EXPLORING MACHINE LEARNING CLASSIFICATION ALGORITHMS FOR CROP CLASSIFICATION USING SENTINEL 2 DATA

Neetu* and S. S. Ray

Mahalanobis National Crop Forecast Centre, DAC&FW, Pusa Campus New Delhi-110012, India
(neetu.ncfc, shibendu.ncfc)@nic.in

Commission III, WG III/10

KEY WORDS: Crop Classification, Random Forest, SVM, CART, Sentinel data, Google Earth Engine

ABSTRACT:

Crop Classification and recognition is a very important application of Remote Sensing. In the last few years, Machine learning classification techniques have been emerging for crop classification. Google Earth Engine (GEE) is a platform to explore the multiple satellite data with different advanced classification techniques without even downloading the satellite data. The main objective of this study is to explore the ability of different machine learning classification techniques like, Random Forest (RF), Classification And Regression Trees (CART) and Support Vector Machine (SVM) for crop classification. High Resolution optical data, Sentinel-2, MSI (10 m) was used for crop classification in the Indian Agricultural Research Institute (IARI) farm for the Rabi season 2016 for major crops. Around 100 crop fields (~400 Hectare) in IARI were analysed. Smart phone-based ground truth data were collected. The best cloud free image of Sentinel 2 MSI data (5 Feb 2016) was used for classification using automatic filtering by percentage cloud cover property using the GEE. Polygons as feature space was used as training data sets based on the ground truth data for crop classification using machine learning techniques. Post classification, accuracy assessment analysis was done through the generation of the confusion matrix (producer and user accuracy), kappa coefficient and F value. In this study it was found that using GEE through cloud platform, satellite data accessing, filtering and pre-processing of satellite data could be done very efficiently. In terms of overall classification accuracy and kappa coefficient, Random Forest (93.3%, 0.9178) and CART (73.4%, 0.6755) classifiers performed better than SVM (74.3%, 0.6867) classifier. For validation, Field Operation Service Unit (FOSU) division of IARI, data was used and encouraging results were obtained.

1. INTRODUCTION

Crop type classification is very important for crop production estimation and there is a huge demand for accurate and timely information about the crop types [1, 2 and 14]. Machine learning techniques were used for crop identification, early in year 2011 and the authors mentioned that not much comparisons have been made between the main machine learning algorithms RF (Random Forest), ANN (Artificial Neural Network) and SVM (Support Vector Machine) [4, 7 and 13]. Non-parametric machine learning algorithm performs better than parametric classifiers such as nearest neighbour or maximum likelihood (ML) [12 and 13]. In a study carried out by Dixon & Candade (2008) using Landsat TM data, similar results were obtained using ANN and SVM while ML (Maximum Likelihood) did not perform well [6]. Many Studies were conducted using Decision Tree (DT) algorithms, classification trees (CT), ANN, SVM and RF and in some studies, it was found that statistically similar accuracies of over 91% were obtained for ANN, SVM and RF [12, 15 16 and 17]. Some studies, also, have shown that SVM achieves a higher level of classification accuracy than either the ML or the ANN classifier, and that the SVM can be used with small training datasets and high-dimensional data [7 and 9]. In the comparison of ANN and ML, Pal & Mather reported non-significant differences in classification accuracy between the two, but for ANN, manual work and computational time effort comes to be much more intense [10,11].

Recently, studies have shown integration of Pixel-Based and Object-Based Algorithms using Sentinel-2 and Landsat-8 Data and automated crop land mapping on Google Earth Engine [8, 15 and 16]. GEE is a platform which can be used for cloud computing and for automated crop classification techniques and crop mapping [18].

The Objective of this study is to explore the ability of different machine learning classification techniques like, Random Forest (RF), Classification And Regression Trees (CART) and Support Vector Machine (SVM) along with the Maximum likelihood classification (MXL) for crop classification using Google earth Engine and ERDAS imagine for IARI farm land using high resolution Sentinel-2 MSI (10m resolution) data and ground truth collected using smartphone based android application.

2. MATERIALS AND METHODS

2.1 Study Area and Satellite Data

The study was carried out in Indian Agricultural Research Institute (IARI), farm (Latitude 28.08^{0}N and Longitude 77.12^{0}E) located in Pusa Campus New Delhi for the study of major crops in the farm land during Rabi season 2016 (from December to March end) (Figure 1).

