Spatial Analysis of Land Use Land Cover Dynamics in the Madurai District Using Sentinel-2 Data and Supervised Learning Algorithms

Naren Raguraman[1], Ashutosh Bhardwaj[2]

[1] International Institute of Information Technology, Bangalore, Karnataka, India - Naren.r.pg@gmail.com
[2] Indian Institute of Remote Sensing, Dehradun, Uttarakhand, India - ashutosh@iirs.gov.in

Keywords: Google Earth Engine, Random Forest, Support Vector Machine, Gradient Tree Boost, Sentinel-2.

Abstract

LULC, or Land Use and Land Cover, refers to the classification and description of different types of land and its usage patterns, including urban areas, forests, agricultural land, etc. In remote sensing, satellite imagery for LULC mapping is becoming more widespread. Numerous studies examine various approaches to improve mapping efficiency and accuracy, highlighting the significance of various data sources, machine learning algorithms, and categorization techniques. This study employs machine learning classifiers, namely Random Forest (RF), Support Vector Machine (SVM), Gradient Boosted Trees (GTB), Classification and Regression Trees (CART), and K-Nearest Neighbors (KNN) for land use and land cover (LULC) classification of Madurai district utilizing Google Earth Engine. The findings reveal the impressive performance of Random Forest, boasting an overall accuracy of 99.01 percent coupled with a commendable Kappa coefficient of 98.68. Conversely. However, amidst these commendable achievements, it’s noteworthy to highlight the nuanced variations observed between the accuracy of training and validation sets. This discrepancy is attributed to the intrinsic intricacies of the learning processes inherent within the algorithms, underscoring the nuanced nature of algorithmic methodologies and their implications on accuracy assessment within spatial analysis frameworks. The generated land use and land cover (LULC) map allows for a comparison between the ground truth data and the surveys conducted to assess issues such as water scarcity and the drying of natural and man-made water bodies.

1. Introduction

The revolutionary technique known as ”geospatial remote sensing” utilizes satellite, aerial, and other sensor-based platforms to gather, process, and visualize data about the Earth’s surface from a distance. This branch of research has profoundly transformed our understanding of the world by providing detailed insights into a wide range of environmental, agricultural, and urban issues. A critical component of geospatial remote sensing is the investigation and understanding of administrative boundaries, which delineate the geographical extent of political or administrative divisions. Understanding the topography of specific administrative boundaries is of paramount importance. These boundaries offer a spatial framework essential for resource management, policy implementation, and governance. Accurate knowledge of administrative boundaries is vital for numerous applications, including natural resource allocation, electoral processes, disaster response, and land-use planning.

Geospatial remote sensing is a powerful tool that enhances our comprehension of the Earth’s surface. It provides crucial insights into administrative boundary research, supporting sustainable development, efficient governance, and well-informed decision-making. As we delve deeper into the complexities of geospatial remote sensing, the emphasis on understanding the topography of specific administrative borders becomes evident as a fundamental element in addressing contemporary challenges and promoting responsible management of Earth’s resources. Land use and land cover (LULC) play a crucial role in global development and climate change. They are intertwined with a nation’s economic and development strategies, reflecting its natural resources, infrastructure, and societal needs. Understanding land use patterns is essential for assessing environmental impacts, predicting future trends, and informing sustainable development policies (Shafiullah et al., 2023).

Various conventional techniques are used to study LULC, including system dynamics, geographic information systems (GIS), numerical studies, and linear programming. These methods allow researchers to analyze spatial data, model land use changes, and assess the socioeconomic factors influencing land use decisions. By integrating environmental and socioeconomic data, researchers can identify the drivers of land use change, evaluate the impacts of different land management strategies, and develop policies to promote sustainable land use practices (Shafiullah et al., 2023). The complexity of land-use/cover change is a critical issue, particularly in tropical regions, where the impacts can be significant. To better understand this complexity, a framework is proposed to provide a more general understanding of the issue. The framework emphasizes the interactions between various drivers and the resulting changes in land use and land cover (Lambin et al., 2003).

