Grenze International Journal of Engineering and Technology, Jan Issue

Comparison of Machine Learning Classification Techniques for Change Analysis in Pauri Garhwal District of Uttarakhand, India.

Purnima Air[1] and Hemant Singh Pokhariya[2]
[1-2]Department of Computer Science and Engineering, Graphic Era (Deemed To Be University), Dehradun, Uttarakhand 248002, India.
Email: airpurnima8@gmail.com, hemantdoon86@gmail.com

***Abstract*— Land use and land cover changes have greatly contributed to global warming and increased urbanization during the last two decades. It is crucial to track alterations of the globe's surface's usage of land as well as cover in order to obtain accurate information about the study region for all development planning scenarios. Land cover categorization is critical for identifying changes, managing development, and tracking and assessing the ground's surface area. Along with RS and GIS methodologies, machine learning algorithms have recently gained popularity in land use and land cover change detection studies. This present study was conducted to detect land use and land cover (LULC) in a heterogeneous landscape of Pauri Garhwal district of Uttarakhand, India by using different machine learning classifiers, i.e. (i) Random forest. (ii) SmileCart (iii) SmileGradientTreeBoost and (iv) SVM. Overall accuracy and kappa coefficient found for RF, SVM, CART and GTB were (0.98, 0.96), (0.97, 0.95), (0.80, 0.75) and (0.89, 0.80) respectively. Results show that Random Forest (RF) provided highest classification accuracy among all four classifier.**

***Index Terms*— Random Forest (RF), Support Vector Machine (SVM), CART, Gradient Tree Boost (GTB) and Land use and land cover (LULC).**

I. INTRODUCTION

One of the most vital resources for humans is land. The fundamental tenet for the advancement of human society is the prudent exploration and use of land resources. Consequently, the problem of land use and land cover changes (LUCC) is a highly useful subject to study. The urbanisation that is connected to the LUCC is equally important to research since it is a representation of human civilization. In order to provide the essential input for environmental planning and ecological management, it is crucial to precisely gather the use of land and land-cover (LULC) information from urban areas [1].

GIS and remote sensing are often used together to produce more accurate and detailed data about the surface the Earth. Remote sensing data can be comprise into GIS to create maps and visualizations that provide a comprehensive view of the Earth's surface. GIS, in turn, that can be used to analyze and interpret remote sensing data and to create models that can predict changes in the Earth's surface over time. Both GIS and remote sensing are important tools in the field of geospatial technology, and their integration has led to many applications in various fields such as natural resource management, disaster response, and urban planning. Remote sensing offers immense promise in land managing resources as a result of its minimal price and extensive time sequence data collection capacity over a broad contiguous area [5]. The most popular technique for determining land cover from

Grenze ID: 01.GIJET.10.1.67

remote sensing is random forest [6]. The Random Forest classifier can be used to categorise heterogeneous circumstances. RF is a simple-to-use classifier given that it just requires the customer to select two values as inputs. There are a lot of limitations on how accurate traditional land cover classification methods can be, as evidenced by numerous research [7]. The parameters of GTB, which use the Infrastructure for SMILE and is employed in the GEE similarly to RF, are the same as those of the original GBT [4]. SVM is one of develop under supervision methods used to address various classification and regression-related concerns. A perfect hyperspace is produced by the classifiers using SVM throughout the learning stage that divides the input datasets less misclassified pixels into a number of classes. For calculating support vectors, the gamma , costing variables, and kernel operations are important factor [2]. Supervised Smile (CART) algorithms were used to categorise remotely sensed photos on the basis of their pixel content. An approach that includes banding was used to choose the top intake picture for classification composing and arithmetic procedures, seasonal composing techniques, and a proceeded class information [3].

