Registration No – ACRS 0167

Assessing the performance of feature selection approach for urban Land Use/Cover Classification in Kolkata Metropolitan Area, India

Prosenjit Barman[1], *Sk. Mustak[2]

[1]Research scholar, Department of Geography, School of Environment and Earth Sciences, Central University of Punjab, Email: prosenjitkm147@gmail.com

[2]Assistant professor, Department of Geography, School of Environment and Earth Sciences, Central University of Punjab, Email: mustak.sk5@gmail.com/ sk.mustak@cup.edu.in

*sk.mustak@cup.edu.in

Abstract: *Feature selection is crucial in machine learning community for land use and land cover classification. The increased level of features in image classification significantly impacts the possibility of classification accuracy. Since the advent of big data, the digital image has expanded, and a significant amount of earth observation data has become freely accessible. Therefore, feature selection is not just about dealing with a vast volume of developing big data; it's also about knowing which features to extract and which are more valuable. Feature selection (FS) seeks to determine the smallest possible number of attributes needed to maintain the class probability distribution as close to the original distribution of all features as is practical. The rigorous feature selection method plays a significant role in reducing the processing time and storage space while it is producing higher accuracy than the initial datasets. The main objective of this study is to examine state-of-the-art feature selection approaches to improve pixel and object-based image classification accuracy. In addition, Planetscope high resolution satellite datasets and a robust Support vector machine (SVM) will be employed for pixel and object-based LULC classification of the Kolkata metropolitan area. The novel feature selection algorithms, e.g., Gain Ratio, information gain, correlation, Fisher's criterion (F-score), Relief will be examined based on accuracy assessment indices, e.g., overall accuracy, kappa, precision, recall, etc. Several spectral (mean and standard deviation of image pixel value), textural (Grey level co-occurrence matrix), and morphological features (area, compactness, density, etc.) will be extracted and fed into the features selection algorithms to obtain the robust features. The best features will be used for pixel and object-based LULC classification pipelines to achieve the best accuracy. This study could be a novel guideline to address the robust feature selection algorithm and best features to map essential urban land use and land cover for urban policy improvement.*

Keywords: Feature selection, Pixel-based classification, Object-based classification, Support vector machine, Machine learning

1. Introduction:

The use of Urban land is the highest level of alteration and modification made by the human on the Earth. It is also a primary reflection of socio-economic function and human activity (B. Chen et al., 2021; Gong et al., 2020). The highest level modification of earth surface widespread effects on climate, biodiversity, food production, public health and living standard (Clinton et al., 2018; Grimm et al., 2008; Seto & Shepherd, 2009).

Urbanization is one of the most important aspects of the contemporary globalization era (Wang et al., 2021). Over half of the global population resides in urban areas, which comprise only 3% of the earth's land surface (Liu et al., 2014). According to the United Nations, in 2019 it is anticipated that more than half around 4.3 billion people of the world’s lives in urban areas and it will reach around 9.8 billion in 2050. Among this population, more than twice which is around 6.7 billion people (66%) will live in urban areas (Henderson, 2003; UN-DESA, 2014). The prime reason for urbanization from regional to a global scale is the rapid conversion of rural land to urban land (Gao & O’Neill, 2020; Seto et al., 2011). Rapid urbanization is fostering various forms of social and economic development (Fan et al., 2018). Numerous issues arise in urban areas as a result of social and economic development, including traffic jams, urban sprawl, pollution, depletion of natural resources, and ecological crises (J. Chen et al., 2016; Lu et al., 2021). However, the large-scale details of the land use map are limited to the metropolitan city level (T. Hu et al., 2016). So urban land use is essential to understanding the diverse urban functionality and challenging socio-economic issues of urban areas, supporting urban planning, urban sustainability analysis, and sustainable development (Seto & Pandey, 2019). The land use classification mapping standard on data source, data availability, methodological approach, classification scheme varies from city to city and country to country (B. Chen et al., 2021; Gong et al., 2020). The source of this kind of differentiation are differentiate complex built up to high level semantic labels (Zhang et al., 2018; Zhang et al., 2019), financial support and skills of mapping personnel (Gong et al., 2020), secure of spatial and temporal explicit with very high-resolution datasets (B. Chen et al., 2021). The real-time precision of the land use category is critical to the availability of data for the planet's dynamic monitoring and management. It gets more challenging to create an accurate, current real-time urban land use map due to the continued increase in urbanization (B. Chen et al., 2021). In order to monitor and assess urban development, whether it be sprawl or compact, current and comprehensive information regarding the various land use classes in metropolitan areas is necessary. The urban land use map is the ultimate way to address various socio-economic, environmental challenges caused by urbanization in city scale, regional level, national level and global level. All the challenges demarcate the importance of developing a robust and cost-effective data model framework. It also helps in approaching to derive an accurate and real time urban land use classification map.

Land use and land cover data are crucial for various geospatial applications such as environmental management, flood risk modeling, urban planning, and sustainable development to support the SDG (Banzhaf et al., 2017; Patino & Duque, 2013; United Nation., 2015; Schulz et al., 2021). According to Riggan and Weih (2009) the land use and cover map arranges spatial data about the various visible features on Earth's surface, including vegetation, built-up areas, crops, and other land uses like waste, fallow, and agricultural land. Any covering of the earth's surface, including water, plants, bare soil, and urban infrastructure, is referred to as "land cover". According to Use & Anderson, (2017) land use is the term used to describe the usage of land for uses other than agriculture, such as recreation, wildlife habitat etc. The identification of land cover establishes the baseline from which monitoring activities can be conducted.

