Analysis of Land Cover Type Using Landsat-8 Data

V. Samuktha, M. Sabeshnav, A. Krishna Sameera, J. Aravinth, S. Veni

Analysis of Land Cover type using Landsat-8 data

Samuktha V[1], Sabeshnav M[1], Krishna Sameera A[1], Aravinth J[1], and Veni S[1]

[1]Department of Electronics and Communication Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India - 641112
samukthav10@gmail.com,j_aravinth@cb.amrita.edu

Abstract. Classification of images attributes to categorizing of images into various predefined groups. A particular image can be grouped into several diverse classes. Examining and ordering the images manually is a tiresome job particularly when they are abundant and therefore, automating the entire process using image processing and computer vision would be very efficient and useful. In this study, the Classifier and Regression trees (CART) algorithm is used to create a classifier model that classifies a region based on the feature specified. The Google Earth Engine (GEE) platform is utilized to conduct the study. The Tier 1 USGS Landsat 8 surface reflectance dataset is employed and is sorted according to the cloud cover. The features are then extracted and are merged to obtain a feature collection. This input imagery is further sampled using particular bands from the Landsat imagery to get a renewed feature collection of training data and the classifier model is trained using the CART Algorithm. An accuracy assessment is further performed to determine the exactness of the proposed model and the results are plotted using a confusion matrix. By applying the CART algorithm for image classification, an accuracy of 83% is achieved which was found to be better than the existing results.

Keywords: USGS · Landsat 8 · GEE · CART algorithm · Reflectance.

1 Introduction

Today, with the escalating necessity, capriciousness and the advancing demands of technologies like artificial intelligence, disciplines such as machine learning, and its subspaces have achieved enormous propulsion. The applications demand tools, such as classifiers, which support an immense volume of data, interpret them and derive features that are propitious. These classification methods aim in categorizing the pixels of a digital image into various classes [1].Usually, classification is implemented with the multi-spectral data and the spectral specimens existing in the features of every pixel is utilized for grouping. The primary intention of the classification of images is to distinguish and mark the features in an image and possibly plays the most essential role in digital image interpretation. Object classification is a complicated job and hence, image classification has a major part in the domain of computer vision. Classification is an art of labeling images into numerous categories. A particular picture can be grouped into several classes and the automation of the entire process of comparing and classifying images would surely make things easier than manual work. There are many real-time implementations which comprise computerized image design, extensive audio-visual databases, face recognition via social networks, and several other applications [2, 3] which require classifiers to obtain high accuracy. Image Classification usually involves the following steps - Initially, we perform the Image pre-processing. Pre-processing of images generally involves the examination of the image, Resizing and Data Augmentation methods such as Gray scaling of images, Gaussian Blurring, Reflection, Equalization, Rotation, and Translation of images [4–7]. This step is succeeded by the extraction of features and training the model. This is a vital process where the analytical or machine learning techniques are applied in classifying the attractive attributes of the picture and extracting features that might be unprecedented over a distinct class, and this will improve the classification model in distinguishing the various classes. This method, known as model training, is the process where the classifier model learns the features from the dataset. These features are then applied to the classification stage for object detection. The detected objects are grouped into predetermined groups by employing suitable grouping techniques that correlate the image and the target patterns [5].

Many classification algorithms are used in image classification and these algorithms can be broadly categorized based on the type of classification techniques the algorithms apply. Supervised classification is predicated on the basis that the pixel specimens of an image, which represent specific classes, can be selected by the user. The image processing tools are then steered by these pixels to utilize these training sites for classifying all additional pixels in the picture [8]. Once every information class has been statistically characterized, the image is classified by performing a reflectance measurement for each pixel and selecting the signatures it relates the most. Classification algorithms and regression techniques are utilized by the supervised classifiers to develop predictive models. Few such algorithms which employ supervised classification are logical regression, random forests model, decision trees, support vector machine (SVM) classifiers, convolutional neural networks, Naive Bayes and k-nearest neighbours [9–11].

