Deforestation Analysis using Unsupervised Clustering and Satellite Images

K. Pradeep Mohan Kumar, Rushan Mukherjee, Mayank Dewangan

Abstract**:** ***Expansion of farmland and unplanned land encroachments have increased as the earth's population boomed, which has led to uncontrolled deforestation across the world. Deforestation and industrialization have given rise to global warming, causing mayhem in the current ecosystem. The weather patterns are disrupted, and natural calamities are occurring more frequently. The after-effects of these events have to lead to dramatic losses in flora and fauna. Even though a large part of India's land is urbanized, there are many protected areas in specific parts of the country that represent significant vegetation that has been affected if we observe from the scale of a subcontinent. In this paper, we aim to trace the deforestation in the Sundarbans from 1988 to 2019. This period is essential as a lot of industrialization and population boom happened during this time. We have selected the Sundarbans because mangroves are a natural defense to cyclones and also provide shelter to a plethora of living organisms. Our study area covers more than 7000 square kilometers of the Indian Sundarbans. The satellite images are from Landsat-5, Landsat-7, and Landsat-8, which are orthorectified. We make use of open-source software like Quantum GIS (QGIS), Google Earth Engine (GEE), and Google Colab, which is a Python IDE in this project. We make use of the K-means clustering, which is an unsupervised learning algorithm. Here, we have described a method to analyze deforestation accurately using low-cost techniques, which can be used by underdeveloped nations and private organizations to help in the fight against illegal deforestation.***

Keywords**:** ***Deforestation, K-means Clustering, Machine learning, NDVI.***

I. INTRODUCTION

Geographic information system (GIS) frameworks are used for gathering, regulating, and understanding data. GIS integrates many types of data despite being based on the science of geography. It analyzes spatial locations and organizes them into layers of information. These are then transformed into visualizations using maps and 3D scenes. We plan to integrate GIS with Machine Learning classification techniques to conduct our project.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author

K. Pradeep Mohan Kumar*, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India. Email: pradeep.k@ktr.srmuniv.ac.in

Rushan Mukherjee, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India. Email: rushanmukherjee@gmail.com

Mayank Dewangan, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India. Email: mayankdewangan119@gmail.com

Unsupervised classification is a sub branch of machine learning classification algorithms which tries to find patterns from unlabeled data.

This project will be a testament of interoperability between the above mentioned domains to analyze real world scenarios like deforestation. Deforestation is the result of rapid industrialization of rural land of India. To combat deforestation is to combat global warming and reduce pollution levels of the whole world. The earth just lost a huge chunk of its forest land in the Amazon because of illegal mining and human negligence which caused wildfires. The results of our study will not only help government reduce crucial time in identifying "stressed" areas prone to deforestation but will also provide a consolidated map of the locations where the forests are blooming again .Our information can also be used by NGOs who are willing to do grass-root level work of educating people not to destroy forests illegally. In this project, we aim to make a common architecture for deforestation analysis of India and also making it open source so that anyone can access the information and learn from it.

A. Study Area

The Sundarbans is the biggest mangrove forest region in the country, covering approximately one million hectares (2.47 million acres) in India and Bangladesh. In the tropics and subtropics some of the most carbon-rich and productive forests are concentrated along the coastlines. Mangrove habitats, consist of salt-tolerant shrubs and trees, play a crucial role in erosion and flood protection, fisheries preservation, carbon reservation, biodiversity conservation as well as nutrient cycling, and even supplying natural coastal defence against storm surges by minimizing wave and wind strength. Such environment monitoring depends on satellite data to assess the present condition. From 2001 to 2012, the planet lost 192,000 hectares (474,000 acres) of mangroves, a gross decline since 2000 of 1.38 per cent (or 0.13 per cent annually). It is a miniscule rate of depletion compared to the pace of tropical deforestation, which from 2000 to 2012 stands at a record of 4.9 per cent (or 0.41 per cent annually). Advances in vegetation cover satellite imagery give fresh ways to track tree transition with much greater accuracy and precision than ever before. Annual revisions to the global tree cover depletion data provide an incentive for reliable tracking of forest transition in mangrove habitats. And satellite imagery mining records provide the ability to chart recent shifts, providing insight to the mangrove depletion situation prior to 2000. Although previous studies have predicted mangrove forest destruction prior to 2000, due to variations in observational methodology we chose not to equate our estimates with those results. Deforestation has been a global issue for decades, largely because of our growing rising population and raising resource demand. Mangrove forests are seriously endangered by anthropogenic exploitation and deforestation because they depend on very limited environments for survival.