The IARI farm land of the Institute is spread over an area of about 500 hectares (approx. 1250 acres). The mean maximum temperature during winter (November-March) ranges from 20.1 ^{o}C to 29.1 ^{o}C and the mean minimum temperature from 5.6^{o}C to 12.7^{o}C. During winter, generally small amount of rainfall

*Corresponding author. This study was carried out as Ph.D. work of the corresponding author from Nirma University

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-W6-573-2019

(about 63 mm) is received. In Rabi season, major crops grown are wheat, rapeseed and mustard, vegetables and horticultural crops such as onion, carrot and potato etc. Satellite data, Sentinel -2 MSI was used for the study. Best Cloud free date (5 Feb 2016) was selected using the automatic filter during 1 January 2016 to 31 March 2016 in Google Earth Engine during the peak vegetative growth of wheat and rapeseed and mustard crop. The bands used for the study are B8-NIR, B4-Green, B3-Red.

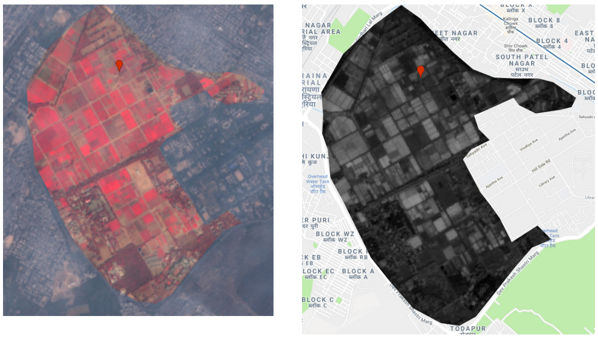
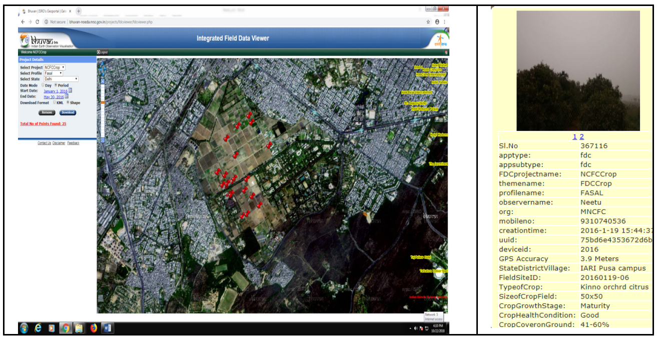


Figure 1.Study area (False Color Composite (left) and NDVI (right) in Google image of IARI farm, Pusa New Delhi

2.2 Ground Truth Data

Field information or Ground Truth Data were collected using smart-phone based android application developed by NRSC (ISRO) under the project FASAL (Forecasting Agricultural outputs using Space Agro-meteorology and Land based observations) of Department of Agriculture, Cooperation and Farmer's Welfare. The ground truth was collected during the month of January 2016 in IARI farm. Around 25 Points with major crop information like crop field area, crop sowing and harvesting information, soil condition, crop growth stage with other crops in the field with latitude and longitude information were collected. Major crops were found as wheat, rapeseed & mustard and vegetables during the season.



Sl.No	367116
apptype:	fdc
appsubtype:	fdc
FDCprojectname:	NCFCCrop
themename:	FDCCrop
profilename:	FASAL
observername:	Neetu
org:	MNCFC
mobileno:	9310740536
creationtime:	2016-1-19 15:44:3
uuid:	75bd6e4353672d6b
deviceid:	2016
GPS Accuracy	3.9 Meters
StateDistrictVillage:	IARI Pusa campus
FieldSiteID:	20160119-06
TypeofCrop:	Kinno orchrd citrus
SizeofCropField:	50x50
CropGrowthStage:	Maturity
CropHealthCondition:	Good
CropCoveronGround:	41-60%