One of the most accurate and frequently applied supervised learning techniques are tree-based algorithms. Predictive modeling with higher efficiency, greater stability, and ease of interpretation are some of their advantages. Models such as decision trees, random forest, and gradient boosting are commonly applied in a variety of data science puzzles. Therefore, it's very effective to acquire the knowledge of these algorithms and implement them while modeling [12, 13].

This study implements the decision tree algorithm, which is also known as the CART Algorithm is used to perform the classification of multispectral images.

Landsat 8 dataset is acquired from USGS and is used for experimental analysis. It employs the concept of supervised learning and has a predefined objective. It is mostly applied in decision making which works on a non-linear basis and has a simple linear decision aspect. They are versatile for resolving any query at hand - classification or regression [14]. We have performed all our study on the platform of Google Earth Engine (GEE). It is a cloud-based web application that allows scientists, researchers and developers to discover corrections, chart trends and quantify variances on the Earth's surface by blending the multi-petabyte directory of satellite imagery and geospatial datasets with planetary-scale examination capacities. It provides a global-scale insight and allows ready-to-use datasets. It makes use of a simple, yet powerful API and presents convenient tools to the users. We have made use of this platform to build a classification model using the CART Algorithm. Our model can classify the features of a given area into vegetation, water bodies, fallow land, and other areas respectively.

The remaining sections are established as follows: Section 2 describes the various works related to this paper. Section 3 presents a detailed description of the methodology of the entire study. It explains the concepts of dataset selection, feature extraction, sampling the imagery, and training the classifier. Section 4 illustrates the results obtained during the process of the study and Section 5 provides the concluding remarks of the study.

2 Related Work

Classifying agricultural lands using remote sensing is a well-studied and implemented idea. Previous works have implemented the classification of cropland and fallow land using multispectral data from sensors with lower resolution like MODIS (250m (bands 1-2) 500m (bands 6) 1000m (bands 8)). Zhuoting Wu and Prasad S. Thenkabail [15] have used MODIS data to classify the cropland. Kyle Pittman [16] also used MODIS data to map cropland in his work. Making use of the most advanced satellites will give us more accurate results. Using LANDSAT 8 (30m (bands 1-5, 7) 60m (bands 3-7) 15m (bands 8)) products has an edge over other works which hasn't been used.
Jeena Elsa George, J Aravinth, and Veni S [17] calculated Top of Atmospheric reflectance manually for land surface temperature whereas we used the LANDSAT 8's TOA product which makes our solution more suitable to implement in real life applications.

Work based on cropland classification mainly uses algorithms like Support Vector Machine (SVM) for classification. Amit Kumar Bhasukala [18] in his work used the SVM algorithm in the classification process. Jhinzhong Kang [19] also used SVM to classify cropland and fallow land. This work uses the CART algorithm in the classification phase which gives us some advantages over previous works such as: not relying on data distribution, no overgrowth in the decision tree. The CART Algorithm has an edge over the other classification techniques and algorithms. Apart from its high accuracy, the CART Algorithm, which predominantly works based on Decision trees can perform multiclass classification. It also provides the most model interpretability compared to the other algorithms which makes it more preferable. The features taken in the CART Algorithm have a non-linear relation which helps in not affecting the performance of decision trees.The fact that this algorithm can handle both numerical and categorical data which is most desirable for performing the classification. Both numerical and categorical data can be handled by the algorithm, making it desirable for performing the classification. M. Tugrul Yilmaz [20]in his work obtained an accuracy of just 70 % using Decision tree classification.

From these studies it was observed that (i) lower resolution sensors were used for input data (ii) Manual computation of TOA has increased computational complexity (iii) Most of the classification models were highly dependent on SVM. To overcome these limitations, an attempt is made to incorporate the CART algorithm for obtaining accurate results in classifying satellite imagery and the right selection of the input data also plays an indispensable role. Using the proposed methodology, it is possible to attain a classification accuracy as high as 82.305 % compared with the existing techniques used by various other authors.