According to various studies, the LULC method of classification utilising moderately and smaller satellite images has a number of spectral and geographical constraints that reduce its accuracy [35]. To overcome the aforementioned obstacles and produce high-precision LULC pictures, researchers have started utilising machine learning methods. The setup of the ML model, training data, and input parameters all affect performance, therefore not all ML approaches necessarily result in a high-precision LULC map. Numerous research have been done so far on classifying land uses using machine learning algorithms [36], however the effectiveness of the models is not fully assessed. The primary objective of this article is to compare the performances of various potential classifiers in multitemporal Landsat TM data-based change analysis. By comparing the different classification techniques we can have the best classification techniques which could be used for change detection analysis. Four classifiers, including (i) Random forest (ii) SVM (iii) SmileGradientTreeBoost and (iv) SmileCart for example were employed in this study.

II. Study Area And Data Used

A. Study area

Uttarakhand's pauri garhwal district, a captivating district, is located in northern india's garhwal region(Fig.1). Pauri garhwal has a picturesque location that entices visitors with its natural beauty. It is located amongst the award-winning himalayan ranges. In addition to rudraprayag in the east, chamoli in the northeast, dehradun in the northwestern, tehri garhwal in the west, and almora and nainital in the south, it has borders with a number of neighbouring districts. The district is conveniently accessible as it is only around 102 kilometres (63 miles) from the capital city of dehradun. With national highway 119 passing through the district and providing convenient access to major cities and towns, pauri garhwal is well-connected through road networks.

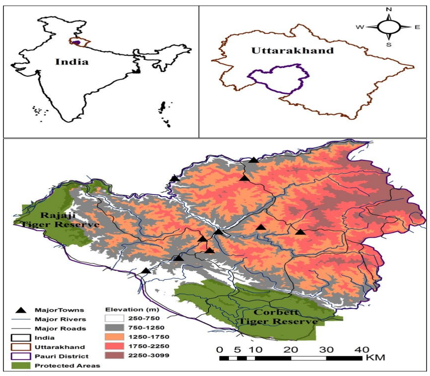

Fig 1. The study region is the PAURI GARHWAL district in Uttarakhand, India.

B. Data used

The USGS Earth Explorer provided the LANDSAT 8 ("OLI+TIRS") image of 2021 for the month of January. These images were put in Google Earth Engine (GEE) for LULC classification. Methodology.

III. METHODOLOGY

A. Methods

Fig. 2 show the methodology used .The LANDSAT image for 2021 was accessible from USGS (http://earthexplorer.usgs.gov/). The Indian administrative district-wise boundary data was used to create the portion of the data (study area). With the utilisation of techniques utilising machine learning, such as Support Vector Machine and Random Forest, SmileCart, and SmileGradientTreeBoost, these picture were divided into five separate LU/LC classes in Google Earth Engine, including water bodies, cities, deserted land, farms, and woods. Trained was carried out on seventy percent of the training samples for all four classifiers, while testing was carried out on thirty percent the overall accuracy (OA), producer accuracy, and user accuracy methodologies, accuracy were evaluated.

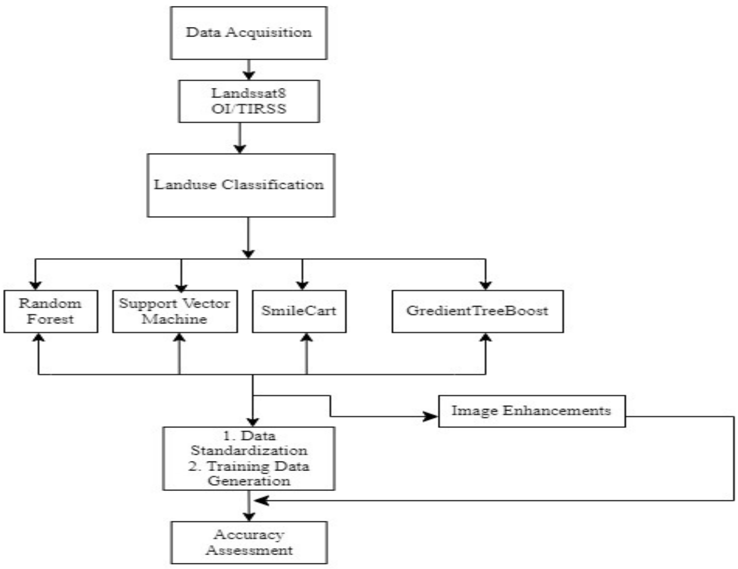

Fig 2. Methodology used for the research.

