Spatio-temporal classification and prediction of land use and land cover change for the Vembanad Lake system, Kerala: a machine learning approach

Parthasarathy Kulithalai Shiyam Sundar1 · Paresh Chandra Deka1

Received: 1 June 2021 / Accepted: 25 October 2021 / Published online: 12 November 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
Land use and land cover (LULC) change has become a critical issue for decision planners and conservationists due to inappropriate growth and its effect on natural ecosystems. As a result, the goal of this study is to identify the LULC for the Vembanad Lake system (VLS), Kerala, in the short term, i.e., within a decade, utilizing three standard machine learning approaches, random forest (RF), classification and regression trees (CART), and support vector machines (SVM), on the Google Earth Engine (GEE) platform. When comparing the three techniques, SVM performed poor at an average accuracy of around 82.5%, CART being the next at accuracy of 87.5%, and the RF model being good at the average of 89.5%. The RF outperformed the SVM and CART in almost identical spectral classes such as barren land and built-up areas. As a result, RF-classified LULC is considered to predict the spatio-temporal distribution of LULC transition analysis for 2035 and 2050. The study was conducted in Idrisi TerrSet software using the cellular automata (CA)–Markov chain analysis. The model’s efficiency is evaluated by comparing the projected 2019 image to the actual 2019 classified image. The efficiency was good with more than 94.5% accuracy for the classes except for barren land, which might have resulted from the recent natural calamities and the accelerated anthropogenic activity in the area.

Keywords Google Earth Engine · Random forest · Support vector machines · Classification · Classification and regression trees · Land use and land cover · CA–Markov chain analysis · LULC prediction

Introduction
The natural and anthropogenic activities worldwide influence the land cover, resulting in modifying its landscapes and the subsequent dynamics of natural processes (Silva et al. 2020). Monitoring and assessing urban growth aid in the planning and utilization of natural resources for the near future. Anthropogenic processes have altered almost half of the Earth’s land surfaces (Tayyebi and Pijanowski 2014). These changes are called land use and land cover (LUC) changes. Over the last few decades, economic prosperity and population growth have resulted in unplanned urbanization and industrialization to meet livelihood and job needs (Jose and Dwarakish 2020). These enormous increases in need boosted the demand for critical infrastructures such as water supplies, sewage services, and recreational activities. It also causes road congestion, pollution, climate change–related problems, urban floods, and urban heat island (UHI) impacts (Jose and Dwarakish 2021; Saxena et al. 2021). As a result, LULC change is regarded as a critical environmental issue of global significance. Major causes of biodiversity loss and related habitat loss are human-induced effects (Elias et al. 2021) such as urbanization, erosion, overgrazing, and resulting land degradation (Abijith et al. 2021). Besides, nature also contributes to this alteration (Halmy et al. 2015; Lambin 1997). The complex interaction of the factors like policy management, human needs, environment, culture, and economics results in changing LULC. Alteration in the LULC can significantly alter the water quality as the increase in urban and agriculture results in nitrates and phosphates in the freshwater (Álvarez-Cabria et al. 2016; Krishnaraj and Deka 2020).
Rapid urbanization in different parts of cities has been a severe threat to developed countries such as India, Indonesia, Malaysia, and Sri Lanka. India is the second most populous country in the world, with 1.38 billion people. According to the Composite Water Management Index (CWMI) report released by Niti Aayog in August 2019 under the Government of India (GOI) stated that many cities in India, such as Delhi, Bengaluru, and Chennai, may face “Day Zero” in the upcoming years (Abijith et al. 2020). Due to the country’s rapid population growth and subsequent urbanization, policies for planning, analyzing, and tracking land use transitions are required to meet the people’s basic needs. As a result, enormous areas of forest cover are being converted to other land uses, causing severe soil erosion. Such rapid soil erosion will lead to a catastrophic flood that will affect the residents downstream. Consequently, sustainable LULC is critical for long-term livelihood and environmental improvement (Mishra et al. 2020).

Role of geospatial technology

Geospatial technology is a significant finding of humanity that has evolved since prehistoric times. The main types of geospatial technology involve remote sensing (RS), geographical information system (GIS), and Global Positioning System (GPS). RS is an excellent tool for regular monitoring and assessing the LULC change in the natural ecosystem, the evolution of landforms by analyzing the geomorphological changes. Likewise, it has played a major role in geography, geology, and environment for the researchers and scientists (Ghosh et al. 2017). It is also more cost-effective and less time-consuming than the traditional method. It has very long spectral bands, and good spatial resolution helps distinguish a significant change in land cover (Abijith et al. 2020). RS and GIS gained much attention from the government and the public due to their ability to map the vulnerabilities during the disaster on a larger scale. The primary goal of GIS is to make it useful to the public and improve their community’s needs (Parthasarathy and Deka 2019).

One of the most extensively utilized sources for analysis is satellite images. In 1972, Earth Resources Technology Satellite (ERTS)-1 was launched and renamed “Landsat” in 1975. Landsat has launched eight series of satellites and has data contiguity for almost 50 years. As of now, Landsat 7 and Landsat 8 are functional. This series of satellites has become one of the important long-term freely available data for the civilian purpose and has been used widely in fields like coastal monitoring (Parthasarathy et al. 2020), LULC (Shi and Yang 2015), vegetation phenology (Senf et al. 2017), and hydrology (Abijith et al. 2020). Thus, Landsat offers a deeper understanding of LULC changes for better decision-making and resource management. Several methods for detecting change using remotely sensed data have been established during the last three decades (Hua et al. 2014; Jat et al. 2017; Rienow and Goetzke 2015; Saxena and Jat 2020; Serasinghe Pathiranage et al. 2018).

