

MAPPING VEGETATION AND MEASURING THE PERFORMANCE OF MACHINE LEARNING ALGORITHM IN LULC CLASSIFICATION IN THE LARGE AREA USING SENTINEL-2 AND LANDSAT-8 DATASETS OF DEHRADUN AS A TEST CASE

Anubhava Srivastava^a, Shruti Bharadwaj^{a,*}, Rakesh Dubey^a, Vinarma Bhushan Sharma^a, Susham Biswas^a

^a Dept. of Computer Science and Engineering, Rajiv Gandhi Institute Of Petroleum Technology, Jais, Amethi, Uttar Pradesh, India
(pgi18001, pgi17001, pgi19001, pgi19002, susham)@rgipt.ac.in

KEY WORDS: Google Earth Engine, Landsat, Sentinel, Machine learning Algorithm, Classification

ABSTRACT:

In recent years, the data science and remote sensing communities have started to align due to user-friendly programming tools, access to high-end consumer computing power, and the availability of free satellite data. In particular, publicly available data from the European Space Agency's "Sentinel" and American Earth observation satellite "Landsat" missions have been used in various remote sensing applications. Google Earth Engine (GEE) is such a tool that publicly allow the use of these available datasets, there is a large amount of available data in GEE, which are being used for computing and analysing purpose. In this article, we compare the classification performance of four supervised machine learning algorithms: Classification and Regression Tree (CART), Random forests (RF), Gradient tree boosting (GTB), Support vector machines (SVM). The study area is located at 30.3165° N, 78.0322° E near the Himalayan foothills, with four land-use land-cover (LULC) classes. The satellite imagery used for the classification were multi-temporal scenes from Sentinel-2 and LANDSAT-8 covering spring, summer, autumn, and winter conditions. Here we collected a total of 2084 sample points in which 536, 506, 505, 540 points belong to urban, water, forest and agriculture points respectively. which were divided into training (70%) and evaluation (30%) subsets. Accuracy was assessed through metrics derived from an error matrix, for accuracy measurement we use confusion and Cohen's kappa calculation method. We have calculated CART (Accuracy 93.52% and Kappa coefficient 91.36%), Random Forest (Accuracy 95.86% and Kappa coefficient 94.48%), Gradient Tree Boost (Accuracy 95.33% and Kappa coefficient 93.37%), Support Vector Machine (Accuracy 73.54% and Kappa coefficient 76.28%) for Landsat 8 data sets and CART (Accuracy 89.24% and Kappa coefficient 85.64%), Random Forest (Accuracy 91.45% and Kappa coefficient 88.59%), Gradient Tree Boost (Accuracy 87.71% and Kappa coefficient 83.58%), Support Vector Machine (Accuracy 84.96% and Kappa coefficient 79.99%) for Sentinel2 data sets. Further analysis for accuracy and machine learning algorithm are discussed in result section.

1. INTRODUCTION

The objective of this study is to test advanced image classification techniques on the cloud-based platform Google Earth Engine for mapping vegetation (Azpiroz et al., 2021) and land use types (Urban, Forest, Water, Agriculture) in different data sets (Landsat-8 and Sentinel 2) over the Dehradun region, and analyze spatial distributions of land cover classes with the help of Remote sensing mechanism. Remote Sensing provides such environment in which researcher easily perform multiple operation on images. Remote sensing is a powerful tool which help in study digital image on many platforms. There are many platforms like ArcGIS, QGIS, Erdas, Elwis are available which are being used in study the remote sensed image. But above software (platform) are required many more scaling technique that are sometime very difficult to new user, Some time user have to switch between application to application for finding remote sensed and classified image of any study area. Google Earth Engine is easily overcome this problem by providing petabyte of data and allow users to free access it. Google Earth Engine a planetary platform that provides open access of satellite imagery from various sensors. It has analytical (Tsai et al., 2018) ability also to understand large amount of datasets.

Google Earth Engine (Kumar et al., 2018) provide computation facility also, it understand two programming languages java script and Python, it has own inbuilt library function to execute instruction. We can perform classification (both supervised and unsupervised) as well clustering operation on the satellite imagery also. Google Earth Engine (Mutanga et al., 2019) provides users

*Corresponding author.