Overall, studying land use and land cover is critical for understanding the complex interactions between human activities and the environment. By employing a combination of analytical techniques and data sources, researchers can gain valuable insights into the drivers of land use change and develop strategies to mitigate its negative impacts on the environment and society (Shafiullah et al., 2023). Recent estimates highlight several key changes, including changes in cropland, agricultural intensification, tropical deforestation, pasture expansion, and urbanization. However, there are still unmeasured land-cover changes that need to be addressed. Climate-driven modifications to land cover further complicate the issue, as they interact with other land-use changes. Land-use change is influenced by resource scarcity, market opportunities, policy interventions, loss of adaptive capacity, and changes in social organization and attitudes (Lambin et al., 2003). It is important to note that changes in land use and land cover can significantly impact ecosystem goods and services. These changes can, in turn, feed back into

the drivers of land-use change, creating a complex and interconnected system that requires careful management and planning (Lambin et al., 2003).

The main objective of this paper is to understand the topography of the study area and to generate Land Use and Land Cover (LULC) maps using multiple supervised machine learning classifiers. By employing various algorithms and comparing their performance based on several accuracy metrics, this study provides valuable insights into the characteristics and effectiveness of each classifier. This comparative analysis aims to identify the most suitable algorithm for accurate LULC classification, enhancing our understanding of the study area’s land use and cover patterns.

2. Study Area

This paper aims to provide a comprehensive understanding of the intricate topographical characteristics within the Madurai district, nestled in the southern expanse of Tamil Nadu, covering an expansive area spanning 3,710 square kilometers (Shafiullah et al., 2023). The Madurai district, which is centered on the city of Madurai, is situated in Tamil Nadu, India’s southern region. The district is located between latitudes 9°30’ N and 10°30’ N and longitudes 77°00’ E and 78°30’ E. The districts of Dindigul to the north, Sivaganga to the east, Virudhunagar to the south, and Theni to the west encircle it. The map of the study area is shown in Figure 1 below:

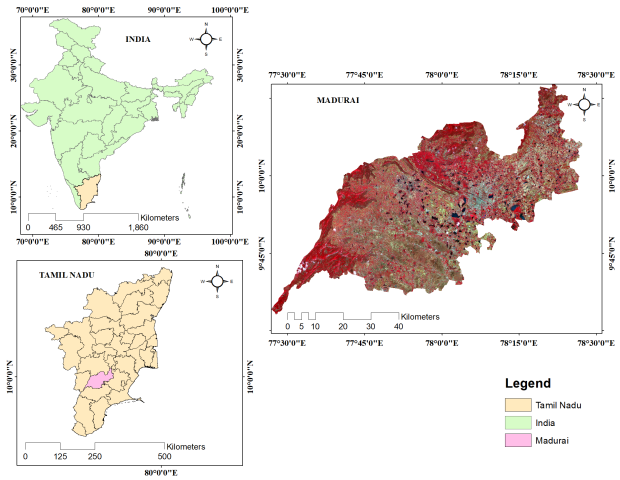

Figure 1. LULC of Madurai Generated by Gradient Tree Boost (GTB)

The Madurai district has a primarily flat geography, with a few hills areas in the west. The district has a tropical climate, which features hot summers, mild winters, and heavy rainfall from October to December during the northeast monsoon season. The average annual rainfall is approximately 850 mm, and the average annual temperature is between 25°C and 35°C. The Madurai district features a variety of land cover and use patterns, such as forests, urban areas, agricultural plains, and arid areas. The main crops used for agriculture are cotton, pulses, millets, and paddy. The most noticeable areas of urban expansion are those surrounding Madurai, a significant center of education, commerce, and culture in southern Tamil Nadu.