3 Methodology

Fig. 1 depicts the block diagram of the proposed system. It has the following stages: (i) Selection of dataset and Region of Interest (ROI) (ii) Image Preprocessing (iii) Feature extraction and (iv) Classification. As presented in Fig. 1, Tier 1 USGS Landsat-8 surface reflectance dataset acquired using the OLI/TIRS sensors is selected. As a step towards preprocessing, the datasets are sorted based on the cloud cover. The images comprise of five Visible and Near Infrared (VNIR) bands and two Short Wave Infrared (SWIR) bands. It is further treated to orthorectified surface reflectance. The region of interest encompasses a small town in the district of Erode along the banks of the river Kaveri. This region was particularly chosen, as this expanse includes a variety of different landforms making it optimal for training and classification.

The features are extracted by utilizing the point marker tool available in GEE. The feature extraction is performed based on a single label called land cover and is assigned different class numbers to the various features extracted. The features extracted in this study are vegetation, water, fallow and others which include urbanized areas, roads, etc. Bands B2, B3, B4, B5, B6, B7, and B10 are selected from the Landsat image for training and are then used to obtain a feature collection of training data by sampling the input imagery. The CART Algorithm, along with the extracted features is used to train the classifier model. The trained model is used to classify the image. For better visualization, a color palette is applied to display the images based on the corresponding color that has been assigned to it in the feature collection.

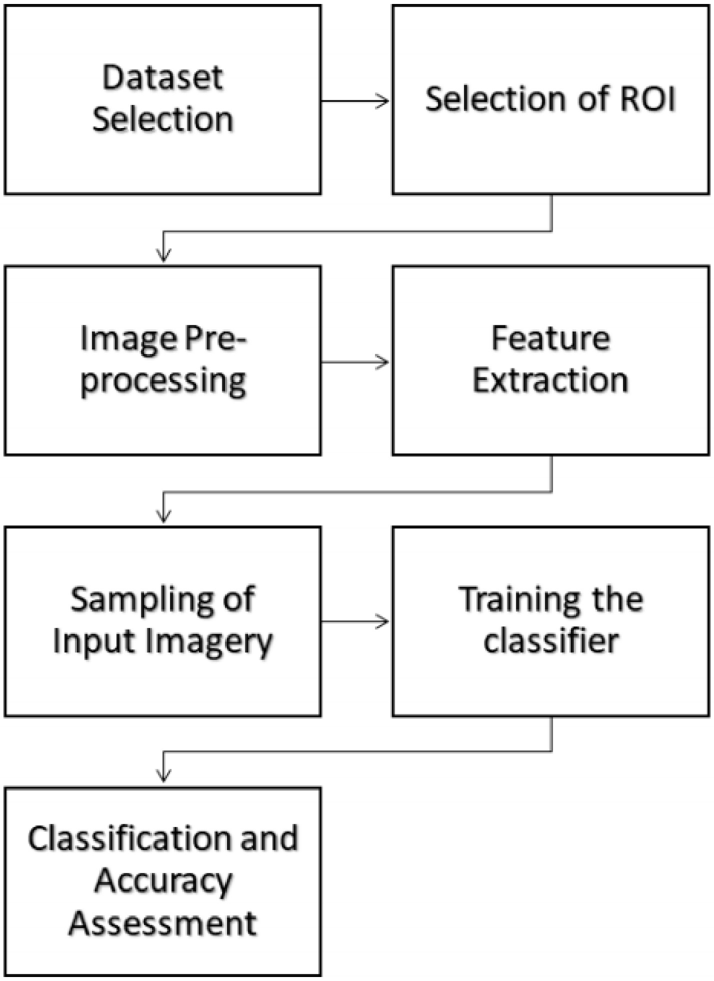

Fig. 1. The Block Diagram for the proposed methodology.

