JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING TECHNIQUES TO PREDICT FOREST COVER BY USING CARTOGRAPHIC VARIABLES

Yada Sindhura[1], Golla Usha Rani[2], Joseph Fathima Rosme[3]
[1,2,3]Assistant Professor, Department of Computer Science & Engineering
Gurunanak institutions Technical Campus, Hyderabad, Telangana, India

Abstract: Natural resource planning is an important aspect for any society. Knowing forest cover type is one of the Forest land is highly required for developing ecosystem management. Any changes that occur in ecosystem should be carefully noticed to avoid further loss. This model is helpful in noticing the changes occurred due to heavy floods or any other calamities which affected the forest land. A machine learning Algorithm is used to predict the forest cover type using the cartographic variables. The approach is to predict the forest cover type using the cartographic variables like aspect, slope, soil type, wilderness area etc. Various Data mining techniques such as decision tress, random forest, regression trees, and gradient boosting machines are used for prediction of the forest cover type.

Keywords:regression,random,wilderness,gradient,prediction,calamites,ecosystem.

I. INTRODUCTION

The Forest Cover Type Prediction is a supervised multi-class classification problem addressing the prediction of the forest cover type using cartographic features. The earlier models developed for prediction of Forest Cover type used cross fitting along with reduction of the variables used for training the dataset. This system splits the entire data into 10 equal groups. For each group ki, the system would train models by conglomerating the other 9 groups before testing on ki. It then averages the percent error on each testing set ki to determine the percent error for the model. It requires creating and regenerating the whole process every time we need to test the cover type. Possibly, as data are descriptive and there is some relation between attributes. Process Modeling is used to visually represent what a system is doing. Random Forest model performed effectively. The model is designed using machine learning which can further be utilized for other types of predictions. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. High correlated features are extracted and used as important features while modeling. Prediction of cover types at different time stamps gives us an analysis of the geological changes at forest level. On the first observation, the fact that four wilderness-area and forty soil type features are mutually exclusive raised the idea to combine them to make out two features. In machine learning literatures these raw features are called one-hot encoded and then process those to combine them back together and see if this might change the models. The collection of cartographic data pertains to terrain mapping. It is ultimately useful for applying topographic correction to satellite images in remote sensing or as background information in scientific studies. Testing performs a very quality role for assurance and for ensuring the ability of software. The results of testing are used latter on during maintenance also. Our model processes the trained set collected from Kaggle records "https://www.kaggle.com/c/forest-cover-type-prediction/data".An Exploratory Data Analysis (EDA) is done to make the data ready for application of the Machine Learning Algorithm choose to apply. The proposed system contains plots presented in a graphical format for easy understanding of user by using Seaborn and Matplot. The model developed is used in future for various attributes for prediction of forest-cover type. This will minimize the need to train the data every time we require to test for an instance.

II.LITERATURE SURVEY

A learning model to improve the effectiveness of prediction in focused web crawlers is proposed in (Safran et al., 2012). They used Naïve Bayesian method for prediction, the authors have used only four attributes such as parent page, URL keywords, anchor words, and adjacent text for computation. The work given in (Pavani and Sajeev, 2017) proposed model, which helps in extracting significant and hidden pages by merging similarity information related to rank and semantic. The information contained within the thresholds are considered for computation. The prediction can be further improved by incorporating master slave architecture in parallel.

In (Zhong et al., 2015) authors discuss two steps process for the prediction of essential proteins. In initial stage, the authors have constructed a feature space, which show the collection of twenty-six feature. Subsequently, they used centrality measures such as BC, CC, DC, WDC and ION etc. to consider only biological and topological features. Later, to generate feature rank list they used support vector machine with recursive feature elimination (SVM-RFE) and pearson correlation coefficient (PCC). This work uses only PPI network for computation, the prediction can be further increase by incorporating gene expression data.

In (Kathuria et al., 2018) uses a machine learning technique for identification of unknown protein structures. They used Amide frequencies and RF classifier for prediction of protein secondary structure. The result shows that the model performs better in amides dataset. ROC curve and area have been used for validation of model. Multi classification techniques can be involved during secondary protein structure prediction to achieve more accuracy.