3. Material and Methodology

To create the land use and land cover (LULC) map, five primary classes were selected: built-up areas, water bodies, trees, agricultural land, and fallow/barren land. These classes were chosen based on the prevalent land cover types in the study area. The selection criteria followed Level 1 of Anderson’s Land Use and Land Cover (LULC) classification standards, which provides a broad categorization of land cover types. This classification scheme ensures that the resulting LULC map is both comprehensive and consistent with established standards, facilitating meaningful analysis and comparison with other studies or datasets (Anderson et al., 1976). Table 1 illustrates the samples used for the classification process.

Classes	Sampling	
	Point	Rectangle
Built-up	25	125
Water	25	125
Trees	25	125
Agriculture	25	125
Fallow/Baren Land	25	125

Table 1. Input sample for Classification.

A systematic sampling strategy is used to improve the model’s accuracy and robustness. Using this method, 150 samples are chosen from rectangular sections of the study area for each class. Then, a 7:3 split of these samples is made between training and validation sets, guaranteeing that 70% of the examples are used to train the model and the remaining 30% are set aside for confirming its functionality. Because the model is trained on a representative and diverse dataset and its performance is carefully assessed on a separate validation set, the methodical and balanced distribution aids in producing accurate and impartial findings. This tactic raises the generalization capacity of the model as well as the overall precision and dependability of the classification outcomes.

3.1 Harmonized Sentinel-2

Harmonized Sentinel-2 MSI (MultiSpectral Instrument) Level-2A satellite images, renowned for their high-quality and calibrated data, were instrumental in the classification process. These images, captured between January 1, 2023, and March 30, 2023, provided a comprehensive view of the study area.

For visualization and sample selection, the Red, Green, and Blue (RGB) bands (B4, B3, and B2) from Sentinel-2 were utilized. These bands are essential for creating visually appealing and informative images that aid in understanding the landscape’s characteristics and features.

To perform the land use and land cover (LULC) classification with precision, the False Color Composite (FCC) bands (B8, B4, and B3) were specifically chosen.

The FCC bands enhance the ability to differentiate between various land cover types, such as vegetation, water bodies, and built-up areas, based on their unique spectral signatures. Table 2 below provides a detailed description of the bands used in the Harmonized Sentinel-2 MSI, highlighting their wavelengths and specific applications in remote sensing analysis.

Band	Wavelength (nm)	Resolution	Description
B1	443	60m	Coastal aerosol
B2	490	10m	Blue
B3	560	10m	Green
B4	665	10m	Red
B5	705	20m	Vegetation red edge
B6	740	20m	Vegetation red edge
B7	783	20m	Vegetation red edge
B8	842	10m	NIR - Near Infrared
B8A	865	20m	Vegetation red edge
B9	945	60m	Water vapor
B10	1375	60m	SWIR - Cirrus
B11	1610	20m	SWIR
B12	2190	20m	SWIR

Table 2. Harmonized Sentinel-2 Bands;
Source: Google Earth Engine Data Catalog: Harmonized Sentinel-2 MSI

A variety of classification algorithms were utilized to categorize the land use and land cover (LULC) in the study area. Each classifier underwent distinct learning stages using the same set of samples, ensuring a fair comparison. By comparing the outcomes, we aim to identify the most effective algorithm among Random Forest (RF), Support Vector Machine (SVM), Gradient Tree Boosting (GTB), Classification and Regression Trees (CART), and K-Nearest Neighbors (KNN). These algorithms were chosen for their proven effectiveness in remote sensing and their ability to handle complex classification tasks. The disparities in the outcomes will provide conclusive insights into selecting the most suitable algorithm for accurately classifying the LULC in the study area.

3.2 Random Forest:

Random forests, also known as random decision forests, are an ensemble learning technique that builds many decision trees during the training phase for problems including regression, classification, and other applications. The random forest's output for classification problems is the class that most trees choose. The random forest classifier is made up of several tree classifiers, each of which generates a classifier using a random vector sampled separately from the input vector and assigns a unit vote to the class that it believes is most likely to correctly classify an input vector (Pal, 2005, Breiman, 2001).

The random forest classifier is made up of N trees, where N is the user-defined number of trees that need to be developed. Each dataset case is handed down to each of the N trees to classify a new dataset. In that instance, the class with the most out of N votes is selected by the forest (Pal, 2005).