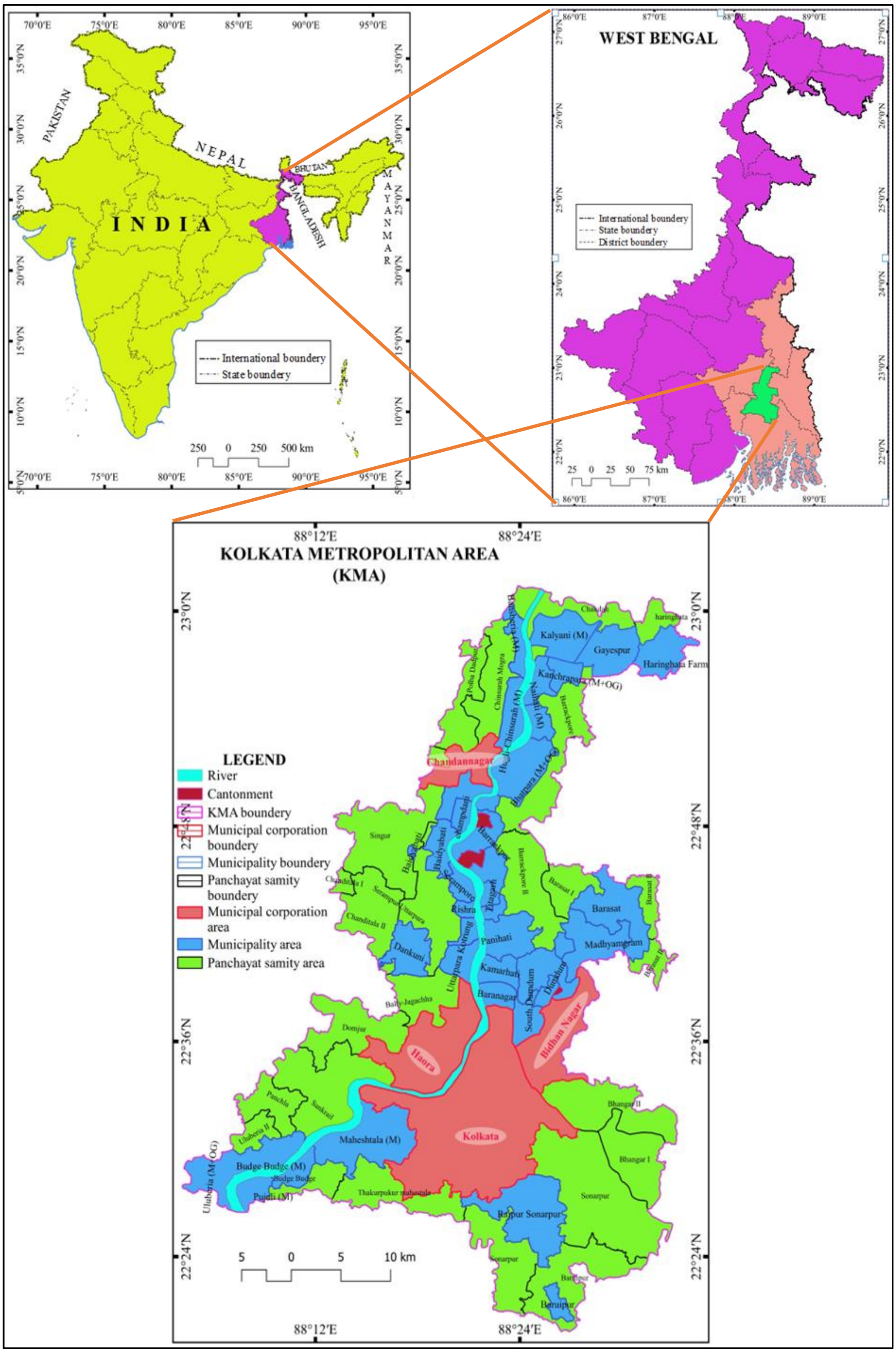
India, one of the most developing countries shares a very significant characteristics features of urbanization. In 2001 India’s total population was 1027 million. Out of this population 285 million population (27.81%) lived in urban area while 742 million population lived in rural area and in 2011 the urban population became 31.16 percent (Census of India 2011). This high population growth and related urbanization creates a challenge for the sustainable development of cities. So, it’s very important to understand the factors, dynamic spatiotemporal development of cities. The urbanization affects the surrounding valuable natural landscape such as wetland, open Space, green space. The conversion of impervious surface impacts on ecosystem, biological diversity, climate etc. which creates various negative affects like heat island (Xu, 2007). Hence, for better evaluation the urban development needs recent and clear elaborate information on multiple land use in an urban area. The provision of information for the dynamic, monitoring, and management of the earth depends heavily on the real-time accuracy of land use maps. In land use classification discipline, a fresh opportunity for more precise and extensive Land use mapping is created by time series feature extraction and machine learning approaches (Rosier et al., 2022). However there has not a complete knowledge about the distribution, pattern and composition of detailed land use type in urban India. In recent years more efforts has been done to map the individual cities using remote sensing datasets (Das et al., 2021; Kantakumar et al., 2019; Paul et al., 2021). But the remote sensing data with social sensing data, Point of Interest data, open street map have not been used to map urban land use map in Urban India.

A brief overview of the literature indicates that the fields of land use prediction, classification, and model development have made extensive use of machine learning approaches. Additionally, it compares the performance of several models in Lulc classification. Because of its durability, high accuracy, and transferability, Machine learning technology is becoming more popular in the land use classification. It introduces a novel approach for classifying land uses using remote sensing technology. The researcher did not use high resolution satellite images, socio-economic features, machine learning algorithms, and ensemble machine learning models for land use land cover classification. This is a substantial advancement in the field of remote sensing research and can be regarded as a state of the art. Therefore, the primary goal of this research is to evaluate the machine learning models and their ensemble model in land use and land cover classification and categorize the detailed land use and land cover map of the Kolkata metropolitan area.

The objective was taken in this study is to assess the feature selection approach of land use and land cover mapping in Kolkata metropolitan area.

2. Study Area:

Kolkata metropolitan area, commonly referred to as Greater Kolkata, is the country's third largest. There are 37 municipalities and four municipal corporations in this region, with Kolkata serving as the primary hub. The whole city area consists of the districts of Kolkata, Howrah, Hooghly, Nadia, and the north and south 24 Paragons. According to the *Kolkata Metropolitan Development Authority*, there are 14.11 million people living in this metropolitan area, which has a total area of 1886.67 km^2. The population density is 7480 persons per square kilometer. The urban area of Kolkata stretches from 88° 02´ E to 88° 32´ E in latitude and from 22° 19´ N to 23° 01´ N in longitudinal direction. Within the boundaries of the Kolkata Metropolitan Development Authority (KMDA), the 1851.41 square kilometer Kolkata Metropolitan Area is made up of three municipal corporations, forty-nine municipalities, and twenty-eight panchayat samity (Figure 1).



Source: Prepared by author

Figure 1. Study area map of Kolkata metropolitan area (KMA)

3. Datasets & Methodology:

The remote sensing datasets, socioeconomic datasets, night light datasets, and built-up height datasets have all been used to conduct the necessary urban land use land cover mapping in the Kolkata metropolitan area. The datasets were gathered from NOAA, Planet, GHSL layer, and cloud platform Google Earth Engine.

Table 1. Description of satellite data

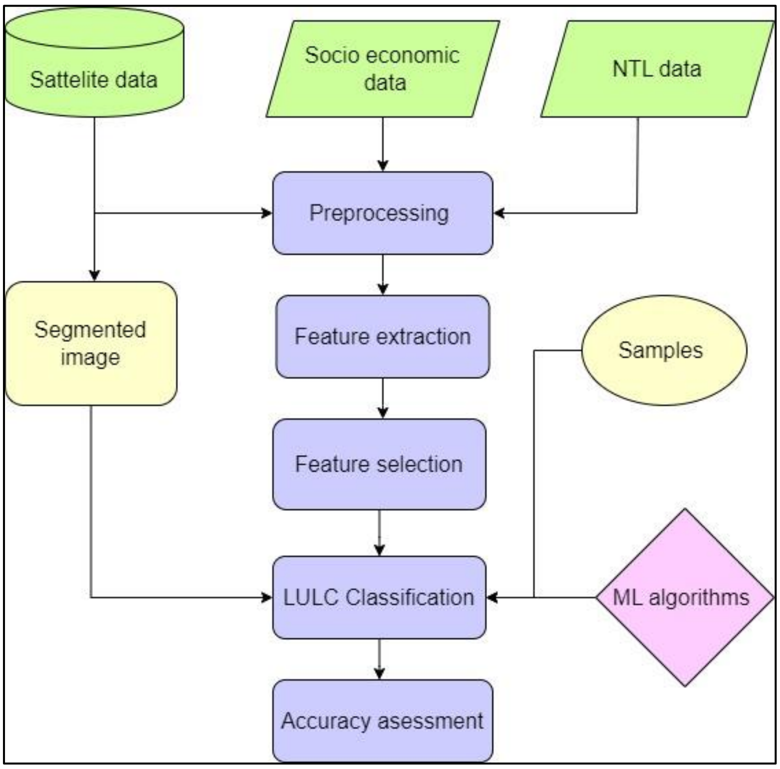
Bands	**Spectral bands**	**Spectral range**	**Resolution**		
			Spatial (meters)	**Temporal**	**Radiometric**
B1	Coastal Blue	431 – 452 nm	3	Daily	16 bits
B2	Blue	465 – 515 nm			
B3	Green I	513 – 549 nm			
B4	Green	547 – 583 nm			
B5	Yellow	600 – 620 nm			
B6	Red	650 – 680 nm			
B7	Red Edge	697 – 713 nm			
B8	NIR	845 – 885 nm			

Source: (Planet Labs PBC, 2023)

In this research study the entire methodological flowchart has been shown in figure 2. The figure shows the entire process to classify urban land use and land cover in KMA.