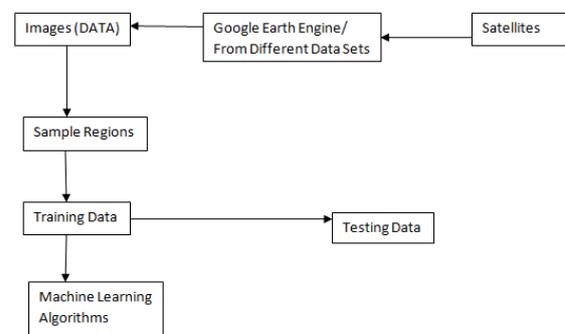


Figure 1: Working Culture Of GEE

to easy access of many datasets like MODIS (Moderate Resolution Imaging Spectroradiometer), SENTINEL, LANDSAT (Land Satellite) etc. Google Earth Engine have capability of processing these large amount of data sets by using its code editor tool. Fig 1 easily describe working culture of GEE. It is clearly mention in fig 1 that there is no need to go user physically on the study area during his processing time. training and testing both are performed on chosen sample points. The rest of the paper is organized as follows, Study Area and Used Data sets are discussed in in Section II Methodology is presented in Section III. Section IV presents the result and its comparative analysis between all the data sets and classification algorithms. Finally in Section V we present the concluding remarks.

2. STUDY AREA AND DATA

2.1 STUDY AREA

Study area is located between the Song river, a tributary of the Ganga on the east, and the Asan river, a tributary of the Yamuna on the west, in the Doon Valley of the Himalayan foothills. The city is known for its beautiful environment and milder climate, and it serves as a gateway to the surrounding area. The city is situated on Geo location 30.3165° N, 78.0322° E at a height of 640 metres (2100 feet) above sea level. Nearly 3088 km² area was examined for the research purpose, which includes land cover classes such as water, Agriculture (Low Vegetation), forest(high Vegetation) and built-up (Urbanization) vegetation.

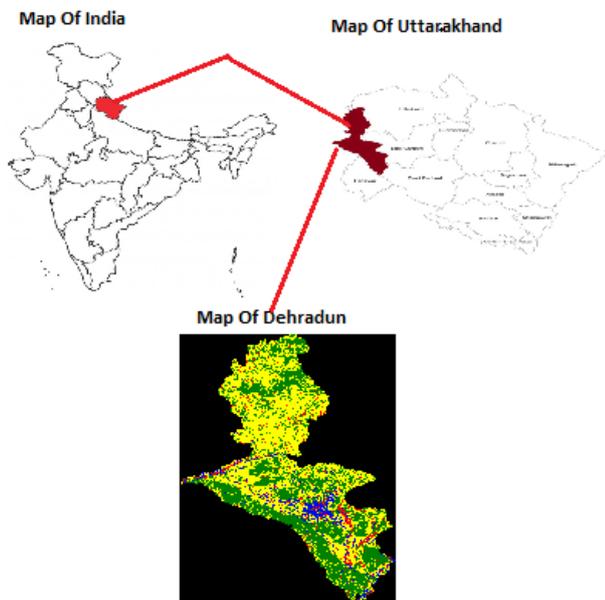


Figure 2: Study Area :- Dehradun

2.2 DATA

As discussed in previous part google earth engine have capability of storing petabytes of data from different satellite program. User can easily use these data sets for monitoring and classification purpose. Well known data sets in GEE(Imagery) dictionary are MODIS, Sentinel and Landsat. In our classification approach we are using version of sentinel and landsat (sentinel-2 and landsat-8) data sets.

2.2.1 LANDSAT : Landsat is joint program of usgs and these satellite have accurate and precise data from year 1972. The landsat data is presented in Google Earth Engine in form of surface reflectance, top of atmospheric corrected reflectance and in various used form, from there we can easily computed ndvi, evi, nbr etc. Landsat data is subcategorized in landsat 1 to landsat 8 directories where landsat1 is carried data from 23-07-1972 to 07-01-1978 and these data sets having five band images b4, b5, b6,b7 and BQA(Quality Assessment Band) where b4, b5, b6 is having pixel size 60 meter, b7 and bqa are having pixel size 30meter and 60 meter respectively, landsat2 is carried data from 12-01-1975 to 26-02-1982 and these data sets having five band images

b4, b5, b6,b7 and BQA(Quality Assessment Band) where b4, b5, b6 is having pixel size 60 meter, b7 and bqa are having pixel size 30meter and 60 meter respectively, landsat3 is carried data from 05-03-1978 to 31-03-1978 and these data sets having five band images b4, b5, b6,b7 and BQA(Quality Assessment Band) and pixel size of these data sets is same as landsat1 and landsat2, landsat4 is carried data from 16-07-1982 to 14-12-1993 and these data sets having five band images b1, b2, b3,b4 and BQA(Quality Assessment Band) , landsat5 is carried data from 01-03-1984 to 21-01-2013 and these data sets having five band images b1, b2, b3,b4 and BQA(Quality Assessment Band) where b1,b2,b3,bqa have pixel size 60 meter and b4 has pixel size 30meter.