Figure 2. Ground Truth Data collected over the study area in January 2016

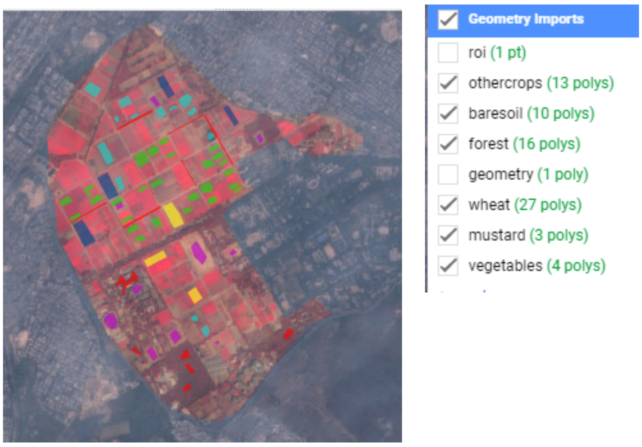
2.3 Google Earth Engine

In this study, the Google Earth Engine (GEE) was used which is an advanced cloud-based platform for geospatial and remote sensed data analysis. In GEE, petabytes–scale archives of publicly available remote sensing imagery (Sentinel-1/2, Landsat-8/MODIS) and other data (different composite products) are available. GEE has a computational infrastructure of Google for parallel processing of geospatial data and also APIs for Javascript and Python for visualization of analysis and online Integrated Development Environment (IDE) for rapid prototyping and visualization of complex spatial analyses using the Javascript API (code editor). (https://developers.google.com/earth-engine/) are available. Many advanced machine learning algorithms for classification such as, Random forest, Naïve Bayes, Classification And Regression Trees (CART) and Support Vector Machine (SVM) are available in GEE platform [18].

2.4 Methodology

The methodology of this research comprised the following steps:

- Access the Sentinel -2 MSI data, on Google Earth Engine Platform using filtering the region of Interest i.e. Study area
- Filtering the image by selecting the best cloud free date available from December 2015 to February 2016 and clipping the image with the study area as desired (as geometry). Mostly IARI farm area was selected without any urban land except those inside the boundary.
- True Color image, False Color image and NDVI computed image were used for identifying the crop features.
- Feature Collections were generated by selecting the training polygons based upon the ground truth collected using smartphone based android application.
- All total 5 Classes were selected, such as Wheat, Rapeseed and Mustard, Vegetables, Baresoil and Other crops.
- Training Data sets were created based upon the feature Collection generated as shown in Figure3.
- After training datasets generation different machine learning algorithms (CART (Classification And Regression Trees), RF (Random Forest), SVM (Support Vector Machine) were used for classifying the image using GEE.



☑ Geometry Imports
□ roi (1 pt)
☑ othercrops (13 polys)
☑ baresoil (10 polys)
☑ forest (16 polys)
□ geometry (1 poly)
☑ wheat (27 polys)
☑ mustard (3 polys)
☑ vegetables (4 polys)

Figure 3. Feature space overlaid on FCC of January, 2016

Apart from this, Sentinel -2 data (5 Feb 2016) was classified, based upon the training signatures generated using ground truth data collected and major crop classes, such as wheat, mustard and vegetables, using MXL classification on ERDAS imagine.

3. RESULTS AND DISCUSSION

3.1 Classification

Classified images are shown using different classifiers as mentioned above using GEE in the code editor.

3.1.1 Classification And Regression Trees (CART): CART is an advanced technique based on Decision Tree (DT) classifier, which is built from a set of training data. The advantage of CART is that it is simple in understanding, visualizing and interpreting. Both numerical and categorical data can handle in CART. The disadvantage is that it can create over complex tree which do not generalize the data, can be called as over fitting. It is also considerable sensitivity to the training datasets, so that a small change to the training data can result in very different set of subsets [3,4 and 18]. Based upon the feature collection and training sets CART classification was done using leaves: 202, maxDepth: 10 training points: 3028 in GEE. Decision tree was created using training points in data. For CART classification, code was written in the code editor of GEE [18]. Classified image was generated for five classes, such as wheat, rapeseed and mustard, vegetables, Bare soil and other crops as shown in figure 4.

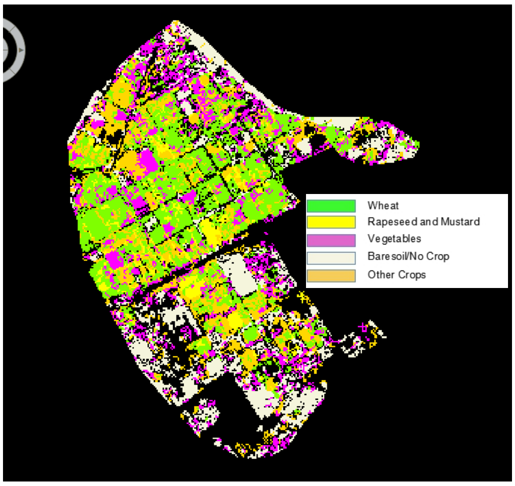


Figure 4. Classified image obtained by classifying Sentinel-2 imagery using CART in GEE