Characteristics of Google Earth Engine

Google Earth Engine (GEE) is a multi-petabyte collection of geospatial data co-located with high-performance and intrinsically parallel computation service. In recent years, it has been in RS big data processing spotlight. GEE is an interactive development environment (IDE) to enable rapid visualization and analysis controlled through an application program interface (API) accessed via the Internet. It contains large publicly available datasets of satellite and aerial imagery in both optical and non-optical wavelengths. It consisted of the most freely available RS imagery of entire Landsat archives and Sentinel-1 and Sentinel-2 (Gorelick et al. 2017). The major advantage of GEE is that it provides large sets of data, including land cover, environmental variables, weather, and climate forecast, which is already pre-processed to access the data. Besides, the raw data is also pre-processed, cloud removed, and mosaicked in the GEE to reduce the computational time. Earth Engine code editor is handled by the client libraries handled via JavaScript and Python (Gomes et al. 2020). GEE utilizes the parallel processing technique using MapReduce architecture. It is a method to process a large amount of data into several smaller chunks in several machines. Thus, the data processed as several chunks were recompiled for the result. As the data can be accessed through the API, it is less labor-intensive and requires less storage space to save the data, and also, its simple yet effective architecture does not require high-power computing machines (Noi Phan et al. 2020). Thus, the use of GEE has significantly increased within the RS community (Tamiminia et al. 2020).

Machine learning algorithms

Random forest classifier

A random forest (RF) classifier is a multi-decision tree ensemble classifier that creates many decision trees using a random selection of training samples and variables. In recent years, these ensemble-learning approaches have been frequently applied in the RS. It is introduced by Breiman (2001) that combines K binary classification and regression trees (CART) as it is a nonparametric classifier, so no statistical assumptions have to be made before the distribution of the data. Input sample subgroup used to build each tree obtained from the bootstrapping. The input variable splits into subgroups based on the testing from the individual algorithm in each tree. RF trees are different from the Decision Trees by selecting each node
in a subgroup of the input variables in a random manner and built without pruning (Pelletier et al. 2016). Hence this is a recursive process till the samples are similar or the splitting no longer enriches the model in each subgroup.

**Support vector machines**

Support vector machines (SVM), introduced by Cortes and Vapnik (1995), are based on statistical learning theory. By minimizing the empirical risk and confidence interval of learning derived from the systemic risk minimization hypothesis, it aimed to achieve strong generalization capability. It is an efficient and robust algorithm for both classification and regression. The SVM concept uses the support vectors at the class domain’s edges to create hyperplanes between classes in feature space. The model seeks optimal hyperplane to separate the classes at maximum margins (Shi and Yang 2015). SVM was designed to handle linearly separable classes by bilinear classification, but hyperplane may not be located between the two cases in most cases. In such cases, the model converts the highly inseparable data into a higher dimension or even infinite-dimensional feature space to separate it linearly (RagHAVENDRA and DEKA 2014). In theory, the error penalty, which allows for misclassification, substantially impacts SVM classification accuracy (USTUNER et al. 2015).

**Classification and regression trees**

CART is a simple binary decision tree classifier developed by Breiman et al. (1984). It is linked by a sequence of nodes, each of which is divided into two branches, ultimately leading to leaf nodes, which represent class labels in classification trees and continuous variables in regression trees. The node splitting continues until a threshold condition is reached. CART determines which input features will provide the best split at each node by using the Gini Impurity Index. It is one of the most extensively used LULC classifiers because of its classification accuracy and performance, despite a known tendency to overfit. When compared to multi-layer neural networks, the fundamental advantage of this architecture is that classification choices can be considered as a white box system with simply understood and interpreted input–output relations (Mather and Tso 2016). CART uses the cross-validation technique for pruning, which eliminates branches, and their removal will not affect the results beyond a certain threshold. This may result in a fall in accuracy for training data classification and the loss of certain information, but it also results in an improvement in accuracy for unknown data (Shao and Lunetta 2012).

**Cellular Automata–Markov chain analysis**

As the population grows, so does the need for land. As the demand for land increases, it eventually increases urban. This further increases urban areas on LULC, causing disturbance in the ecosystem affecting sustainability (Aburas et al. 2018). Thus, cellular automata (CA)–Markov chain model is used to understand the factors affecting the spatio-temporal distribution of LULC and to predict the future LULC changes. It is one of the most widely used models to predict the LULC change (Ozturk 2015). This model was created using Idrisi TerrSet’s Land Change Modeler (LCM). To forecast the change in LULC, this hybrid model combines the CA with the Markov chain model (Aburas et al. 2018). As the Markov chain model is stochastic, it predicts one cell transformation to another, i.e., the transition probability of the cell. However, the disadvantage of the Markov chain model is that it does not consider the effect of the neighboring cells over the other. Thus, it lacks spatial modeling capability (Ozturk 2015). CA model considers only the neighboring cell of interest for the estimation of the future. Thus, both the models combined to make the CA–Markov model analyze the spatio-temporal changes in the land cover. It takes two time periods of LULC (i.e., earlier and later dates) as the input to analyze the change in trend. This software aids in analyzing and developing models in the event of stable land cover as opposed to a rapidly changing environment. LCM facilitates the comparison of LULC categories and the net change observed by every class and the contributor to the net change observed by every other LULC category (Hamad et al. 2018). Predicting LULC transition has been used in several applications, including environmental planning by modeling rural development and urban growth, identifying conservation target areas and establishing alternate conservation strategies, analyzing the dynamics of changing agriculture, and simulating rangeland dynamics under various climate change scenarios (Halmy et al. 2015).