The work in (Zhang et al., 2016) used chaos game concept for prediction of protein secondary structure from the given sequence of proteins. The accuracy of structure depends upon the likeness of protein data, this issue can lead to unwanted structure prediction. They used a time series technique, feature vector of 36 dimension and chaos game representation (CGR) to overcome this issue. The prediction accuracy can be further increased by incorporating random tree learning techniques

III.EXISTING SYSTEM

To predict forest cover type four different techniques are applied Regression, Random Forest, Decision tree and GBM and their accuracy and performance has been compared.

IV.PROPOSED SYSTEM

The proposed system contains plots presented in a graphical format for easy understanding of user by using Seaborn and Matplot. The model developed is used in future for various attributes for prediction of forest cover type. This will minimize the need to train the data every time we require testing for an instance. The proposed system results in exploring the data before training the model, more accuracy rates for the future prediction, and the forest cover type can be monitored time to time.

On the first observation, the fact that four wilderness-area and forty soil type features are mutually exclusive raised the idea to combine them to make out two features. In machine learning literature these raw features are called one-hot encoded and then process them to combine them back together and see if this might change the models.

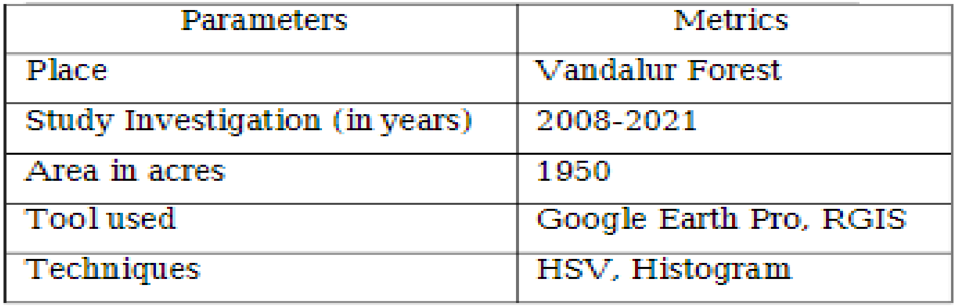
V.METHODOLOGY

In this research, satellite images and forest cover area of vandalur has been carried out of datasets. While performing the analysis, the forest cover area may contain buildings parks, lakes etc. with respect to that need to predict the land degradation from 2008 to 2021.

A. Area of Study

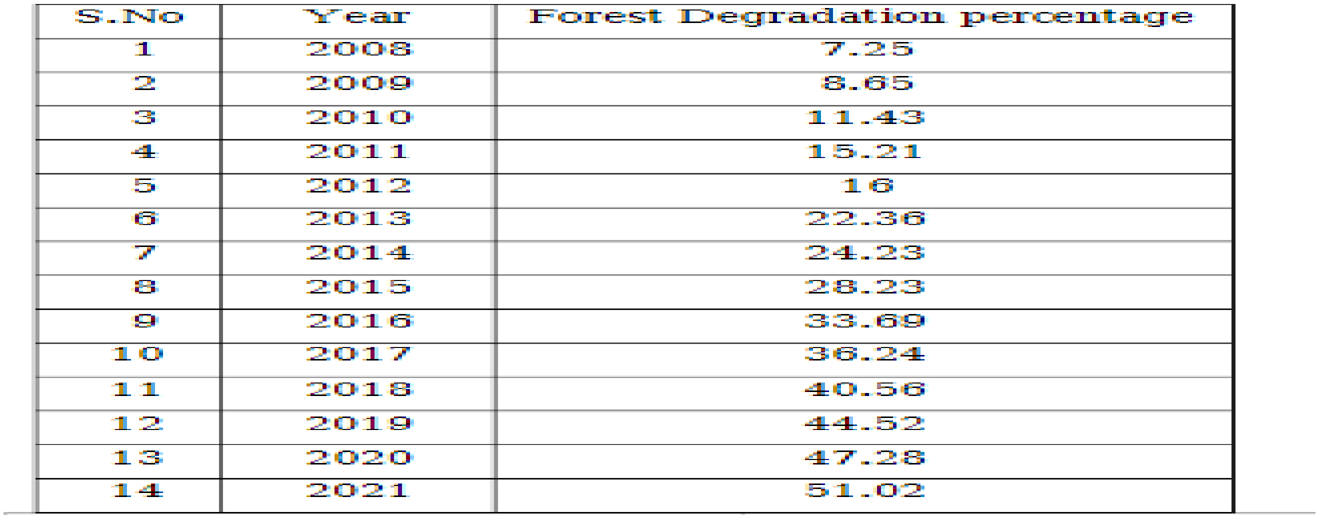
The investigation was undertaken in vandalur forest area. Satellite pictures are collected using Google Earth Pro and then evaluated using image processing algorithms. The detailed information of the vandalur forest cover area are given in table 1. Land satellite data was collected in vandalur forest leveraging the Google engine for this investigation. From 2008 through 2021, satellite pictures of land are collected.