3.1.2 Random Forest: Over fitting problem in Decision Tree classifier as mentioned above is overcome by Random Forests as it constructs an ensemble of Decision Trees. The number of decision trees is set as 10 to create per class and the number of variables per split is square root of the number of the variables [3, 4 and 18]. The training data sets have been used for generating the decision tree. Code was written for random forest classification in the code editor of GEE [18] using training sets generated as mentioned above. Figure 5 shows the classification output of Random Forest classifier.

3.1.3 Support Vector Machine (SVM): SVM is very popular technique for solving problems in classification and regression. In support vector machines the classification problem solves through the concept of margin, which is defined as the smallest distance between the decision boundary and any of the samples. The decision boundary is chosen to be the one for which the margin is maximized. In this case margin is the perpendicular distance between the decision boundary and closest of the data points. Support vectors are the data points, which determines the location of this boundary [3, 4 and 18]. Kernel type used is 'RBF (radial bias function), gamma: 0.0002 and cost: 1. When margin increases, it leads to the particular choices of the decision boundary. For SVM classification code was written in the code editor of GEE [18]. Figure 6 shows the classification output of SVM classifier.

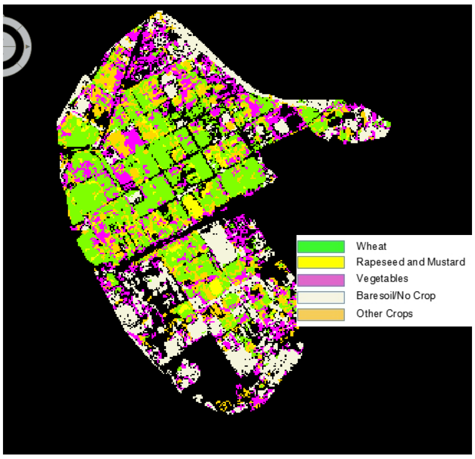


Figure 5. Classified image obtained by classifying Sentinel-2 imagery using Random Forest in GEE.

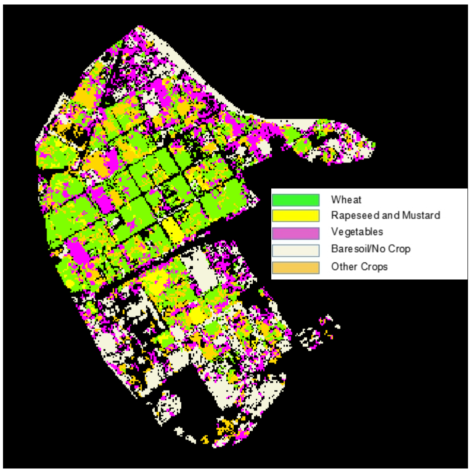


Figure 6. Classified image obtained by classifying Sentinel-2 imagery using SVM in GEE

3.1.4 Maximum Likelihood Classification: MXL classification is one of most the popular classification techniques. Under this, the classes are identified based upon the maximum likelihood of the pixel, belonging to a particular class [15]. Training signatures were generated using ground truth sites collected form bhuvan server. Crop signature profiles were also generated (Figure 7) and signatures were merged using signature separability and major crop classes were selected as wheat, mustard and vegetables along with other classes as other crop, baresoil, and orchard etc. Figure 8 shows the classification output of MXL classifier.

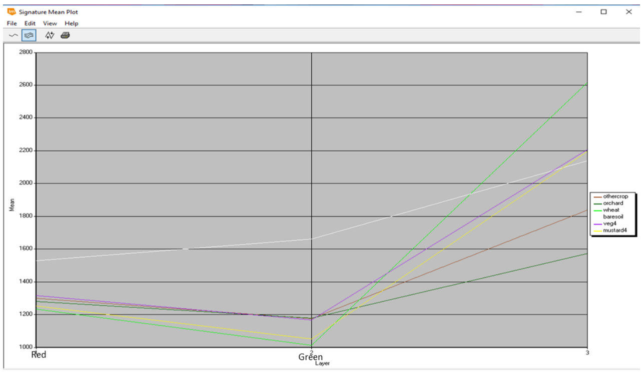


Figure 7. Crop signatures for wheat (green light), mustard (yellow), vegetables (purple), orchard (green dark), other crops(brown) and baresoil (white) in Sentinel -2 data