4 Results and Discussions

4.1 Dataset Selection

We have taken the Tier 1 USGS Landsat 8 dataset for our study. Landsat 8 dataset contains totally 11 bands namely coastal, blue, green, red, NIR, SWIR 1, SWIR 2, pan, Cirrus, TIRS 1, TIRS 2. Bands 1 − 7 and 9 has the resolution of 30 meters and band 8 has the resolution of 15 meters and bands 10 and 11 has 100 meter as their resolution but resampled to 30 meters. The data format of the Landsat 8 data is GeoTIFF with 16-bit pixel values. Surface reflectance

images which are atmospherically corrected obtained from the Landsat Operational Land Imager (OLI) and the Landsat Thermal Infra-Red Scanner (TIRS) sensors are included in this dataset [17].
These images comprise of five visible and near infrared (VNIR) bands and two short wave infrared (SWIR) bands. They are further treated to ortho-rectification of surface reflectance, and the two thermal infrared (TIR) bands are treated to temperature brightness ortho-rectification as shown in Fig. 2.

Managing Data Imbalancing The arrangement of Land Cover (LC) classes is often imbalanced with some majority LC classes dominating upon minority classes. Although standard Machine Learning (ML) classifiers can deliver high accuracies for majority classes, they comprehensively fail to present reasonable accuracies for minority classes. This is essentially due to the class imbalance problem. In our study, a hybrid data balancing technique called the Partial Random Over-Sampling and Random Under-Sampling (PROSRUS), was applied to solve the class imbalance issue. Unlike many data balancing techniques which attempt to fully balance datasets, PROSRUS uses a partial balancing procedure with hundreds of fractions for a majority and minority classes for balancing datasets. For this, time-series of Landsat-8 along with several spectral indices was used within the Google Earth Engine (GEE) cloud platform. It was discerned that PROSRUS performed better than numerous other balancing methods and improved the precision of minority classes without affecting the overall classification accuracy.
The PROSRUS method blends the two well-known data-level balancing methods, ROS [21] and RUS [22]. ROS, a simple oversampling technique, randomly duplicates samples from minority class(es) to balance the distribution of classes. Balancing an original imbalanced dataset completely utilising this method could create overfitting of the classifier due to the duplication [23]. On the other hand, RUS randomly eliminates samples from the majority classes to fit the data distribution. The principal deficiency of a fully balancing dataset using RUS is that it may miss relevant data [24]. The hybrid method used in our paper not only takes the advantages of both ROS and RUS but also restricts their limitations by analysing 200 different fractions in the balancing design.

4.2 Selection Of Region Of Interest

We have taken our region of interest in the Erode district of Tamil Nadu, India as shown in Fig. 3. The region bounded by the four coordinates are considered as follows:
[77.6975208136198, 11.388427392274773],
[77.72799070802898,11.388427392274773],
[77.72799070802898,11.411985832912757],
[77.6975208136198, 11.411985832912757], We have particularly chosen these coordinates, as this expanse of the area encompasses several different landforms making it optimal for training and classification. After obtaining the images

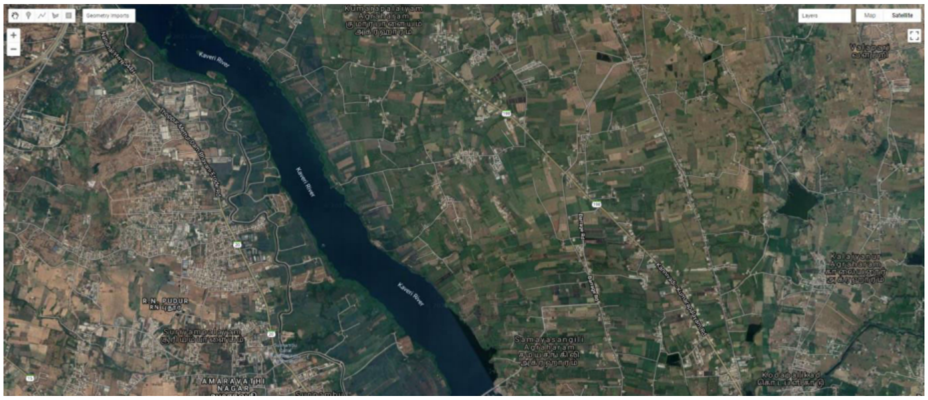

Fig. 2. Original Image Data from Study Area

with the lowest cloud cover, we apply the false-color composite to the image and proceed with the subsequent steps.