The usage of decision-tree-based learning algorithms among the finest classifiers. Decision-tree bands have traditionally been favoured because of their higher precision. One method for mixing decision trees is random forest construction, while another is boosting. The machine learning algorithms CART decision-tree, RF, GTB, and SVM are introduced in this part with a brief background overview.

a. CART

A decision classification tree that makes simple decisions and regression analyses easier is called a Classification and Regression Tree (CART) [23]. CART moves by separating nodes up depending on a predetermined threshold, using all of the intermediate nodes, until it reaches the terminal nodes, but one of the additional group sets, trees are created utilising this method's further grouping of the input data. In this project, the "classifier.smileCart" strategy from the Google Earth Engine library was employed. Leo Breiman is credited with creating the gradient boosting machine [24].

b. Random Forest (RF)

The utilisation of Google Earth Engine is permitted for map [16] as well as the unrestricted use of conventional methods including Support vector machines (SVM), categorization and regression trees (CART), and random forests (RF) in GEE [17,19–21]. Numerous decisions trees (DTs) make form the tree-based ensemble classification algorithm known as random forest. The number of DTs to be developed in the RF model and the quantity of variables that are to be used at each DT's breaking site must both be determined [18].Several data subsets form the beginning of the training set of data that were chosen at randomly and replaced (bootstrap samples) were employed to train the ensemble's predictive forests. Furthermore, the DT vertices of the ensemble are partitioned using random subsets of the k predictors. Because of those random events, each tree model in the ensemble is probably distinct. After each of the group's tree models have been run, the DTs in the ensemble cast a unanimous vote to determine the ultimate classification result [22].

c. Smile Gradient Tree Decision Boosting (GTB)

Gradient boosting methods were subsequently created by J.H. Friedman [25]. Gradient-boosting algorithms are frequently employed with CART when the decision models are poor learners. The degree to which each foundation learner is a good match is enhanced by converting Friedman's concept into gradient boosting methods.

d. *Support vector machine (SVM)*

Finding the ideal hyperplane is the goal of the non-parametric, contemporary machine learning statistical technique using support vector machines (SVM) [8] that divides the information gathered from multidimensional features into distinct, predetermined clusters that are compatible with the information to be used for training. Because they may yield improved classification accuracy even with a little training sample, SVM-based approaches are particularly appealing , when it comes to RS [9]. For a variety of tasks, including activities relating to biophysics, land utilisation and surface area, and geomorphological changes, several other studies have shown that SVM classification is more effective than other pixel-based techniques [10-15].

B. Classification

The classification method reveals the numbers of pixels in a picture are properly identified by a classifier. Using a Random Forest, Support Vector Machine, SmileCart, and SmileGradientTreeBoost classifier, March LANDSAT images of the year 2021 were separated into five distinct LULC groups. For the purpose of training machine learning models, which we generated practise data for each Google Earth land cover type. It is determined how accurate the model is by using the testing data, and its efficiency is evaluated using the simulated data.

C. Evaluation of accuracy

The accuracy of the categorization was impacted by methods, processes, time, and space [26, 27, 28]. Numerous studies [29] utilising various LULC classifiers discovered small to large variances in the classification accuracy of LULC. Depending on their accuracy, several classifiers' efficacy was assessed. Overall precision, defined as the proportion of testing data that is successfully categorised, is the primary statistic used to assess the precision and potency of all classifiers. A little disagreement between the precision in this instance demonstrates the results of the classifications utilised assessment in this inquiry. It is not simply the classifier that affects a LULC classification's performance by place and time. Changes in the atmosphere, the surface, and the light might be to blame [30]. A set of data was produced using stratified random sampling [31]. Using stratified selection, the collection of data is divided into groups based on the attributes that make up each group. By using divided sampling, the dataset or strata are split up based on the attributes' qualities. Each stratum's samples have a certain amount in common after being chosen at randomly from among the kinds or strata through stratification. Getting sample data with great efficiency is the major goal of the stratified sampling technique. The common consensus is that accurate representations may be easily extracted with stratified sampling [32]. These datasets for sampling were split into sets for instruction and approval. Thirty percent from all datasets were utilised as evaluation datasets, while the remaining seventy percent were used for training. The selection of datasets were gathered using certain auxiliary datasets, including Google Street View, as a guide. A few of these accuracy assessments were used to assess the correctness of each categorised map for 2021. ArcGIS Pro and Google Earth Engine both provide several ways to examine the accuracy of multiple LULC classifiers.