To summarize, it was found that GEE aids in the study of LULC shift in a cost-effective and time-consuming manner, and it is commonly used in the literature (Agarwal and Nagendra 2019; Gomes et al. 2020; Noi Phan et al. 2020; Sidhu et al. 2018; Tamiminia et al. 2020; Tassi and Vizzari 2020; Xing et al. 2021). The CA–Markov chain model helps in determining the future land cover change and their land use patterns for decision-makers to provide sustainable development (Aburas et al. 2018; Ansari and Golabi 2019; Bose and Chowdhury 2020; Faichia et al. 2020; Fu et al. 2018; Ghosh et al. 2017; Gidey et al. 2017; Halmy et al. 2015; Hamad et al. 2018; Leta et al. 2021; Ozturk 2015).

In this work, LULC was classified in the GEE platform using three commonly used nonparametric classification machine learning algorithms: (i) RF, (ii) SVM, and (iii) CART. We attempted to understand the performance of these
models using the same collection of training and validation points. Then, using the CA–Markov chain model, we forecast the future spatio-temporal LULC transition analysis and comprehend the changes in LULC for the years 2035 and 2050.

**Materials and methods**

**Study area**

The field of study comprises six watersheds, including Periyar, Muvattupuzha, Meenachil, Manimala, Pamba, and Achenkovil, draining into the Vembanad Lake system (VLS) in Kerala, India (Fig. 1). It falls in six districts, namely Ernakulam, Idukki, Kottayam, Alappuzha, Pathanamthitta, and Kollam. It is located in between the latitude of $9°1'9''$ N to $10°20'22''$ N and the longitude of $76°16'47''$ E and $77°24'43''$ E comprising an area of 12,183 km². An area of 398.12 km² lies below the mean sea level (MSL), and 763.23 km² is located below 1 m MSL. The eastern part of the study consists of Western Ghats, whereas the Arabian Sea binds the western part. The area in and around the VLS flooded recently in the 2018 and 2019 Kerala floods. Lots of built-up and agricultural lands were flooded completely, causing lots of damage to the resources, in which the state took quite a lot of damage due to this unprecedented rainfall. It has a moist, tropical climate, with an average of 150 rainy days each year. The southwest summer monsoon, as well as the northeast winter monsoon, brings rain to this state. From May to August, the southwest monsoon season takes place, whereas between September and November, the northeast winter monsoon arrives, bringing cool winds with it. As a result, the state receives around 3000 mm of rainfall on average (CWC 2018). The mean annual temperature in the state’s coastal lowlands is around 25–27.5 °C, compared to 20–22.5 °C in the eastern highlands. Charnockites, charnockitic gneisses, and pyroxene-bearing granulite in the Western Ghats and central areas of the study characterize the geology of the study. The sedimentary deposits of the Neogene and Quaternary periods dominate the western sections of the sample. Alluvial deposits from the recent past can be found along the coast.

**Data sources and preparation**

The data prepared from various sources are shown in Table 1. Landsat 7 and Landsat 8 Top of Atmosphere (TOA)
images are used for the classification in the GEE environment. The cloud cover is set to less than 30% in the imagery, and the classification is predominantly focused in the premonsoon season (February–May) of the study (Table S1). Landsat 7 was chosen for 2009, while Landsat 8 was chosen for 2013, 2015, 2017, and 2019. Digital elevation model (DEM) is prepared using the ASTER data in the ArcGIS environment. Slope and stream networks are derived from the DEM using hydrology tools in the ArcGIS environment. OpenStreetMap is an open licensed map to create a free editable map worldwide that allows users to download all the features like road, river, and streams to the desired vector file. Hence, road layer is downloaded from the OpenStreetMap. The built-up land and forest are extracted from the classified 2019 land cover using the Raster tool in ArcGIS. Public and industrial (PAI) areas are prepared from the Google Earth Pro comprising schools, grounds, hospitals, hotels, religious places, industrial and commercial places, railway stations, bus stations, airports, and other frequently accessible places by the public like markets and theaters. Population density map is derived from the Census of India (2011). According to the Census of India (2011), this state has the highest population density in the country. Mean annual rainfall is collected from Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) in the GEE environment. All the data are converted to the projected coordinate system WGS 1984 UTM Zone 43 N and fed into the CA–Markov model.

**Land use and land cover classification**

GEE helps analyze the data at a planetary scale and helps the researchers and developers detect changes in the map trends and quantity differences in the Earth’s surface. Landsat 7 and Landsat 8 TOA reflectance images are available in the environment, where they are calibrated using rescaling parameters proposed by Chander et al. (2009). The images before and after the years were used to replace and supplement the images that had been obscured by clouds and fog, and the most pixel image composites were created. In the GEE, median ee.Reducer function is used to determine the median value of each pixel in an image collection for a specified timeline in order to reduce the collection to a single image for each year. The obtained image from the image collection is combined with the Normalized Difference Vegetation Index (NDVI) (Fig. S1), Modified Normalized Difference Water Index (MNDWI) (Fig. S2), Normalized Difference Built-up Index (NDBI) (Fig. S3), and Bare Soil Index (BSI) (Fig. S4) to further enhance and differentiate the classes in the classification. In recent years, machine learning techniques for high-precision classification have been developed in the area of RS.

LULC is categorized using three machine learning techniques in the GEE environment: RF, SVM, and CART. The RF classifier has attracted the interest of the RS community (Abijith and Saravanan 2021; Belgiu and Drăgu 2016; Gislasen et al. 2006; Hamad et al. 2018; Mahdianpari et al. 2018). RF algorithm in GEE allows the users to set the following arguments: the number of decision tree to create (ntree), the number of variables per split, the fraction of input to bag per tree, and the maximum number of leaf nodes in each tree, in which ntree is the most essential input (Noi Phan et al. 2020). The RF algorithm is given as input using the following code in GEE: ee.Classifier.smileRandomForest(ntree, variablesPerSplit, minLeafPopulation, bagFraction, maxNodes, seed), where an empty random forest classifier is created.

