Prediction of land-change using machine learning for the deforestation in Paraguay

Max Muller[1], Shweta Vincent[2], Om Prakash Kumar[3]
[1]Karlsruhe Institute of Technology, Germany
[2]Department of Mechatronics Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, India
[3]Department of Electronics and Communication Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, India

Article Info	ABSTRACT
	Northwestern Paraguay is being deforested at a very rapid rate. This article studies the rate of deforestation that has happened in this area using satellite images of LandSAT5 and LandSAT7. The rate of the deforestation is detected from 1986 to 2011, graphically using which future prediction is made. The images of LandSAT8 are used to validate the prediction made until 2018. An extrapolation of the graph shows that the process of deforestation is already 3 years ahead of its forecast. An overall accuracy of 98% has been achieved using this technique. The root mean squre error (RMSE) is around 0.011.

Corresponding Author:

Shweta Vincent,
Department of Mechatronics Engineering,
Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, India.
Email: shweta.vincent@manipal.edu

1. INTRODUCTION

In times of global warming, prediction tools for change of landscape and land-use are essential for showing and surveying the consequences of global warming. Landscape changes could occur due to active manmade processes such as deforestation or growth of cities or could be passive natural such as droughts or rise in seawater level. These changes serve as critical information for decision makers of policies related to the conservation of the environment. Machine learning has revolutionized the detection and prediction modelling for changes in land cover features in recent times. Volumes of literature is available to study and adapt from, to detect and predict changes in land features. Various machine learning algorithms and procedures have been used by researchers worldwide in order to achieve this prediction.

An ensemble of Decision Tree and Random Forest algorithms has been widely made use of for classification problems of land features and creation of land maps. Gauci A. et al [1] have presented an article that focuses on the usage of DSLR images of the Maltese islands to perform an automatic land cover mapping of the region. They have attempted to map different forms of vegetation that is present in this area. They have performed supervised classification to detect the forms of vegetation present in the area. Their classification of types of vegetation is validated with *in situ* data and proves to be in accordance with the *in situ* data. Cheun J. et. al [2] have utilized airborne hyperspectral images for classification of ecotopes.

They have investigated the usage of the adaboost and random forest classification algorithms perform ecotope mapping of different types. Their predicted classifications are in good sync with the available *in situ* data. Hansen M. et al. [3] have also presented an article that validates the usage of decision tree classifiers for usage in land cover mapping. Punia M. et al. [4] have utilized the IRS-P6 AWiFS data on decision tree classifiers for land use and land cover mapping over Delhi. An overall classification accuracy of around 91% has been achieved by the authors. Rodreiguez-Galiano V. F. et al. [5] have presented yet another research article stressing on the feasibility of Random Forest algorithm for classification of land cover mapping.

The random forest algorithm enables them to achieve an accuracy of around 92% for a complex terrain data set. Effective mapping of urban areas with an accuracy of 93% has been achieved by Schneider A. et al. [6] using an ensemble of decision tree classifiers. Stumpf A. et al. [7] have used the random forest algorithm along with object-oriented analysis for mapping of landslide images procured using very high-resolution images. They have been able to achieve accuracy between 73% to 87% for four test sites. Waske B. et al. [8] have innovated and created a software platform named imageRF for analysis of remotely sensed images using the random forest algorithm. Regression models of machine learning also find their application in predicting the change in land cover or land usage based on prior data sets. Extensive research has been carried out in this area. Nurwanda A. et al. [9] have observed the satellite images of Bogor in Indonesia to study and predict changes in land usage and land temperature. They have used a multi-layer perceptron model and Markov Chain predictions to arrive upon prediction models for the city. Mathew A. et al. [10] have generated a similar prediction model as [9] for the city of Chandigarh, India. They have demonstrated the superiority of the support vector regression model over the artificial neural network model to predict land surface temperature for the city of Chandigarh. A regression model has been created and utilized for the assessment of exposure to air pollution by Morley D. W. et al. [11]. Silva A. C. O. [12] have used the Bayesian network approach for prediction of the amount of deforestation in the Amazon rain forest area. Sales M. et al. [13] have created a geostatistical model for prediction of rate of deforestation for a short term. They have also used the Amazon forest data set as their reference. They have been able to achieve an accuracy of 90% in predicting the next area of deforestation in the Amazon.