IV. Results And Discussion

A. Classification of land use and cover

Supervised and unsupervised categorization, as well as other standard land cover classification systems, has been found to be affected by a number of problems, including the fact that each classification is area-dependent and has a distinct precision for specific land cover class [33].Using the Random Forest, Support Vector Machine, Smile Cart, and Smile Gradient Tree Boost classifiers , LANDSAT images of the year 2021 sorted into five land cover groups with the use of the incredibly powerful online cloud computing tool Google Earth Engine (GEE). Urban, water, barren, forest, and agricultural areas are the five types of land cover. Simulated maps of categorised output for the years 2021 are shown in Figures 2, 3, 4 and 5.

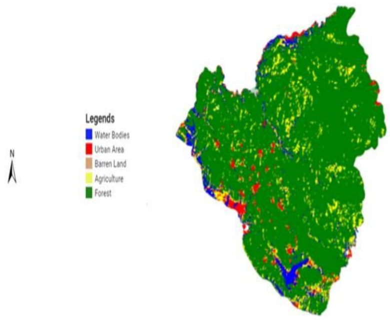

Fig.2 Classified map 0f 2021 by using Random Forest classifier.

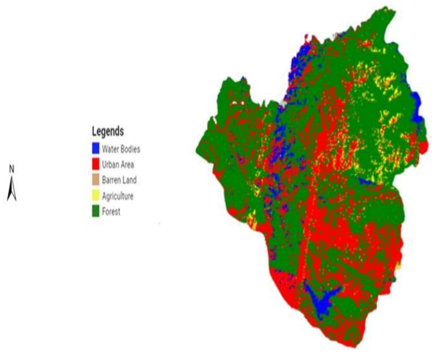

Fig.3 Classified map of 2021 by using Support vector machine classifier.

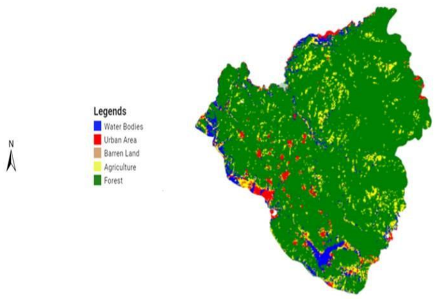

Fig. 4 Classified map of 2021 by using SmileCart classifier.

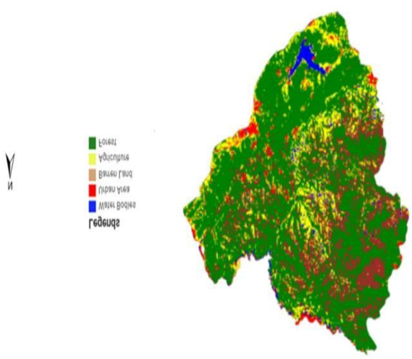

Fig.5 Classified map of 2021 by using Smile Gradient Tree Boost.

B. Overall accuracy assessment

The contingency method, commonly known as the contingency matrix (confusion matrix), is the most used technique for estimating the degree of accuracy. The contingency matrix, according to [34], includes some information, there are overall accuracy, kappa coefficient, user accuracy and producer accuracy. Estimators of total accuracy include the accuracy of the producer and the accuracy of the user. In contrast to user accuracy, which is the precision visible from the user's perspective of the map, producer accuracy refers to the precision noticed from the producer's perspective of the geographical area. The overall accuracy and kappa coefficient results, which were computed using the confusion matrix, are shown in Table I, and they show that RF performed better than all the classifiers for 2021 and classes, with an average OA of 98% with kappa coefficient 96%. However, it has been shown that in 2021, only SVM slightly outperformed RF. Despite the fact that the average OA for SVM is almost identical to that of RF (98%), RF is still preferred over SVM due to its superior performance in terms of class accuracy metrics and consistency across classes and Landsat sensors as opposed to SVM, which had inconsistent results over time. SVM has overall accuracy with 97% and kappa coefficient with 95%. The classifier that performed the most consistently was GTB, It had a third-place finish and an average OA of 89%.with kappa coefficient 80%. CART was the classifier that performed the least consistently, with an average OA of 80% and kappa coefficient 75%. In Fig. 11 Percentage of four classifiers with their respective kappa coefficient and overall accuracy are shown. In Table II the performance measures for each LU/LC class in year 2021, when RF, SVM, CART and GTB Algorithm were employed.