3.1 Feature extraction:

A survey of the literature on mapping urban land use and cover serves as the foundation for feature selection and conceptualization. The study utilized spectral, textural, and geometric satellite image features and other data for object-based and pixel-based image classification. Land cover mapping was accomplished by applying spectral and textural features in object-based image classification, or OBIA. Geometric and socioeconomic variables were used into OBIA to map land uses. The census features were used to classify the rural and urban built up in a pixel-based image classification. The features that were extracted were displayed on a graph.



Source: Prepared by author

Figure 2. Methodological framework

3.1.1 Spectral features:

Mean of spectral band: Extracting the mean bands from Planetscope bands allowed us to determine the mean spectral response of different urban land cover items. In object-based image analysis, the mean of the bands was used as a feature to categorize urban land cover.

NDVI and NDWI: Spectral features are superior to simple spectral band features in providing a contextualized understanding of an object's land cover classification (Gong et al., 1992; Tolentino & de Lourdes Bueno Trindade Galo, 2021). Equation 1 & 2 was utilized in this study to extract the Normalized Difference Vegetation Index (NDVI) and Normalized Difference Weighted Index (NDWI) from the satellite image.

$$NDVI = \frac{(Band4 - Band3)}{(Band4 + Band3)} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \quad (1)$$

$$NDWI = \frac{(Band5 - Band4)}{(Band5 + Band4)} \dots \dots \dots \dots \dots . \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots (2)$$

3.1.2 GLCM features:

An object's grey level cooccurrence matrix (GLCM) is a second order statistical textural feature (Zadeh et al., 2022). The possibility of discovering two different Gray levels in adjacent pixels was investigated via GLCM. It computes texture values across the cooccurrence matrix and evaluates the correlation between pairs of pixels (Haralick et al., 1973). For this investigation, eight GLCM texture features were calculated. At the object level, the GLCM characteristics mean, variance, homogeneity, contrast, dissimilarity, entropy, second moment, and correlation were obtained.

$$\text{GLCM-Mean (MEA)} = \sum_{ij=0}^{N-1} i\left(p_{ij}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\ldots\ldots.\ldots.(3)$$

$$\text{GLCM-Variance (VAR)} = \sum_{ij=0}^{N-1} p_{ij}(i - MEA)^2 \ldots\ldots\ldots\ldots\ldots..\ldots\ldots...(4)$$

$$\text{Homogeneity} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(P(i,j) / \quad 1 + (i-j)^2) \ldots\ldots\ldots\ldots\ldots\ldots\ldots..(5)$$

$$\text{Contrast} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(P(i,j) \quad (i-j)^2) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(6)$$

$$\text{Dissimilarity} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(P(i,j)| \quad (i-j)|) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(7)$$

$$\text{Entropy} = -\sum_{i=1}^{N_g}\sum_{J=1}^{N_g}\left(P(i,j) \quad log\left(P(i.j)\right)\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(8)$$

$$\text{Angular second moment} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(P(i,j)\right)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9)$$

$$\text{Correlation} = \sum_{i,j=0}^{N-1} P_{i,j} \quad \frac{(i-\mu_i) \quad (j-\mu_j)}{\sigma_i \quad \sigma_j} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(10)$$

3.1.3 GLDV features:

The total of the GLCM diagonals is the Grey Level Difference Vector (GLDV), which is used to determine the absolute difference between neighbors (Aguilar et al., 2012). This study used equation to generate three GLDV metrics (Laliberte & Rango, 2009).

$$GLDV_{mean} = \sum_{i,j=0}^{N-1} V_k - \quad N^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(11)$$

$$GLDV_{contrast} = \sum_{i,j=0}^{N-1} P_k \quad (i-j)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(12)$$

$$GLDV_{entropy} = \sum_{i,j=0}^{N-1} P_k \quad (-lnP_k) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(13)$$

3.1.4 Geometry features:

Geometric characteristics are utilized to group the things according to size and shape. Geometric characteristics were essentially used to classify land use. Regarding dimensions, form, and spatial arrangement, each urban land use is distinct (Sandborn & Engstrom, 2016).

3.1.5 Socio economic features:

Socioeconomic information was gathered in order to categorize the developed regions into rural and urban areas. Geometric mean was used to project socioeconomic data, including population, density, and non-agricultural worker data, in order to determine urban and rural built-up regions as defined by the census definition for 2021 (Lal, 2020). It was suggested that the population forecasting equation 14.

$$P_n = \quad P\left(1+\frac{I_G}{100}\right)^n \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (14)$$

Where, IG = Geometric mean (%); P = present population; n = Number of decades

3.1.6 Night light features:

Built-up area was extracted by means of the night light features. Every segment's night light characteristics were taken out and used to classify the land cover. Basically, the night light data collected from urban street lights, business lights, and residential lights (X. Hu et al., 2019). The metropolitan region was extracted using the night light attributes.

3.2 Feature selection:

Feature selection (FS) seeks to determine the smallest possible number of attributes needed to maintain the class probability distribution as close to the original distribution of all features as is practical (Hall & Holmes, 2003; Venkatesh & Anuradha, 2019). Pedergnana et al., (2013) state that feature selection is an essential stage in the classification process since it improves the process and lowers dimensionality by removing redundant data. Gain Ratio, information gain, and correlation are three FS filter-based techniques that were used in this study. The filter method works independently of the classifier. The correlation between the variable and classes is assessed using the correlation approach. It was anticipated by this filter approach which features were significantly corelated with the classes.

The gain ratio is an extension of information gain measures. It measures the gain ratio with regard to the target class in order to evaluate features. The gain ratio uses divided information to apply a sort of normalization to information gain (Tolentino & de Lourdes Bueno Trindade Galo, 2021). The gain ratio is defined as (equation 15)

$$\text{GainRatio (D)} = \frac{Gain\ (A)}{SplitInfo_A(D)} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots..\dots\dots\dots\dots (15)$$

According to Tolentino & de Lourdes Bueno Trindade Galo, (2021) information gain (equation 16) is the difference between the initial and updated knowledge requirements based on the percentage of classes.

$$\text{Gain (A)} = \text{Info(D)} - \text{Info}_A\text{(D)}\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots...(16)$$

3.3 Machine learning algorithms selection:

The classical machine learning algorithms Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Artificial Neural Network (ANN) and ensemble ML model were implemented in Object based image analysis based on the literature.

Support Vector Machine (SVM): SVM is a data-driven method for addressing the classification issue. A hyperplane or collection of hyperplanes is constructed by the SVM to classify all inputs in a high-dimensional space (Gove & Faytong, 2012). The SVM-RBF was utilized to tackle the multi-class classification problem. K-fold cross validation was used to modify the SVM's cost and gamma parameter in order to reduce overfitting and underfitting and enhance classification accuracy for the best possible model fitting (Persello & Bruzzone, 2014). The optimum parameter selection was achieved using SVM with linear, RBF, and polynomial kernels.

Random Forest (RF): An RF classifier is an ensemble classifier consisting of multiple decision trees that utilizes a randomly selected subset of the training set and variables (Belgiu & Drăgu, 2016). A statistical instrument called the margin function calculates the difference between the average number of votes cast for the correct class and the incorrect class.

$$mg(X,Y) = av_k \quad I\big(h_k \quad \big((x) = y\big) \quad -max_{j\neq Y} \quad av_k I(h_k \quad (x)) \quad -j\big) \quad \dots (17)$$

The development of a high-accuracy model involved parameter adjustment using the number of estimators and leaves. The ensemble model employs bagging to generate multiple predictors and aggregate their votes for a final decision. The bagging of random forest categorizes the number of decision trees from a specific subset of training data (Chowdhury, 2024).

