 pixels the area is 400 pixels. If the input image is geo referenced, the area is in the map units of the input image. For a segmented object with 20 x 20 pixels, where the input image pixel resolution is 2 meters, the total area is 1600 square meters (400 pixels x 2 meters x 2 meters).
Solidity	A shape measure that compares the area of the polygon to the area of a convex hull surrounding the polygon. The solidity value for a convex polygon with no holes is 1.0, and the value for a concave polygon is less than 1.0. Solidity = Area / area of convex hull
Major_Length	The length of the major axis of an oriented bounding box enclosing the polygon. Values are map units of the pixel size. If the image is not geo referenced, then pixel units are reported.
Minor_Length	The length of the minor axis of an oriented bounding box enclosing the polygon. Values are map units of the pixel size. If the image is not geo referenced, then pixel units are reported.

In our proposed system, the feature selection is made by regarding the budget level that only necessary features are extracted for further classifying it.

3.6 Random forest Decision tree classification

The basic structure of the decision tree however, consists of one root node, a number of internal nodes and finally a set of terminal nodes. The data is recursively divided down the decision tree according to the defined classification framework. DT used in this system are known to produce results of higher accuracies in comparison to traditional approaches such as the ‘‘box’’ and ‘‘minimum distance to means’’ classifiers but the performance of DTs can be affected by a number of factors including: pruning and decision thresholds, which is improved in this system by adding the improved Random forest method.

Random forest consist of a collection of trees, wherein each tree is grown by random independent data sampling & feature splitting, producing a collection of independent identically distributed trees. The resulting classifiers are robust, are easy to train, and yield strong generalization performance. The resulting classifiers in our system are robust, are easy to train, and yield strong generalization performance of LC and LU identification.

The random forests is mainly use for maintaining the classification in case of prediction-time budget constraints, which presents a major challenge. The two main advantage of RF is,

- It do not account for feature acquisition costs.
- It provides a clear diversity amongst trees developed

In our context the classifier f is a random forest T, consisting of K random trees $D_1, D_2, ...D_k$ that are learnt on training data. Suppose example/label pairs (x, y) are distributed as

$(x, y)d \sim H$. The goal is to learn a classifier f from a family of functions F that minimizes expected loss subject to a budget constraint,

$$\min_{f \notin F} E_{xy}[L(y, f(x))], s,t. \quad E_x[C(f,x)] \le B \qquad (1)$$

Where $L(y, \hat{y})$ is a loss function, $C(f,x)$ is the cost of evaluating the function of f on example x and B is a user specified budget constraint.

In this paper, we assume that the feature acquisition cost, $C(f,x)$ is a modular function of the support of the features used by function f on example x, that is acquiring each feature has a fixed constant cost. We can then minimize the empirical loss subject to a budget constraint,

$$\min_{f \notin F} 1/n \sum_{i=1}^{n} [L(y_i, f(x_i)], s,t. \ 1/n \sum_{i=1}^{n} [C(f, x_i)] \le B \qquad (2)$$

Consequently, the expected cost C for an instance x during prediction-time can be written as follows:

$$E_f \left[E_x \left[C\left(f, x \right) \right] \right] \le \sum E_{Dj} \left[E_x \left[C\left(D_j, x \right) \right] \right] \qquad (3)$$

Where, in the RHS we are averaging with respect to the random trees. As the trees in a random forest are identically distributed the RHS scales with the number of trees. This upper-bound captures the typical behaviour of a random forest due to the low feature correlation among trees.

The features present in the data of any given input image or sample is strictly depends on the budget and the arms present in a tree. The higher and lower values present in a tree determines the structure of tree. Normally, the design of tree algorithm consist of an unsymmetrical structure in them due to the higher and lower random values present in the left and right nodes. It generates the node based on their cost or weight, which is applied to each of the features. The symmetrical and un- symmetrical structure of tree is directly depends on the feature values applied to each arm or node.

The aim behind Random Forests used in this system is to generate multiple little trees from random subsets of data. In that way, each of those small trees gives some group of ill-conditioned classifiers. Each of them is capturing different regularities since random subset of the instances are in the interest. At the extreme randomness, it adjusts the nodes from random subset of the features as well. In this way feature based randomness is also implied in this method. After creating the number of trees in a random way, more cluttered decision boundaries are likely to obtain from it. In addition weight is added to each of the nodes to calculate the cost of the decision tree.

When the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This out-of-bag data helps to get an unbiased running estimate of the errors present in the classification, as nodes are added to the forest. It is also used to get calculation of variable and its importance. After each tree is built, all of the data are run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees. Proximities are used in replacing missing data, locating outliers, and producing illuminating low-dimensional views of the data.