CART is one of the most widely used supervised machine learning models for evaluating the effectiveness of the detailed and automated LULC classification approach (Pan et al. 2021; Shetty et al. 2021). The important tuning parameter for CART in GEE is the maximum number of leaf nodes in each tree. The CART algorithm is given as input using the code ee.Classifier.smileCart(maxNodes, minLeafPopulation, bagFraction, maxNodes, seed), where an empty CART classifier is created.

The SVM, a novel algorithm based on statistical learning theory, has not exploited much within the RS community (Adam et al. 2014; Adelabu et al. 2014; Shi and Yang 2015). SVM provides many tuning parameters in the model, but the most important parameters are kernel type, gamma value in

| Sl. No | Parameter        | Data source        | Software used | Resolution (raster)/feature type (vector) |
|--------|------------------|--------------------|---------------|-----------------------------------------|
| 1      | LULC             | Landsat 7 & Landsat 8 Path: 143 and 144 Row: 53 and 54 | GEE | 30 |
| 2      | DEM              | ASTER              | ArcGIS | 30 |
| 3      | Slope            | ASTER              | ArcGIS | 30 |
| 4      | Road             | OpenStreetMap      | ArcGIS | Line |
| 5      | Built-up land    | LULC               | ArcGIS | 30 |
| 6      | Stream           | ASTER              | ArcGIS | Line |
the kernel function, and cost (C) parameter (Yang 2011). An SVM classifier is created using the code `ee.Classifier.libsvm(decisionProcedure, svmType, kernelType, shrinking, degree, gamma, coef0, cost, nu, terminationEpsilon, lossEpsilon, oneClass)` in GEE. As a result, we aimed to test, investigate, and compare the performance of these machine learning algorithms in the classification of LULCs.

**Classification accuracy**

The classification is prepared by giving the region of interest (ROI), for the selected image in the form of point and polygons. For providing a better ROI, 15–35 pixels in each sample were selected. All three models were given the same collection of training and validation points as input to examine their results closely. It is classified into 5 major classes like waterbody, built-up land, vegetation, barren land, and forest. Every class was trained with 75–95 ROIs for classification and 60–75 ROIs for validation. It also ensured that the information was normally distributed and spectrally pure. According to Lillesland and Kiefer (1979), a minimum of 50 samples per class is required as a rule of thumb. Random stratified sampling is used to determine accuracy, with the minimal number of observations put in each segment at random.

Once the classification was done using the above machine learning techniques, error matrix has been generated for each year to identify the accuracy of the classifications. The degree to which the results are close to values accepted as true is referred to as accuracy. Thus, from the error matrix, various matrices such as the overall accuracy, consumer accuracy, producer accuracy, and kappa statistics (Damtea et al. 2020) are identified.

The consumer accuracy of each class depends on the number of correctly classified pixels in the class to the number of pixels belonging to this class in the classification. Simultaneously, the producer accuracy depends on the number of correctly classified pixels to that of the number of pixels belonging to each class in the reference data. In the notation similar to that of Cohen (1960), the kappa coefficient ($\kappa$) is estimated. The reduction of errors by the classification classes was proportional to that of the error by the completely random class (Forghani et al. 2007; Tassi and Vizzari 2020). The magnitude of $\kappa$ usually lies between $-1$ and $+1$. The values more than $+0.5$ indicate it is in good agreement with the classification (Taati et al. 2015). The best-performing model is analyzed. Then, the classified model is further future predicted to analyze the spatio-temporal change in the model.

**CA–Markov model**

CA–Markov analysis has been carried out using the Idrisi TerrSet software in the LCM module. The LULC map 2009 and 2015 are given as the earlier land cover image and later land cover image, respectively. The transition probabilities from 2009 to 2015 and the 2015 LULC base map were used to model the 2019 LULC using the CA–Markov approach. CA assesses the contiguity configuration as well as transition probabilities (Hamad et al. 2018). The CA–Markov model, which consists of the basic LULC layer, transition potential areas for future change formed by the Markov chain model, and transitional potential layers for LULC such as road network, built-up regions, and stream network, is used to determine the appropriate transition from one class to the other (Halmy et al. 2015).

The potential for transition probability is calculated based on the area of each LULC class (preferably barren land, forest, and vegetation) that can be converted into the built-up land (Fig. 2). The transition of these areas was divided into

![Fig. 2](https://example.com/fig2.png)  
**Fig. 2** Potential for the transition from barren land (1), forest (2), and vegetation (3) to built-up respectively
the number of time steps (as an iteration) given by the simulation, which provided the areas to be transitioned each iteration. Land suitability maps show the most suitable pixel to each LULC class. Down-weighted pixels are those that are further apart from the pixel to be converted. These neighborhood rules were defined using the contiguity filter of $5 \times 5$ pixels. As a result, each cell is surrounded by a $5 \times 5$ matrix space that defines the neighborhood of each land class. The standard contiguity filter is given as

$$
\text{Contiguity filter}_{5 \times 5} = \begin{bmatrix}
0 & 0 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 \\
0 & 1 & 1 & 1 & 0 \\
0 & 0 & 1 & 0 & 0
\end{bmatrix}
$$

Assignment of the pixels in the future depends on the pixel’s suitability to a specific LULC class. This simulation continues till each pixel of the LULC class is iterated.

The LULC layer of 2009 and 2015 is given as the input to predict the LULC for 2019. This is performed to prove the model’s accuracy under the given conditions. Then, the model is further used to predict the future spatio-temporal changes of LULC change. The overall methodology for the study is shown in Fig. 3.

