Predicting Urban Land Cover Using Classification: A Machine Learning Approach

Tanush Jadhav
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed University), Pune, India
ORCID - 0009-0001-3295-3311

Tanishq
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed University), Pune, India
ORCID - 0009-0004-0604-8379

Spoorthi Jagadish
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed University), Pune, India
ORCID - 0009-0001-0247-4344

Mayur Gaikwad
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed University), Pune, India
mayur.gaikwad@sitpune.edu.in

Shivali Wagle
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed University), Pune, India
shivali.wagle@sitpune.edu.in

Ruchi Jayaswal
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed University), Pune, India
ruchi.jayaswal@sitpune.edu.in

Shruti Patil
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed University), Pune, India
headaiml@sitpune.edu.in

2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC) | 979-8-3503-2614-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/R10-HTC57504.2023.10461930

***Abstract*—The classification of urban land cover is a crucial step in comprehending the evolution of the urban environment and its effects. We are providing a comparative analysis of machine learning methods for classifying urban land cover using remote sensing data in this research. The study utilises the Urban Land Cover dataset which is obtained from the UCI Machine Learning Repository, which contains high-resolution images taken of urban areas. Also, comparison of several popular machine learning classification algorithms, such as Decision Tree classifier, Random Forest classifier, Support Vector Machine classifier (SVM/SVC), XGBoost classifier, K-Nearest Neighbors classifier (KNN) and Ridge classifier is done along with their accuracy scores. This comparison shows that the Random Forest algorithm outperforms the other machine learning algorithms with an overall accuracy of 91.38% after removing the outliers and using Grid Search CV to tune the hyper-parameters.**

***Index Terms*—Urban Land Cover Classification, Machine Learning Algorithms, Remote Sensing Data, Comparative Analysis, MAUP, OBIA, Random Forest.**

I. Introduction

Urban land cover classification plays a crucial role in understanding the advancement of urban environments and their impact on the ecosystem, resource management and urban planning. Remote sensing data, obtained from satellite or other aerial platforms, provides valuable information about the earth's surface for a detailed analysis, monitoring and detection [1]. It helps for accurately mapping and monitoring urban land cover.

Urban land cover information is a crucial resource for urban management and planning. It can be derived from high-resolution aerial or satellite imagery and has various applications, including mapping green space and impervious surfaces, and also updating building footprint.Geographic Information System (GIS) data, this GIS data has been widely used for the analysis of land-based sustainability [9]. However, extracting accurate land cover information from high-resolution data is quite a complex task. The high degree of spectral variability within land cover classes, caused by factors such as angle of the sun, gaps in tree canopies and shadows, can significantly reduce the accuracy of traditional pixel-based image classification methods [5]. This happens because of the modifiable areal unit problem (MAUP), which causes a mismatch between pixels and real-world objects of interest [4].

This paper aims to explore the challenges and techniques involved in extracting urban land cover information from high-resolution data, while focusing on addressing the MAUP. This paper will also be investigating the potential of new techniques such as object-based image analysis (OBIA) and machine learning algorithms in improving the classification accuracy of the model [4]. The extraction of land cover information from

979-8-3503-2614-7/23/$31.00 ©2023 IEEE

remote sensing images can be quite a challenging task due to the arbitrary size of pixels, which makes it difficult to correspond them with real-world objects [2]. To tackle this problem, the approach of Geospatial object-based image analysis has been used in several studies. The OBIA approach segments the image into homogeneous regions before classification, and the attributes of these segments are used for classification instead of using the attributes of single pixels. This approach can help reduce within-class spectral variability, it also incorporates spatial and contextual information, and reduces the sensitivity of classification to the modifiable areal unit problem (MAUP).

For this research, the aim is to develop a machine learning approach for the prediction of urban land cover using the Urban Land Cover dataset. We will also be exploring the various machine learning algorithms used for classification. These algorithms include Decision Tree classifier, Random Forest classifier, XGBoost classifier, Ridge classifier, Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) classifier which will help us to identify the most accurate and efficient method that can be used for urban land cover classification [5].

The goal of our research is to develop an accurate land cover classification model which can then be deployed in practical applications, such as resource management, environmental monitoring and urban planning. The results of this study will add to the body of knowledge in the field of urban land cover classification using remote sensing and machine learning, and they may also have applications in decision-making, sustainability, and urban management.