Hyperspectral imaging has gained momentum due to the intricate details of imaging provided by state-of-the-art systems. The challenge faced by researchers is to extract information available through these images effectively. Vishvanathan S. et al. [14] have presented a detailed article on the fundamentals of the usage of hyperspectral images, their denoising techniques and classification techniques. Zhang Y. [15] has presented the development of a novel algorithm for improving the spatial resolution of hyperspectral resolution based on the mixing of the model that is observed and the spatial model. Mohamed A. B. et al. [16] have showcased the success of a technique of spectral unmixing for improving the image resolution in hyperspectral images. Bierniraz J. et al. [17] have showcased the need for unmixing of spectral signatures due to low spatial resolution of hyperspectral images. Their proposed algorithm unmixes the spectral bands of the image without affecting the spatial resolution of the image. Patro R.N. et al. [18] have presented yet another technique for improvement of spatial resolution of hyperspectral images using a Gaussian filtering technique. Their technique yields better classification results using the SVM classifier. Effects of deforestation have also been studied in [19-21]. Our research article presents a study of the deforestation in Paraguay specifically, the Northwest of the town of Filadelfia [22-25]. The deforestation of the tropical forest is a major issue due to:
− the loss of land for many species
− endangering the last indigenous people with no contact to our modern societies
− the destruction of massive CO_2-stores
− the release of greenhouse gases in form as barbecue coal sold in European discounters certified by FSC (The Forest Stewardship Council sets standards for responsible forest management with the slogan "Forests for all forever").

This research article aims at analysing the trend in the deforestation in Filadelfia from 1980 till 2010 using LandSAT5 images. This trend is extrapolated and the predictions made are tested using LandSAT7 and/or LandSAT8 images from 2010 till 2018 using a rectified regression line. This activity enables to predict the rate of deforestation in future and create policies to curb the rate of deforestation in the area to raise an awareness in the global community. Section 2 of this article presents the methodology for data collection and pre-processing of data. Section 3 presents the prediction model and discussion on the same. Section 4 conlcudes the article.

2.   METHOD: DATA COLLECTION AND PRE-PROCESSING

Satellite images of the region Filadelfia have been acquired using the Land Viewer tool provided by earth observing system (EOS) [17]. The main requirement of the images is to have a resolution that allows

a clear distinction between tropical forest and other features of the landscape, such as rivers, fields or villages. ESO [17] provides on its webpage the tool, *Land Viewer*, to view and download images of regions of the Earth acquired by specific satellites. Amongst them, the LandSAT5 TM (thematic mapper) is chosen to fulfil our requirements of high resolution imagery. *Land Viewer* provides images of specific areas over nearly 30 years (1984-2011). The images can be downloaded in different resolutions. A free trial allowed the best resolution of 240 m/pixel which a suitable resolution for our analysis. The colour code can be changed, according to the requirements. For our analysis, the colour code *Colour Infrared (Vegetation)* with the band B4, B3 and B2 has been used to clearly indicate the vegetation differences.

The Landsat 5 program stopped in January 2013 and the last images (of high quality) of the region of concern using the *Land Viewer* were acquired until 2011. To further collect data of the same region with an equivalent resolution, the Landsat7 TM has been used for data from 2011 to 2017. Here, the colour code *Color Infrared (Vegetation)* contains bands B4, B3 and B2. The colour code of both satellites (LandSAT 5 and LandSAT 7) are identical. In addition to LandSAT7, LandSAT8 has been used for a further validation. The colour code *Color infrared (Vegetation)* contains the following bands: B5, B4 and B3. The vegetation ratio is such a parameter which depicts the change in tropical forest cover in a region. It is defined as the ratio of pixels representing tropical/rain forest ($n_{tropical\_forest}$) and total number of pixels ($n_{total} - n_{frame}$) in the image and is shown in (1).