**Results and discussion**

**LULC using GEE**

The LULC for 2009, 2013, 2015, 2017, and 2019 were classified using RF, SVM, and CART in GEE, as shown in Figs. 4, 5, and 6, respectively. The indices help classify the images accurately by different means. The NDVI and MNDWI are widely used to extract vegetation, forest, and waterbodies whereas the NDBI and BSI aid in the classification of built-up and barren lands, respectively (Liu et al. 2020).

Previous studies of Cánovas-García et al. (2017), Ghimire et al. (2012), and Noi Phan et al. (2020) stated the performance of the RF model is good at the $n_{\text{tree}}$ of 100. However, according to the current study, the $n_{\text{tree}}$ of 30 functioned better and had a greater accuracy than the $n_{\text{tree}}$ of 100. For 2009, this model accurately categorized the waterbody and built-up and barren lands than the forest and vegetation types. Using Landsat 8 images, the differentiation between forest and vegetation was clear and precise. This may be attributed to the Landsat 8’s (level 1 products) higher radiometric resolution, i.e., twice the predecessor Landsat 7 (Abuzar et al. 2014; Mancino et al. 2020). Table 2 shows the consumers and producers’ accuracy for RF classification. The consumers and producers’ accuracy for RF classification are relatively high for the forest classes, whereas in the built-up areas, it is low.

Over the image, an SVM model is trained with the same collection of ROI and validation points. The importance of the SVM settings has a significant impact on classification results (Huang et al. 2002; Shi and Yang 2015). The kernel type, cost parameter, and gamma value are the critical parameters in the model. Hence, kernel type is assigned with radial basis function (RBF), cost parameter ($C$) is assigned with the value of 100, and gamma value is given as 0.143 (Kavzoglu and Colkesen 2009; Yang 2011). Table 3 shows the consumers and producers’ accuracy in SVM. In the built-up and barren land classes, the consumers’ accuracy is so low that the former is classified as the latter.

The CART model which was carried out with the same training and testing samples for the respective years performed considerably better than the SVM. The maximum number of leaf nodes is set to default. In 2009, the shallow waterbodies in the southern side of the lake is misclassified as the built-up area due to the identical pattern in the Landsat 7 images. Table 4 reveals that, with the exception of built-up areas, the CART model’s consumers and producers’ accuracy is comparable to that of the RF classification in all the years.

Table 5 shows the performance of RF, SVM, and CART algorithms for each year. In 2009, both RF and CART performed equally well in the classification. CART, on the other hand, did not perform better in the other years, resulting in a very modest change in the OA and $k$. RF is more suited to multi-class problems so that it can handle small differences in classification. The SVM model shows some misclassification in the region around the waterbodies. During the dry season, the classification is accurate despite RF’s outperforming SVM in terms of efficiency. When the three approaches were compared, SVM performed poorly with an average accuracy of about 82.5%, CART came second with an accuracy of 87.5%, and the RF model performed well with an accuracy of 89.5%. In almost comparable spectral classifications, such as barren land and built-up regions, the RF outperformed the SVM and CART. As a result, the RF algorithm–based classification is used for further analysis.

Figure 7 shows the change in area from the year 2009 to the year 2019 (Table S2). The waterbody area decreased steadily from 2013 to 2017, but after the flooding, the area of the waterbody increased significantly in the study. It has also been observed that the vegetation has declined by 8% between 2017 and 2019. There is a boom in the built-up area from the year 2009 to the year 2019. This is also the main reason for the study, as Kerala cities have the highest rate of urbanization. In a decade, the urban area has grown by almost 97%. Figure 7 shows the rapid increase in barren land in the year 2019.
may be attributed to the combined effect of the anthropogenic activities and the state’s floods in 2018 and 2019, which caused landslides. It accounts for a 125% increase in barren land from 2009 to 2019, resulting in decreased vegetation cover and forest cover combined. Thus, the RF-classified image in GEE gives a very effective insight into the LULC study transition over a decadal period.
The classified image based on the RF algorithm is then used to forecast and interpret the LULC change in 2019, 2035, and 2050 using the CA–Markov analysis in LCM.

**Future prediction using the CA–Markov model**

LULC changes are influenced by various driver variables such as DEM, slope, population density, mean annual rainfall, distance from the road, built-up, stream network, forest, and PAI areas (Fig. 8). The DEM and slope reveal that the built-up class has a higher probability in areas with a lower elevation and a lesser degree of slope. Distance from the road, distance from the built-up area, and distance from the stream are the driver variables that signify the likelihood of a rise in the built-up classes that are very close to the road, current built-up area, and stream, respectively, and those that are farther apart. In this analysis, the distance from the road, DEM, slope, and distance from built-up areas were generated and used to predict the increase in built-up areas but the resulting map failed to perform better. The inclusion of distance from the stream to the model increased the forecast map’s accuracy for 2019. Thus, it reveals that the urban area is heavily reliant on the river valley, as many residents depend on agriculture. In addition to the model, some of the parameters like distance from forest and PAI areas, mean annual rainfall, and population density made the model still more reliable in the prediction of the future trends of the LULC and the resultant map shows high accuracy in all the classes except barren land.

The association between the driver variables and the distribution of LULC classes in later year image (2015) has been derived by finding Cramer’s $V$. Although it is an imprecise fishing tool, higher Cramer’s $V$ shows that the variable’s possible explanatory value is good. However, it does not guarantee good results because it does not account for the mathematical constraints of the modeling technique utilized and the relationship’s complexity. Cramer’s $V$ for the driver variables is given in Table 6. $V < 0.3$ shows the weak association, $V \in [0.3–0.5]$ shows the medium association, and $V > 0.5$ shows the strong association. Mean annual
rainfall shows a weak association with the classification. Distance from built-up areas, PAI areas, and forest and slope show the medium association whereas DEM, distance from road, stream, and population density show potentially high Cramer’s V.