TABLE I Study area



Parameters	Metrics
Place	Vandalur Forest
Study Investigation (in years)	2008-2021
Area in acres	1950
Tool used	Google Earth Pro, RGIS
Techniques	HSV, Histogram

TABLE II Percentage of Forest Degraded



S.No	Year	Forest Degradation percentage
1	2008	7.25
2	2009	8.65
3	2010	11.43
4	2011	15.21
5	2012	16
6	2013	22.36
7	2014	24.23
8	2015	28.23
9	2016	33.69
10	2017	36.24
11	2018	40.56
12	2019	44.52
13	2020	47.28
14	2021	51.02

The table 2 is categorized as follows with year and how much percentage of land cover area has been reduced from 2008 to 2021.

Algorithm:

Step 1: Google Earth Pro Engine satellite data used
Step 2: Sorting the data by year using GIS mechanism
Step 3: Data preprocessing using parameters
Step 4: Using image processing to forecast
Step 5: To characterize the green parts employing HSV
Step 6: Image retrieval determines precision.

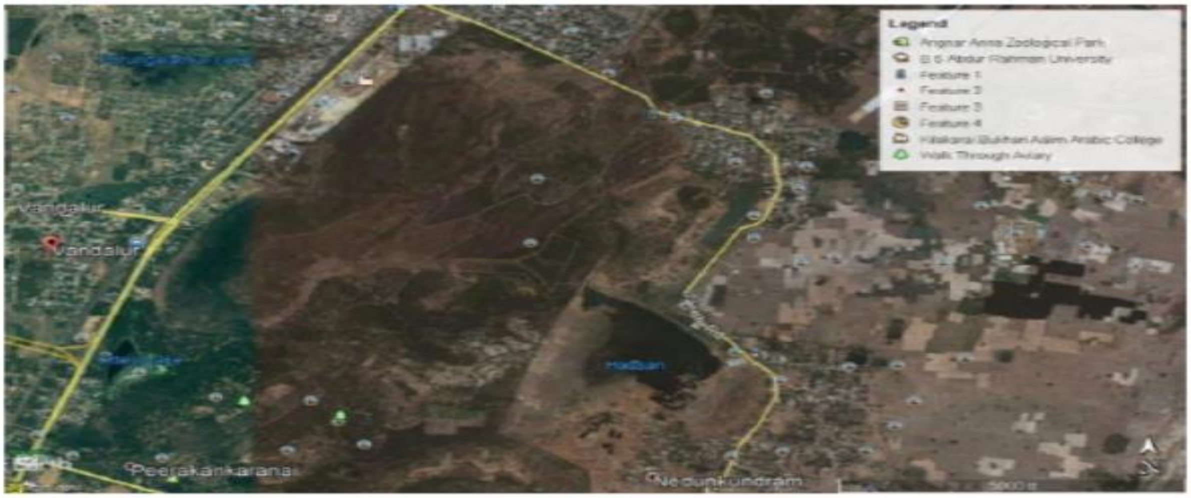


Fig. 2 2007 Forest Land region

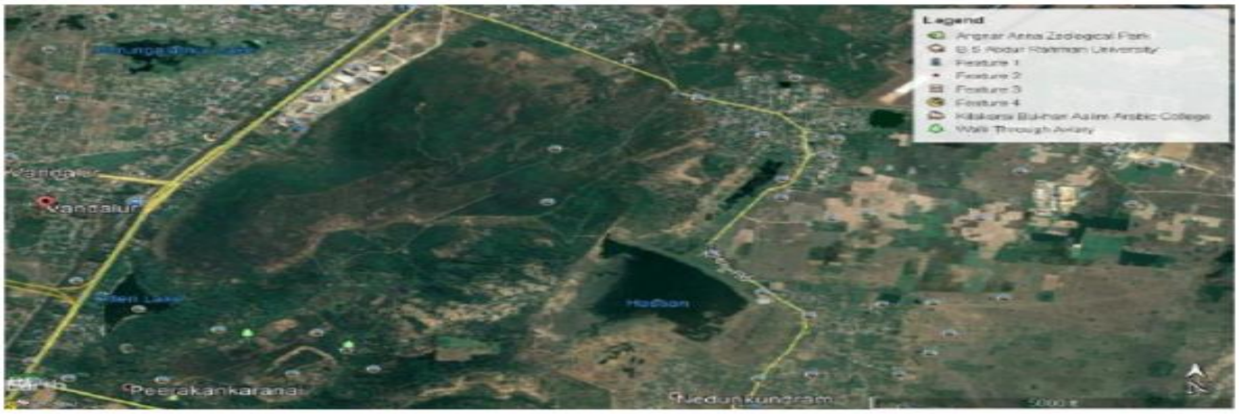