Composite	Landsat Band	Application
Natural- Colour	3/2/1	General-Rcognisable to the naked eye
Natural- like	7/4/2	Rcognisable but clearer than natural colour
Colour Infrared	4/3/1	Agriculture Areas (Bright Red)
False Colour	5/4/1	Agriculture Areas (Bright Green)
False Colour	7/5/3	Urbanization
False Colour	4/5/3	coastal and wetland areas

Table 1: Band Combination For Landsat Datasets

Landsat7 and landsat8 data sets are widely updated data sets and these data sets have data availability very up-to-date, landsat 7having data from January 1993 to present date and these datasets area carrying basically three types of data, surface reflectance, Top of atmospheric and raw images these images having band value b1, b2, b3,b4,b5, and b7 and data having 30 meter resolution in these bands, any user can easily compute burn area index(from red and nearer bands),Enhanced vegetation Index(From Near IR,red and blue bands), normalize difference vegetation index(from near IR and red bands), Normalize burn ratio thermal(from Near IR, Mid IR and thermal bands).Normalize difference snow index(From mid IR and green bands),Normalize difference water index(from near IR and second IR bands) from these bands, landsat8 is last datasets available in landsat community landsat8 data available from 11-04-2013 to present date and data having b1,b2,b3,b4,b5, b6, b7 and sr_qa_aerosol for aerosol attribute and these bands having ultra blue, blue, green, red, near infrared shortwave, infrared1 and infrared 2 colors respectively.

Composite	Sentinel Band	Application
Natural- Colour	4/3/2	General-Rcognisable to the naked eye
Natural- like	12/8/3	Rcognisable but clearer than natural colour
Colour Infrared	8/4/3	Agriculture Areas (Bright Red)
False Colour	11/8/2	Agriculture Areas (Bright Green)
False Colour	12/11/4	Urbanization
False Colour	8/11/4	coastal and wetland areas

Table 2: Band Combination For Sentinel Datasets

2.2.2 SENTINEL : Sentinel is other dataset collection that is offered by Copernicus program under the head of European Space Agency. Sentinel data sets is sub categorized in sentinel 1A, sentinel 1B, sentinel 2A, sentinel 2B, sentinel 3 and sentinel 5P in all these sentinel data sets sentinel 1A and sentinel 1B is collection of weather radar images, sentinel 2A and sentinel 2B is collection of high resolution optical images, sentinel

3 is collection of images that is uses for climate and environmental monitoring and sentinel 5P is collection of images used for air quality indexing. Sentinel -1 data is presented in ortho-corrected form, these data sets are updated daily and ingested in GEE within two days of its availability, data is ingested in gee after three level of preprocessing, thermal noise removal, radiometric calibration and terrain correction and finally terrain corrected values are converted into decibel via logarithmic function. Sentinel-1 data is presented in 4 Bands VV, HH, VH and HV. In these four bands VV and HH are single Polarization band, other two VH and HV are dual band(cross polarization) each bands having pixel size of 10 meter. Availability of these data sets is from 03-10-2014 to till present date. Sentinel 2 is another data sets presented in GEE provided by Copernicus land monitoring system, S2 having frequency of five days and provide image of wide swath and high resolution, these datasets are computed by sen2cor(A processor for sentinel-2 level-2A) and available by 28-03-2013 to present day. S2 datasets containing B1, B2,B3, B4, B5, B6, B7, B8, B8A, B9, B11, B12, AOI, AOT, WVP, SCL, TCI-R, TCI-B, MSK-CLDPRB, MSK-SNWPRB, QA10, QA20,QA60 data band. Sentinel-3 is another type of datasets generally used for ocean and land water instrument, availability of these data sets by 18-10-2016 to present date. Resolution of data is up to 300m, and data is updated in 2 day frequency. S3 data is containing 0a01- radiance to 0a21- radiance data band. Sentinel 5P provide data at 0.01 arc degree and subcategorized in eight part on the basis of their use, these eight part are Sentinel-5P(UV Aerosol, Cloud, Carbon Monoxide, Formaldehyde, Nitrogen dioxide, Ozone, Sulphur Dioxide and Methane) these different data sets are used for estimation and evaluation quality in atmosphere.