The accuracy of the model is evaluated with the crosstab module in the Idrisi TerrSet to analyze the agreement of predicted 2019 image to that of the actual 2019 image. It is found that the predicted image of 2019 lies in very good agreement with the classified image of 2019 with a $\kappa$ value of being 84.90%. Hence, Table 7 compares the predicted map of 2019 to the classified map of 2019, and the accuracy for each class in LULC is evaluated. The LULC classes have a high correlation, with more than 95% confidence precision in the waterbody, built-up area, and forest cover. The vegetation cover is 94% compared to the barren land, which is the least accurate than the actual 2019 LULC map, 47.21%. This is because the study area was one of the most severely damaged and devastated areas during the 2018 and 2019 Kerala floods (Parthasarathy et al. 2021). These heavy floods caused landslides and erosion along the river banks in the Western Ghats section (Jacinth Jennifer and Saravanan 2021). This may be one of the key explanations for more barren land than predicted by the model.

The areal changes in the predicted LULC of 2019, 2035, and 2050 are shown in Fig. 9. Following 2019, there is a gradual growth in built-up and barren lands in 2035 and 2050 (Table S3). The estimated barren land area for 2019 is 383.7 km², while the total area is 812.7 km². For the same forecast rate, the annual rise in barren land accounts for just 446.67 km² and 448.84 km², respectively. There is a slight reduction in the built-up area from 555.16 km² in 2019 (predicted) to 552.36 km² in 2019 (actual). The estimated growth in a built-up area for 2035 and 2050 is 633.64 km² and 684.01 km², respectively. In the long run, there is a considerable decrease in vegetation cover.

Figure 10 shows the spatio-temporal prediction of the LULC maps for 2019, 2035, and 2050. The map depicts the rise in the built-up area along the stream until 2035; in 2050, only the density of built-up rises rather than moving...
Fig. 6 LULC classification using CART

Table 2 Consumers’ accuracy and producers’ accuracy using RF for each year of LULC

| LULC       | Consumers’ accuracy (%) | Producers’ accuracy (%) |
|------------|-------------------------|-------------------------|
|            | 2009  | 2013  | 2015  | 2017  | 2019  | 2009  | 2013  | 2015  | 2017  | 2019  |
| Waterbody  | 100   | 88.46 | 96.51 | 94.7  | 100   | 100   | 80.7  | 87.37 | 96.43 | 94.8  |
| Built-up land | 95    | 76.66 | 72.37 | 89.55 | 83.33 | 91.57 | 92    | 87.3  | 80   | 85.53 |
| Vegetation | 75.6  | 94.82 | 93.33 | 87.18 | 97.53 | 97    | 93.21 | 96    | 100  | 97.53 |
| Barren land | 88.78 | 92.54 | 97.8  | 85.29 | 85.77 | 72.04 | 85.87 | 81.48 | 85.29 | 86   |
| Forest     | 89.35 | 90.14 | 89.36 | 91.12 | 92.65 | 88.23 | 94.58 | 93.82 | 88.25 | 91.35 |

Table 3 Consumers’ accuracy and producers’ accuracy using SVM for each year of LULC

| LULC       | Consumers’ accuracy (%) | Producers’ accuracy (%) |
|------------|-------------------------|-------------------------|
|            | 2009  | 2013  | 2015  | 2017  | 2019  | 2009  | 2013  | 2015  | 2017  | 2019  |
| Waterbody  | 100   | 82    | 94    | 87.5  | 94.66 | 100   | 71.92 | 98.36 | 97    | 92.2  |
| Built-up land | 90.18 | 63.8  | 91.17 | 89.65 | 82.09 | 83.13 | 92    | 58.49 | 64.67 | 68   |
| Vegetation | 76.86 | 80    | 55.81 | 60.19 | 74    | 67.74 | 53.22 | 80.57 | 89.17 | 88   |
| Barren land | 88.78 | 92.54 | 97.8  | 85.29 | 85.77 | 72.04 | 85.87 | 81.48 | 85.29 | 86   |
| Forest     | 89.35 | 90.14 | 89.36 | 91.12 | 92.65 | 88.23 | 94.58 | 93.82 | 88.25 | 91.35 |
spatially. Thus, the spatio-temporal trend indicates that by 2050, almost 685 km² of land will be transformed to a built-up area, with a density significantly higher than the current LULC.

To simulate the transitions in this study, all driving variables were employed. From Table 6, it is evident that Cramer’s V for distance to road and population density is above 0.6 showing the close reliability of two layers with the transition potential maps. Population density and rainfall have a direct relationship with the LULC change. Distance from PAI areas showed the significant improvement in the model as it is one of the most accessible regions by the people in their daily life such as schools, colleges, and religious places. The elevation and slope are recognized as
the imperative topographic factors affecting LULC change. Topography has effects on the spread and extent of urban distribution and forest and barren land conversion to agricultural land. It is also found that the deforestation decreases with the increase in the slope. Other driving factors like distance from stream, distance from built-up area, and distance from roads also play a role in land use change since they make it easier for inhabitants to access the basic commodities. Further, some drivers are not only limited to local specific issues; rather, they are regional, national, and global
issues. Only the trend of anthropogenic presence on the land affects the expected LULC transition. It can vary mainly due to natural disasters such as the 2018 and 2019 floods, which increased the amount of barren land. The current pattern of LULC transition is primarily due to development practices, and it may result in land deterioration, resulting in pollution, a decline in groundwater quality, tension in coastal areas, and so on. These predicted maps can be seen as a prospectus for stakeholders to better understand the impact of land use patterns on land cover. It can help them demonstrate an interest in environmental planning and decision-making for potential land use management and sustainable land cover utilization.