3.3 Support Vector Machine:

Support vector machine (SVM) is a technique for supervised machine learning that finds the best line or hyperplane in an N-dimensional space to maximize the distance between each class to classify data. The goal of SVMs, which are based on statistical learning theory, is to locate decision boundaries in a way that results in the best possible class separation (Pal, 2005, Vapnik, 1995).

The SVMs choose the single linear decision border with the largest margin between the two classes. The distance between the nearest points of the two classes to the hyperplane added together is the definition of the margin (Pal, 2005, Vapnik, 1995).In the beginning, SVMs were created to solve binary (two-class) problems. The right multi-class technique is required when working with numerous classes. For the multi-class problem, strategies like ”one against one” and ”one against the rest” are frequently used (Pal, 2005, Cristianini and Shawe-Taylor, 2000).

3.4 Gradient Tree Boosting:

Gradient Tree Boosting is an effective boosting approach that turns multiple weak learners into strong learners. It uses gradient descent to train each new model to minimize the loss function, such as mean squared error or cross-entropy of the preceding model. Each time around, the algorithm calculates the gradient of the loss function about the current ensemble's predictions, trains a new weak model to minimize this gradient, and repeats the process. The procedure is then continued until a stopping requirement is satisfied after the new model's predictions have been added to the ensemble. It is a group of potent machine-learning methods that have demonstrated significant effectiveness in a variety of real-world settings. They can be easily tailored to meet the specific requirements of an application, such as being trained to take into account various loss functions (Natekin and Knoll, 2013).

3.5 Classification and Regression Trees:

The Classification and Regression Tree (CART) algorithm, a type of decision tree methodology, is extensively employed for both regression and classification tasks. This technique leverages supervised learning, which involves utilizing labeled data to make predictions on new, unlabeled data. By systematically splitting the training dataset into distinct classes, CART seeks to minimize the variance within each subset. The decision tree grows through a process known as binary recursive partitioning, wherein the dataset is iteratively divided into smaller, more homogenous groups based on the maximum and minimum variances of the variables. This approach ensures that each resulting subset is more uniform in terms of the target variable, enhancing the model's predictive accuracy (Bittencourt and Clarke, n.d.).

3.6 K-Nearest Neighbors:

The k-nearest neighbors (KNN) technique, a non-parametric supervised learning classifier, relies on the concept of proximity to classify or predict the grouping of an individual data point. Unlike the Support Vector Machine (SVM), which is a fast learner and determines the decision boundary using the training set before considering any pixels with an unknown class, KNN is considered a lazy learner. This means that KNN does not create a decision boundary during training. Instead, it simply stores the training data and waits until it is presented with an unknown data point to make a classification decision. When an unknown pixel needs to be classified, KNN assesses the stored training pixels and classifies the new pixel based on the majority class of its nearest neighbors. This on-the-fly decision-making process is what distinguishes KNN from more eager learners like SVM. (Hamilton et al., 2018).

3.7 Accuracy assessment:

Four accuracy assessments were employed to evaluate the produced land use and land cover (LULC) map. A crucial component of any classification project is accuracy assessment. It makes a comparison between the identified image and another source of data that is regarded as reliable or ground truth.

Overall accuracy represents the likelihood that an individual data point in a test set is correctly classified. This metric reflects the proportion of correctly predicted instances among the total instances in the test dataset, providing a comprehensive measure of a model’s performance.

$$OverallAcuracy = \frac{(TP + TN)}{(P + N)}, \tag{1}$$

where TP = True Positive
TN = True Negative

The Kappa coefficient measures the degree of agreement between the classifications of two datasets collected on separate occasions. This statistic accounts for the agreement occurring by chance, providing a more robust evaluation of consistency and reliability between the datasets.

$$Kappa = \frac{P_0 - P_e}{1 - p_e}, \tag{2}$$

where P_0 = Observed Agreement
p_e = Expected Agreement

The Producer’s Accuracy represents class-wise accuracy from the point of view of the map maker. The Producer’s Accuracy is calculated by dividing the number of correctly classified samples of class c by the number of samples with the true labels of class c.