The proposed RF algorithm of decision trees attempts to provide a symmetrical structure of tree with the application of further budgeting principle in them. The weight and symmetrical structure is based on the length of the number of features available in the samples. The problem of un-symmetriness of the other existing tree algorithms are considered as the main phenomena in this system design and produces successful node segmentation. After each tree is built, all of the data are run down the tree, and budget are computed for each pair of cases. If two cases occupy the same terminal node, their budget is increased by one. At the end of the result, the budget values are normalized by dividing the number of trees and the nodes are adjusted. The values are simultaneously increased and decreased to maintain the equal level of left and right node. Even though it may increase the time of classification, it produces an optimal output of generated trees based on their feature (attributes) values. The Budget Based Random Forest given below, Sample value consists of Training and Validation values for X, Y and length of m.

Pseudo code:

Algorithm- BUDGET RF

```
procedure BUDGET RF (F; B; C; S)
    $T \leftarrow \emptyset$ ; $S \leftarrow (x_1, y_1), ......., (x_m, y_m)$, $F \leftarrow Features$ , $root \leftarrow 0$
    $B \leftarrow Budget$  $C \leftarrow Cost$
    Train $[T, root] \leftarrow$ Random Forest (S, F)
    for $i \in 1, ....., m$ do
        if  $root<C$
            $root = root - B$
        else if $root>C$
            $root = root + B$
        else
            $root = root$
        end
        Train $[T, root] \leftarrow$ Random Forest (S, F)
    end for
$T \leftarrow T \cup T$
end function

function Random Forest (S, F)
        for $i \in 1, ....., m$ do
            $S^{(i)}$ ← A bootstrap sample from S
            $t_i$ ← Randomized Tree Learn( $S^{(i)}$ , F)
            $T \leftarrow T \cup \{t_i\}$
        end for
return T
end function

function Randomized Tree Learn(S, F)
        At each node:
            f ← very small subset of F
                Split on best attributes f into left and right arm
                if $f \langle gain(f)$
                    declare $root$ ; $f \rightarrow leftarm$
                else if $f \geq gain(f)$
                    declare $root$ ; $f \rightarrow rightarm$
                end
return The learned tree , $root$
end function
```

The random forest algorithm works as follows: for each tree in the forest, we select a bootstrap sample from S where $S^{(i)}$ denotes the i^{th} bootstrap. We then learn a decision-tree using a randomized selection of decision-tree learning algorithm. The algorithm is modified as follows: at each node of the tree, instead of examining all possible attributes-splits, initially we randomly select some subset of the attributes $f \subseteq F$. Where, F is the set of attributes. The node then splits on the best attributes in f rather than F . Gain (f) is represented as follows,

$$Gain(f) = root(f) - \min \cos t(f) \tag{4}$$

Here root is a threshold to split the leaf, based on attribute. So the equation (5) defines root as follows,

$$root(f) = \frac{\max(f) - \min(f)}{2} \tag{5}$$

In practice f is much, much smaller than F deciding on which attributes to split is oftentimes the most computationally expensive aspect of decision tree learning. By narrowing the set of attributes using Budget, we drastically speed up the learning of the tree. Each roots are validate through S bootstrap samples and decision are made at the time of root node selection in every branches. So the asymmetric branches will increases the cost value C . Symmetric structure is based on the number of classes in attributes or number of variant in attributes. When C is increased, RF root will reduced on the basis of Budget. C is get decreased, RF root will increased on the basis of Budget. Budget is the collection of number of rows and columns in the attributes. Hence, collection of budget is named as cost, the generation of cost is based on the equation (6) and (7),

$$C = \sum_{i=1}^{m} \sum_{j=1}^{n} B(i, j) \tag{6}$$

$$B = \begin{cases} B & 0 < B < 1 \\ 0 & otherwise \end{cases} \quad (7)$$

Where, i is represented as number of objects, j is specified as number of attributes.

In a tree structure, nodes have parent and children, root node is the first node or parent node, and it only has children nodes. Splitting of node operation is based on below equation,

Where, node is specified as $gain(f)$,

$$f = \begin{cases} left & if\ f < gain(f) \\ right & otherwise \end{cases} \quad (8)$$

Target classes are named in numeric values like water bodies' class1, Forest region class2, urban region class 3, and barren land class 4. The experimental verification of the proposed RF based decision tree algorithm and their results are further discussed below for the identification.

4. Experimental Result and Discussion

The analysis process of urbanization using the proposed Improved Random Forest DT algorithm provides increased accuracy of identifying the LC and LU regions than the all other approaches developed earlier. The spectral and texture attributes are calculated for further differentiating the band levels found in the image sample.

4.1 Experimental data collection

In this research, an experiment is made on two images containing all the attributes required to classify, which consist of two random years 2007 and 2013 from the database USGS Earth Explorer website respectively shown in Fig.1, 2. The USGS Earth explorer containing the spectral view of Cloud Free Landsat satellite data of Hyderabad region is considered.

Table 1: Details of Landsat data collected

No.	***Image taken on***	***Satellite/ sensor***	***Reference system/ path/ Row***
1	February 2017	Landsat5 /TM	WRS-2/144/48
2	February 2013	Landsat7 /ETM+	WRS-2/144/48

The below two images Fig 1 and 2 shows the spectral view of the selected two random years 2007 and 2013.