The largest mangrove region of about 198,000 km2 worldwide was in 1980. The overall region was less than 150,000 km2 in 2000, a decline of 25 per cent from 1980. This indicates that the average pace of deforestation over these twenty years ranged from 1-per cent. The key causes for mangrove forest depletion are urbanization, aquaculture and over-exploitation of wood, shrimp, and shellfish.

B. Recent Advances

The latest developments in satellite imaging and processing technologies have increased the quality of low or even no cost satellite pictures, with sufficiently high temporal and spatial resolutions. This, in addition, extended the limits of forest observation capabilities, allowing for the creation of fairly medium to high-global land cover mappings to facilitate cost-forest surveillance. Likewise, modern cloud-such as Google Earth Engine (GEE) also created the capacity to store petabytes of satellite imagery on a global scale, as well as algorithms for controlled image processing with free study access. These developments enable data collection and processing on a much larger scale and also in fairly inaccessible areas with inadequate data collection capability at higher precision. Based on these global strategies we utilize related methods in a regional environment to develop a land cover and measure forest change through Landsat image time series analysis.

II. RELATED WORKS

[1] The registered correctness of the arrangement procedure were sensible which can be clarified by the way that the all-out number of accurately ordered pixels was high .The related basic property externalities required at nearby, local and worldwide scales, require the observing of land use elements crosswise over forested scenes in creating future techniques and approaches concerning rural enhancement, common woodland protection and monoculture tree estates. The managed arrangement (Maximum probability) on two dates of IRS (LISS III) satellite information was performed to evaluate the change in land usage from 1998 to 2010.The utilization of remote detecting information has been instrumental in observing the changing example of vegetation crosswise over assorted scenes.

[2]The study surveyed deforestation in woods environment of Rudraprayag area situated in Indian Himalaya. This investigation built up a spatial model for deforestation vulnerability utilizing recurrence proportion model in GIS condition. The outcomes showed that the physical variables (incline and height) and anthropogenic components (good ways from street, closeness of backwoods to the settlement and agrarian vicinity to timberland) have emphatically quickened deforestation. The investigation region encountered lost about 112.5 km square woods region during 1990 and 2015. Change of woods into rural land has been the prime driver of deforestation.

[3]The accessibility of the recorded information of Landsat images as the result of the innovation of Google Earth Engine point towards a significant improvement in the mechanism for checking and analysing area use change over enormous geographic areas. This investigation effectively builds up a territorial scale examination and decides the classes and the circulation of land spread in Savannah River. It also distinguished the spatial and the worldly difference in the land spread that happened as a result of the revisions in land usage in the course of recent years in the Savannah River bowl. Change of land use was seen prevailing in the backwoods zones during all interims of the investigation timeframe.

[4]LULC pictures were acquired from Landsat-8 and Sentinel-2 informational indexes utilizing pixel based MLC technique, and the outcomes were assessed utilizing exactness evaluation by 400 arbitrary focuses. Overall exactness and Kappa coefficient for Landsat-8 implied LULCC and Sentinel-2 determined LULCC were 83.91 %, 0.78 and 88.74 %, 0.85 respectively because of precision evaluation. In spite of the fact that it appears that Sentinel-2 speaks to LULC superior to Landsat-8 by and large, this circumstance can change if distinctive grouping techniques and insights are utilized. Albeit generally speaking precision for Sentinel-2 inferred LULC is superior to Landsat-8 determined LULC.

[5] Divided legitimate timberland limits of state-claimed woods and community/private woodland were utilized for the appraisal of the backwoods spread in 2005 and 2011.The limits were mapped and confirmed on VHR picture. The depicted limits were utilized to survey the LC and woodland spread change from 2005 to 2011 exclusively for the state-possessed woodland and community/private timberland.