$$vegRatio = \frac{n_{tropical\_forest}}{n_{total} - n_{frame}} \tag{1}$$

Figure 1 shows the region of Paraguay which has been considered for our analysis. Figure 2 presents a sample image of the 92nd day of the year 2000. The pixels which need to be classified are coloured and the black pixels (marked as 'Pixel of frame') are ignored. On importing the image into MATLAB software, a 3 dimensional matrix (N_x x N_y x 3) for the image gets created. It has entries in the x and y axes for each pixel with 3 colour entries in RGB (ranging between 0 to 255) for the corresponding pixel. All pixels for the frame which are black show and colour intensity below 3. These are ignored for further processing as they do not contain any information of interest.

Figure 1. Area of concern-northwestern Paraguay

Consider the image of the 92nd day of the year 2000 as shown in Figure 2. The pixels of the image which have to be classified are coloured and the black frame is to be ignored (marked as 'Pixel of frame'). A matrix with dimensions N_x x N_y x 3 is created when an image is imported into the MATLAB software. Every pixel is allotted positions on the x-axis and y-axix with 3 colour entries for every corresponding location. These colour entries stand for the intensity of the colours RGB (Red-Green-Blue) and have values

between 0 and 255. All frame pixels (black) show colour intensities below 3. These are ignored for further computations.
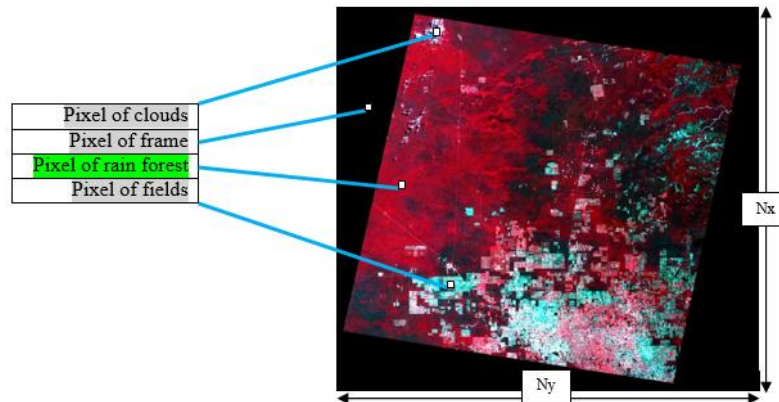

Figure 2. Example of classification of the image (92nd day of the year 2000) in color infrared (B4, B3, and B2) of LandSAT5

The classification of every pixel as either a tropical forest or not is binary classification problem. In multiple iterations and images, it has been found that there is a threshold for pixel intensity classification of tropical forests for LandSAT5. After every pixel has been classified, some individual pixels didn't fulfil the criterion of the threshold- these are called stand-alone pixels. From Figure 2, one can observe that most fields are about 10 x 10 pixels (approximately 3 km x 3 km) large and easily identifiable by the naked eye. In order to find these individual pixels, the classification matrix is being smoothed by multiple iterations so that no stand-alone pixels appear. Figure 3 outline the procedure to fit a polynomial to the calculated vegetation ratio images for each year. MATLAB software fits a second order polynomial to the points as shown in Figure 4. The x-axis value of the graph has been calculated by the year added with the day of capture divided by the total of days in a usual year as shown in (2).

$$x_{axis} value = n_{year} + \frac{n_{day\,of\,capture}}{n_{total\,days\,in\,a\,year}} \tag{2}$$
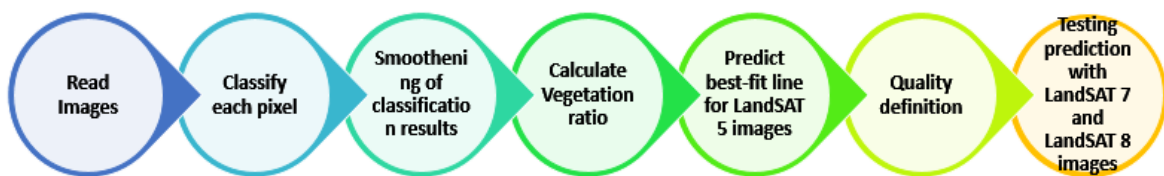