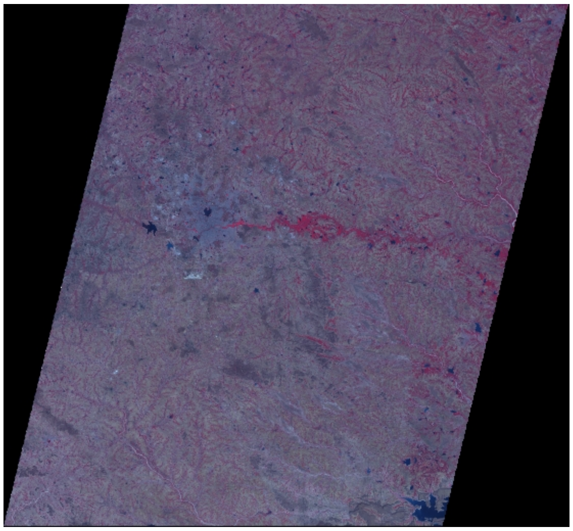


Fig. 1 Ariel view of 2007 satellite image

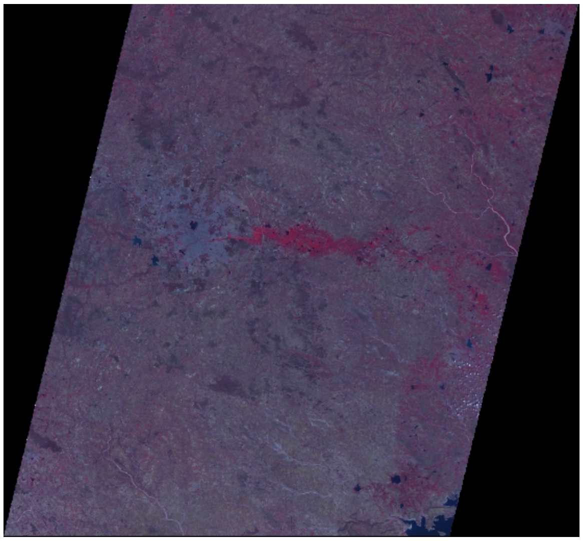


Fig. 2 Ariel view of 2013 satellite image

The spectral training images Fig. 1, 2 are passed into a series of process such as pre-processing of training and test data, Identification of classes, Segmentation of the data, and classification results, which are all performed for finding the accuracy of detecting the LC and LU data.

4.2 Class selection

In this experiment, Land cover regions are denoted by the Built up area and Barren land, whereas the Land use regions are denoted by the Vegetation and Water bodies that is applicable for both the 2007 and 2013 satellite images.

4.3 Pre-processing of training and test data

The intensity values are detected from the Fig. 1, 2 based on the RGB value calculation method, which are used for adjusting the pixels of the attributes present in the image. The enhanced image is provided with a legible quality of RGB existence and a noise free data. In fig 3, the improved spectral view of 2007 year is represented, likewise the enhancement is performed for Fig. 2 of 2013 spectral image.

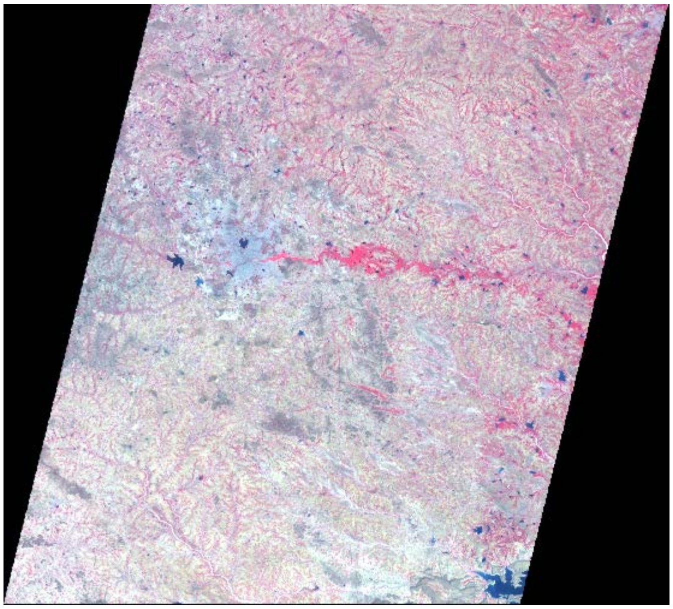


Fig. 3 Enhanced spectral image of 2007

4.4 Segmentation results

The attributes classes are segmented into four bands denoting the spectral intensities of the RGB values such as Near Infra-red (Near IR), Short IR1, Short IR2, and Thermal IR shown in Fig. 4, 5, 6, 7, 8, 9,10,11. From these four spectral bands, two bands may not be considered by our proposed method. The short IR1 and short IR2 from 2007 and 2013 is considered less, whereas the other bands are analyzed to follow up classification.