[6]The related basic property externalities required at nearby, territorial and worldwide scales, require the observing of land use elements crosswise over forested scenes in creating future systems and strategies concerning rural enhancement, regular woodland protection and monoculture tree estates. The figured correctness of the arrangement procedure were sensible which can be clarified by the way that the all-out number of effectively grouped pixels was high.

III. PROPOSED MODEL

Our proposed model will incorporate existing techniques and seek to improve them with the help of new and emerging methodologies and algorithms:-

1. Make use of Google Earth Engine code editor, an online tool used to collect, select and download mosaicked images.
2. These images are taken from Landsat 5, 7 and 8 satellites launched by NASA (National Aeronautics and Space Administration).
3. We have taken images with a cloud cover percentage of 0.5 %, so that there is no obstruction of view and we are presented with a clean image.
4. Each mosaicked image downloaded comprises on one band each namely Red and Near Infrared band.

5. Next in line is to use QGIS software to combine these band images and conduct NDVI (Normalized Difference Vegetation Index) calculations.
6. Once NDVI band is determined, we give the images a specific color palette as prescribed by GIS scientists i.e. Red, Yellow and Green. Explanation of said color palette is given in later sections.
7. In this step we use Python IDE and apply K-means unsupervised clustering algorithm to find percentage of green pixels (denoting forests).
8. The final step is to calculate the area of forest loss, and plot necessary graphs to display and analyze the results.

IV. METHODOLOGY

A. Normalized Difference Vegetation Index

It represents the relation of red visible light (Pure, usually absorbed by chlorophyll in the plant) and NIR wavelength (dispersed by the mesophyll structure of the leaf). Using the following equation, the result from the Near-Infrared(NIR) band and RED band is drawn, the corresponding NDVI image includes values from -1 to 1. The value above 0.7 includes trees, 0.5-includes plants and 0.3-comprises shrubs and the negative value contains snow, mountains and other scattered structures. The NDVI is calculated as:

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{RED}}{\mathrm{NIR} + \mathrm{RED}}$$

B. K-means Clustering

Clustering is a process by which a collection of data is grouped into a particular category. K-means clustering algorithm divides a set of data into a set of k data groups. It classifies a given data set into dislocated clusters k. K-means algorithm consists of two steps. In the first phase it calculates the centroid k. The second step deals with the algorithm taking each point from the corresponding data point to the cluster which has the closest centroid. There are various techniques for determining the distance from the nearest centroid and Euclidean distance is one of the more widely known methods. When the grouping is done, it recalculates each cluster's current centroid. Depending on that centroid, a current Euclidean distance is determined between every centre and every data point. Then the minimum Euclidean distance points in the cluster is determined. Each cluster in the partition is defined in terms of its member elements and centroid.

The sum of distances from all the points in the cluster is minimized for each cluster at a point known as **the centroid**. Therefore k-means clustering is an iterative algorithm. In this, it minimizes the number of distances between each point as well as its cluster centroid for all clusters. Let's consider a picture with *x*×*y* resolution and also the picture must be clustered into k clusters.

Let *p(x, y)*be defined as input pixels for the cluster and Ck be defined as the centre of the clusters. The algorithm for *k*-means clustering is following as:

1. Initialize the number of clusters k, with centre Ck.

2. Calculate the Euclidean distance d, between the centre and each pixel of an image.

Euclidean Distance: $d = ||p(x,y) - Ck||$

3. Every pixel is assigned to the nearest centre based on the distance d.

4. New position of the centre is recalculated.

New Centre Position:$Ck = \frac{1}{k}\sum_{y\in Ck}\sum_{x\in Ck} p(x,y)$

5. Repeat until tolerance satisfied or error value.

6. Image formed by reshaping the cluster pixels.

Since the first centroid is randomly chosen, different initial centres can get different outcomes. Therefore the main centre should be consciously chosen such that we get segmentation of our desires. Computational complexity of the algorithm depends on the number of data units, the number of determined clusters and the size of the iterations.

C. Architecture Diagram

The Flow Chart of our methodology can be seen in the figure given below.