A. Motivation

A global phenomenon that has significant implications for the environment, society, and the economy is called Urbanization. Analysing and understanding urban land cover patterns becomes increasingly important for urban planning, environmental monitoring, and resource management while cities continue to grow and expand. Remote sensing data such as aerial or satellite imagery, provides a well-supplied source of information for studying urban land cover. [1] A comprehensive and diverse collection of urban land cover data is offered by the Urban land cover dataset.

The motivation for conducting a research study on this topic was to better understand and analyse the dynamic and complex nature of urban landscapes. The Clustering algorithms, a type of unsupervised machine learning techniques, provides a promising approach for identifying patterns within the urban land cover data without relying on predefined class labels. We can potentially uncover hidden patterns and structures, identify similar land cover types, and gain insights into the spatial distribution and dynamics of different types of land cover found in urban areas by applying clustering algorithms to the dataset.

II. Description of Dataset

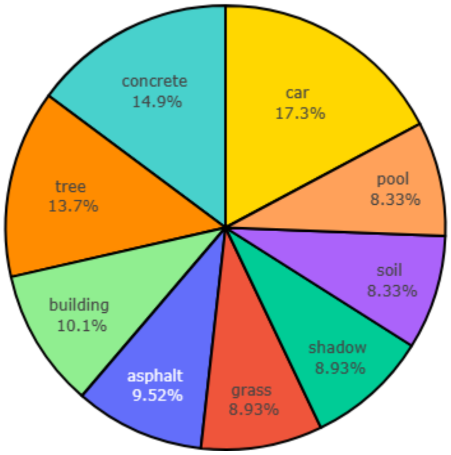

Fig. 1. Distribution of Classes

The Urban land cover dataset used for our research has been obtained from the UCI repository, which consists many publicly available datasets. This dataset contains multi-spectral imagery taken from the Sentinel-2 satellite, that spans a spatial extent of approximately 140 square kilometres. Sentinel-2 data for these types of classification is extremely useful, since the data from sentinel-2 can help in achieving high accuracy as it is provides high spatial resolution [11]. The spatial resolution of the image is 10 meters, allowing fine grained and detailed analysis of urban land cover patterns. Classes such as concrete, pool, soil, tree, shadows, buildings, cars, grass and asphalt are the land cover labels included in the dataset which represent common urban features that are the basic needs in urban planning, environmental monitoring, and resource management. A total of 148 features, which basically is 21 features taken over 7 different land sizes is including in the dataset. Spectral Bands, indices, and other derived values were the features included that provide a rich source of information for land cover classification. Each Image in the dataset is a 256x256 RGB image, where RGB values represent the colour intensities of the red, green, and blue channels. This dataset has been created by manually annotating each image, to label the distinguishable land cover types present in the image. These Annotation have been performed by experts in the field. Multiple annotators annotated each of the images to ensure consistency and accuracy of all the labels. This dataset is relevant for research studies which focuses on machine learning, remote sensing, and urban planning providing a rich and diverse source of information for evaluating and developing land cover classification models, examining different machine learning algorithms, and investigating pre-processing techniques for optimizing model performance. This dataset can also be used for studying urban environmental impacts, urban sustainability, and urban management planning. The dataset has been extensively used for urban land cover mapping tasks which includes developing and evaluating image classification algorithms. And this dataset is specifically useful for evaluating the performance of algorithms in challenging scenarios, where multiple classes are present in the same image and where the classes are visually identical to each other.