### Conclusion

GEE has a powerful scripting language that works in tandem with its cloud infrastructure and user-friendly API. It aided in the efficient study of the classification based on machine learning. From the analysis, the RF algorithm outperformed CART and SVM model in terms of accuracy. This demonstrates the RF algorithm’s ability to tackle multi-class classification and outperform the other two models. The machine learning application in the CA–Markov model of

#### Table 6. Cramer’s V for the driver variables

| Driver variable      | Cramer’s V |
|----------------------|------------|
| DEM                  | 0.5816     |
| Slope                | 0.3695     |
| Distance from road   | 0.6040     |
| Distance from built-up area | 0.4290 |
| Distance from stream | 0.5620     |
| Population density   | 0.6135     |
| Distance from PAI areas | 0.4562 |
| Distance from forest | 0.3974     |
| Mean annual rainfall | 0.2187     |

#### Table 7. Accuracy assessment of the predicted 2019 LULC with the actual 2019 LULC

| LULC           | 2019 actual (km²) | 2019 predicted (km²) | Accuracy (%) |
|----------------|-------------------|----------------------|--------------|
| Waterbody      | 1054.38           | 1034.14              | 98.08        |
| Built-up land  | 552.36            | 555.16               | 99.49        |
| Vegetation     | 6437.12           | 6782.62              | 94.63        |
| Barren land    | 812.70            | 383.70               | 47.21        |
| Forest         | 3326.70           | 3427.63              | 96.97        |

#### Fig. 9. Graphical visualization of areal change in the actual LULC 2019 with that of the predicted LULC for 2019, 2035, and 2050
Idrisi TerrSet software helps analyze the spatio-temporal change prediction of the LULC using the LCM module. Except for barren land, this model had more than 94% accuracy for all LULC classes. The lower prediction accuracy may be influenced by an increase in both anthropogenic and natural disasters. This might be a potential limitation of the analysis since the spatio-temporal variation is limited to the current trend of transition in the LULC, which can differ over time. This research can also be investigated using high-resolution images such as Sentinel to observe the evolving trend. As of now, the Sentinel collection of satellite data is not available for more than a half-decade; this may be the potential reach of the analysis.

As urbanization in developed countries becomes uncontrollable and unsustainable, determining the spatial pattern and patterns of urban development is a crucial problem to solve, to achieve a sustainable environment. As a result, there is a need to raise public consciousness about the dramatic improvements and degradation of natural environments. In this regard, the research was carried out to examine the transition in the urban environment. Since the study area was seriously impacted by the Kerala floods in 2018 and 2019, this study will help society understand the changes in the LULC due to anthropogenic and flood-related factors. This will assist planners in doing the requisite precautionary plans for the current and potential urban growth trend for sustainable development. The transition study in the LULC groups reveals a dramatic growth in the built-up area and barren land, followed by a steady decline in vegetation and forest cover.

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/s11356-021-17257-0](https://doi.org/10.1007/s11356-021-17257-0).

**Acknowledgements** The authors would like to acknowledge Dr. Subrahmanya Kundapura for his valuable suggestions and timely help, and would also like to acknowledge the Department of Water Resources and Ocean Engineering, National Institute of Technology Karnataka Surathkal, India, for providing the infrastructural support.

**Author contribution** Participation of P.K.S.S. includes data collection, analyzing the results, writing the article, and revising the reviewers’ comments. Participation of PCD includes supervision, conceptualization, and reviewing and editing the article.

**Data availability** The data generated or analyzed during this study are included in this published article and its supplementary information files.

**Declarations**

**Ethics approval and consent to participate** The authors confirm that this article is an original research and has not been published or presented previously in any journal or conference in any language (in whole or in part). Consent to participate is not applicable.

**Consent for publication** Not applicable.

**Competing interests** The authors declare no competing interests.

**References**

Abijith D, Saravanan S (2021) Assessment of land use and land cover change detection and prediction using remote sensing and CA Markov in the northern coastal districts of Tamil Nadu. Environ Sci Pollut Res, India. [https://doi.org/10.1007/s11356-021-15782-6](https://doi.org/10.1007/s11356-021-15782-6)

Abijith D, Saravanan S, Jennifer JJ, Parthasarathy KSS, Singh L, Sankrithi R (2021) Assessing the impact of damage and government
response toward the cyclone Gaja in Tamil Nadu, India. In: Disaster Resilience and Sustainability. Elsevier, pp 577–590

Abijith D, Saravanan S, Singh L, Jennifer JJ, Saranya T, Parthasarathy KSS (2020) GIS based multi-criteria analysis for identification of potential groundwater recharge zones - a case study from Ponnaniyaru watershed, Tamil Nadu, India. HydroResearch 3:1–14. https://doi.org/10.1080/11648/J.AIRS.2020.02.002

Aburas MM, Abdullah SH, Ramli MF, Ash’ari AR, Ahmad M (2018) Simulating and monitoring future land-use trends using CA-Markov and LCM models. IOP Conf Ser: Earth Environ Sci 169(1):012050

Aburas MM, Sheffield K, Whitfield D, O’Connell M, McAllister A (2014) Comparing inter-sensor NDVI for the analysis of horticulture crops in South-Eastern Australia. Am J Remote Sens 2(1). https://doi.org/10.11648/J.AIRS.20140201.11

Adam E, Mutanga O, Odindi J, Abdel-Rahman EM (2014) Land-use/cover classification in a heterogeneous coastal landscape using RapidEye imagery: evaluating the performance of random forest and support vector machines classifiers. Int J Remote Sens, Taylor & Francis 35(10): 3440–3458

Adelabu S, Mutanga O, Adam E (2014) Evaluating the impact of red-edge band from RapidEye image for classifying insect defoliation levels. ISPRS J Photogramm Remote Sens, International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) 95: 34–41