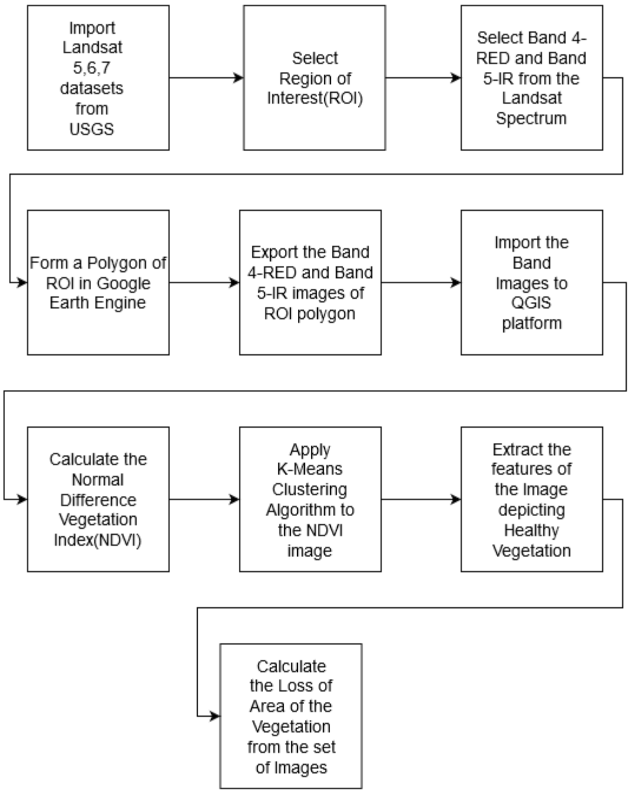
Import Landsat 5,6,7 datasets from USGS
Select Region of Interest(ROI)
Select Band 4-RED and Band 5-IR from the Landsat Spectrum
Form a Polygon of ROI in Google Earth Engine
Export the Band 4-RED and Band 5-IR images of ROI polygon
Import the Band Images to QGIS platform
Calculate the Normal Difference Vegetation Index(NDVI)
Apply K-Means Clustering Algorithm to the NDVI image
Extract the features of the Image depicting Healthy Vegetation
Calculate the Loss of Area of the Vegetation from the set of Images

Fig.1.The Process flow of the methodology

D. Dataset Properties

The Landsat system is Earth's longest operating effort to collect satellite imagery. Formerly known as Earth Resources Technology Satellite. Landsat satellite photos are stored in the U.S. and are a valuable tool for climate change studies and implementation in agriculture, geography, education, forestry, development work and aerial observation at numerous Landsat receiving stations across the world, and can be accessed via the website (USGS). The 8 spectral bands of Landsat 7 have their spatial resolution ranging from 15 to 60 meters whereas the temporal resolution of Landsat 7 is 16 days. The images taken by Landsat are easy to download as they are split into scenes.

Each Landsat scene is about 185 kilometers long and 185 kilometers wide. Landsat 7 data has eight spatial resolution spectral bands varying from 15 to 60 meters with a temporal resolution of 16 days.

Landsat images are typically fragmented for convenience in transferring images. A Landsat coverage is about 185 by 185 kilometers.

Table.1. Properties of the Dataset

Features	**Landsat-5**	**Landsat-7**	**Landsat-8**
Acquisition Date	19 February,1988	28 February, 2000	23 January, 2019
Number of Pixels	50176px		
ROI Polygon(Longitude, Latitude)	88.23ºE, 22.16ºN, 88.23ºE, 21.47ºN, 89.12ºE, 21.47ºN, 89.12ºE, 22.16ºN		
Band Combination	Near Infrared = Band 4 , Red = Band 3		Near Infrared = Band 5 , Red = Band 4
Format	TIFF	TIFF	TIFF

E. Classification Assessment

Every Landsat picture comprises of 9 spectral bands. Combining these factors result in each pixel possessing a spectral designation. RGB bands generate contrasts in colors between different forms of land cover. These can be differentiated by unaided vision and these bands form a significant part of the spectral designation of any pixel. Including only the RGB will also narrow the spectrum of wavelengths.

The Near Infrared (NIR) band with wavelength 0.7 µm–0.9 µm contains details on the green coefficient of the reflected surface that enables us classify live and stable plants, while the shortwave infrared (SWIR) bands with wavelengths of 0.9–2.5 µm contribute details on surface water quality that is helpful in separating the mangrove from the delta where vegetation is more likely to be present. Atmospheric disturbances don’t affect higher wavelengths.