Fig. 3 2009 Forest Land region

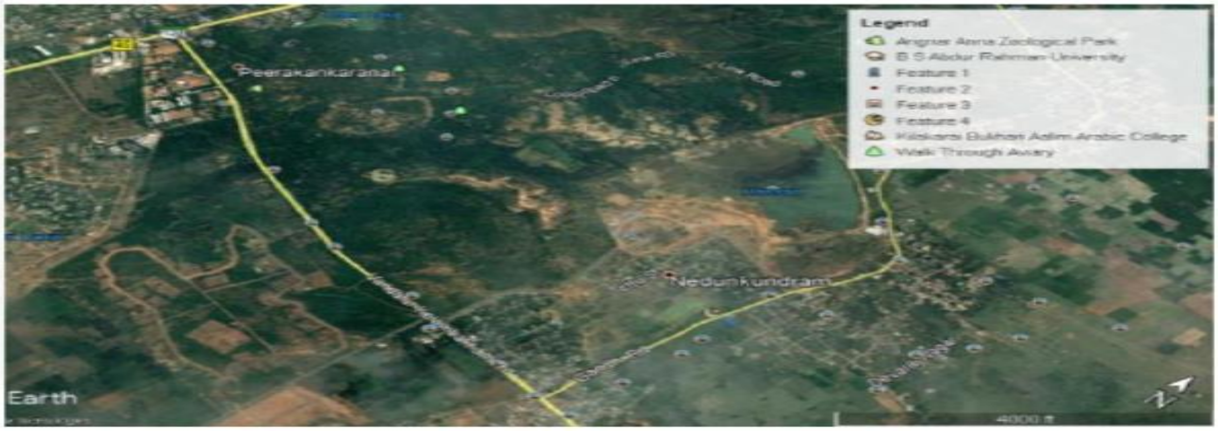


Fig. 4 2010 Forest Land region

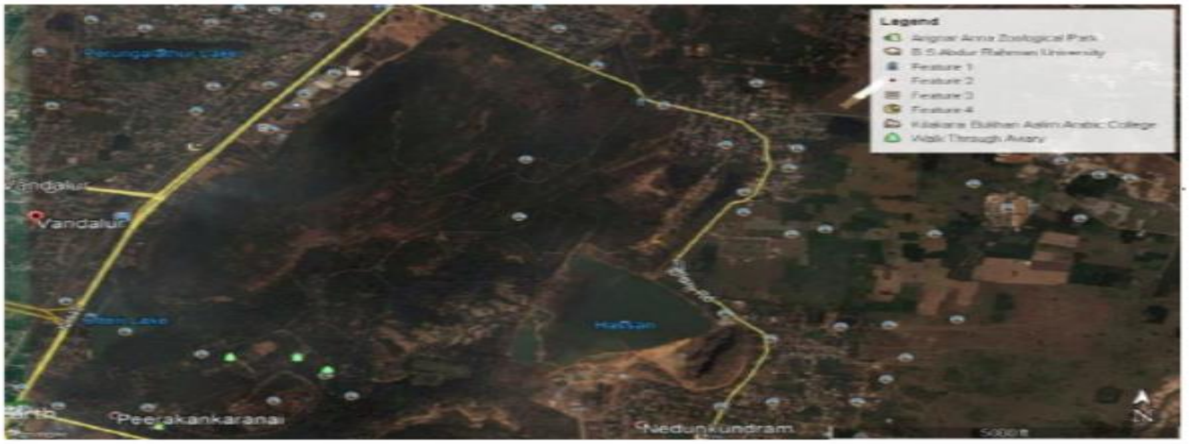


Fig. 5 2011 Forest Land region

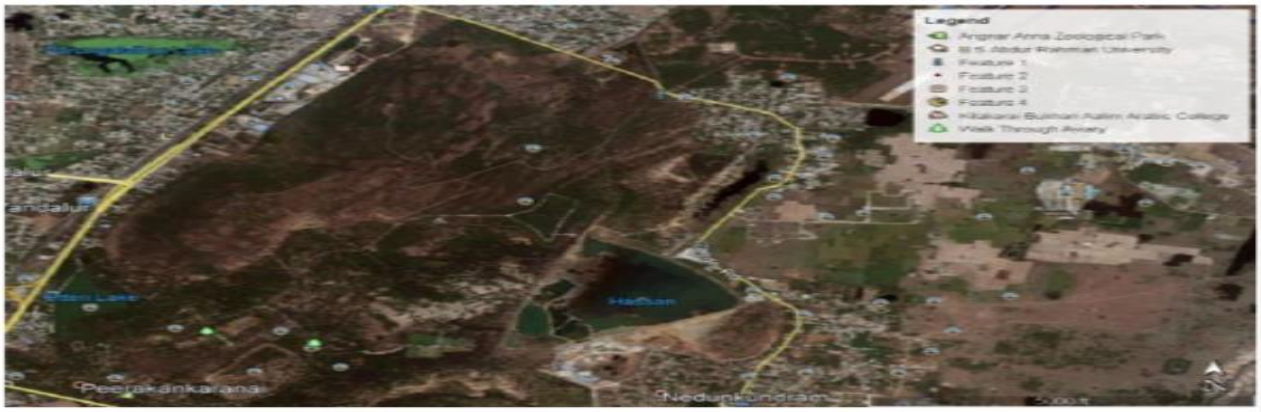


Fig. 6 2012 Forest Land region

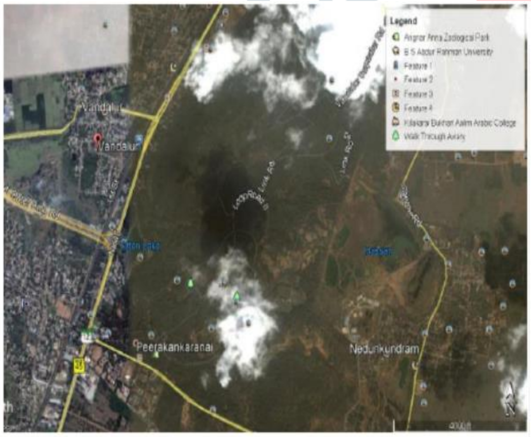


Fig. 7 2013 Forest Land region

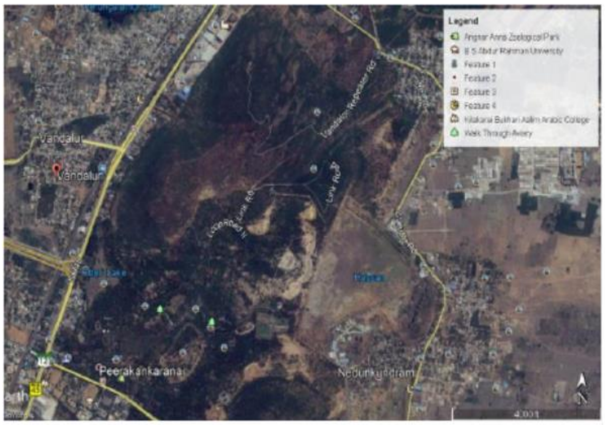


Fig. 8 2014 Forest Land region

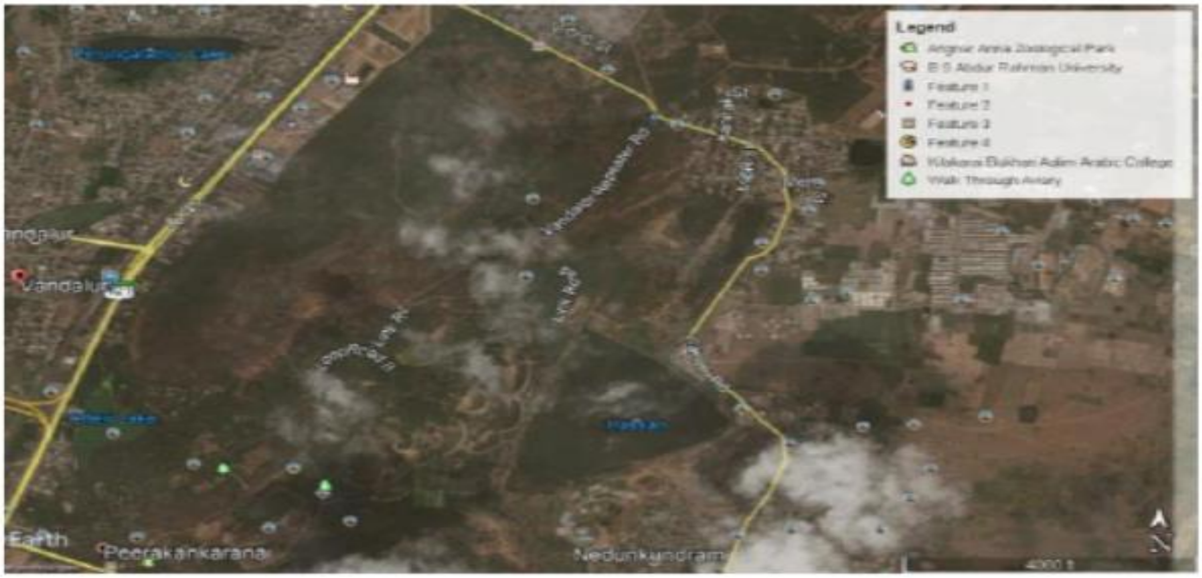


Fig. 9 2015 Forest Land region

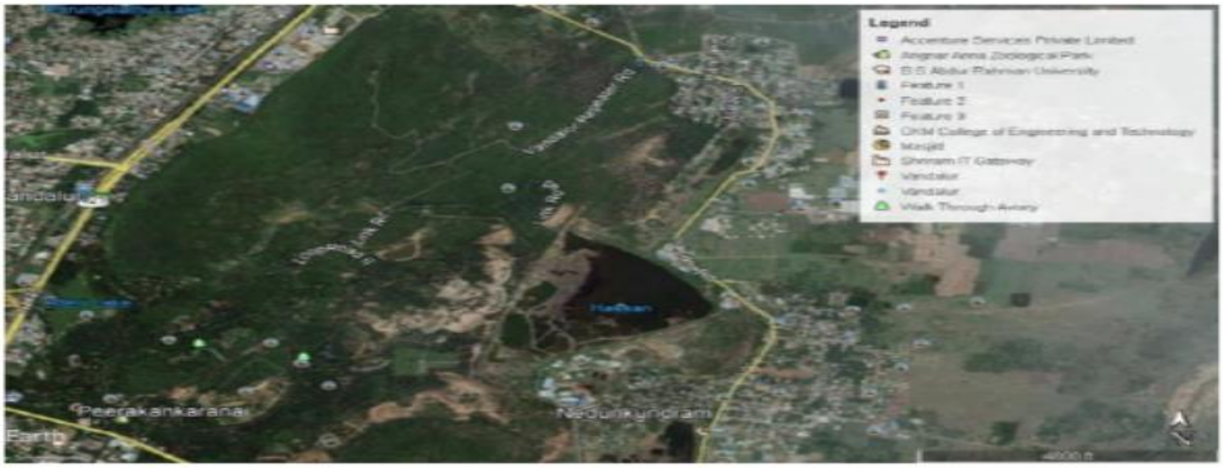