Agarwal S, Nagendra H (2019) Classification of Indian cities using Google Earth Engine. J Land Use Sci, Taylor & Francis 14(4–6): 425–439

Álvarez-Cabría M, Barquín J, Peñas FJ (2016) Modelling the spatial and seasonal variability of water quality for entire river networks: relationships with natural and anthropogenic factors. Sci Total Environ 545–546:152–162

Ansari, A., and Golabi, M. H. (2019). Prediction of spatial land use change based on LCM in a GIS environment for desert wetlands – a case study: Meighan Wetland, Iran. Int Soil Water Conserv Res, Elsevier B.V 7(1): 64–70

Belguy M, Drágu L (2016) Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 114:24–31

Bose A, Chowdhury JR (2020) Monitoring and modeling of spatio-temporal urban expansion and land-use/land-cover change using markov chain model: a case study in Siliguri Metropolitan area, West Bengal, India. Modeling Earth Systems and Environment, Springer International Publishing 6(4): 2235–2249

Breiman L (2001) Random forests. Mach Learn. Springer 45(1):5–32

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and Regression Trees. Routledge, Boca Raton

Cánovas-García F, Alonso-Sarría F, Gomariz-Castillo F, Oñate-Valdivieso F (2017) Modelling the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery. Comput Geosci, Elsevier Ltd 103: 1–11

Census of India (2011) https://censusindia.gov.in/2011-Common/CensusData2011.html. Accessed 15 Aug 2021

Chander G, Markham BL, Helder DL (2009) Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. Remote Sensing of Environment, Elsevier 113(5): 893–903

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement, Sage PublicationsSage CA, Thousand Oaks 20(1): 37–46

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 1995 20:3, Springer 20(3): 273–297

CWC (2018) Study report Kerala floods of August 2018. Trivandrum

Damte W, Kim D, Im S (2020) Spatiotemporal analysis of land cover changes in the Chemoga Basin, Ethiopia, using Landsat and Google Earth images. Sustainability 2020, 12; 3607, Multidisciplinary Digital Publishing Institute 12(9): 3607

Elías T, Shiril KG, E.V. K (2021) Physical model studies on damage and stability analysis of breakwaters armoured with geotextile sand containers. Geotext Geomembr 49(3):604–618. https://doi.org/10.1016/j.geotexmem.2020.12.001

Faichia C, Tong Z, Zhang J, Liu X, Kauzaa E, Ullah K, Al-Shaihah B (2020) Using RS data-based CA–Markov model for dynamic simulation of historical and future LUCC in Vientiane. Laos Sustain (Switzerland) 12(20):1–20

Forghani A, Cechet B, Nadimpalli K (2007) Object-based classification of multi-sensor optical imagery to generate terrain surface roughness information for input to wind risk simulation. In: IEEE International Geoscience and Remote Sensing Symposium. pp 3090–3095

Fu X, Wang X, Yang YJ (2018) Deriving suitability factors for CA-Markov land use simulation model based on local historical data. J Environ Manag. Academic Press 206: 10–19

Ghimire B, Rogan J, Galiano V, Panday P, Neeti N (2012) An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA. GISci Remote Sens, Taylor & Francis Group 49(5): 623–643

Ghosh P, Mukhopadhyay A, Chanda A, Mondal P, Akhand A, Mukherjee S, Nayak SK, Ghosh S, Mitra D, Ghosh T, Hazra S (2017) Application of cellular automata and Markov-chain model in geospatial environmental modeling–a review. Remote Sens Appl: Soc Environ, Elsevier B.V 5(October 2016): 64–77

Gidey E, Dikinya O, Segbeyo R, Segosebe E, Zenebe A (2017) Cellular automata and Markov chain (CA_Markov) model-based predictions of future land use and land cover scenarios (2015–2033) in Raya, northern Ethiopia. Model Earth Syst Environ 3:1245–1262

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pattern Recogn Lett 27(4):294–300

Gomes VC, Queiroz FR, Ferreira KR (2020) An overview of platforms for big earth observation data management and analysis. Remote Sens 12(8):1–25

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017) Google Earth Engine: planetary-scale geospatial analysis for everyone. Remote Sens Environ 202:18–27

Halmy MWA, Gessler PE, Hicke JA, Salem BB (2015) Land use/land cover change detection and prediction in the north-western coastal desert of Egypt using Markov-CA. Appl Geogr, Elsevier Ltd 63: 101–112

Hamad R, Balzter H, Kolo K (2018) Predicting land use/land cover changes using a CA-Markov model under two different scenarios. Sustainability 10(10):3421

Hua L, Tang L, Cui S, Yin K (2014) Simulating urban growth using the SLEUTH model in a coastal peri-urban district in China. Sustainability (switzerland) 6(6):3899–3914

Huang C, Davis LS, Townshend JRG (2002) An assessment of support vector machines for land cover classification. Int J Remote Sens, Taylor & Francis Group 23(4): 725–749

International Geoscience and Remote Sensing Symposium. pp 049.

Jat MK, Choudhary M, Saxena A (2017) Application of geo-spatial techniques and cellular automata for modelling urban growth of a heterogeneous urban fringe. Egyptian J Remote Sens Space Sci, National Authority Remote Sens Space Sci 20(2):223–241

Jose DM, Dwarakish GS (2020) Uncertainties in predicting impacts of future land use and land cover scenarios (2015–2033) in Vientiane. Laos Sustain (Switzerland) 12(20):1–20

Jose DM, Dwarakish GS (2021) Bias Correction and Trend Analysis of Temperature Data by a High-Resolution CMIP6 Model over a Tropical River Basin. Asia-Pacific J Atmos Sci. https://doi.org/10.1007/s13143-021-00240-7