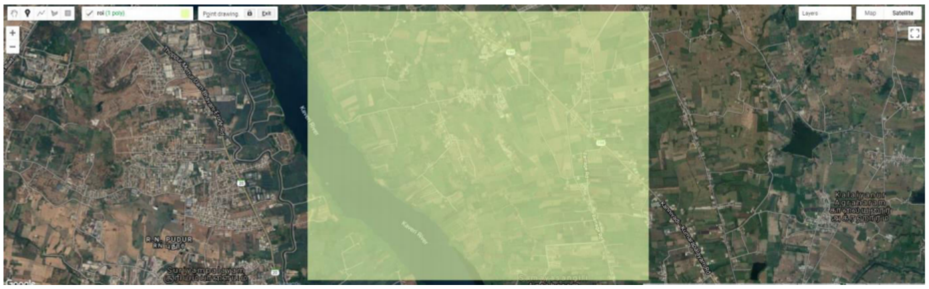

Fig. 3. The Selected region of interest

4.3 Feature Extraction

The next step is to extract the features from our region of interest. Here, the point marker feature available in the Google Earth Engine (GEE) is utilized to perform this task. The features extracted are named accordingly and are imported as a feature collection and are given different properties. The features of vegetation zones, water bodies, fallow lands and other urban areas are extracted are shown in Fig. 4,5,6, and 7 respectively. The features are then merged under a single property, named land cover as shown in Fig. 8. After performing the feature extraction, around 621 elements are obtained in the feature collection. After retrieving the images with the least cloud cover, we use the false-color composite to the image and continue with the subsequent steps.