Artificial Neural network (ANN): A supervised classifier machine learning system with multiple hidden layers is called an artificial neural network (ANN), sometimes known as a multilayer perception network (MLP) (Alshari et al., 2023). The backpropagation technique was used to generate the hidden layer and identify the underlying pattern in the data (Chowdhury, 2024). The feed forward neural network is the most efficient neural network for pattern recognition in the classification of remote sensing images (Iqbal & Aftab, 2019). The input and hidden layer were connected by a weight matrix. Whereas b1 is the bias, Z2 is the weighted sum of the same input and the hidden layer. So, the equation of the neuron is

$$Z^2 = W^1 .a^1 + b^1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots .. (18)$$

$$a^2 = g(Z^2) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \ldots\ldots\ldots (19)$$

Equation is used in ANN architecture to calculate weight. the hidden layer's output layer, which is produced by a different matrix and the necessary activation function using equation

$$Z^3 = W^2 .a^2 + b^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (20)$$

$$a^3 = g(Z^3) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots .. (21)$$

The loss function defines the difference between the input and the predicted value.

$L(x, a^3)$……………………………………………………..(22)

3.4 Performance based on quantitative indices:

The quantitative performance of classifier models was assessed using indices including the F1 score for precision, recall, and accuracy evaluation. The precision (equation 23) value represents a positive forecast that shows the amount of expected settlement locations that really occur. It aids in defining the model's dependability. The number of true settlement points that were correctly anticipated to be settlement points, or the actual positive value, is the recall (equation 24) value. For a forecast result to be considered reasonable, it must possess perfect precision and a recall value of one or 100%. F-score (equation 25) illustrates the predictive value accuracy. A thorough grasp of precision and recall can be obtained by utilizing the harmonic mean of precision. It is calculated using the weighted average of recall and precision (Li et al., 2011; Rudiastuti et al., 2022b).

$$\text{Precision} = \frac{\text{TP}}{\text{TF} + \text{FP}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (23)$$

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (24)$$

$$\text{F} - \text{score} = 2 \times \frac{Precision \times \text{Recall}}{\text{Precision} + \text{Recall}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (25)$$

Where, TP=true positive, TN= true negative, FP=false positive and FN =false negative and observed agreement = overall accuracy.

4. Results and Discussion:

The metropolitan area of Kolkata is mapped with regard to land cover and land use. These maps show the detailed types of land use and land cover in the Kolkata metropolitan area. The three distinct categories of urban built-up areas are residential, commercial, and industrial. It's really hard to tell them apart from the surrounding flora that's associated with other built-up regions because of their poor reflectance. The rural built-up region is not divided into subcategories since it is difficult to differentiate between it and the surrounding vegetation due to its sparsely population density (Huang et al., 2020).

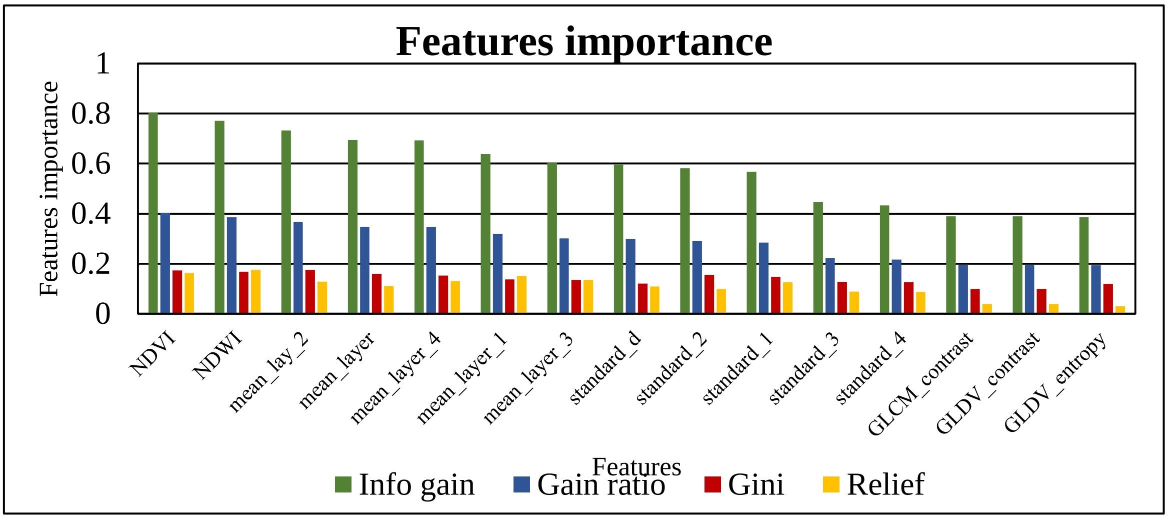
4.1 Extracted and normalized features:

Spectral, textural, and geometric aspects of the segmented images have been retrieved, enhancing the capacity to discriminate between various land cover and land use classifications. This is especially true for the land use category that distinguishes between built-up residential and commercial regions, resulting in a comprehensive classification of urban land use and land cover that yields acceptable outcomes. The optimal features for urban land cover and land use classification have been chosen using the combined and normalized features. Twenty-two features were extracted in total for the classification of land cover while ten features for urban land use based on the literature review. The retrieved characteristics were highly relevant in differentiating the land cover and land use groups.

4.2 Importance of feature categories and selection of best features:

4.2.1 Feature selection in Land cover classification:

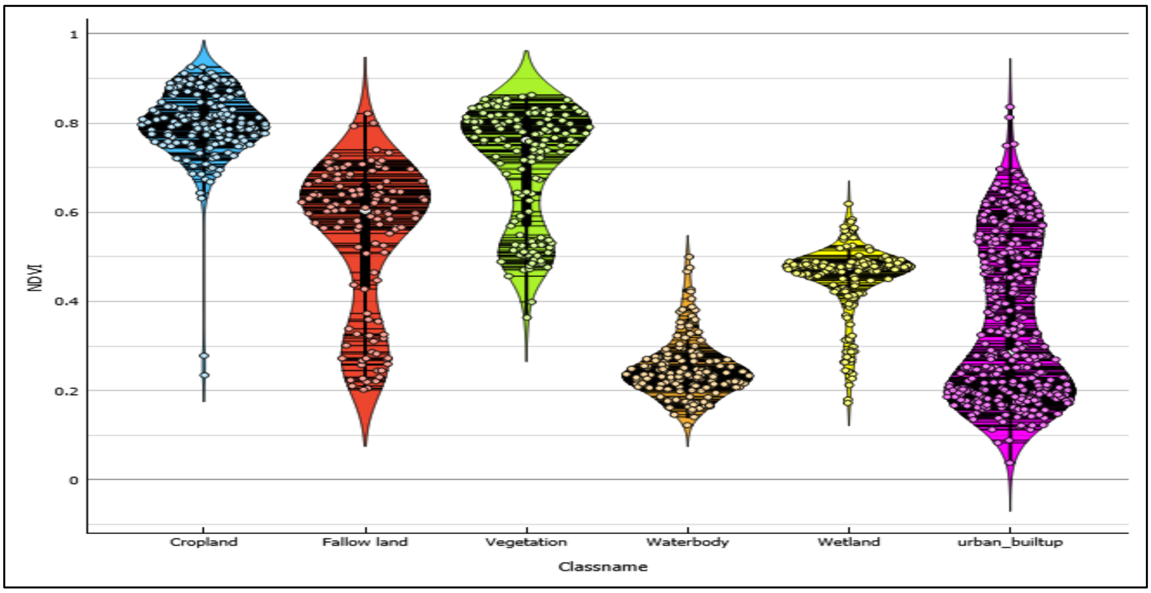
Each feature has an associated relevance value that is calculated by considering the matrix of the feature selection method. Several sets of traits are identified by the four-feature selection process as being essential for distinguishing between LULC classes. Each FS technique yields feature significance values for urban land cover that vary from 0.02 to 0.7. The qualities that appear to most successfully aid in class discrimination can be ranked using the FS technique. The classification of land cover is highly affected by spectral features like NDVI and NDWI. A violin graph has been made to represent the distribution range of the NDVI and NDWI index (figure 4 & 5). All metrics utilized in the quantification of importance consider each class's separability, and some qualities are also used to provide a more comprehensive picture of a single class (Tolentino & de Lourdes Bueno Trindade Galo, 2021). A feature's capacity to discriminate between the different classes in the dataset is assessed using the feature ranking technique (Chandrashekar & Sahin, 2014). The highest accuracy was achieved with the top 15 features according to the rank of four feature selection approach. The 15 features are the most suitable for the land cover categorization since the accuracy obtained by selecting 15 features exceeds the accuracy attained by selecting all features.