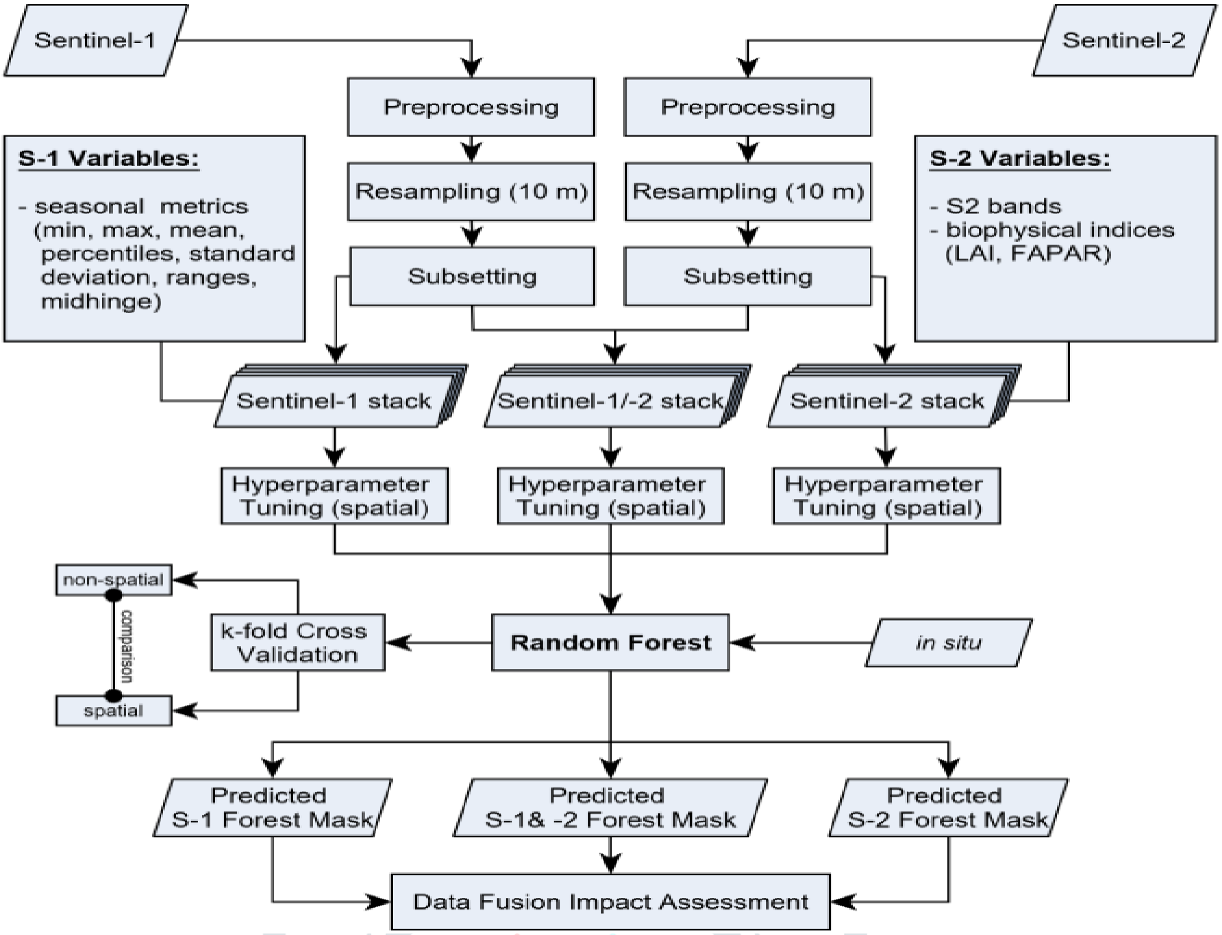

Legend
Accenture Services Private Limited
Arignar Anna Zoological Park
B S Abdur Rahman University
Feature 1
Feature 2
Feature 3
CKM College of Engineering and Technology
Masjid
Shriram IT Gateway
Vandalur
Vandalur
Walk Through Aviary
Peerakankarana
Nedunkundram
Earth
4000 ft

Fig. 10 2016 Forest Land region

VI.SYSTEM ARCHITECTURE


Sentinel-1
Sentinel-2
Preprocessing
Preprocessing
S-1 Variables:
- seasonal metrics (min, max, mean, percentiles, standard deviation, ranges, midhinge)
Resampling (10 m)
Resampling (10 m)
S-2 Variables:
- S2 bands
- biophysical indices (LAI, FAPAR)
Subsetting
Subsetting
Sentinel-1 stack
Sentinel-1/-2 stack
Sentinel-2 stack
Hyperparameter Tuning (spatial)
Hyperparameter Tuning (spatial)
Hyperparameter Tuning (spatial)
non-spatial
comparison
spatial
k-fold Cross Validation
Random Forest
in situ
Predicted S-1 Forest Mask
Predicted S-1& -2 Forest Mask
Predicted S-2 Forest Mask
Data Fusion Impact Assessment