Figure 3. Algorithm to predict the rate of deforestation the years to come

In Figure 4 it can be observed that some of the vegetation ratio points (marked inside the black box) deviate from the fitted regression line by a considerable amount. This indicates that a method needs to be devised to identify the quality of data obtained from the images of LandSAT5 for the presented region. Multiple ways of identifying the quality of the data are possible. To identify the accuracy of the regression line, a typical RSME (root square mean error) value is calculated. However, since the data points may vary due to the quality of the classification threshold and the quality of the picture (e.g. presence of clouds or smoke from deforestation by fire, as shown in Figure 5, the procedure of the exact treatment of the data points needs to be discussed. The rainforest is not able to recover within a year in a way that the vegetation rate raises by over 0.07% as the data points would suggest as shown in Figure 6.
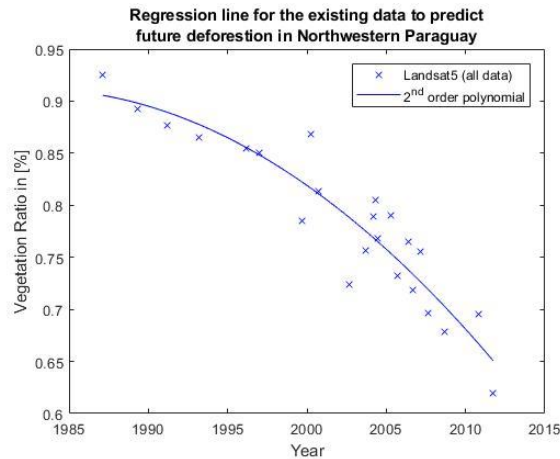
Figure 4. Vegetation ratio of Landsat5 images with 2nd order polynomial regression line

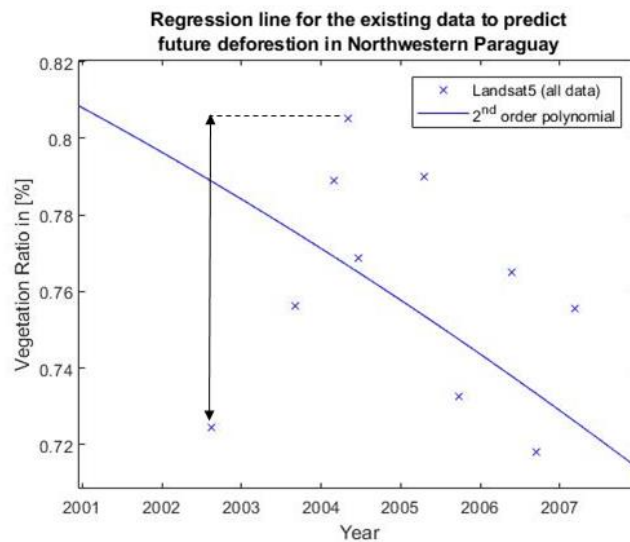Figure 5. Deforestation by fire clearing visible on Landsat5 image of 2008

Figure 6. Detailed view of the data for LandSAT5 for 2001-2008 indicating the big uncertainty of the data

Observing the images very closely, one could even go that far and assume that there is no significant reforestation process to recover the loss of rainforest in the region observed. Since all images represent the same area and by identifying the quality of the classification by simply looking at each image, it might make sense to work only with the "best" data points and predict a curve from these "best" data points only. The largest amount of the variance of the data points is due to the uncertainty of the threshold, not of the actual change of vegetation ratio of the image. The correlation coefficient of two random variables is a measure of their linear dependence as shown in (3). If each variable has N scalar observations, then the Pearson correlation coefficient is defined as:

$$\rho(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{A_i-\mu_A}{\sigma_A}\right)\left(\frac{B_i-\mu_B}{\sigma_B}\right) \tag{3}$$

Where μ_A and σ_A are the mean and standard deviation of A, respectively, and μ_B and σ_B are the mean and standard deviation of B." In Figure 7, the coefficients of correlation between the data points (set A) and the prediction line (set B with same x-value) are for all data and best data 0.931 and 0.987 respectively. The RMSE (4) are for all data and best data 0.027 and 0.011 respectively.