Source: Prepared by author

Figure 3. Feature importance for land cover classification

The features were ranked using the four features selection method. The relevance of the top 15 selected traits is shown in a bar graph (figure 3). The four methods for delineating the land cover are depicted in the figure, where the primary classification criteria were NDVI and NDWI.

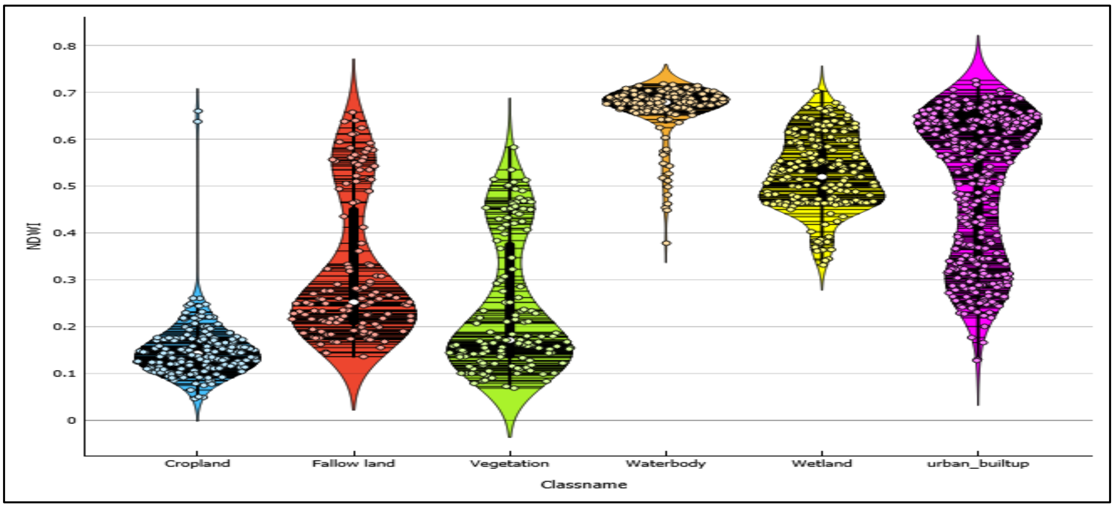


Source: Prepared by author

Figure 4. NDVI data distribution in land cover classification

According to the variable importance, the most important variables in the level 1 land cover categorization are the NDVI, NDWI, mean, and standard deviation of the image bands. It makes sense that the majority of the variations between urban land cover classes can be explained by the vegetation, water index, and spectral mean value of band that characterized the farmland, fallow land built up, and wetland land cover classes. Spectral and textural data collected from satellite-based observations through geospatial big data can be used to expose the distribution, pattern, and composite of urban land cover types through a multidimensional lens.

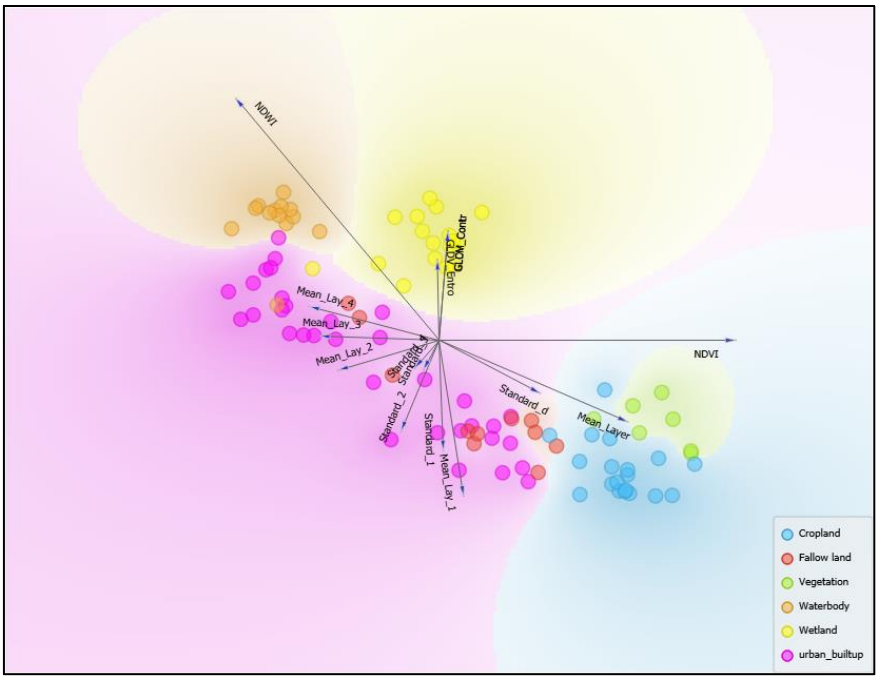
Feature selection has made use of Anova, Gini, gain ratio, and information gain. These methods make clear why specific standards are essential for classifying the urban landcover. The feature sensitivity or significance for classifying land cover has been presented to aid in the selection of the most noteworthy features and their importance.



Source: Prepared by author

Figure 5. NDWI sample distribution

The Free Viz diagram shown above is one such technique for analyzing the probable significance of the chosen features in order to ascertain how features and land cover classes interact (figure 6). The Free Viz optimization method states that the angle between the arrows in a graphic reflects the connection between the features, while the length of an arrow in the graphic signifies the magnitude of the characteristics (Nigam et al., 2021).



Source: Prepared by author

Figure 6. free Viz diagram to show the relationship of different features

The distribution statistics of the selected attributes were computed and plotted using a range of diagrams. Machine learning algorithms, including RF, SVM, DT, ANN, and ensemble model, have been used to classify land cover using the best selected features.

4.2.2 Experiment on training sample size and best parameter:

The significance of the training sample size lies in the fact that the Hughes phenomenon has an impact on the classifiers' accuracy, which makes it crucial. Various training sample sizes have been tested with the ML algorithm parameters. Table 2 shows the training sample size and accuracy assessment of object-based land cover and land use classification.

Table 2: Sample size and accuracy assessment of OBIA land cover and land use classification