3. METHODS

3.1 Land-Cover Classification Algorithms

We have divided our study in two part, first we apply classification algorithms on developed data (Landsat and sentinel) and other is calculating and comparing accuracy of each algorithm. Machine learning, in its most basic form, employs pre-programmed algorithms that learn and improve their operations by analysing incoming data and making predictions within a reasonable range. These algorithms tend to make increasingly accurate predictions as additional data is fed into them. Although there are a few different ways to do it, they can be categorised into groups of machine learning algorithms, according to their intended use, into three primary categories as well as the manner in which the underlying computer is being taught. These are the three categories: supervised, unsupervised, and semi-supervised. A labelled dataset is used in supervised machine learning techniques. For our classification purpose we are using supervised machine learning algorithm here. The underlying algorithm is initially trained using the training dataset. The unlabeled test dataset is then fed into the trained algorithm, which categorises them into similar groups. CART (Breiman et al., 1984) stands for Classification and Regression trees. In 1984, CART was introduced by Breiman and it builds trees for both classification and regression and it is based on Hunt's algorithm. Based on binary splitting of the attributes CART generates the classification tree. CART is not similar to other Hunt's based decision tree algorithm because it is also used for analysis of regression by using regression trees. CART adopts a non backtracking approach that is greedy approach in which decision tree are generated in top-down manner. CART algorithm is very illustratable model that is a person who was not having any statistical analysis also can easily interpret the CART model. The Input for CART model is: (1) Data partition, P, which is

collection of training instance and their related class labels; (2) Attribute list, the collection of attributes (3) Attribute Selection Method, a way to decide the splitting scale that "best" dividing up the data set into individual classes. Algorithm uses Information gain and Gini index for creating decision tree. Information gain is based on information theory, used for computing best attributes that has maximum information about class. Random Forest (Ho, T. K., 1995) is an another example of machine learning algorithm uses ensemble approach for maintaining decision tree, Each decision tree predict a class result and which class result have most vote is uses as root for the tree, this technique is implemented at every level and on the basis of result we classify our data. Random Forest first discussed in 1995 by Tim Kam Ho using random subspace method. Gradient Tree Boost (Friedman, J.H., 2001) is another classifier that uses decision tree for classification in gradient tree boost each predictor corrects its predecessor's error; its base learner is CART. GTB is basically used for eliminating bias error. There are the several classifier and they have different properties. Support vector machine (Cortes C. Vapnik, V, 1995) is one of the most effective classifiers among all those, which are linear. By using Support vector machine we are able to handle certain cases where there is non-linearity by using non linear basis function or these are called kernel function. Support vector machine is so popular because it has a clever way to prevent over fitting and we can work with relatively large number of feature without requiring too much computation. The main aim of Support vector machine is to make the decision boundary and finest line that can divide t-spatial space into group so that we can simply tag the new input instance in the correct category in the future. Support vector machine select the nearest points that help in creating the decision surface. These nearest points are support vectors. In this algorithm, we want a classifier that maximize the partition between the points and decision surface.

3.2 Accuracy Assessments

The classification achieved using the techniques outlined above does not always produce ideal results. As a result, the classified image contains numerous errors as: cluster labeling after unsupervised classification, incorrect labeling of training areas, indistinguishable classes and band correlation, an inaccurate classification technique, and so on. The accuracy of a map produced using remotely sensed data is assessed by comparing it to another map obtained from some other source. The landscape is constantly changing. As a result, the ground reference should be collected as near to the date of remote sensing data collecting as possible. The development of a classification error matrix is one of the most frequent ways of representing classification accuracy (confusion matrix or contingency table). The first step in creating an error matrix is to find ground reference test pixels or a sample collection from which an error matrix can be created. In this sense, there are numerous mathematical approaches. A minimum of 50 samples of each land use land cover class should be included, according to most experts. If the study region is big or there are more than 12 land use land cover classifications, the sample size should be 75 to 100 points. The following processes can be used to sample data: random, systematic, stratified random, stratified systematic unaligned, and cluster sampling. Here we are getting data from google earth engine, that is very accurate, precise and up to date so we do need to go and collect data physically. For our classification purpose we have divided land cover classes into four we collected total of 2084 sample points in which 536, 506, 505, 540 points belong to urban, water, forest and agriculture points respectively. Which were divided into training (70%) and evaluation (30%) sub sets.