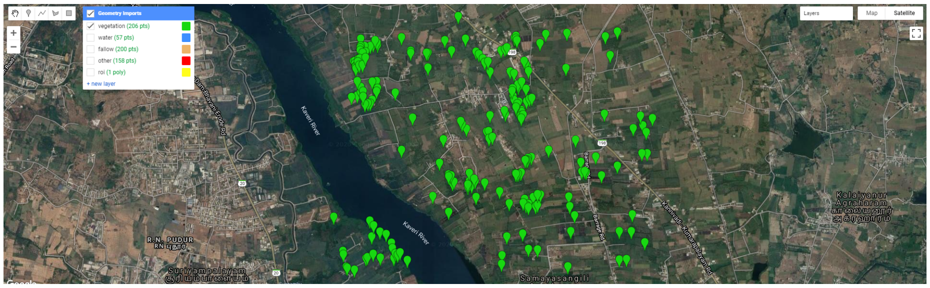

Fig. 4. Extracted Features Of Vegetation

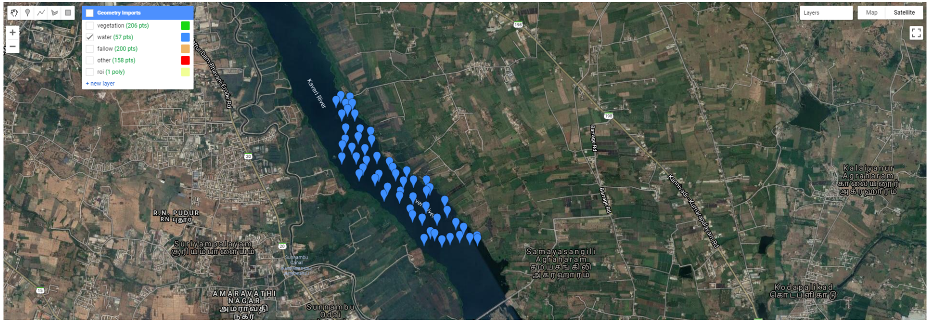

Fig. 5. Extracted Features Of Water Bodies

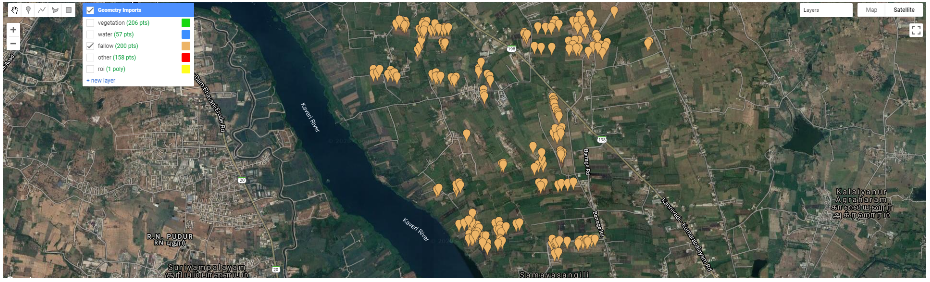

Fig. 6. Extracted Features Of Fallow Lands

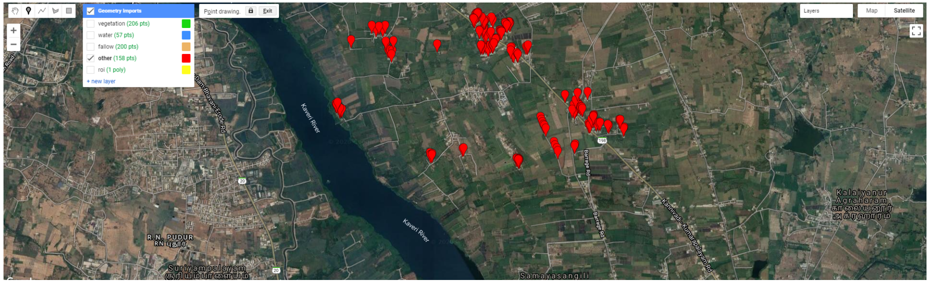

Fig. 7. Extracted Features Of Other Areas

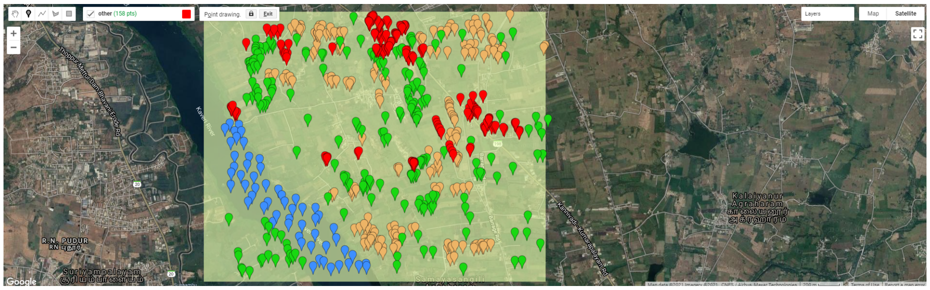

Fig. 8. Sampling the input imagery

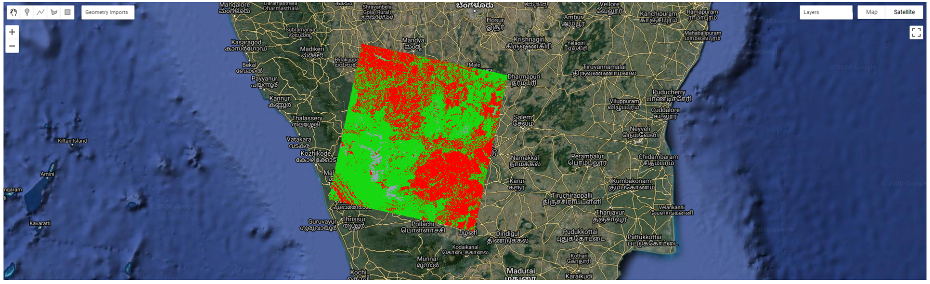

Fig. 9. Trained Classifier Model

4.4 Sampling the Input Imagery

Bands B2, B3, B4, B5, B6, B7, and B10 are selected from the Landsat image for training the image and then use these bands for sampling the input images to get a feature collection of training data.