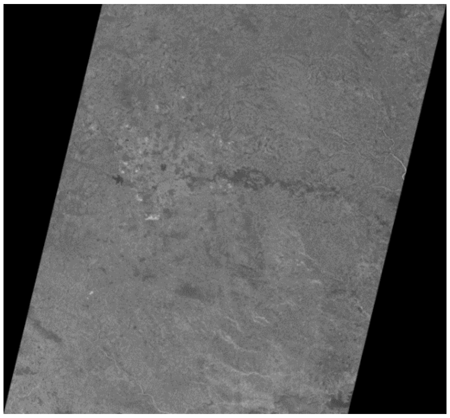


Fig. 4 Near IR representation of 2007

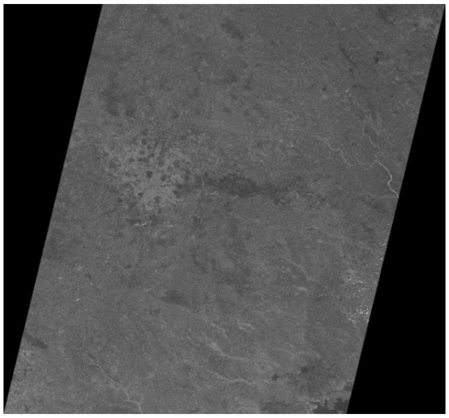


Fig. 5 Near IR representation of 2013

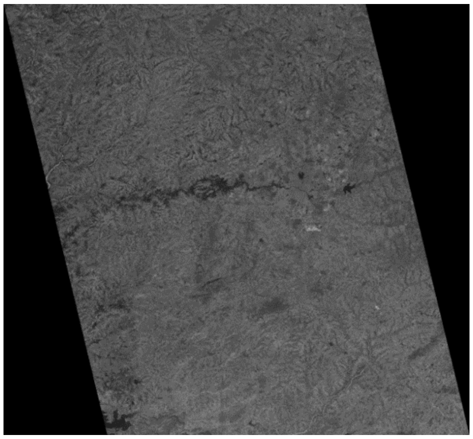


Fig. 6 Short IR1 representation of 2007

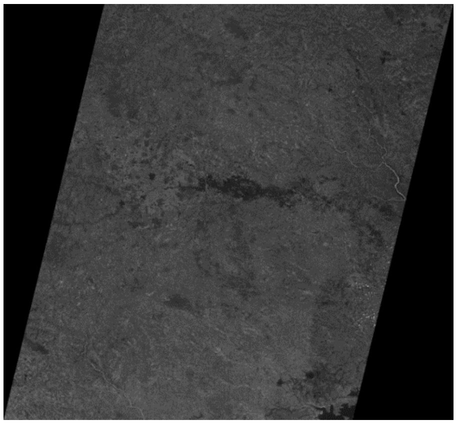


Fig. 7 Short IR1 representation of 2013

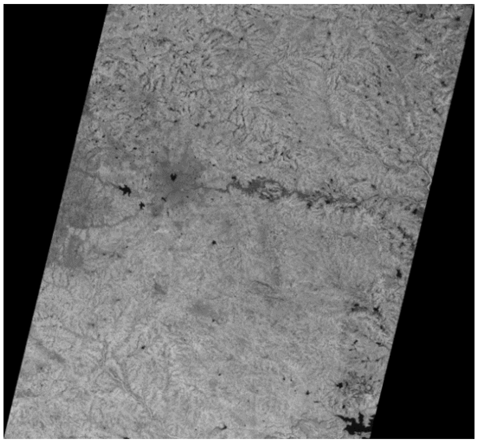


Fig. 8 Short IR2 representation of 2007

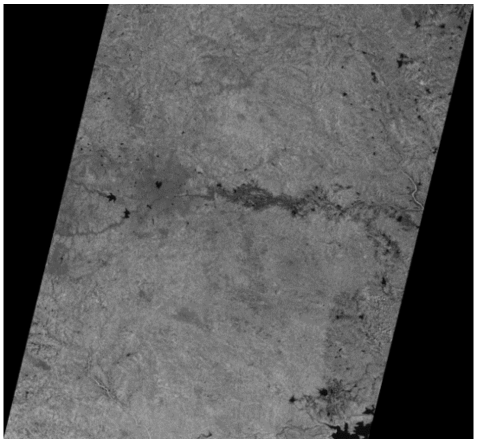


Fig. 9 Short IR2 representation of 2013

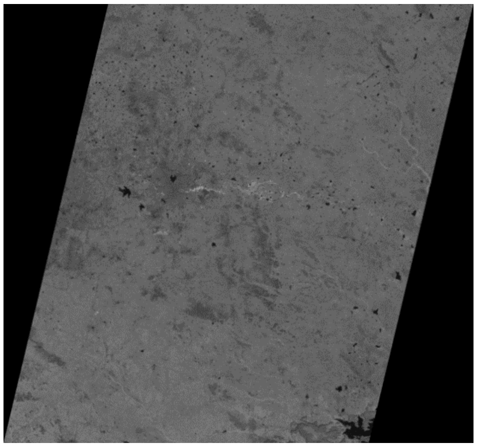


Fig. 10 Thermal IR representation of 2007

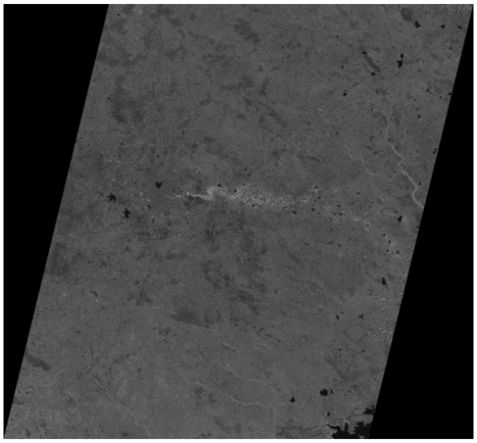


Fig. 11 Thermal IR representation of 2013

As a result, the derived bands were used as the dependent variables whereas the land cover classes were used as the independent variables.

4.5 Classes

In addition, the spectral band results are identified and the attributes classes are separated from it, showing the waterbodies, forest, urban lands and bare lands. The below figs. 12, 13, 14, 15, 16, 17, 18, 19 represents the fine definition of the attributes classes for 2007 and 2013 discussed above.