TABLE I. PERFORMANCE MEASURE FOR EACH LU/LC CLASS IN 2021, WHEN RF, SVM, CART AND GTB ALGORITHM WERE EMPLOYED.

Classifier	Random Forest		Support vector machine		SmileCart		GradientTreeBoost	
Land cover	*User accuracy*	*Producer accuracy*	*User accuracy*	*Producer accuracy*	*User accuracy*	*Producer accuracy*	*User accuracy*	*Producer accuracy*
Water bodies	0.91	0.98	0.89	0.96	0.87	0.81	0.85	0.91
Urban	0.92	0.95	0.90	0.92	0.90	0.82	0.94	0.89
Agriculture	0.96	0.91	0.91	0.91	0.84	0.73	0.72	0.80
Forest	0.98	0.97	0.94	0.97	0.82	0.75	0.90	0.79
Barren land	0.96	0.78	0.92	0.76	0.91	0.84	0.8	0.6
Overall	0.98		0.97		0.80		0.89	
Kappa	0.96		0.95		0.75		0.80	

TABLE II. OVERALL ACCURACY AND KAPPA COEFFICIENT FOR FOUR CLASSIFIER.

Classifier	Overall accuracy %	Kappa coefficient %
Random Forest	98%	96 %
SVM	97%	95%
SmileCart	80 %	75 %
GTB	89%	80%

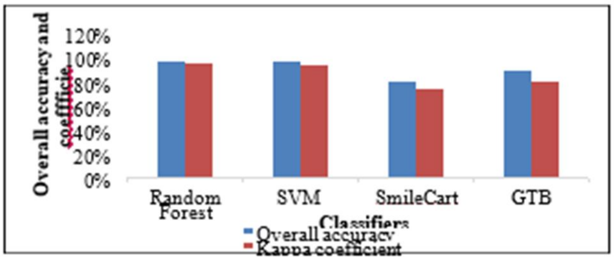

Fig. 11 Percentage of four classifiers with their respective overall accuracy and kappa coefficient.

V. CONCLUSION

The accuracy of the machine classifiers CART, RF, GTB, and SVM for the year 2021 in Pauri Garhwal, Uttarakhand, India, was first examined in this work.The aim of the paper is compare the precision of the four techniques for categorization (Random Forest, support vector machine, SmileCart, and GradientTreeBoost) used to categorise satellite photos of India's Uttarakhand state's Pauri Garhwal district. The results showed that for LULC, The best overall accuracy was demonstrated by RF of 98%. The best overall accuracy was attained by RF of 98%, followed by SVM (97 %), GTB (89%) and finally CART (75%).Employing the average trademark plot, it was possible to classify the characteristics with identical spectrum reflectance measurements more precisely by the use of Random Forest with minor mistakes, which was significantly better than other classification techniques, according to the confusion matrix-based preciseness assessment. For numerous categories, both the general precision of classification and the individual categorization accuracy, RF outperformed other classification algorithms. This research will assist in measuring the shift in land use and provide a foundation for subsequent decision-making.

DECLARATIONS

The researchers state their innocence and are not aware of any internal or external concerns that may have seemed to have impacted the study described in this work.

REFERENCES

[1] Fan, F., Weng, Q. and Wang, Y., “Land use and land cover change in Guangzhou, China, from 1998 to 2003, based on Landsat TM/ETM+ imagery”, Sensors, vol. 7(7), pp.1323-1342, 2007.