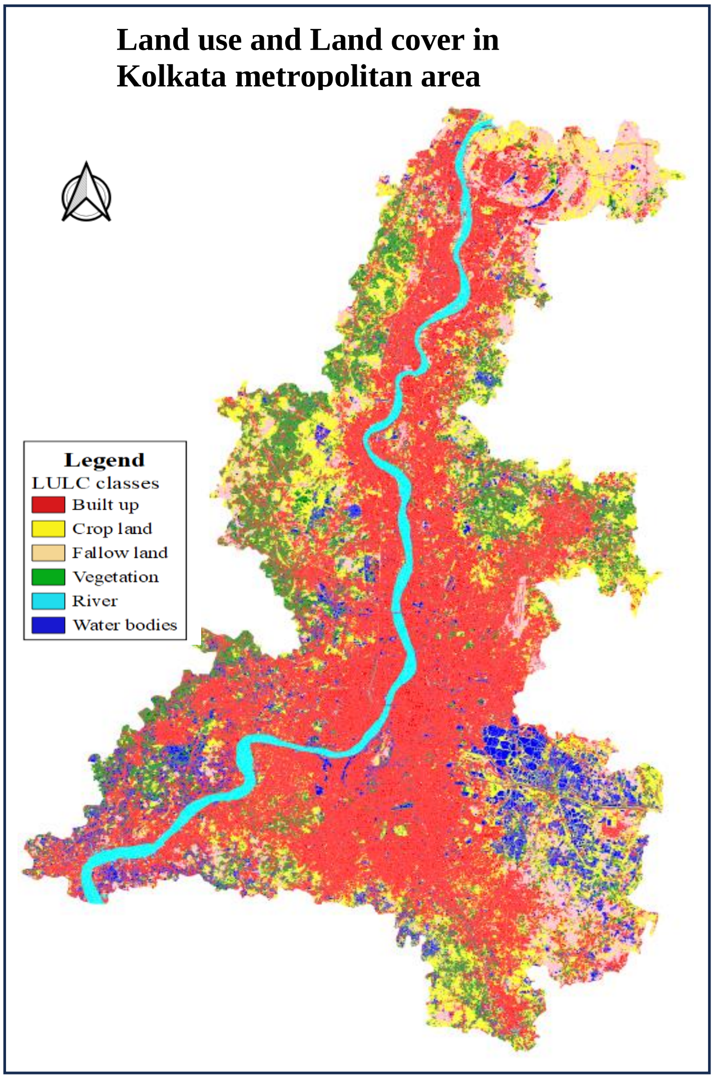
Size of training sample	Cross validation accuracy for urban land cover					Size of training sample	Cross validation accuracy for urban land use			
	SVM	RF	ANN	DT	Ensemble		SVM	RF	ANN	ensemble
470	0.91	0.85	0.90	0.82	0.89	320	0.89	0.82	0.89	0.88
586	0.89	0.91	0.89	0.84	0.90	400	0.89	0.88	0.92	0.89
704	0.91	0.92	0.88	0.90	0.89	478	0.88	0.89	0.91	0.89
821	0.91	0.91	0.90	0.87	0.91	558	0.88	0.88	0.90	0.90
938	0.93	0.93	0.89	0.89	0.91	638	0.89	0.87	0.93	0.89
1056	0.94	0.92	0.91	0.88	0.92	718	0.88	0.94	0.96	0.90

Source: Prepared by author

The experiment shows that the absence of the optimal training sample has a major effect on overall accuracy. Consequently, the result shows that by reducing the influence of the Hughes phenomenon, identifying land use and cover using training sample sizes in different ratios enhances overall accuracy (Mboga et al., 2017; Mustak, 2018).

4.2.3 OBIA based land cover classification:

Object-based image classification has been utilized for the classification of land cover using machine learning and ensemble model. Level 1 classification of the object-based land cover has been completed with a 94% overall accuracy and kappa coefficient of 0.96. The built-up land cover class has the highest producer accuracy (99%) and user accuracy (96%). In contrast, the vegetation has the lowest producer accuracy (84%), and user accuracy (86%). Above 80% producer and user accuracy in the land cover classifications indicates strong classification performance (table 2). The confusion matrix illustrates how easily built-up parcels and crops can be confused for fallow land, vegetation, and other land cover types (figure 8A). In the urban land cover class, vegetation had the lowest score for user accuracy, and producer accuracy is low on fallow land. The relatively low accuracy of the land cover classifications could have two possible causes. First off, there are many less examples of these land cover classifications than there are of the built-up class. Second, because the textural and spectral properties of the different land cover classifications are similar, machine learning algorithms have trouble differentiating between them. Figure 8B shows the train and cross validation accuracy of ensemble model in land cover classification. The PAs for agricultural and fallow land are significantly lower than the UAs, indicating that there are more omission errors than commission errors. These omission errors were caused by pixels of crop and fallow land being wrongly labeled as vegetation because to their similar spectral reflectance.



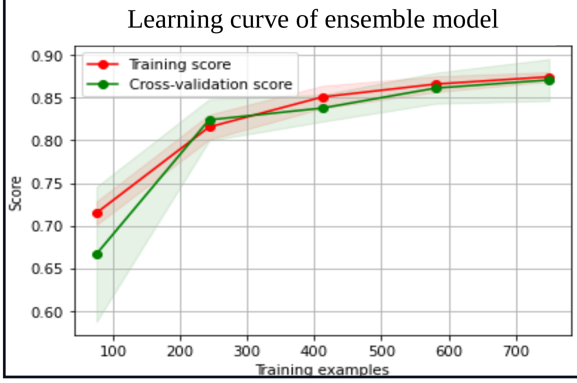
Source: Prepared by author

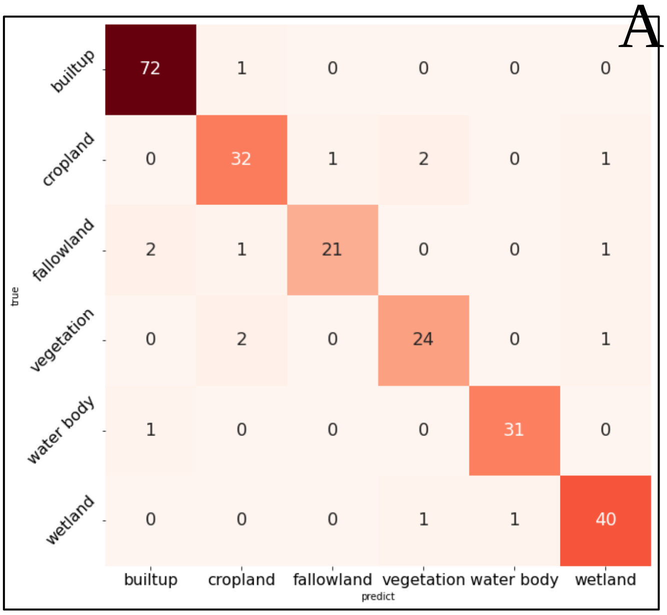
Figure 7. LULC map of KMA

Different approaches have been taken in object-based image classification with regard to sample size, recommended best features set, and parameters of corresponding machine learning algorithms. Additionally, the results show that whereas producer accuracy is high, user accuracy in built-up and wetland areas is poor. It makes clear that most land uses are wrongly classified as other land cover groups, but wetlands and built-up areas are less commonly misclassified as such. Overall findings suggest that the complexity of vegetation, fallow land, and crops is one of the causes of lower overall accuracy. The primary reason for the misinterpretation of vegetation and agriculture is the shared characteristics between built-up and fallow land. The most crucial factor to take into account when classifying urban land cover is the mix of classification algorithms and classification approach, since observations have shown to influence accuracy.

Table 2. user and producer accuracy of ensemble model in land cover classification

Land cover classes	User accuracy	Producer accuracy
Built up	0.96	0.99
Cropland	0.91	0.89
Fallow land	0.95	0.84
Vegetation	0.86	0.89
Waterbodies	0.97	0.94
Wet land	0.91	0.95





B

5. Conclusion and Recommendation:

Source: Prepared by author

Figure 8. A. Confusion matrix of ensemble model in land cover classification, B. learning curve of ensemble model describe the training accuracy and cross validation accuracy in land cover classification, table describe the produce r and user accuracy of each class

main procedures can be used to accomplish urban land use and land cover mapping: collecting data from multiple open sources, segmenting data at different resolutions, extracting features for the mapping of land use and cover, gathering samples, and experimenting with individual and group model mapping. The importance of data, the optimal feature set, classification strategies, and the training sample ratio in pixel- and object-based land use and cover classification are all systematically explained in this article. The approaches and findings may change as a result of mapping multi-scale important urban land use and land cover categories. The feature selection approach showed that while spectral and textural indices have a greater contribution to distinguishing the land cover classes, geometrical characteristics have been used to categorize land use classification based on feature significance. While the NDVI and NDWI are the most promising features in land covet classification, built-up height and compactness are the promising elements to establish the land use classes. The accuracy of the results showed that the object-based classification methodology offers a superior distinction of fundamental land use land cover categories, and that the multi resolution segmentation method is very effective in generating the segments required for the feature classification. OBIA offers strong classification performance and functions well with high resolution datasets. Object-based image classification is a useful technique for extracting specific information that makes it possible to classify land use and land cover. While some errors have been found in the residential, commercial, and industrial classes, the wet land class has been detected almost perfectly. The built-up class has been identified as the source of the biggest mistakes. Using open-source machine learning and scientific data processing makes it simple to experiment with different parameters and algorithms and determine which categorization method works best for a certain application. When it comes to classifying urban land use and land cover using high-dimensional feature sets, the multi-stacking ensemble model performs better than the individual model, despite the individual model performing well in the multi-model comparison.