Using descriptive and inferential statistics, summative assessments of prediction performance can be computed from the error matrix

ID	Land Cover Class	Color	Number Of Samples
1	Urban	0000ff	536
2	Water	008000	506
3	Forest	ff0000	505
4	Agriculture	ffff00	540

Table 3: Training Points & Color legend of landcover classes

(Kohavi, R. and Provost, F.,1998). By using confusion matrix we calculated omission error, commission error, overall accuracy, user's accuracy, producer's accuracy and kappa coefficient (Cohen J, 1960). Omission error is pixels that pertain to the real class but aren't classified into it, Commission error represents pixels that are classified to a class but pertain to another class, Overall accuracy indicates the overall accuracy of the classification. The total number of successfully categorized pixels divided by the total number of reference pixels yields this value. The drawback of this measure is that it will not tell us all about how well individual classes are classified. Producer and user accuracy are two frequently used measurements of class accuracy that are based on omission and commission error. Producer's accuracy refers to the likelihood of a particular characteristic of the ground being classed as such. It's calculated by multiplying the number of pixels correctly classified in each category by the number of pixels sampled for that category, user's accuracy refers to the likelihood that a pixel designated as a specific class in the map actually belongs to that class. It's calculated by dividing the number of pixels correctly classified in this category by the number of pixels categorized in this category. Cohen's kappa coefficient is a discrete multivariate accuracy assessment method. Pixels are randomly assigned to classes in the categorization process, resulting in a percentage accurate value, obviously, pixels are not assigned at random during picture classification, but there are statistical techniques that attempt to account for random chance's role when evaluating classification accuracy. The resulting Kappa measure accounts for chance agreement in classification and indicates how much better the classification performed as compared to the likelihood of randomly assigning pixels to their right groups.

Cohen kappa is calculated as :

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

Where P_o = Observed proportional agreement,

P_e = Expected proportion of agreement.

P_o and P_e is calculated by :

$$P_o = \frac{1}{N} \sum_{j=1}^k f_{jj}, \quad (2)$$

$$P_e = \frac{1}{N^2} \sum_{i=1}^k r_i c_i, \quad (3)$$

$$r_i = \sum_{j=1}^k f_{ij}, \forall i, \quad (4)$$

$$c_j = \sum_{i=1}^k f_{ij}, \forall j, \quad (5)$$

Where f_{ij} defines the number of cases that the first observer assigned a particular case to category i and the second to j . r_i and c_j the row and column totals for category i and j

4. RESULTS AND DISCUSSION

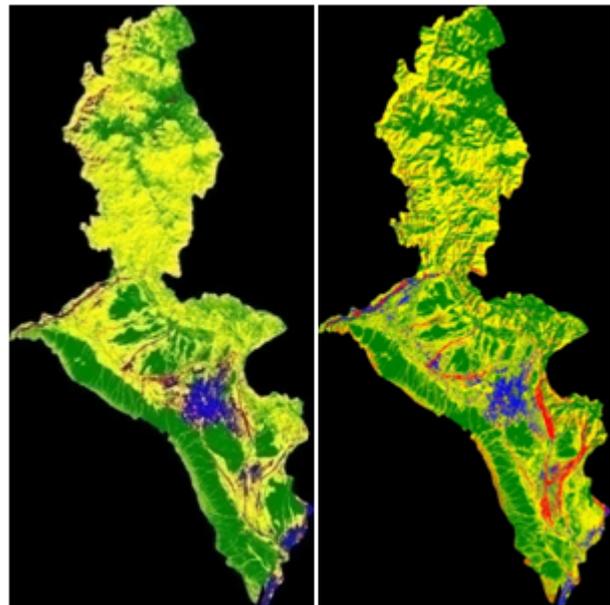


Figure 3: Collection of output Data generated by Sentinel-2 and LANDSAT-8 at resolution 30m by CART