[2] Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y.A. and Rahman, A., “Land-use land-cover classification by machine learning classifiers for satellite observations—A review”, Remote Sensing, vol. 12(7), pp.1135, 2020.

[3] Praticò, S., Solano, F., Di Fazio, S. and Modica, G., “Machine learning classification of mediterranean forest habitats in google earth engine based on seasonal sentinel-2 time-series and input image composition optimization”, Remote Sensing, vol. 13(4), pp.586, 2021.

[4] Pham, H.T., Nguyen, H.Q., Le, K.P., Tran, T.P. and Ha, N.T., “Automated Mapping of Wetland Ecosystems: A Study Using Google Earth Engine and Machine Learning for Lotus Mapping in Central Vietnam”, Water, vol. 15(5), pp.854, 2023.

[5] Chandra, G., Defourny, P., & Surendra, S., “Land cover characterization and mapping of continental Southeast Asia using multi-resolution satellite sensor data”, International Journal of Remote Sensing, vol. 24(21),pp. 4181–4196, 2003.

[6] Jin, Y., Liu, X., Chen, Y., Liang, X., “Land-cover mapping using Random Forest classification and incorporating NDVI time-series and texture: A case study of central Shandong”, International journal of remote sensing, vol. 39(23), pp. 87038723, 2018.

[7] Maxwell, A.E., Warner, T.A., Fang, F., “Implementation of machine-learning classification in remote sensing: An applied review”, International Journal of Remote Sensing, vol. 39(9), pp. 2784-2817 2018.

[8] Cortes, C. and Vapnik, V., “Support-vector networks”, Machine learning, vol. 20, pp.273-297, 1995.

[9] Mantero, P., Moser, G., & Serpico, S. B., “Partially supervised classification of remote sensing images through SVMbased probability density estimation”, IEEE Transactions on Geoscience and Remote Sensing, vol. 43, pp. 559–570, 2005.

[10] Melgani, F., “Contextual reconstruction of cloud-contaminated multitemporal multispectral images”, IEEE Transactions on Geoscience and Remote Sensing, vol. 44, pp. 442–455, 2006.

[11] Tang, S., Chen, C., Zhan, H., & Zhang, T., “Determination of ocean primary productivity using support vector machines”, International Journal of Remote Sensing, vol. 29, 6227–6236, 2008.

[12] Andermann, C. and Gloaguen, R., “Estimation of erosion in tectonically active orogenies. Example from the Bhotekoshi catchment, Himalaya (Nepal)”, International Journal of Remote Sensing, vol. 30(12), pp.3075-3096, 2009.

[13] Cao, X., Chen, J., Matsushita, B., Imura, H. and Wang, L., “An automatic method for burn scar mapping using support vector machines”, International Journal of Remote Sensing, vol. 30(3), pp.577-594, 2009.

[14] Knorn, J., Rabe, A., Radeloff, V. C., Kuemmerle, T., Kozak, J., & Hostert, P., “Land cover mapping of large areas using chain classification of neighboring Landsat satellite images”, Remote Sensing of Environment, vol. 113, pp. 957–964, 2009.

[15] Knudby, A., LeDrew, E., & Brenning, A., “Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques”, Remote Sensing of Environment, vol. 114, pp. 1230 – 1241, 2010.

[16] Sun, Z.; Xu, R.,Du, W., Wang, L.; Lu, D., “High-Resolution Urban Land Mapping in China from Sentinel 1A/2 Imagery Based on Google Earth Engine”, Remote Sensing, 11(7), pp.752, 2019.

[17] Hird, J.N., DeLancey, E.R., McDermid, G.J. and Kariyeva, J., “Google Earth Engine, open-access satellite data, and machine learning in support of large-area probabilistic wetland mapping”,Remote sensing, vol. 9(12), pp.1315, 2017.