These higher wavelengths are used to build indicators that detect vegetated and deteriorating land cover. Greater wavelengths are seldom influenced by ambient changes. Thus they are used to create indicators used to distinguish vegetated and barren land.

The Surface-data provide important details for atmospheric corrections. When we were picking photographs with low cloud cover percentage. It may influence the sunlight absorption and, thus, the reflectance values of the wavelengths of the bands. As part of the prediction data, the information from these bands is used to provide evidence from pseudo-patterns which may influence the other bands ' reflectance values.

The Landsat picture bands can be used to measure correlations capable of capturing different categories of land cover, which are instrumental in detecting vegetation and soil changes. Plain images give poor results in comparison to when vegetation indicators are used in tandem with machine learning algorithms.

V. RESULTS AND DISCUSSION

The major classes of palette were represented in correspondence to the Mangrove forest and the Delta. The Vegetation area from the Landsat satellite image is been analyzed by QGIS software. The palette of the NDVI images consists of five different colors representing the characteristics of the map components. The images of ROI from Near Infrared spectrum and Red spectrum are shown in image below.

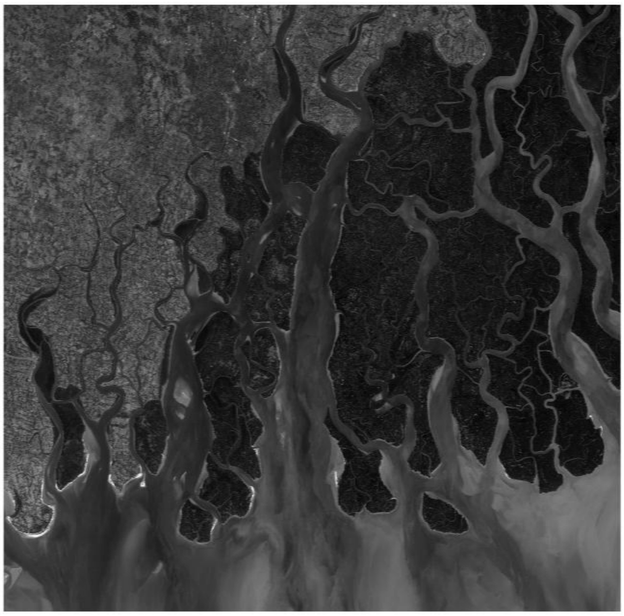

Fig.2.The Image of Sundarbans in Landsat-5 Band-3

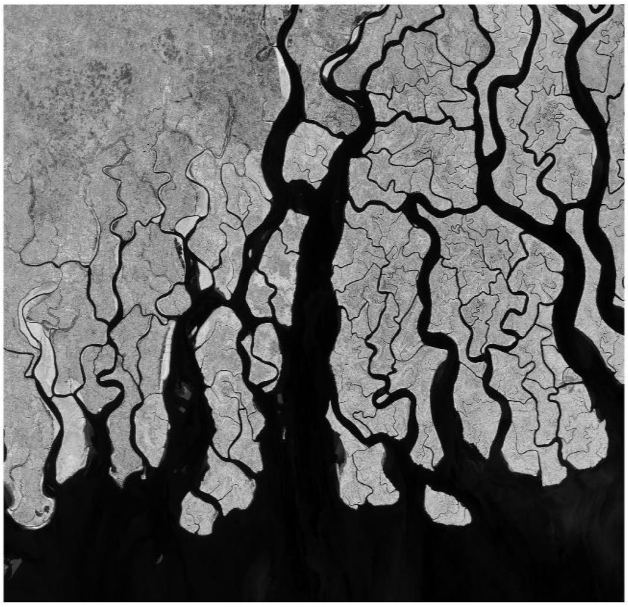

Fig.3.The Image of Sundarbans in Landsat-5 Band-3

The analysis of vegetation is done by extracting the IR and R spectrum of the Region of Interest from the Landsat 5, Landsat 7 and Landsat 8. The NDVI is been implemented upon these images to calculate Normal Difference Vegetation Index of the respective area. The NDVI highlights the reflectance of the vegetation matter into the image. The resulting image has a set of different values from -1 to +1.