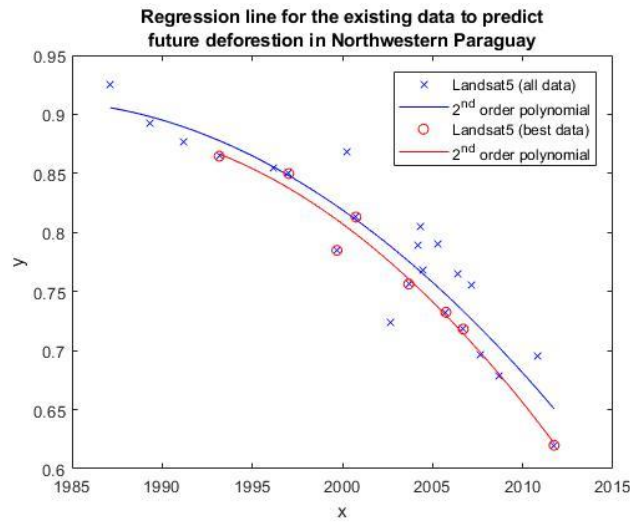$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(A_i - B_i)^2}{N}} \qquad (4)$$

Figure 7. Comparison of prediction line for all and "best" LandSAT5 data

3.   TESTING RESULTS AND DISCUSSIONS

In order to test the predicted 2nd-order polynomial as shown in Figure 7, data of LandSAT7 has been gathered. The goal is, to find out, if the prediction of the vegetation ratio for the years 2011-2017 fits the actual vegetation ratio received through the images of LandSAT7 taken from 2011-2017. The quality of images of LandSAT7 differs from that of LandSAT5. Though the resolution is identical, dark strips appear on the sides of the images visible at every image of LandSAT7 for most regions. The challenge is to smoothen out these strips with an algorithm. This is done by checking every set of 3 x 3 pixels for its third lowest colour intensity. Setting the colour of all 3 x 3 pixels to this specific colour will decrease the resolution. This procedure has been tested also on LandSAT5 images to test the change of vegetation ratio by this algorithm hoping for a minimum change. Applying this algorithm on the LandSAT5 images, the maximum difference of the calculated vegetation ratio of the images with and without prior smoothing algorithms is of 1.6%. This is shown in (5). Figure 8 show the vegetation ratio points of LandSAT 7 images on the same plot of Figure 7.

$$\Delta VegRatioLS5 = \frac{|VegRatioLS5_{PriorSmoothing} - VegRatioLS5_{Original}|}{\frac{VegRatioLS5_{PriorSmoothing} + VegRatioLS5_{Original}}{2}} * 100 \qquad (5)$$
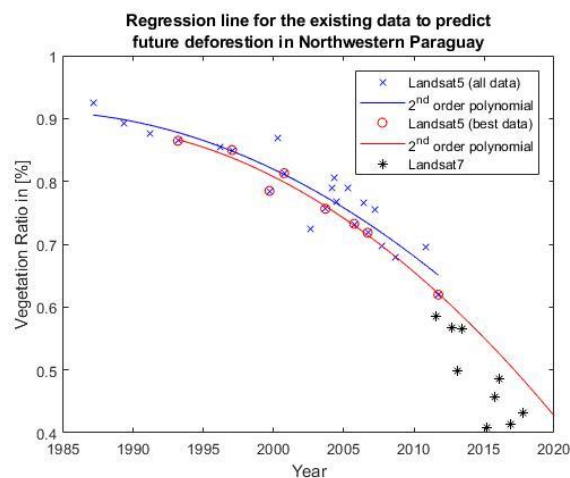

Figure 8. Location of LandSAT7 data compared to prediction line of "best" data of LandSAT5

Prediction of land-change using machine learning for the deforestation in Paraguay (Max Muller)