$$Producer's Accuracy = 100\% - O_E, \tag{3}$$

where O_E = Omission Error

The User’s Accuracy is the accuracy from the point of view of a map user, not the map maker. The user’s accuracy essentially tells use how often the class on the map will actually be present on the ground. This is referred to as reliability.

$$User's Accuracy = 100\% - C_E, \tag{4}$$

where C_E = Commission Error

4. Results and Discussion

The generated land use and land cover (LULC) map for the study area is depicted in the following figures. These figures provide a detailed visual representation of the LULC classes, which include built-up areas, water bodies, trees, agricultural land, and fallow/barren land, as classified using the harmonized Sentinel-2 MSI Level-2A satellite images. The classification was performed using multiple algorithms, namely Random Forest (RF), Support Vector Machine (SVM), Gradient Tree Boosting (GTB), Classification and Regression Trees (CART), and K-Nearest (KNN).

4.1 Land Use And Land Cover (LULC):

The Land Use and Land Cover (LULC) map of Madurai, generated using the Random Forest (RF) algorithm, is displayed in Figure 2 below:

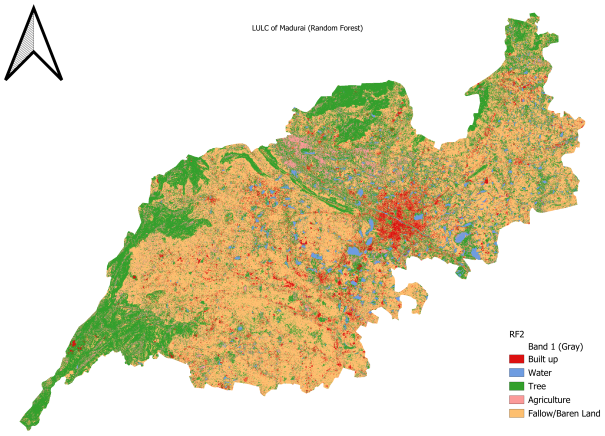

Figure 2. LULC of Madurai Generated by Random Forest (RF)

The Land Use and Land Cover (LULC) map of Madurai, generated using the Support Vector Machine (SVM) algorithm, is displayed in Figure 3 below:

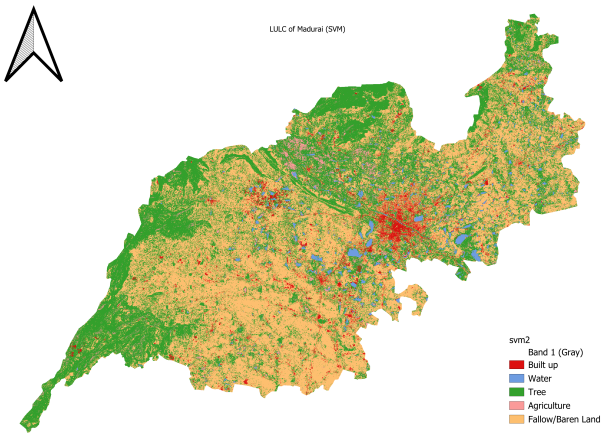

Figure 3. LULC of Madurai Generated by Support Vector Machine (SVM)

The Land Use and Land Cover (LULC) map of Madurai, generated using the Gradient Tree Boost (GTB) algorithm, is displayed in Figure 4 below:

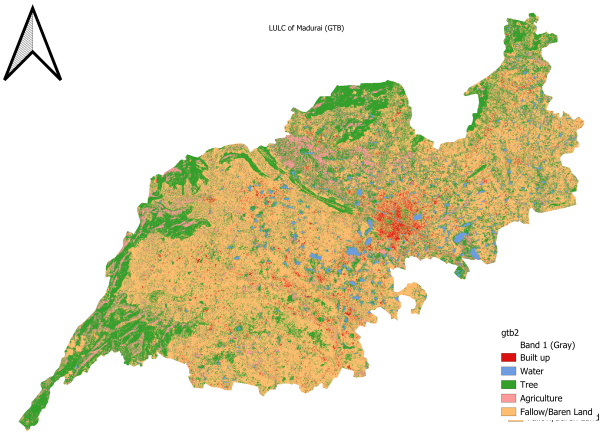

Figure 4. LULC of Madurai Generated by Gradient Tree Boost (GTB)