VII.RESULTS

The predictions are based on the presence of color on the satellite photos, which helps to categories the results. The HSV model outputs and histogram outputs are given in the following figures (13-19). The data has been discovered by expectations that each year, a certain amount of land is occupied by buildings, parks, and other uses. By 2030, more than 75% of the vandalur forest's land cover area would have been lost. As a result of the investigation, it was determined that research institutes, schools, and colleges occupied 50% of the total land cover area.

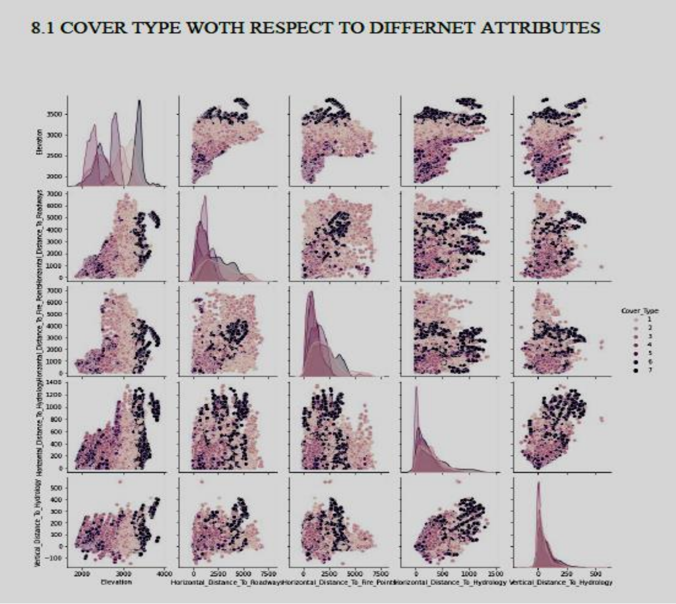


Table 2. Results

Algorithms Used	Accuracy %
1) Logistic Regression	19.4%
2) Decision Trees	35.6%
3) GBM	60%
4) Random Forest	74.8%

VIII.CONCULSION

For the Forest Cover Type prediction task, models based on trees/forests had a good accuracy, achieving 86% of correct predictions on test set. Possibly, as data are descriptive and there is some relation between attributes, Random Forest model performed effectively. We designed the model in machine learning which can further be utilized for other types of predictions. High correlated features are extracted and used as important features while modeling. Prediction of cover types at different time stamps gives us an analysis of the geological changes at forest level. We have built a model which trains itself only once and generates test results for any number of instances of test dataset. Statistical data is given manually to the model to fetch results.

IX.FUTURE SCOPE

- Rapid forest cover detection can be extended for Monitoring changes in the forest cover such as decision-making, forest planning and management, climate change studies and wildlife habitat.
- The More Dataset in the forest cover type and other related datasets will help yielding more accuracy than produced and can cover more domains which have similar data fields and helps to ease the situation

X.REFERENCES

[1]. Cheng, K., Wang, J.: Forest type classification based on integrated spectral-spatial-temporal features and random forest algorithm - a case study in the Qinling mountains. Forests **10**(7), Article no. 559 (2019)

[2]. Badulescu, L. (2017). Data mining classification experiments with decision trees over the forest covertype database. In *21st International Conference on System Theory Control and Computing (ICSTCC), 236*.

[3]. Immitzer, M.; Vuolo, F.; Atzberger, C. First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe. Remote Sens. 2016, 8, 166.

[4]. Mentch, L., & Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. The Journal of Machine Learning Research, 17(1), 841–881.

[5]. Kishore, R.R., Narayan, S.S., Lal, S., Rashid, M.A.: Comparative accuracy of different classification algorithms for forest cover type prediction. In: 2016 3rd Asia- Pacific World Congress on Computer Science and Engineering (APWC on CSE), pp. 116–123 (2016).

[6]. Xue, J.-H., & Hall, P. (2015). Why does rebalancing class-unbalanced data improve auc for linear discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(5), 1109–1112.

[7]. Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", second edition Morgan Kaufmann publisher.