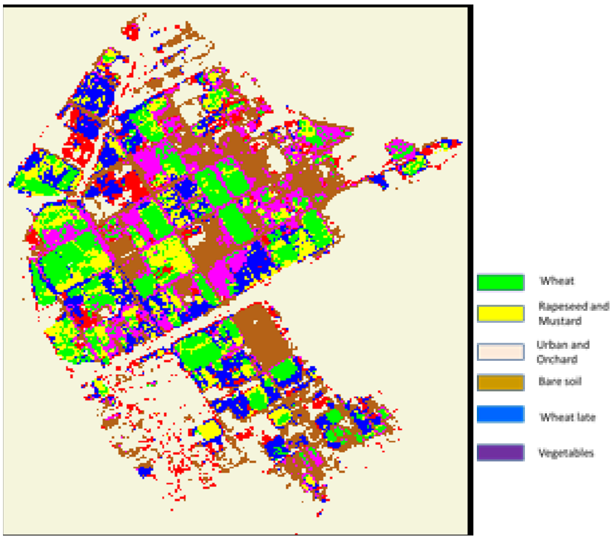
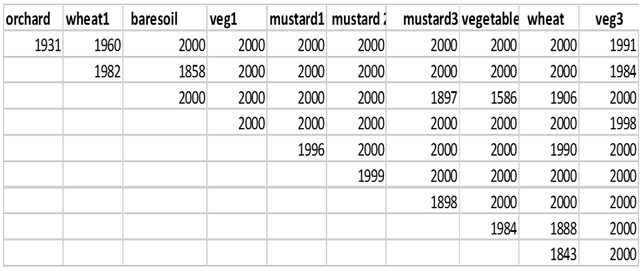


Figure 8. Separability analysis between the classes (above) and classified image obtained by classifying Sentinel-2 imagery using MXL classification (below)

3.2 Separability Analysis:

In this analysis, separability between the signatures were computed and classes were merged, for major crops. Wheat, mustard and vegetables were clearly separable (> 1900), but vegetables classes were not highly separable (< 1700) and were found to be difficult to classify.



orchard	wheat1	baresoil	veg1	mustard1	mustard 2	mustard3	vegetable	wheat	veg3
1931	1960	2000	2000	2000	2000	2000	2000	2000	1991
	1982	1858	2000	2000	2000	2000	2000	2000	1984
		2000	2000	2000	2000	1897	1586	1906	2000
			2000	2000	2000	2000	2000	2000	1998
				1996	2000	2000	2000	1990	2000
					1999	2000	2000	2000	2000
						1898	2000	2000	2000
							1984	1888	2000
								1843	2000

3.3 Accuracy Assessment

Accuracy (User’s, Producer’s and Overall) have been calculated for MXL, CART, Random Forest and SVM classification. Confusion matrix for different classifiers are shown in Table 1, 2 and 3. Major classes such as Wheat, Mustard and Vegetables have been taken in accuracy assessment, in addition to the other classes such as Urban and Orchard, Baresoil and Other Crops. It was observed that in case of MXL, overall accuracy was 73.5% whereas using machine learning algorithms overall accuracy was found to be, CART- 73.4%, RF-93.3% and SVM -74.3% (Table 4). In addition to this F parameter was computed for all study classes [14] using different classifiers. A comparison of OA, the crop classes such as wheat, mustard and vegetables possessed a specific behaviour and their F value ranged for all classes from 64 % to 93% using machine learning algorithms whereas in MXL, F value ranges from 33% for vegetables, and 57% to 70% for mustard and wheat respectively. In signature separability analysis, it was shown that vegetables classes were mixed with the other classes Moreover, RF performed better than CART and SVM classifiers.