The positive values reflect healthy vegetation and the negative values reflect the water bodies.The image shown below is of Sundarbans in the year of 1988. The image has a total of 50176 pixels. The total area of the region covered in the image is 7128.211 square Kilometers.

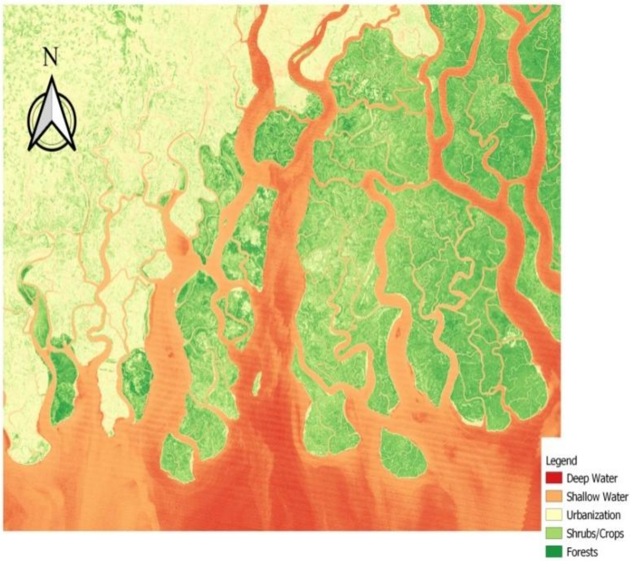

Fig.4.The NDVI of Sundarbans in 1988, Landsat-5 image

The Mangroves of Sundarbans in 1988 had a high density of healthy vegetation and areas that were undisturbed by any encroachments. The Amount of Mangroves is about 8562px out of the 50176px. The healthy vegetation is about 1216.3172 square kilometers. The Percentage of the vegetation cover is 17.0639%.The decade of 1980's had a considerably low population and demand when compared from today's standards.

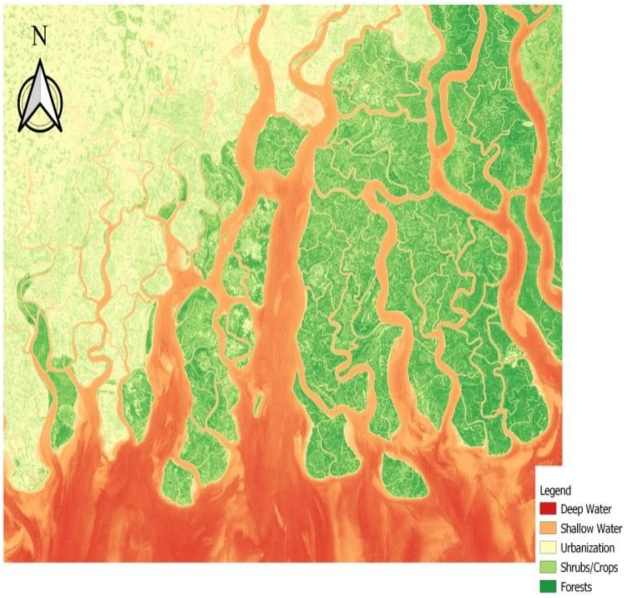

Fig.5.The NDVI of Sundarbans in 2000, Landsat-7 image

Within 12 years there was a huge increase in demand and changes in the industrial policy in the country. The amount of healthy vegetation decreased to 7555px out of the 50176px. The Mangroves shrunk to 1073.2629 square kilometers. The percentage of the vegetation cover decreased about 2% to 15.0569%. The rate of urbanization has increased so much that in only a decade about 2% of Asia's Biggest Mangroves were diminished.