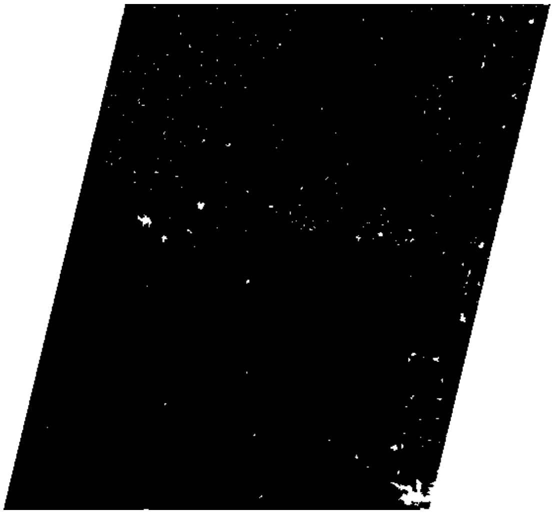


Fig. 12 Water bodies representation of 2007

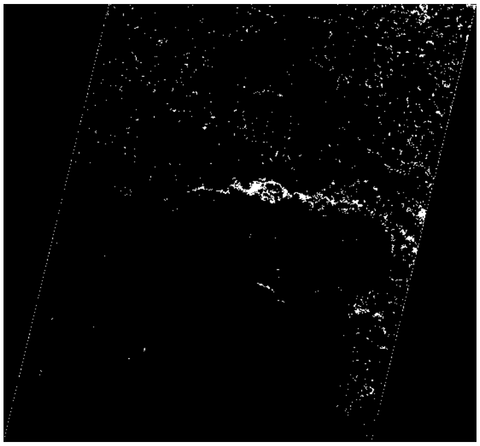


Fig. 14 Forest representation of 2007

Fig. 13 Water bodies representation of 2013

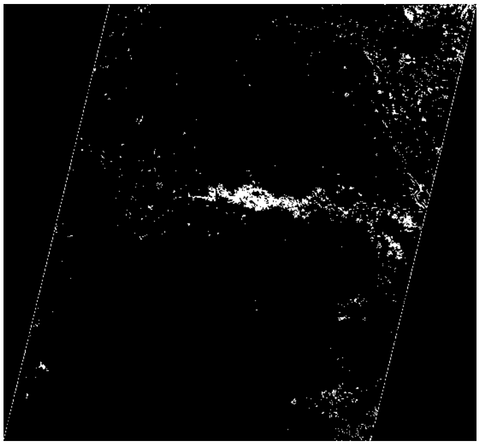


Fig. 15 Forest representation of 2013

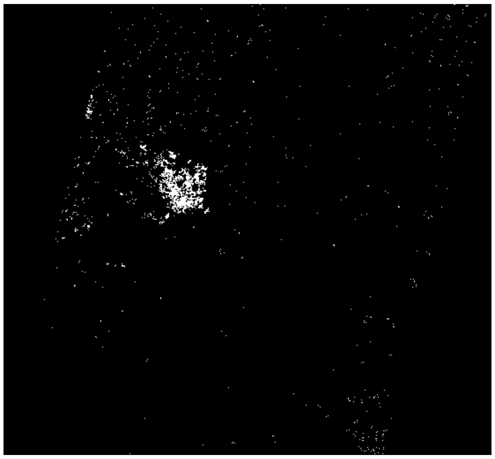


Fig. 16 Urban land representation of 2007

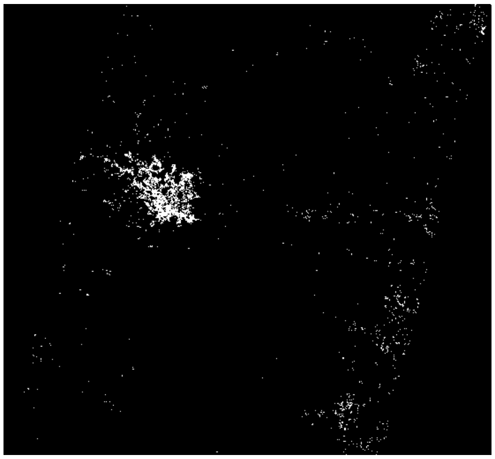


Fig. 17 Urban land representation of 2013

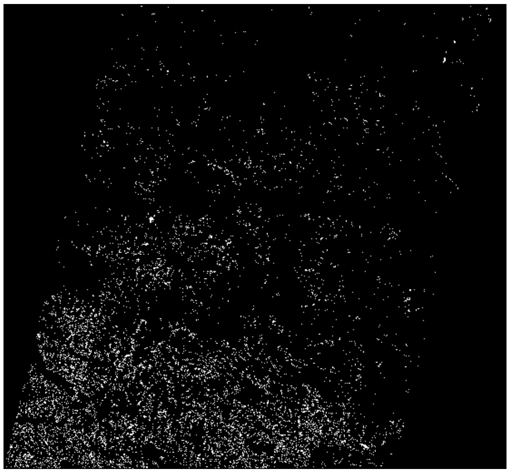


Fig. 18 Bare land representation of 2007

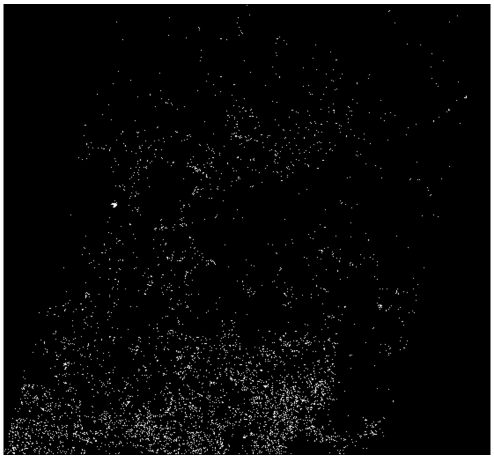