To avoid the inaccuracy of LandSAT7 images due to the black shades on the sides, LandSAT8 has been found not to contain black strips. A LandSAT8 image was chosen due to a good and clear quality of the image and hence for its ability for clear classification. Even though still some parts on the right middle and upper parts are not clear, the classification errors even more or less out. The result is the following shown in Figure 9. Image in Figure 8 of LandSAT8 from 2016 (day 236) has a vegetation ratio of 43%. The predicted vegetation ratio due to Landsat5 best images (red line) is of 51%. Using the following method to calculate the percentile difference, we get a prediction error of 16%, which is very significant. Hence, the deforestation ratio raised even higher than the predicted 2nd order. This is computed using (6). The validation of the prediction model described is shown in Figure 10. Figure 11 shows prediction of the year rain forest vanishes in the area observed.

$$\Delta VegRatio2016 = \frac{|VegRatio2016_{predicted} - VegRatio2016_{Landsat8}|}{\frac{VegRatio2016_{predicted} + VegRatio2016_{Landsat8}}{2}} * 100 \qquad (6)$$
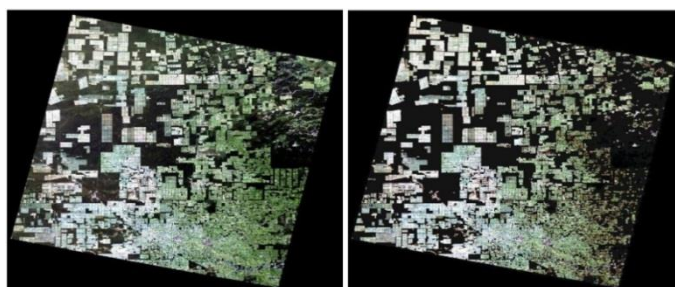

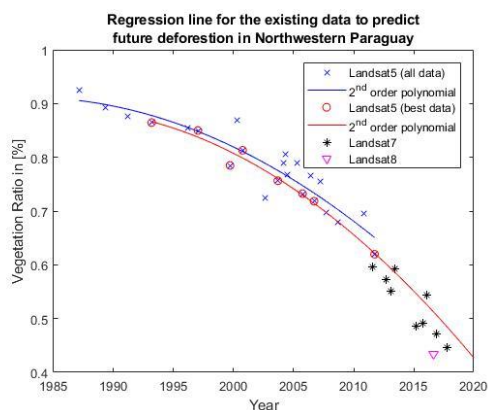Figure 9. Classification results of LandSAT8 image

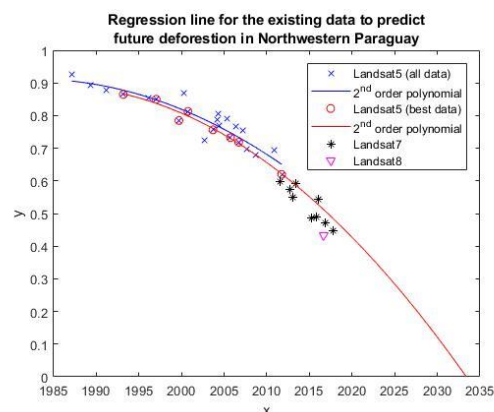Figure 10. Validation of data prediction from LandSAT5 by "true" data of Landsat8 image

Figure 11. Prediction of the year rain forest vanishes in the area observed

4.   CONCLUSION

In this paper, the images of the satellites LandSAT5, LandSAT7 and LandSAT8 have been used to determine the amount of rainforest in a specific region in North-western Paraguay. In order to analyse the change of rainforest (vegetation ratio), images of 1986 to 2018 have been processed by an algorithm to classify each pixel of the image as displaying rainforest or not. By the aid of MATLAB curve fitting toolbox, the data till 2010 showed to fit very well on a 2nd-order polynomial curve. This implies that the deforestation of the rainforest is increasing in speed. The deforestation rate, meaning the absolute value of the slope of the 2nd-order polynomial curve is not constant but the deforestation rate is growing. In addition, it was found out, that during the last decade 2010th, the deforestation rate increased more rapidly as forecasted by the 2nd-order polynomial curve. In case the curve of LandSAT5 would be true, all rain forest could be gone latest by 2035. The result of the LandSAT8 image suggests an even quicker disappearance than 2035. By 2018, the data received by LandSAT8 shows that the deforestation process is already 3 years ahead of its forecast. This is represented in Figure 11.