Actual Vs Predicted	Urban and Orchard	Bare soil	Other crops	Wheat	Mustard	Vegetables	Row Total
Urban and Orchard	369	16	14	40	19	11	469
Baresoil	18	384	0	12	0	1	415
Othercrops	4	0	363	96	18	43	524
Wheat	5	3	121	591	47	75	842
Mustard	1	0	19	45	234	4	303
Vegetables	21	0	74	91	10	288	484
Column total	397	403	517	875	328	422	3037

Table1. Confusion Matrix using CART Classification

Actual Vs Predicted	Urban and Orchard	Baresoil	Other crops	Wheat	Mustard	Vegetables	Row Total
Urban and Orchard	440	6	3	6	0	5	460
Baresoil	7	399	0	9	0	0	415
Othercrops	5	0	487	19	5	8	524
Wheat	7	1	30	797	5	2	842
Mustard	4	0	3	19	275	2	303
Vegetables	5	0	19	32	2	426	484
Column total	463	406	523	882	287	443	3028

Table2. Confusion Matrix using RF Classification

Actual Vs Predicted	Urban and Orchard	Bare soil	Other crops	Wheat	Mustard	Vegetables	Row Total
Urban and Orchard	363	13	20	29	13	22	460
Baresoil	21	382	0	8	0	4	415
Other crops	7	2	344	119	20	32	524
Wheat	12	10	84	646	26	64	842
Mustard	5	0	15	62	216	5	303
Vegetables	20	0	72	83	10	299	484
Column total	408	407	463	947	285	426	3028

Table 3. Confusion Matrix using SVM Classification

	MXL					CART				
Crop	PA	UA	F	OA	Kappa	PA	UA	F	OA	Kappa
Wheat	66.5	73.2	69.7			67.5	70.2	68.8		
Mustard	54.2	60.6	57.2	73.5	0.709	71.3	77.2	74.2	73.4	0.676
Vegetables	33.1	33.9	33.5			68.2	59.5	63.6		
	SVM					RF				
Crop	PA	UA	F	OA	Kappa	PA	UA	F	OA	Kappa
Wheat	68.2	76.7	72.2			90.4	94.7	92.5		
Mustard	75.8	71.3	73.5	74.3	0.687	95.8	90.8	93.2	93.3	0.919
Vegetables	70.2	61.8	65.7			96.2	88	91.9		

PA-Producer's accuracy, UA- User's accuracy, OA- Overall accuracy, F-Measure and Kappa-Kappa Coefficient

Table 4. Comparison of accuracies of MXL, CART, SVM and RF based classification

3.4 Validation

The crop area was estimated using crop classification methods as defined in earlier sections. The actual crop sowing data information was used for comparison and validation with the different classifiers for wheat, mustard and vegetables. The actual crop sowing data was obtained from FOSU (Farm Operational Service Unit) of IARI for the year 2016 for the Rabi Crops, in which wheat and mustard are major crops and then the vegetables and horticulture crops. In this study, for major crops, wheat, mustard and vegetables area estimates were compared and almost similar kind of results were obtained using machine learning methods, except SVM in which wheat was observed slightly higher may be due to the kernel and cost function. For vegetables and horticulture crops, MXL did not perform well as machine learning classifiers used were able to match the actual crop sown area as shown in Figure 9. Among machine learning techniques, accuracy in RF was found better than CART and SVM.

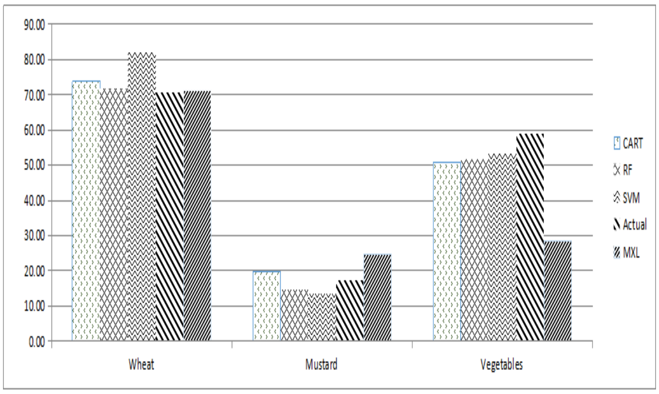


Figure 9. Crop area obtained by classifying Sentinel-2 imagery using different classifiers

4. CONCLUSION

Google Earth Engine is a very powerful geospatial tool for the applications using remote sensing data, with already in-built various satellite data, for which downloading of the data is not required. In this study, MXL classification has been done using ERDAS software and CART, Random Forest and SVM have been used using GEE. Classification accuracies were compared and it was found that overall accuracies ranged from 73% to 93%, in which Random forest performed better than MXL, SVM and CART. For wheat, and mustard crops, all the classifiers were able to discriminate and obtained the accuracy more than 68% in all crop cases, but using MXL, accuracy was higher than 54 % in wheat and mustard but in case of vegetable MXL classification did not perform well. Using Machine learning algorithms, especially using GEE, vegetables could also be discriminated, with good accurcay.