Fig. 19 Bare land representation of 2013

4.6 Classification results

The DT based classification is applied in the segmented images to derive the LC and LU region view that is mentioned in the below Figs. 20, 21, 22, 23 of the years 2007 and 2013.

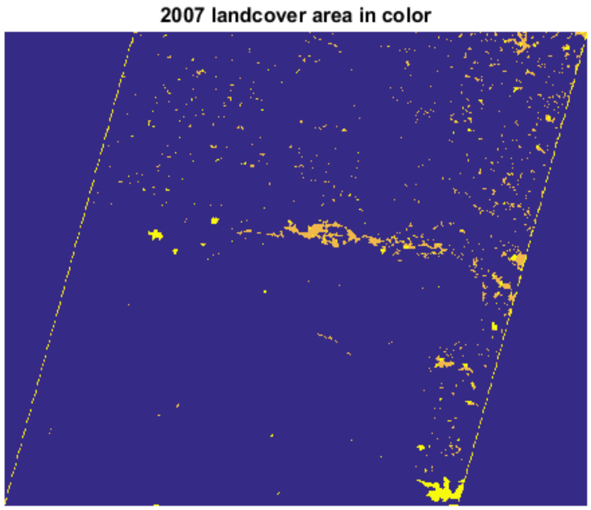


Fig. 20 LC representation of 2007 spectral image

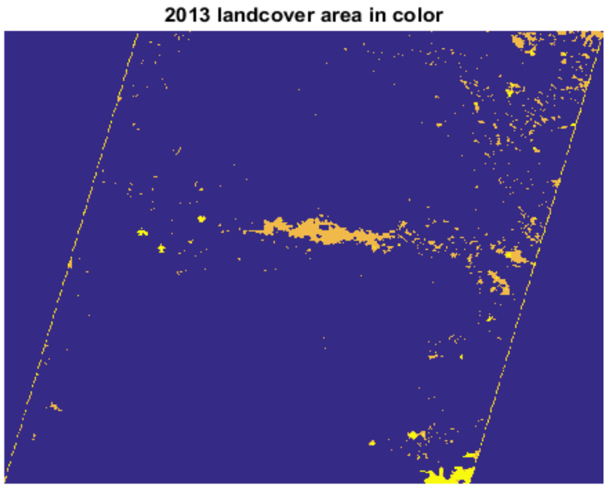


Fig. 21 LC representation of 2013 spectral image

The LC and LU is identified by applying this technique with the implementation of Random forest and it provides a good accuracy than the other algorithms.