REFERENCES

[1] Gauci, A., Abela, J., Austad, M, Cassar, L.F., Adami, K.Z, "A Machine Learning Approach for Automatic Land Cover Mapping from DSLR images of the Maltese Islands," *Journal of Environmental Modelling & Software*, vol. 99, pp. 1-10, 2018.

[2] Cheun, J., Chan, W., Paelinkx, D, "Evaluation of Random Forest and Adaboost tree- based Ensemble Classification and Spectral Band Selection for Ecotope Mapping using Airborne Hyperspectral Imagery," *Journal of Remote Sensing of Environment*, vol. 112, no. 6, pp. 2999-3011, 2008.

[3] Hansen, M., Dubayah, R., Defries, R, "Classification trees: An Alternative to Traditional Land Cover Classifiers," *Remote Sensing Letters*, vol. 17, pp. 1075-1081, 1995.

[4] Punia, M., Joshi, P. K., Porwal, M. C, "Decision Tree Classification of Land Use Land Cover for Delhi, India using IRS-P6 AWiFS data," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5577-5583, 2011.

[5] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J. P, "An Assessment of the Effectiveness of Random Forest Classifer for Land-Cover Classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93-104, 2012.

[6] Schneider, A., Friedl, M. A., Potere, D, "Mapping Global Urban Areas using 500-m data: New Methods and Datasets based on 'Urban Ecoregions," *Journal of Remote Sensing of Environment*, vol. 114, no. 8, pp. 1733-1746, 2010.

[7] Stumpf, A., Kerle, N, "Combining Random Forest and Object-oriented Analysis for Landslide Mapping form High Resolution Imagery," *Proceedia Environmental Sciences*, vol. 3, pp. 123-129, 2011.

[8] Waske, B., van der Linden, S., Oldenburg, C., Jakimow, B., Rabe, A., Hostret, P, "ImageRF- A User-oriented Implementation for Remote Sensing Image Analysis with Random Forests," *Journal of Environmental Modelling and Software*, vol. 35, pp. 192-193, 2012.

[9] Nurwanda, A., Honjo, T, "The Prediction of City Expansion and Land Surface Temperature in Bogor City," *Indonesia. Journal of Sustainable Cities and Society*, vol. 52, pp. 101772, 2020.

[10] Mathew, A., Sreekumar, S., Khandelwal, S., Kumar, R, "Prediction of Land Surface Temperatures for Surface Urban Heat Island Assessment over Chandigarh city using Support Vector Regression model," *Journal of Solar Energy*, vol. 186, pp. 404-415, 2019.

[11] Morley, D. W., Gulliver, J, "A Land use Regression Variable Generation, Modelling and Prediction Tool for Air Pollution Exposure Assessment," *Journal of Environmental Modelling & Software*, vol. 105, pp. 17-23, 2018.

[12] Silva, A. C. O., Fonesca, L. M. G., Korting, T. S., Escada, M. I. S, "A Spatio-Temporal Bayesian Network Approach for Deforestation Prediction in an Amazon Rainforest Expansion Frontier," *Journal of Spatial Statistics*, vol. 35, pp. 100393, 2020.

[13] Sales, M., de Bruin, S., Herold, M., Kyriakidis, P., Souza Jr., C, "A Spatioemporal Geostatistical Hurdle Modle Approach for Short-term Deforestation Prediction," *Journal of Spatial Statistics*, vol. 21, no. A, pp. 304-318, 2017.

[14] Vishvanathan, S., Soman, K. P., Hassaballah, M, "Hyperspectral Images: Fundamentals and Advances," *Studies in Computational Intelligence*," vol. 110, pp. 1-24, 2019.