4.5 Training the classifier

The classifier model is now trained using the extracted features by applying the CART algorithm. The trained model is used to classify the image. To have a better visualization, a color palette is incorporated to view the images based on the corresponding color that has been assigned in the feature collection as shown in Fig. 9. As the succeeding step, Land Use Land Cover Mask (LULC2010) is applied to classify the entire study area according to the colors specified in the color palette.

An accuracy assessment to determine the exactness of our model is conducted. Initially, a column of random uniforms was added to the dataset and the random function of GEE is used to split the datasets into testing and training datasets respectively. 70% of the dataset is used as the testing dataset and 30 % as the training dataset. The model is trained using the training dataset and is tested with the testing dataset. The Confusion matrix, a tool usually used to estimate the performance of machine learning problems, is adopted to plot the results as an error matrix, to determine the accuracy of the model. The confusion matrix obtained by this classifier is presented in Table1.

Table 1. The Confusion matrix of classified features upon the specific ROI for accuracy assessment

Class	**Accuracy**	**Precision**	**Recall**	**F1Score**
Vegetation	84.77%	0.95	0.79	0.86
Waterbodies	100%	1	1	1
Fallow land	87.24%	0.63	0.78	0.69
Other	92.59%	0.68	0.94	0.79

The image is classified based on the supervised classification method and to calculate the efficiency of the classification, the dataset collection has been split into a test set and training set where the training dataset records for 70 per cent of the images and is used to train the classifier. The rest of the images were taken as the testing dataset. To compute the accuracy of the trained model, it is executed with the testing dataset. Then it classifies the pixels of the test set based on the features that it has been trained to classify. The overall accuracy obtained in classifying the four classes of features is 82.305%
José M. Peña-Barragán, Moffatt K. Ngugi, Richard E. Plant, Johan Six [25] have carried out the assessment of crops by applying Object based Image Analysis (OBIA) along with various vegetation indices and crop phenology and obtained an overall accuracy of 79% and our work proved to be more robust in classifying the features in terms of accuracies, by applying the same classification model Decision Tree as they did.

5 Conclusion and Future Works

An improved method of multi-spectral image classification was attempted by using the CART algorithm, which leads to promising results as compared to the existing techniques. The Tier 1 USGS Landsat 8 surface reflectance dataset, which is a multispectral dataset, is employed and is ordered according to the cloud cover. The features were extracted and merged to achieve a collection of features. This input imagery is additionally sampled using the particular bands from the Landsat imagery to obtain a renewed feature collection of training data, and the classifier model is trained using the CART algorithm. An accuracy assessment is further performed to determine the exactness of the model developed and an overall accuracy of 82.305% was achieved. From this study, it is observed that the use of decision tree based algorithms enhanced the performance of the classification model with increased accuracies. This model can be employed in the fallow land classification as a future work.