The Land Use and Land Cover (LULC) map of Madurai, generated using the Classification and Regression Tree (CART) algorithm, is displayed in Figure 5 below:

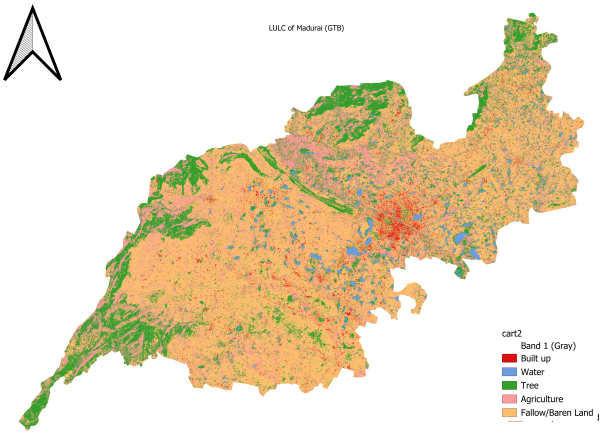

Figure 5. LULC of Madurai Generated by Classification and Regression Tree (CART)

The Land Use and Land Cover (LULC) map of Madurai, generated using the K-Nearest Neighbor (KNN) algorithm, is displayed in Figure 6 below:

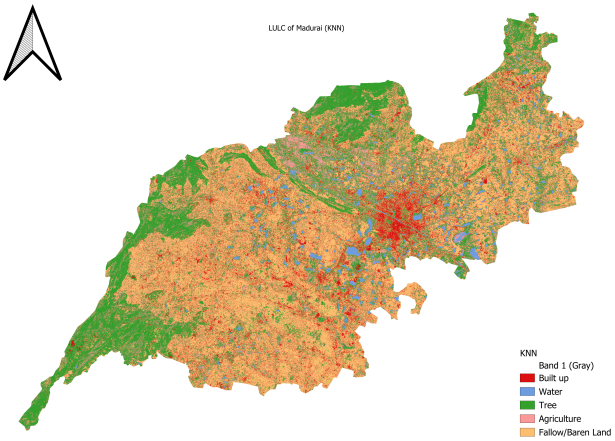

Figure 6. LULC of Madurai Generated by K-Nearest Neighbor (KNN)

4.2 Accuracy Assessments

The accuracy of the classification was evaluated using multiple metrics: overall accuracy, Kappa coefficient, Producer’s Accuracy, and User’s Accuracy, offering a comprehensive assessment of the classification results. Table 3 below presents the area covered by each land use and land cover (LULC) class in the study area, based on the generated LULC map.

	Built-up	Water	Trees	Agriculture	Fallow
RF	254.8	118.2	1204.6	370.2	1819.3
SVM	33.4	12.8	1347.8	10.1	2362.4
GTB	112.2	102.8	1185.2	433.1	1933.7
CART	109.1	99.3	774.6	771.3	2012.6
KNN	455.5	127.1	1147.2	428.1	1609.1

Table 3. Land cover area (Km^2) for each class for different classifiers.

Table 4 below includes the overall accuracy and Kappa coefficient for the classification of each Land Use and Land Cover (LULC) class:

Classifier	Accuracy Assessment	
	Overall Accuracy(%)	Kappa coefficient
RF	99.01	98.68
SVM	82.40	76.17
GTB	76.72	69.31
CART	76.35	69.09
KNN	92.47	89.92

Table 4. Overall Accuracy and Kappa Coefficient for different classifiers.