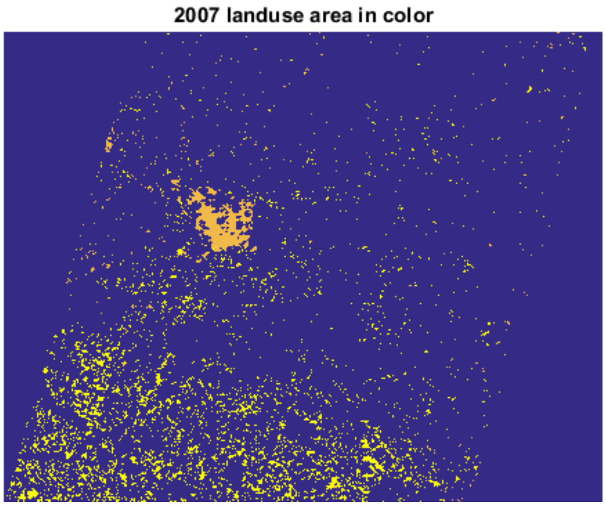


Fig. 22 LU representation of 2007 spectral image

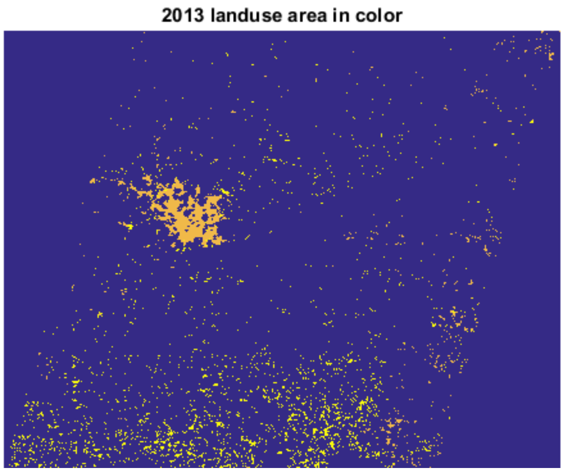


Fig. 23 LU representation of 2013 spectral image

The classified images of the two years can be enough to calculate the area of different land cover and land use and also it helps to observe the changes that are taking place in the range of data. The classified images obtained after pre-processing and Random forest DT classification that showing the land use and land cover of the Hyderabad city are given in the following figures Fig. 3, 4and 5, 6.

The below table predicts that the difference between the training and test data found in 2007 and 2013 images in the form of performance accuracy, in which accuracy is divided for existing, proposed and user perspective and the overall accuracy and kappa statistics is also calculated as shown in table 2.

Table 2: Classification Accuracy Assessment Report

LC &LU class	*2007*			*2011*		
	Existing Accuracy	*Proposed Accuracy*	*User's Accuracy*	*Existing Accuracy*	*Proposed Accuracy*	*User's Accuracy*
Water	98%	98%	100%	97%	97.2%	100%
Forest	88%	90%	75%	89%	91.1%	60%
Urban	28%	26%	71.43%	34%	35%	100%
Barren land	67%	67.06%	85%	78%	79.2%	75%
Overall Accuracy	71.33%	71.33%	71.33%	73.23%	73.73%	73.73
Kappa Statistic	0.49			0.56		

The accuracy of all the LC and LU classes are improved in our proposed system than the system and improved overall accuracy is also achieved, which is clearly represented in table 2.

4.7 Analysis report

The distribution of LC &LU classes in the image should not be same all the time rather it will get changed regarding climate and weather condition. Generally, the water bodies region sometimes tend to form as barren land when it is summer and vice versa for rainy season. Likewise the barren land sometimes form as forest area abruptly. The experimental results in our system also tend to change according to such natural phenomena. An assumption of climate is made in this research and a report is generated to denote the pixel occupancy of the classified LC and LU classes of 2007 and 2011, which is measured in hectares shown in table3.

Table 3: Analysis report of 2007 & 2013 land cover classes in hectares

LC &LU class	*2007(ha)*	*2013(ha)*	*2007-2013(ha)*
Water	5011.2	3323.17	-1688.04
Forest	9622.89	13200.06	+3577.17
Urban	6016.63	8665.1	+2648.47
Barren Land	23272.75	12311.06	-10961.68
Land Cover	15964.95	18743.08	+2778.125
Land Use	36441.62	18743.08	-17697.97

The calculation of LC and LU regions are given as LC= Forest + water, LU= Urban + barren land. The above mentioned table and calculation method depicts that, the land cover (LC) is of 85.0% because the water bodies is 66.3% and forest is 72.9%. Likewise the LU region becomes as 51.4% as the urban use is 69.4% and barren land is 52.8%. These generated value is for particular assumption of weather condition considered by us in this research. On the other hand the value of water and forest may become positive if the barren lands are occupied by more water or more forest.

5. Conclusion

The paper is a source that reproduces the Urbanization assessment and analysis data with its improved RFDT algorithm based on budget calculation. The spectral and texture attributes are the major components constitutes the spatial feature system, is calculated and the band level differentiation was implemented to provide an improved classification system. The algorithm has been experimented by comparing it with the earlier methods and the reports are tabulated. The classification results of LC and LU regions from the tabulation, shows the rate of urbanization and different attributes of water, forest and barren lands in the years 2007 and 2013 as well. It is concluded by stating that the urbanization in Hyderabad region and its detection accuracy is improved by the proposed technique to a great extent.