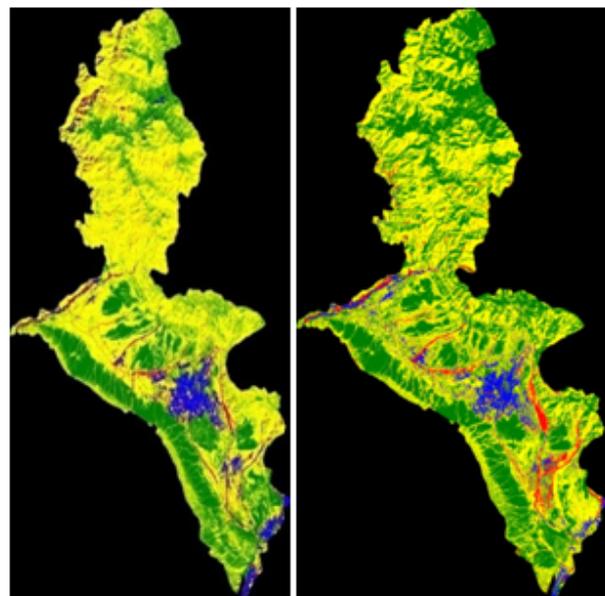


Figure 4: Collection of output Data generated by Sentinel-2 and LANDSAT-8 at resolution 30m by Random Forest

As discussed in earlier section we have performed classification approach on landsat and sentinel data sets in this classification approach we divided total study area in to four land use land cover class, Area covered by blue(0000ff) is denoted as urban, area covered by green (ff0000) is denoted as forest, area covered by yellow (ffff00) is denoted as agriculture and area covered by red (008000) is denoted as water area. Figure 3 shows the classi-

■ URBAN ■ FOREST ■ WATER ■ AGRICULTURE

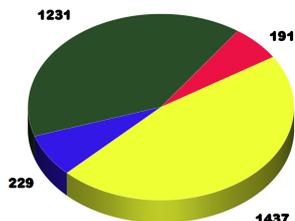


Figure 5: Area of each landcover classes by CART

■ URBAN ■ FOREST ■ WATER ■ AGRICULTURE

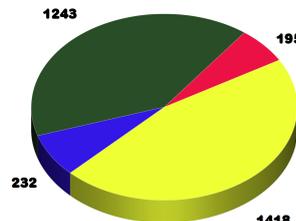


Figure 8: Area of each landcover classes by GTB

■ URBAN ■ FOREST ■ WATER ■ AGRICULTURE

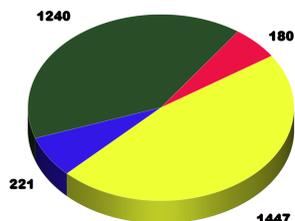


Figure 6: Area of each landcover classes by RF

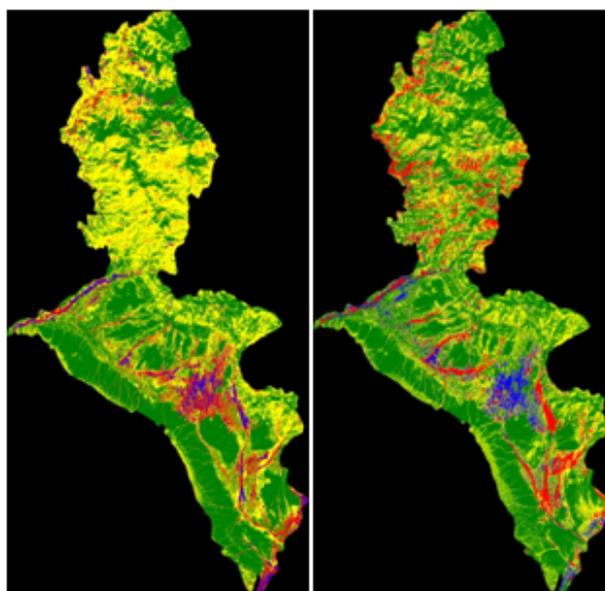


Figure 9: Collection of output Data generated by Sentinel-2 and LANDSAT-8 at resolution 30m by Support Vector Machine

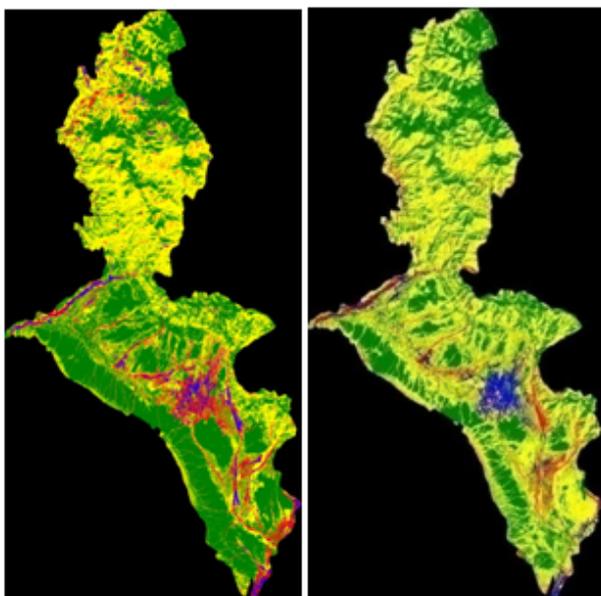


Figure 7: Collection of output Data generated by Sentinel-2 and LANDSAT-8 at resolution 30m by Gradient Tree Boost