This study was done not only to explore the machine learning algorithms but to enhance the capabilities of high-resolution data, Sentinel-2 using GEE. In future, the study may be extended using the multiple dates of satellite data along with microwave data.

ACKNOWLEDGEMENTS

This research work was carried as part of the Ph.D. work of the first author. The first author is very thankful to Nirma University and Mahalanobis National Crop Forecast Centre (for providing the opportunity to work. The authors are thankful to NRSC (ISRO) for developing the android based software for the collection of ground truth data. The authors are thankful to ESA (Sentinel 2) and the developers of the Google Earth Engine. The authors are thankful to Field Operations Service Unit (FOSU) division of IARI for providing the field data for validation purposes.

REFERENCES

1. Andrii S., Mykola L., Nataliia K. Alexei N. and Sergii S. 2017. Exploring Google Earth Engine Platform for Big Data

Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping. *Front. Earth Sci.* 5(17).

2. Report 2015-16, ICAR–Indian Agricultural Research Institute, ISSN 0972-6136, 206. url. http://www.iari.res.in/

3. Bishop C. M. 2006. Pattern Recognition and Machine Learning. *Springer*. pp. 1-758.

4. Breiman, L. 2001. Random forests. Mach. Learn. 45, 5–32.

5. C. Huang, L. S. Davis & J. R. G. Townshend. 2010. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23, pp. 725-749.

6. Dixon, B. & Candade, N. 2008. Multispectral land use classification using neural networks and support vector machines: one or the other, or both?. *International Journal of Remote Sensing.* 29(4), pp. 1185-1206.

7. Galiano V F R & Rivas C M. 2012. Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and Digital Terrain Models. *International Journal of Digital Earth*. 7, 492-509.

8. Goldblatt, R., You, W., Hanson, G. and Khandelwal, A.K. 2016. Detecting the boundaries of urban areas in India: A dataset for pixel-based image classification in google earth engine. *Remote Sensing*, 8, 28

9. Huang, C., Davis, L. S. and Townshend, J. R. G. 2002. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing,* 23, pp. 725-749.

10. M. Pal & P. M. Mather. 2005. Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*. 26, pp. 1007-1011.

11. M. Pal & P. M. Mather. 2003. An assessment of the effectiveness of decision tree methods for land cover classification. *International Journal of Remote Sensing,* 86, 4, pp. 554-565.

12. Nitze I.,Schulthess U., Asche H. 2012. Comparison Of Machine Learning Algorithms, Random Forest, Artificial Neural Network And Support Vector Machine To Maximum Likelihood For Supervised Crop Type Classification. *Proceedings of the 4th Geobia, Brazil* , pp. 35-40

13. Pandya M., Astha B., Potdar M.B., Kalubarme M.H., Agarwal B. 2013. Comparison of Various Classification Techniques for Satellite Data. *International Journal Of Scientific & Engineering Research*, 4(1)

14. Tilman D., Balzer C., Hill J., and Befort BL. 2011. Global food demand and the sustainable intensification of agriculture. *Proceedings of National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1116437108 108(50) pp.20260-4.

15. Xiong, J., Thenkabail, P.S., Gumma, M.K., Teluguntla, P., Poehnelt, J., Congalton, R.G., Yadav, K. and Thau, D. 2017. Automated cropland mapping of continental africa using google earth engine cloud computing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 126, pp. 225-244.

16.Xiong J.,Thenkabail S.P., Tilton J.C., Gumma M.K., Teluguntla P., Oliphant A., Congalton G.R., Yadav K.,and Gorelick. N. 2016. Nominal 30-m Cropland Extent Map of Continental Africa by Integrating Pixel-Based and Object-Based Algorithms Using Sentinel-2 and Landsat-8 Data on Google Earth Engine *Remote Sensing,* 9, 1065.

17.Yang, C.; Everitt, J. H. & Murden, D. 2011. Evaluating high resolution SPOT 5 satellite imagery for crop identification. *Computers and Electronics in Agriculture.* 75, pp. 347-354.

18. https://developers.google.com/earth-engine/tutorials