References

Aguilar, M. A., Vicente, R., Aguilar, F. J., Fernández, A., & Saldaña, M. M. (n.d.). *Optimizing object-based classification in urban environments using very high resolution geoeye-1 imagery*.

Alshari, E. A., Abdulkareem, M. B., & Gawali, B. W. (2023). Classification of land use/land cover using artificial intelligence (ANN-RF). *Frontiers in Artificial Intelligence*, *5*, 964279. https://doi.org/10.3389/FRAI.2022.964279/BIBTEX

Banzhaf, E., Kabisch, S., Knapp, S., Rink, D., Wolff, M., & Kindler, A. (2017). Integrated research on land-use changes in the face of urban transformations – An analytic framework for further studies. *Land Use Policy*, *60*, 403–407. https://doi.org/10.1016/j.landusepol.2016.11.012

Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/J.ISPRSJPRS.2016.01.011

Census of India 2011. (n.d.).

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/J.COMPELECENG.2013.11.024

Chen, B., Tu, Y., Song, Y., Theobald, D. M., Zhang, T., Ren, Z., Li, X., Yang, J., Wang, J., Wang, X., Gong, P., Bai, Y., & Xu, B. (2021). Mapping essential urban land use categories with open big data: Results for five metropolitan areas in the United States of America. *ISPRS Journal of Photogrammetry and Remote Sensing*, *178*, 203–218. https://doi.org/10.1016/J.ISPRSJPRS.2021.06.010

Chen, J., Zhu, L., Fan, P., Tian, L., & Lafortezza, R. (2016). Do green spaces affect the spatiotemporal changes of PM2.5 in Nanjing? *Ecological Processes*, *5*(1), 1–13. https://doi.org/10.1186/S13717-016-0052-6/TABLES/4

Chowdhury, M. S. (2024). Comparison of accuracy and reliability of random forest, support vector machine, artificial neural network and maximum likelihood method in land use/cover classification of urban setting. *Environmental Challenges*, *14*(November 2023), 100800. https://doi.org/10.1016/j.envc.2023.100800

Clinton, N., Stuhlmacher, M., Miles, A., Uludere Aragon, N., Wagner, M., Georgescu, M., Herwig, C., & Gong, P. (2018). A Global Geospatial Ecosystem Services Estimate of Urban Agriculture. *Earth's Future*, *6*(1), 40–60. https://doi.org/10.1002/2017EF000536

Das, T., Jana, A., Mandal, B., & Sutradhar, A. (2021). S patio-temporal pattern of land use and land cover and its effects on land surface temperature using remote sensing and GIS techniques: a case study of Bhubaneswar city, Eastern India (1991–2021). *GeoJournal 2021 87:4*, *87*(4), 765–795. https://doi.org/10.1007/S10708-021-10541-Z

Fan, P., Ouyang, Z., Duong Nguyen, D., Thuy, T., Nguyen, H., Park, H., & Chen, J. (2018). *Urbanization, economic development, environmental and social changes in transitional economies: Vietnam after Doimoi*. https://doi.org/10.1016/j.landurbplan.2018.10.014

Gao, J., & O'Neill, B. C. (2020). Mapping global urban land for the 21st century with data-driven simulations and Shared Socioeconomic Pathways. *Nature Communications 2020 11:1*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-15788-7

Gong, P., Chen, B., Li, X., Liu, H., Wang, J., Bai, Y., Chen, J., Chen, X., Fang, L., Feng, S., Feng, Y., Gong, Y., Gu, H., Huang, H., Huang, X., Jiao, H., Kang, Y., Lei, G., Li, A., … Xu, B. (2020). Mapping essential urban land use categories in China (EULUC-China): preliminary results for 2018. *Science Bulletin*, *65*(3), 182–187. https://doi.org/10.1016/J.SCIB.2019.12.007

Gong, P., Marceau, D. J., & Howarth, P. J. (1992). A comparison of spatial feature extraction algorithms for land-use classification with SPOT HRV data. *Remote Sensing of Environment*, *40*(2), 137–151. https://doi.org/10.1016/0034-4257(92)90011-8

Gove, R., & Faytong, J. (2012). Machine Learning and Event-Based Software Testing: Classifiers for Identifying Infeasible GUI Event Sequences. *Advances in Computers*, *86*, 109–135. https://doi.org/10.1016/B978-0-12-396535-6.00004-1

Grimm, N. B., Faeth, S. H., Golubiewski, N. E., Redman, C. L., Wu, J., Bai, X., & Briggs, J. M. (2008). Global change and the ecology of cities. *Science*, *319*(5864), 756–760. https://doi.org/10.1126/SCIENCE.1150195/SUPPL_FILE/GRIMM.SOM.REV.PDF

Hall, M. A., & Holmes, G. (2003). Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE Transactions on Knowledge and Data Engineering*, *15*(6), 1437–1447. https://doi.org/10.1109/TKDE.2003.1245283

Haralick, R. M., Dinstein, I., & Shanmugam, K. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-3*(6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314

Henderson, V. (2003). The urbanization process and economic growth: The so-what question. *Journal of Economic Growth*, *8*(1), 47–71. https://doi.org/10.1023/A:1022860800744

Hu, T., Yang, J., Li, X., Gong, P., He, Y., Weng, Q., Koch, M., & Thenkabail, P. S. (2016). Mapping Urban Land Use by Using Landsat Images and Open Social Data. *Remote Sensing 2016, Vol. 8, Page 151*, *8*(2), 151. https://doi.org/10.3390/RS8020151

Hu, X., Qian, Y., Pickett, T. A., & Zhou, W. (2019). *Urban mapping needs up-to-date approaches to provide diverse perspectives of current urbanization: A novel attempt to map urban areas with nighttime light data*. https://doi.org/10.1016/j.landurbplan.2019.103709

Huang, X., Wang, Y., Li, J., Chang, X., Cao, Y., Xie, J., & Gong, J. (2020). High-resolution urban land-cover mapping and landscape analysis of the 42 major cities in China using ZY-3 satellite images. *Science Bulletin*, *65*(12), 1039–1048. https://doi.org/10.1016/j.scib.2020.03.003

Iqbal, A., & Aftab, S. (2019). Computer Network and Information Security. *Computer Network and Information Security*, *4*, 19–25. https://doi.org/10.5815/ijcnis.2019.04.03

Kantakumar, L. N., Kumar, S., & Schneider, K. (2019). SUSM: a scenario-based urban growth simulation model using remote sensing data. *European Journal of Remote Sensing*, *52*(sup2), 26–41. https://doi.org/10.1080/22797254.2019.1585209/SUPPL_FILE/TEJR_A_1585209_SM7916.DOC

Kolkata Metropolitan Development Authority. (n.d.). Retrieved September 29, 2022, from https://kmda.wb.gov.in/

Lal, S. (2020). Mathematical modeling of Population forecasting in India. *Journal of Interdisciplinary Cycle Research*, *12*(1), 275–283.