Table 5 below includes the Producer’s Accuracy for the classification of each Land Use and Land Cover (LULC) class:

	Built-up	Water	Trees	Agriculture	Fallow
RF	98.8	99.6	99.2	97.7	99.7
SVM	80.8	96.2	87.2	59.6	87.8
GTB	60.1	96.4	63.8	76.5	93.7
CART	58.4	96.3	56.4	86.7	95
KNN	90.6	98.9	91.8	86.2	96.3

Table 5. The Producer’s Accuracy (%) for each class for different classifiers.

Table 6 below includes the User’s Accuracy for the classification of each Land Use and Land Cover (LULC) class:

	Built-up	Water	Trees	Agriculture	Fallow
RF	99.6	99.9	98.5	98.8	98.8
SVM	86.0	99.8	75.6	75.7	84.6
GTB	93.1	99.6	79.3	60.8	66.4
CART	93.8	99.7	85.5	57.3	68.3
KNN	93.0	99.7	91.4	87.1	92.7

Table 6. The User’s Accuracy (%) for each class for different classifiers.

https://doi.org/10.5194/isprs-archives-XLVIII-4-2024-367-2024

5. Conclusion

The conclusion is that the accuracy assessment indicates that Random Forest outperforms other classifiers across all accuracy metrics, including overall accuracy, the kappa coefficient, the Producer’s accuracy, and the User’s accuracy (From Tables: 4-6). Further analysis shows that only Random Forest and K-Nearest Neighbors achieve accuracy rates above 90% (From Table 4).

A notable observation from the Producer’s and User’s accuracy metrics is that certain classes in different classifiers exhibit lower accuracy compared to other classifiers for the same class. For example, the Agriculture class in the Support Vector Machine classifier has a Producer’s accuracy of 59.6%, whereas other classifiers achieve accuracy above 75% for the same class (From Table 5). Similarly, from Table 5 Gradient Tree Boost and Classification and Regression Trees have Producer’s accuracy of 60.1% and 58.4% for the built-up class, and 63.8% and 56.4% for the same class, respectively.

This variability in accuracy among classifiers highlights their sensitivity to different classes. While one classifier may perform best for one class (e.g., water), it may perform poorly for another class (e.g., trees). On the other hand, another classifier may achieve a balanced accuracy across all classes.

It is important to acknowledge that the accuracy results are based on the provided input samples, and the ground truth could potentially differ. Therefore, selecting the appropriate classifier is crucial for accurate land use and land cover (LULC) classification, as different classifiers exhibit significant variability in results even when using the same input samples.

References

Anderson, J. R., Hardy, E. E., Roach, J. T., Witmer, R. E., 1976. A land use and land cover classification system for use with remote sensor data. *Professional Paper*. http://dx.doi.org/10.3133/pp964.

Bittencourt, H., Clarke, R., n.d. Use of classification and regression trees (CART) to classify remotely-sensed digital images. *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*. http://dx.doi.org/10.1109/igarss.2003.1295258.

Breiman, L., 2001. Random forests. *Machine learning*, 45, 5–32.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines.

Hamilton, D., Pacheco, R., Myers, B., Peltzer, B., 2018. kNN vs. SVM: A comparison of algorithms. *Fire Continuum-Preparing for the future of wildland fire, Missoula, USA*, 78, 95–109.

Lambin, E. F., Geist, H. J., Lepers, E., 2003. Dynamics of land-use and land-cover change in tropical regions. *Annual review of environment and resources*, 28(1), 205–241.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7. http://dx.doi.org/10.3389/fnbot.2013.00021.

Pal, M., 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1), 217–222. http://dx.doi.org/10.1080/01431160412331269698.

Shafiullah, G., Al-Ruwaih, F., Umar, S. M., 2023. Morphometric analysis using multivariate statistical approaches for the prioritization of potential watersheds in Kuwait. *International Journal of Energy and Water Resources*. http://dx.doi.org/10.1007/s42108-023-00270-z.

Vapnik, V. N., 1995. The nature of statistical learning theory.