■ URBAN ■ FOREST ■ WATER ■ AGRICULTURE

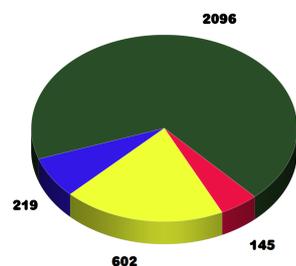


Figure 10: Accuracy generated by classification algorithms over landsat-8

ALGO	CLASS	L8(LANDSAT 8 DATA SET)			
		TOTAL ACCURACY	CONSUMER ACCURACY	PRODUCER ACCURACY	KAPPA ACCURACY
CART	URBAN	0.9352	0.9215	0.9333	0.9136
	FOREST		0.9281	0.9403	
	WATER		0.9415	0.9354	
	AGRICULTURE		0.9514	0.9315	
RANDOM FOREST	URBAN	0.9586	0.9532	0.9659	0.9448
	FOREST		0.9523	0.9756	
	WATER		0.9738	0.9494	
	AGRICULTURE		0.9559	0.9441	
GRADIENT TREE BOOST	URBAN	0.9533	0.9613	0.9942	0.9377
	FOREST		0.9411	0.9171	
	WATER		0.9686	0.9807	
	AGRICULTURE		0.9403	0.9161	
SUPPORT VECTOR MACHINE	URBAN	0.7354	0.8125	0.8924	0.7628
	FOREST		0.8041	0.7777	
	WATER		0.7712	0.7065	
	AGRICULTURE		0.7125	0.7625	

Figure 11: Accuracy generated by classification algorithms over landsat-8 datasets

ALGO	CLASS	S2(SENTINAL DATA SET)			
		TOTAL ACCURACY	CONSUMER ACCURACY	PRODUCER ACCURACY	KAPPA ACCURACY
CART	URBAN	0.8924	0.8954	0.8838	0.8564
	FOREST		0.8947	0.9216	
	WATER		0.8758	0.8639	
	AGRICULTURE		0.9018	0.8963	
RANDOM FOREST	URBAN	0.9145	0.9079	0.9548	0.8859
	FOREST		0.9058	0.9277	
	WATER		0.9469	0.8503	
	AGRICULTURE		0.9041	0.9207	
GRADIENT TREE BOOST	URBAN	0.8771	0.8281	0.9244	0.8358
	FOREST		0.8881	0.9407	
	WATER		0.8911	0.8136	
	AGRICULTURE		0.9172	0.8367	
SUPPORT VECTOR MACHINE	URBAN	0.8496	0.8284	0.9032	0.7999
	FOREST		0.8245	0.8493	
	WATER		0.9236	0.8231	
	AGRICULTURE		0.8382	0.8231	

Figure 12: Accuracy generated by classification algorithms over sentinel-2 datasets

Figure 3 shows the classification output generated by CART classification algorithm over landsat and sentinel data sets, first part of figure 3 shows output generated by landsat-8 data set and another part is output generated by sentinel data sets, classification output shows that area covered by land cover classes urban, agriculture, water and forest is 229 sqkm, 1231 sqkm, 191 sqkm, and 1437 sqkm respectively. Figure 4 shows the classification output generated by Random Forest classification algorithm over landsat and sentinel data sets, first part of figure 4 shows output generated by landsat-8 data set and another part is output generated by sentinel data sets, classification output shows that area covered by land cover classes urban, agriculture, water and forest is 221 sqkm, 1240 sqkm, 180 sqkm, and 1447 sqkm respectively.

Figure 7 shows the classification output generated by Gradient Tree Boost classification algorithm over landsat and sentinel data sets, first part of figure 7 shows output generated by landsat-8 data set and another part is output generated by sentinel data sets, clas-

sification output shows that area covered by land cover classes urban, agriculture, water and forest is 232 sqkm, 1243 sqkm, 195 sqkm, and 1418 sqkm respectively.