[18] Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. and Rigol-Sanchez, J.P., “An assessment of the effectiveness of a random forest classifier for land-cover classification”, ISPRS journal of photogrammetry and remote sensing, vol. 67, pp.93-104, 2012.
[19] Wu, Q., Lane, C.R., Li, X., Zhao, K., Zhou, Y., Clinton, N., DeVries, B., Golden, H.E. and Lang, M.W., “Integrating LiDAR data and multi-temporal aerial imagery to map wetland inundation dynamics using Google Earth Engine”, Remote sensing of environment, vol. 228, pp.1-13, 2019.
[20] Adepoju, K.A. and Adelabu, S.A., “Improving accuracy of Landsat-8 OLI classification using image composite and multisource data with Google Earth Engine”, Remote Sensing Letters, vol. 11(2), pp.107-116, 2020.
[21] Tsai, Y.H., Stow, D., Chen, H.L., Lewison, R., An, L. and Shi, L., “ Mapping vegetation and land use types in Fanjingshan National Nature Reserve using google earth engine”, Remote Sensing, vol. 10(6), pp.927, 2018.
[22] Breiman, Leo., "Random forests.", Machine learning, vol. 45, pp. 5-32, 2001.
[23] Shetty, Shobitha. "Analysis of machine learning classifiers for LULC classification on Google Earth engine” , Master's thesis, University of Twente, 2019.
[24] Breiman, L., “Arcing the Edge”, Berkeley Statistics: Berkeley, CA, USA, 1997; Volume 4, pp. 1683–1702, 1997.
[25] Friedman, J. H., “Greedy function approximation: a gradient boosting machine”, Annals of statistics, pp.1189-1232, 2001.”
[26] Hsu, C.W., Chang, C.C., Lin, C.J., “A Practical Guide to Support Vector Classification”, Technical Report, Department of Computer Science and Information Engineering, University of National Taiwan: Taipei, Taiwan, pp. 1396–1400, 2003.
[27] Camargo, F. F., Sano, E. E., Almeida, C. M., Mura, J. C., & Almeida, T., “A comparative assessment of machine-learning techniques for land use and land cover classification of the Brazilian tropical savanna using ALOS-2/PALSAR-2 polarimetric images”, Remote Sensing, vol. 11(13), pp. 1600, 2019.
[28] Rodriguez-Galiano, V.F. and Chica-Rivas, M., “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, vol. 7(6), pp.492-509, 2014.
[29] Reis, S., “Analyzing Land Use/Land Cover Changes Using Remote Sensing and GIS in Rize, North-East Turkey”, Sensors , vol. 8, pp. 6188–6202, 2008.
[30] Islam, K., Jashimuddin, M., Nath, B., & Nath, T. K. “Land use classification and change detection by using multi-temporal remotely sensed imagery: The case of Chunati wildlife sanctuary, Bangladesh”, The Egyptian Journal of Remote Sensing and Space Science, vol. 21(1), pp. 37-47, 2018.
[31] Manandhar, R., Odeh, I.O. and Ancev, T., “Improving the accuracy of land use and land cover classification of Landsat data using post-classification enhancement”, Remote Sensing, vol. 1(3), pp.330-344, 2009.
[32] Li, C., Ma, Z., Wang, L., Yu, W., Tan, D., Gao, B., Feng, Q., Guo, H. and Zhao, Y., “Improving the accuracy of land cover mapping by distributing training samples”, Remote Sensing, vol. 13(22), pp.4594, 2021.
[33] M. K. Ridd and J. Liu, “A Comparison of Four Algorithms for Change Detection in an Urban Environment”, Remote Sensing of Environment, vol. 63(2), pp. 95-100, 1998.
[34] Rwanga, S. S., & Ndambuki, J. M., “Accuracy assessment of land use/land cover classification using remote sensing and GIS”, International Journal of Geosciences, vol. 8(04), pp. 611, 2017.
[35] Pal, S.; Talukdar, S., “ Assessing the role of hydrological modifications on land use/land cover dynamics in Punarbhaba river basin of Indo-Bangladesh”, Environ. Dev. Sustain, vol. 22, pp. 363–382, 2018.
[36] Dutta, D.; Rahman, A.; Paul, S.K.; Kundu, A. Changing pattern of urban landscape and its effect on land surface temperature in and around Delhi. Environ. Monit. Assess. 2019, 191, 551.