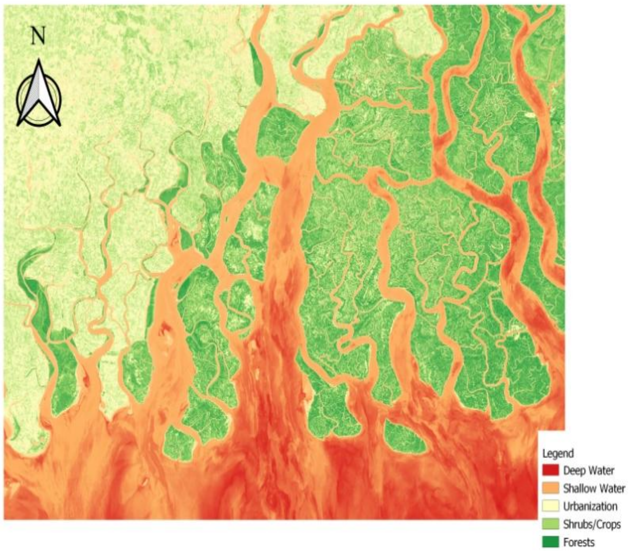

Fig.6.The NDVI of Sundarbans in 2019, Landsat-8 image

The Global trend has influenced the country policies regarding conservation of forests and wildlife. The amount of healthy vegetation increased to 7970px out of the 50176px. The mangroves have been under conserved and the area of the forests have increased to 1132.2177 square kilometers. The percentage of the vegetation cover increased about .8% to 15.840%. The upcoming decade has started to sense the consequences of the loss of these biodiversities and so the measures to facilitate the wildlife are taken and the drastic loss in only a small time gap would take a comparatively large time to recover.

VI. CONCLUSION

A conclusion In this paper, we have discussed about a proposed model which takes commercially available data gathered by Forest Survey of India and predicts the deforestation or primary forest cover loss.The extent of deforestation in the mangroves of Sundarbans has been calculated. K Means Clustering has been applied on the Landsat Satellite Imagery. As can be seen from the NDVI results and calculation, the mangroves and the freshwater swamp forests have dramatically decreased and the water cover's reach has also taken over the parts of the delta which were not flooded in the past years. To validate the future forest cover percentage we use satellite images and classification algorithm to compare the accuracy of the proposed model. Further improvements are possible, where we can improve the accuracy of the model with the help of an improved dataset.

Table.2.Calculation of Forest Cover

Year	Vegetation Area(Square Kilometer)	Forest Cover%
1988	1216.317	17.0639
2000	1073.262	15.0569
2019	1132.217	15.8840

Table.3.Calculation of deforested area after decades

Year	Area Deforested%
1988-2000	11.70%
2000-2019	-5.49 %
1988-2019	6.91%

The loss of vegetation has left the whole region prone to high tides and flash flooding. The native species of the area are very affected and are endangered due to the same. The inferences from this case study calls for the authorities to invest resources and elevate their efforts in devising an action plan to conserve the Sundarbans.