References

[1] Liang, Bingqing, and Qihao Weng. "Assessing urban environmental quality change of Indianapolis, United States, by the remote sensing and GIS integration." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 4, No.1, pp. 43-55, 2011.

[2] Kit, Oleksandr, Matthias Lüdeke, and Diana Reckien. "Texture-based identification of urban slums in Hyderabad, India using remote sensing data." Journal of Applied Geography, Vol.32, No.2, pp. 660-667, 2012.

[3] El Bastawesy, Mohammed, et al. "Detection and Assessment of the Waterlogging in the Dryland Drainage Basins Using Remote Sensing and GIS Techniques." IEEE journal of selected topics in applied earth observations and remote sensing, Vol. 5, No.5, pp. 1564-1571, 2012.

[4] Rao, Dasika P. "A remote sensing-based integrated approach for sustainable development of land water resources." IEEE Transactions on Systems, Man, and Cybernetics, Vol. 31, No.2, pp. 207-215, 2001.

[5] Muthu, Kavitha, and Maria Petrou. "Landslide-hazard mapping using an expert system and a GIS." IEEE transactions on geoscience and remote sensing, Vol. 45, No.2, pp. 522-531, 2007.

[6] Gelagay, Habtamu Sewnet, and Amare Sewnet Minale. "Soil loss estimation using GIS and Remote sensing techniques: A case of Koga watershed, Northwestern Ethiopia." International Soil and Water Conservation Research, Vol. 4, Issue 2, pp. 126–136, June 2016.

[7] Otukei, John Richard, and Thomas Blaschke. "Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms." International Journal of Applied Earth Observation and Geo information, Vol. 12, pp. S27-S31, Feb 2010.

[8] Kim, Kyoungok. "A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree." Pattern Recognition, Vol. 60, pp. 157-163, Dec 2016.

[9] Van de Vlag, Danil E., and Alfred Stein. "Incorporating uncertainty via hierarchical classification using fuzzy decision trees." IEEE Transactions on geoscience and remote sensing Vol. 45, No.1, pp. 237-245, 2007.

[10] Baraldi, Andrea. "Fuzzification of a crisp near-real-time operational automatic spectral-rule-based decision-tree preliminary classifier of multisource multispectral remotely sensed images." IEEE Transactions on Geoscience and Remote Sensing, Vol. 49, No.6, pp. 2113-2134, 2011.

[11] Otukei, J.R. and Blaschke, T., “Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms,” International Journal of Applied Earth Observation and Geoinformation, vol. 12, pp. S27-S31, 2010.

[12] Moustakidis, S., Mallinis, G., Koutsias, N., Theocharis, J.B. and Petridis, V., “SVM-based fuzzy decision trees for classification of high spatial resolution remote sensing images,” IEEE Transactions on Geoscience and Remote Sensing, vol. 50, no. 1, pp. 149-169, 2012.

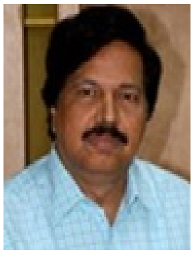
[13] Bakos, K.L. and Gamba, P., “Hierarchical hybrid decision tree fusion of multiple hyperspectral data processing chains,” IEEE Transactions on Geoscience and Remote Sensing, vol. 49, no. 1, pp. 388-394, 2011.

[14] Baraldi, A., Wassenaar, T. and Kay, S., “Operational performance of an automatic preliminary spectral rule-based decision-tree classifier of spaceborne very high resolution optical images,” IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 9, pp. 3482-3502, 2010.

[15] Millán, V.E.G., Sanchez-Azofeifa, G.A. and Malvárez, G.C., “Mapping Tropical Dry Forest Succession with CHRIS/PROBA Hyperspectral Images Using Nonparametric Decision Trees,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 6, pp. 3081-3094, 2015.

[16] Kheir, R.B., Abdallah, C., Runnstrom, M. and Martensson, U., “Designing erosion management plans in Lebanon using remote sensing, GIS and decision-tree modeling,” Landscape and Urban Planning, vol. 88, no. 2, pp. 54-63, 2008.

P. Kalyani received BC.A and M.C.A degrees from Sri Venkateswara University in 2003 and 2006 respectively. She is a Research Scholar in the department of Computer Science, Sri Venkateswara University, Tirupati, A.P, India. She worked as Assistant professor in PCET, Nellore,A.P, India. Her research focus is on Data Mining.



P.Govindarajulu Professor, Dept of Computer Science, Sri Venkateswara University, Tirupati, India. He received his M.Tech, IIT Madras (Chennai), Ph.D. IIT Bombay (Mumbai). His area of research: Databases, Data Mining, ImageProcessing, Intelligent Systems and Software Engineering,Parallel Computing.