References

1. Sunitha Abburu and Suresh Babu Golla. Satellite image classification methods and techniques: A review. *International journal of computer applications*, 119(8), 2015.
2. Ajay Nagesh Basavanhally, Shridar Ganesan, Shannon Agner, James Peter Monaco, Michael D Feldman, John E Tomaszewski, Gyan Bhanot, and Anant Madabhushi. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *IEEE Transactions on biomedical engineering*, 57(3):642–653, 2009.
3. Kwontaeg Choi, Kar-Ann Toh, and Hyeran Byun. Incremental face recognition for large-scale social network services. *Pattern Recognition*, 45(8):2868–2883, 2012.
4. S Sathya, Sundeep Joshi, and S Padmavathi. Classification of breast cancer dataset by different classification algorithms. In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 1–4. IEEE, 2017.
5. Pooja Kamavisdar, Sonam Saluja, and Sonu Agrawal. A survey on image classification approaches and techniques. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(1):1005–1009, 2013.
6. TV Nidhin Prabhakar and P Geetha. Two-dimensional empirical wavelet transform based supervised hyperspectral image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 133:37–45, 2017.
7. Sowmya V Kavitha Balakrishnan and Dr KP Soman. Spatial preprocessing for improved sparsity based hyperspectral image classification. *International Journal of Engineering Research & Technology*, pages 1–5, 2012.
8. Cristina Gómez, Joanne C White, and Michael A Wulder. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116:55–72, 2016.
9. Hetal Bhavsar and Amit Ganatra. A comparative study of training algorithms for supervised machine learning. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(4):2231–2307, 2012.
10. K Rithin Paul Reddy, Suda Sai Srija, R Karthi, and P Geetha. Evaluation of water body extraction from satellite images using open-source tools. In *Intelligent Systems, Technologies and Applications*, pages 129–140. Springer, 2020.
11. S Saravanamurugan, S Thiyagu, NR Sakthivel, and Binoy B Nair. Chatter prediction in boring process using machine learning technique. *International Journal of Manufacturing Research*, 12(4):405–422, 2017.
12. Kwontaeg Choi, Kar-Ann Toh, and Hyeran Byun. Incremental face recognition for large-scale social network services. *Pattern Recognition*, 45(8):2868–2883, 2012.
13. Sonia Singh and Priyanka Gupta. Comparative study id3, cart and c4. 5 decision tree algorithm: a survey. *International Journal of Advanced Information Science and Technology (IJAIST)*, 27(27):97–103, 2014.
14. Sincy V Thambi, KT Sreekumar, C Santhosh Kumar, and PC Reghu Raj. Random forest algorithm for improving the performance of speech/non-speech detection. In *2014 First International Conference on Computational Systems and Communications (ICCSC)*, pages 28–32. IEEE, 2014.
15. Zhuoting Wu, Prasad S Thenkabail, Rick Mueller, Audra Zakzeski, Forrest Melton, Lee Johnson, Carolyn Rosevelt, John Dwyer, Jeanine Jones, and James P Verdin. Seasonal cultivated and fallow cropland mapping using modis-based automated cropland classification algorithm. *Journal of Applied Remote Sensing*, 8(1):083685, 2014.

16. Kyle Pittman, Matthew C Hansen, Inbal Becker-Reshef, Peter V Potapov, and Christopher O Justice. Estimating global cropland extent with multi-year modis data. *Remote Sensing*, 2(7):1844–1863, 2010.
17. Jeena Elsa George, J Aravinth, and S Veni. Detection of pollution content in an urban area using landsat 8 data. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 184–190. IEEE, 2017.
18. Amit Kumar Basukala, Carsten Oldenburg, Jürgen Schellberg, Murodjon Sultanov, and Olena Dubovyk. Towards improved land use mapping of irrigated croplands: Performance assessment of different image classification algorithms and approaches. *European Journal of Remote Sensing*, 50(1):187–201, 2017.
19. Jinzhong Kang, Hongyan Zhang, Honghai Yang, and Liangpei Zhang. Support vector machine classification of crop lands using sentinel-2 imagery. In *2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics)*, pages 1–6. IEEE, 2018.
20. M Tugrul Yilmaz, E Raymond Hunt Jr, Lyssa D Goins, Susan L Ustin, Vern C Vanderbilt, and Thomas J Jackson. Vegetation water content during smex04 from ground data and landsat 5 thematic mapper imagery. *Remote Sensing of Environment*, 112(2):350–362, 2008.
21. Nitesh V Chawla. Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, pages 875–886, 2009.
22. Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.
23. Miriam Seoane Santos, Jastin Pompeu Soares, Pedro Henrigues Abreu, Helder Araujo, and Joao Santos. Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. *ieee ComputatioNal iNtelligeNCe magaziNe*, 13(4):59–76, 2018.
24. Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.
25. José M Peña-Barragán, Moffatt K Ngugi, Richard E Plant, and Johan Six. Object-based crop identification using multiple vegetation indices, textural features and crop phenology. *Remote Sensing of Environment*, 115(6):1301–1316, 2011.