[15] Zhang, Y, "Spatial Resolution Enhancement of Hyperspectral Image based on the Combination of Spectral Mixing Model and Observation Model," *Proceedings of SPIE-The International Society for Optical Engineering*, p. 9244, 2014.

[16] Mohamed, A. B., He, M., Mei, S, "Hyperspectral Image Resolution Enhancement Using High-Resolution Multispectral Image Image Based on Spectral Unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 10, pp. 6574-6583, 2014.

[17] Bieniarz, J., Cerra, D., Avbelj, J., Müller, R., Reinartz, P, "Hyperspectral Image Resolution Enhancement Based on Spectral Unmixing and Information Fusion," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 10, p.5194, 2011.

[18] Patro, R. N., Subudhi, S., Biswal, Pradyut, "Effect of Spatial Enhancement on Hyperspectral Image Classification," vol. 10, no. 11, pp. 1-6, 2019

[19] Karia, J.P., Porwal, M.C., Roy, P.S, "Forest change detection in Kalarani round, Vadodara, Gujarat— a Remote Sensing and GIS approach," *Journal of the Indian Society of Remote Sensing*, vol. 29, p. 129-135, 2001.

[20] Bouvet, A., Mermoz, S., Ballère, M., Koleck, T., Le Toan, T, "Use of the SAR Shadowing Effect for Deforestation Detection with Sentinel-1 Time Series," *Remote Sensing*, vol. 10, p. 1250, 2018.

[21] Shimizu, K., Ota, T., Mizoue, N, "Detecting Forest Changes using LandSAT 8 and Sentinel-1 Time Series Data in Tropical Seasonal Forests. Remote Sensing," vol. 11, no. 16, p. 1899, 2019.

[22] "Vast Tracts in Paraguay Forest Being Replaced by Ranches", 2012, [Online], Available at: https://www.nytimes.com/2012/03/25/world/americas/paraguays-chaco-forest-being-cleared-by-ranchers.html.

[23] Baumann, M., Israel, C., Piquer-Rodriguez, M., Gavier-Pizzaro, G.,Volante, J. N, "Deforestation and Cattle Expansion in Paraguayan Chaco 1987-2012," *Journal of Regional Environmental Change*, vol. 17, no. 4, pp. 1179-1191, 2017.

[24] Hansen, M.C., P.V. Potapov, R. Moore, M. Hancher, S.A. Turubanova, A. Tyukavina, D. Thau, S.V. Stehman, S.J. Goetz, T.R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C.O. Justice, and J.R.G. Townshend, "High-Resolution Global Maps of 21st-Century Forest Cover Change," Journal of Science, vol. 342, no. 6160, pp. 850-853, 2013.

[25] "Nasa's Earth Observing System", 2017, [Online], Available at: https://eospso.nasa.gov/content/nasas-earth-observing-system-project-science-office.

BIOGRAPHIES OF AUTHORS

Max Muller, former bachelor student of mechanical engineering at Karlsruhe Institute of Technology and master student of mechanical engineering at Karlsruhe Institute of Technology, Karlsruhe, Germany, and Pennsylvania State Institute, PA, US, conducted a research project at Manipal Institute of Technology, KA, India. After his master thesis at the German Aerospace Center on type days for energy system modelling, he currently works at the German TSO TransnetBW GmbH.

Shweta Vincent, is working as an Assistant Professor in the Department of Mechatronics at Manipal Institute of Technology affiliated to the Manipal Academy of Higher Education, India. She completed her Ph.D. in 2020 in the field of Antenna design for a GB-SAR system used for landslide monitoring from Karunya Institute of Technology and Sciences, India. Her research areas include Antenna design, Machine Learning and GIS.

Om Prakash Kumar received his Bachelor of Technology degree from Karunya University, India in Electronics and Communication Engineering. He received her Master of Technology degree from Karunya Unviersity, India in VLSI. His research areas of interest include Design of VLSI systems and Antenna Design. He is currently employed with Manipal Institute of Technology, Manipal Academy of Higher Education, India in the department of Electronics and Communication Engineering.