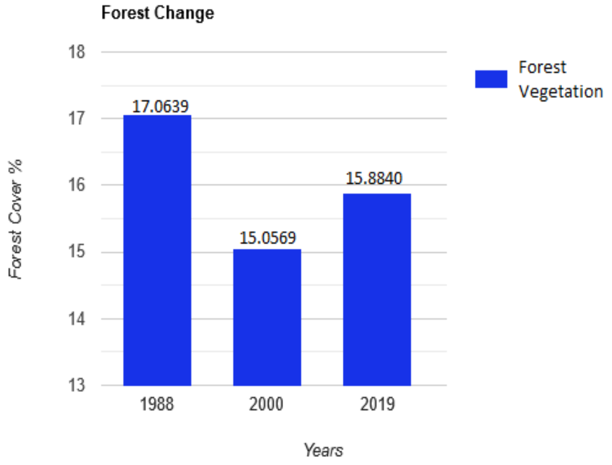

Fig.7.The Forest Cover wrt ROI over the years

REFERENCES

1. Shah, S., Sharma, D. Land use change detection in Solan Forest Division, Himachal Pradesh, India.For.Ecosyst.2, 26 (2015) doi:10.1186/s40663-015-0050-7.
2. Sahana, Mehebub & Hong, Haoyuan & Sajjad, Haroon & Liu, Junzhi & Zhu, A-Xing. (2018). Assessing deforestation susceptibility to forest ecosystem in Rudraprayag district, India using fragmentation approach and frequency ratio model. Science of The Total Environment. 627. 1264-1275. 10.1016/j.scitotenv.2018.01.290.
3. Zurqani, Hamdi & Post, Christopher & Mikhailova, Elena & Schlautman, Mark & Sharp, Julia. (2018). Geospatial analysis of land use change in the Savannah River Basin using Google Earth Engine. International Journal of Applied Earth Observation and Geoinformation. 69. 175-185. 10.1016/j.jag.2017.12.006.
4. Sekertekin, A. , Marangoz, A. M., and Akcin, H. Pixel-based classification analysis of land use land cover using sentinel-2 and landsat-8 data, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-4/W6,91-93,https://doi.org/10.5194/isprs-archives-XLII-4-W6-91-2017, 2017.
5. Lv, Zhiyong, Tongfei Liu, Yiliang Wan, JónAtliBenediktsson and Xiaokang Zhang. “Post-Processing Approach for Refining Raw Land Cover Change Detection of Very High-Resolution Remote Sensing Images.” Remote Sensing 10 (2018): 472.
6. M.D. Behera, P. Tripathi, P. Das, S.K. Srivastava, P.S. Roy, C. Joshi, P.R. Behera, J. Deka, P. Kumar, M.L. Khan, O.P. Tripathi, T. Dash, Y.V.N. Krishnamurthy, Remote sensing based deforestation analysis in Mahanadi and Brahmaputra river basin in India since 1985, Journal of Environmental Management,Volume 206,2018,Pages 1192-1203, ISSN03014797, https://doi.org/10.1016/j.jenvman.2017.10.015.
7. Guigonan Serge Adjognon, Alexis Rivera-Ballesteros, Daan van Soest,Satellite-based tree cover mapping for forest conservation in the drylands of Sub Saharan Africa (SSA): Application to Burkina Faso gazette forests,DevelopmentEngineering, Volume 4, 2019, 100039, ISSN 2352-7285, (https://doi.org/10.1016/j.deveng.2018.100039.).
8. Quy Van Khuc, BaoQuang Tran, Patrick Meyfroidt, Mark W. Paschke, Drivers of deforestation and forest degradation in Vietnam: An exploratory analysis at the national level,Forest Policy and Economics, Volume 90, 2018, Pages 128-141,ISSN 1389-9341, https://doi.org/10.1016/j.forpol. 2018.02.004.
9. Shahzad, N., Saeed, U., Gilani, H. et al. Evaluation of state and community/private forests in Punjab, Pakistan using geospatial data and related techniques. For. Ecosyst.2, 7 (2015) doi: 10.1186/s40663-015-0032-9.
10. Rodrigo Antônio de Souza, Paulo De Marco, Improved spatial model for Amazonian deforestation: An empirical assessment and spatial bias analysis,Ecological Modelling,Volume 387, 2018, Pages 1-9,ISSN 0304-3800,https://doi.org/10.1016/j.ecolmodel.2018.08.015

AUTHORS PROFILE

K. Pradeep Mohan Kumar, Ph.D., is working as Assistant Professor in the Department of Computer Science and Engineering at SRM Institute of Science and Technology, Kattankulathur, Chennai. He is a member of ISTE and CSI as well as reviewer for peer reviewed journals. He has completed his Ph.D. in Computer Science and Engineering in the year 2016. Master’s degree in the same discipline in the year 2007 and Bachelor’s degree in the same discipline in the year 2005. His recent focus has been on research efforts to create awareness upon the recent global climate change.

Rushan Mukherjee, is a senior year student in the Computer Science Department at SRM Institute of Science and Technology, Kattankulathur, Chennai, Faculty of Science, to be completed in the year 2020. His research interests are Geographical Information Systems, Machine Learning, Data Science, and Artificial Intelligence. His recent research focus is to try and develop open-source technologies particularly in the domain of GIS and Machine Learning. He is also a student member of IET (The Institution of Engineering and Technology).

Mayank Dewangan, is currently pursuing his Bachelor’s degree in Computer Science and Engineering from SRM Institute of Science and Technology, Kattankulathur, Chennai to be completed in the year 2020. His area of interest includes Remote Sensing, Geographical Information Systems and Machine Learning. He aims to study various ways open-source software can be used to further improve the existing technologies. He currently has a CGPA of 7.41. He is also a student member of IET(The Institution of Engineering and Technology).