Laliberte, A. S., & Rango, A. (2009). Texture and scale in object-based analysis of subdecimeter resolution unmanned aerial vehicle (UAV) imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(3), 761–770. https://doi.org/10.1109/TGRS.2008.2009355

Li, N., Sepúlveda, N., & Li, N. (2011). IEEE Xplore Full-Text PDF_李玲琪 自聚焦 写波导. *Proceedings of the 2011 IEEE International Conference on Robotics and Biomimetics*, 1343–1348. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7781053

Liu, Z., He, C., Zhou, Y., & Wu, J. (2014). How much of the world's land has been urbanized, really? A hierarchical framework for avoiding confusion. *Landscape Ecology*, *29*(5), 763–771. https://doi.org/10.1007/S10980-014-0034-Y

Lu, J., Li, B., Li, H., & Al-Barakani, A. (2021). Expansion of city scale, traffic modes, traffic congestion, and air pollution. *Cities*, *108*, 102974. https://doi.org/10.1016/J.CITIES.2020.102974

Mboga, N., Persello, C., Bergado, J. R., & Stein, A. (2017). Detection of Informal Settlements from VHR Images Using Convolutional Neural Networks. *Remote Sensing 2017, Vol. 9, Page 1106*, *9*(11), 1106. https://doi.org/10.3390/RS9111106

Mustak, S. (2018). *Evaluating the performance of machine learning algorithms for urban land use mapping using very high resolution. April*.

Nation., U. (2015). THE 17 GOALS | Sustainable Development. In *Sustainable Development*. https://sdgs.un.org/goals#goals%0Ahttps://sdgs.un.org/goals

Nigam, A. K., Ojha, A. A., Li, J. G., Shi, D., Bhatnagar, V., Nigam, K. B., Abagyan, R., & Nigam, S. K. (2021). Molecular properties of drugs handled by kidney oats and liver oatps revealed by chemoinformatics and machine learning: Implications for kidney and liver disease. *Pharmaceutics*, *13*(10), 1–16. https://doi.org/10.3390/pharmaceutics13101720

Patino, J. E., & Duque, J. C. (2013). A review of regional science applications of satellite remote sensing in urban settings. *Computers, Environment and Urban Systems*, *37*(1), 1–17. https://doi.org/10.1016/J.COMPENVURBSYS.2012.06.003

Paul, S., Saxena, K. G., Nagendra, H., & Lele, N. (2021). Tracing land use and land cover change in peri-urban Delhi, India, over 1973–2017 period. *Environmental Monitoring and Assessment*, *193*(2), 1–12. https://doi.org/10.1007/S10661-020-08841-X/TABLES/5

Pedergnana, M., Marpu, P. R., Mura, M. D., Benediktsson, J. A., & Bruzzone, L. (2013). A novel technique for optimal feature selection in attribute profiles based on genetic algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(6), 3514–3528. https://doi.org/10.1109/TGRS.2012.2224874

Persello, C., & Bruzzone, L. (2014). Active and semisupervised learning for the classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(11), 6937–6956. https://doi.org/10.1109/TGRS.2014.2305805

Planet Labs PBC. (2023). *PlanetScope Product Specifications*. https://assets.planet.com/docs/Planet_PSScene_Imagery_Product_Spec_letter_screen.pdf

Riggan Jr., N.D. and Weih Jr., R.C. (2009) A Comparison of Pixel-Based versus Object-Based Land Use/Land Cover Classification Methodologies. Journal of the Arkansas Academy of Science, 63, 145-152. - References - Scientific Research Publishing. (n.d.). Retrieved January 12, 2023, from https://www.scirp.org/reference/ReferencesPapers.aspx?ReferenceID=1526910

Rosier, J. F., Taubenböck, H., Verburg, P. H., & van Vliet, J. (2022). Fusing Earth observation and socioeconomic data to increase the transferability of large-scale urban land use classification. *Remote Sensing of Environment*, *278*(May). https://doi.org/10.1016/j.rse.2022.113076

Rudiastuti, A. W., Lumban-Gaol, Y., Silalahi, F. E. S., Prihanto, Y., & Pranowo, W. S. (2022). Implementing Random Forest Algorithm in GEE: Separation and Transferability on Built-Up Area in Central Java, Indonesia. *JOIV : International Journal on Informatics Visualization*, *6*(1), 74–82. https://doi.org/10.30630/JOIV.6.1.873

Sandborn, A., & Engstrom, R. N. (2016). Determining the Relationship between Census Data and Spatial Features Derived from High-Resolution Imagery in Accra, Ghana. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(5), 1970–1977. https://doi.org/10.1109/JSTARS.2016.2519843

Schulz, D., Yin, H., Tischbein, B., Verleysdonk, S., Adamou, R., & Kumar, N. (2021). Land use mapping using Sentinel-1 and Sentinel-2 time series in a heterogeneous landscape in Niger, Sahel. *ISPRS Journal of Photogrammetry and Remote Sensing*, *178*(March), 97–111. https://doi.org/10.1016/j.isprsjprs.2021.06.005

Seto, K. C., Fragkias, M., Güneralp, B., & Reilly, M. K. (2011). A Meta-Analysis of Global Urban Land Expansion. *PLOS ONE*, *6*(8), e23777. https://doi.org/10.1371/JOURNAL.PONE.0023777

Seto, K. C., & Pandey, B. (2019). Urban Land Use: Central to Building a Sustainable Future. *One Earth*, *1*(2), 168–170. https://doi.org/10.1016/J.ONEEAR.2019.10.002

Seto, K. C., & Shepherd, J. M. (2009). Global urban land-use trends and climate impacts. *Current Opinion in Environmental Sustainability*, *1*(1), 89–95. https://doi.org/10.1016/J.COSUST.2009.07.012

Tolentino, F. M., & de Lourdes Bueno Trindade Galo, M. (2021). Selecting features for LULC simultaneous classification of ambiguous classes by artificial neural network. *Remote Sensing Applications: Society and Environment*, *24*, 100616. https://doi.org/10.1016/J.RSASE.2021.100616

UN-DESA. (2014). 2014 revision of the World Urbanization Prospects. *World Urbanization Prospects, July 2014*, 1–26.

Use, A. L., & Anderson, T. (2017). *Land Use / Land Cover Mapping*. 1–5.

Venkatesh, B., & Anuradha, J. (2019). A Review of Feature Selection and Its Methods. *BULGARIAN ACADEMY OF SCIENCES CYBERNETICS AND INFORMATION TECHNOLOGIES □*, *19*(1). https://doi.org/10.2478/cait-2019-0001

Wang, H., Gong, X., Wang, B., Deng, C., & Cao, Q. (2021). Urban development analysis using built-up area maps based on multiple high-resolution satellite data. *International Journal of Applied Earth Observation and Geoinformation*, *103*, 102500. https://doi.org/10.1016/J.JAG.2021.102500

Xu, H. (2007). Extraction of urban built-up land features from landsat imagery using a thematic-oriented index combination technique. *Photogrammetric Engineering and Remote Sensing*, *73*(12), 1381–1391. https://doi.org/10.14358/PERS.73.12.1381

Zadeh, F. A., Ardalani, M. V., Salehi, A. R., Farahani, R. J., Hashemi, M., & Mohammed, A. H. (2022). *An Analysis of New Feature Extraction Methods Based on Machine Learning Methods for Classification Radiological Images*. https://doi.org/10.1155/2022/3035426

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., & Atkinson, P. M. (2018). An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sensing of Environment*, *216*, 57–70. https://doi.org/10.1016/J.RSE.2018.06.034

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., & Atkinson, P. M. (2019). Joint Deep Learning for land cover and land use classification. *Remote Sensing of Environment*, *221*, 173–187. https://doi.org/10.1016/J.RSE.2018.11.014

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Sk. Mustak, Department of Geography, Central University of Punjab, for his constant support, invaluable guidance, and encouragement throughout the preparation of this research paper. His insights and expertise have been instrumental in shaping the direction of my work. I would also like to extend my heartfelt thanks to the Central University of Punjab and the University Grants Commission (UGC) for providing the financial assistance that made this research possible. Finally, I am grateful to my colleagues, family, and friends for their unwavering support and encouragement throughout this journey.