Figure 9 shows the classification output generated by support vector machine classification algorithm over landsat and sentinel data sets, first part of figure 9 shows output generated by landsat-8 data set and another part is output generated by sentinel data sets, classification output shows that area covered by land cover classes urban, agriculture, water and forest is 219 sqkm, 2096 sqkm, 145 sqkm, and 602 sqkm respectively. There is lot of change is showing in agriculture and forest class compare to another three decision tree based classification algorithm.

Here if we focus on figure 3, figure 4, figure 7 and figure 9 some places have different colour than our define land cover classes, these are the those common geopoints where one or more than one existing landcover classes are collectively present, if we go in more depth like geopoint wise analysis then we can find at a single geolocation only a unique land cover class is present. We have also discuss some common geopoints where points belongs to different landcover class in other classification algorithm, so area covered by each land cover class is changes with classification algorithm, but in fact a common geo point should be fixed in all classification algorithm, so for evaluating our research outcome we check accuracy of above classification algorithm figure 11 and figure 12 shows the accuracy of these algorithm for data generated by landsat-8 and sentinel-2 data set community respectively, here we measure each class accuracy, total accuracy as well as Cohen's kappa accuracy also. now on focusing at figure 11 and figure 12 we find that classification algorithms are producing different accuracy with different data sets. CART is producing .9325 total accuracy and .9136 kappa accuracy with landsat-8 data sets but .8924 and .8524 taotal and kappa accuracy with sentinel-2 data sets. it clearly shows that performance of classification algorithm are changes with the availability of data sets also. If we focus in more depth then we found that each landcover classes accuracy are changes in data sets

5. CONCLUSION

Landuse land cover classification by machine learning algorithm provide percentage of each landcover class in study area. Here algorithm based on decision tree like CART, RF and GTB are performing well in compare to non- decision tree classification algorithm support vector machine. There is lots of space to other researcher in this area for calculating change detection and monitoring other changes in any land cover class.

REFERENCES

- Azpiroz, Izar Osés Fernández, Noelia Quartulli, Marco Olaizola, Igor Guidotti, Diego Marchi, Susanna. (2021). Comparison of Climate Reanalysis and Remote-Sensing Data for Predicting Olive Phenology through Machine-Learning Methods. *Remote Sensing*. 13. 1224. 10.3390/rs13061224.
- Bharadwaj, Shruti Dubey, Rakesh Biswas, Susham. (2022). A Novel Method to Determine the Optimal Location for a Cellular Tower by Using LiDAR Data. *Applied System Innovation*. 5. 25. 10.3390/asi5020030.
- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J.. *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.

Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*. 1960;20(1): 37-46. doi:10.1177/001316446002000104

Cortes, C. Vapnik, V., 1995. Support-vector networks. *Machine learning*, 20(3), pp.273–297.

Dubey, Rakesh Bharadwaj, Shruti Zafar, Itaf Biswas, Susham. (2021). GIS Mapping of Short-Term Noisy Event of Diwali Night in Lucknow City.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189–1232.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282).

Jogarah, Keshav Soopaul, Keshav Beeharry, Yogesh Hurbungs, Visham. (2018). Hybrid machine learning algorithms for fault detection in android smartphones. *Transactions on Emerging Telecommunications Technologies*. 29. e3272. 10.1002/ett.3272.

Kohavi, R. and Provost, F. (1998) Glossary of terms. *Machine Learning—Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*. *Machine Learning*, 30, 271-274. <https://doi.org/10.1023/A:1017181826899>

Kumar, Lalit Mutanga, Onesimo. (2018). Google Earth Engine Applications Since Inception: Usage, Trends, and Potential. *Remote Sensing*. 10. 1509. 10.3390/rs10101509.

Maazouzi, Faiz Bahi, Halima. (2012). Using multi decision tree technique to improving decision tree classifier. *International Journal of Business Intelligence and Data Mining - IJBIDM*. 7. 10.1504/IJBIDM.2012.051712.

Mutanga, Onesimo Kumar, Lalit. (2019). Google Earth Engine Applications. *Remote Sensing*. 11. 591. 10.3390/rs11050591.

Srivastava, Anubhava Ahmad, Parvez. (2016). A Probabilistic Gossip-based Secure Protocol for Unstructured P2P Networks. *Procedia Computer Science*. 78. 595-602. 10.1016/j.procs.2016.02.122.

Tsai, Yu Hsin Stow, Douglas Chen, Hsiang Ling Lewison, Rebecca An, Li Shi, Lei. (2018). Mapping Vegetation and Land Use Types in Fanjingshan National Nature Reserve Using Google Earth Engine. *Remote Sensing*. 10. 927. 10.3390/rs10060927.