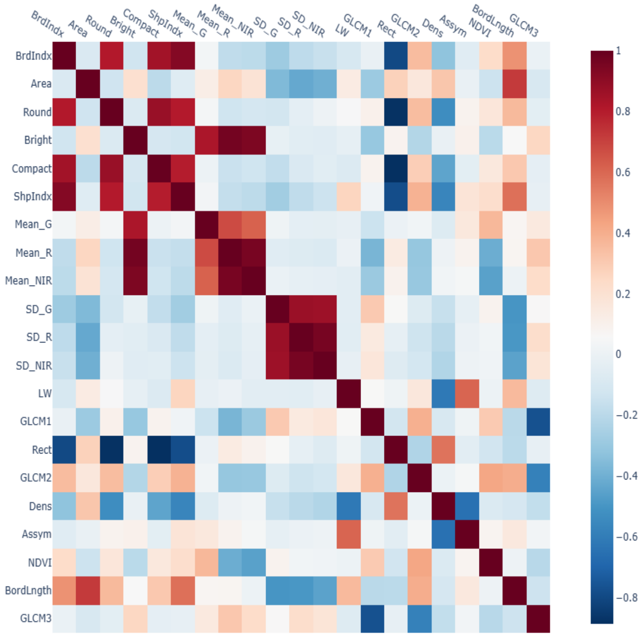

Fig. 2. Correlation between Features

III. Problem Formulation

The problem that is being addressed in this research paper is about using machine learning techniques for accurate classification of urban land cover. Urban areas are characterized by complex and dynamic land cover patterns, which includes different types of natural and built-up features like concrete, pool, soil, tree, shadows, buildings, cars, grass and asphalt. It is crucial for these land cover types to have Accurate and timely mapping for environmental monitoring, urban planning, and resource management. Nonetheless, traditional methods for land cover classification often suffer from limitations such as low accuracy, labour-intensive manual analysis, and inability to capture temporal changes.

IV. Methodology

A. *Data Preprocessing*

First, the dataset was checked to see if there were any null values present, but none were present. Then the statistical summary of the dataset was done and necessary visualizations were done to visualize and understand the dataset. After visualizing our dataset, we proceeded to detect and remove outliers present in our data. This was done by calculating the first (Q1) and third (Q3) quartile, to derive the IQR (InterQuartile Range). In this case, the threshold value was selected as 1.5 times the obtained IQR. Any value that lies below the difference of Q1 and IQR or the sum of Q3 and IQR is deemed to be an outlier.

$$Q1 = [(n+1)/4] \tag{1}$$

$$Q3 = [3(n+1)/4] \tag{2}$$

$$IQR = Q3 – Q1 \tag{3}$$

B. *Feature selection*

As mentioned in the description of the dataset, the dataset consists of 21 main features that have been taken for 7 different image scales (i.e. 20m, 40m, 60m, 80m, 100m, 120m and 140m). So, for the classification model different sets of 21 features have been selected, so as to capture as much information as possible. We have taken the last 21 features of the dataset, which are the features of the image scale of 140m, since the larger the size the more information it can collect for a given class. Also, the average values for each of the 21 features were taken for the 7 different image scales, as it gives an idea of what the center value for each of the 21 features are.

C. *Classification algorithm selection*

Classification is a predictive modeling problem in Supervised Machine Learning that involves categorizing or labeling input data into predefined classes. It is important for the following reasons:

- Prediction and decision making
- Pattern Recognition
- Feature Selection
- Data Organization and Summarizing

Below are the classifiers used in our model to find the best fitting classification algorithm:

- Decision Tree (DT) classifier [14]- This algorithm follows a tree like structure which consists of a root node which branches into decision nodes(sub-nodes) which has leaf nodes which consists of the final outcome of the tree. It follows a top-down recursive divide and conquer method. There are two types of attribute selection methods for selecting the best attributes for the root and decision nodes:
  a) Information Gain (IG) - it is the measure of change in entropy of an attribute.
  b) Gini Index - It is the measure of impurity in a dataset.
- XGBoost classifier [3]- It is an optimised distributed gradient boosting library which is known for scalable machine learning training. It enhances predictions by combining weak models.It efficiently manages missing values without requiring significant pre-processing, which makes it popular for handling huge datasets,in tasks like classification and regression.
- Random Forest (RF) classifier [15] - This algorithm is a type of bootstrap aggregating ensemble method.
- K-Nearest neighbors (KNN) classifier [13] - This is one of the simplest algorithms in machine learning classification and it is robust to noise in the data. It is a distance based classifier. It is non parametric and is a lazy learner, which means that it does not make any assumptions on the data provided and it stores the training data without learning anything from it and it starts classifying once it receives the testing data. But there are a few disadvantages of using this algorithm, such as, it requires

high computational power, it does not handle outliers and has difficulties when the data consists of missing values.

- Support Vector Classifier (SVM) [6] - Support Vector Machines are powerful supervised ML tools used for regression and classification. The objective of SVM is to find an optimal hyperplane in N-dimensional space,maximising the margin between various class data points. SVMs excel in tasks like image/text classification,anomaly detection,utilising nonlinear and high dimensional data effectively.
- Ridge Classifier [10]- This algorithm can be used for both multi-class and binary classification problems. It works by adding a penalty term to the cost function. The penalty term usually is the sum of the squares of coefficients of the features, this ensures that the coefficients are small, which helps in avoiding over-fitting. The loss function for this types of classification is the mean square of loss between predicted and actual values along with the L2 penalty term.

TABLE I
Accuracy Score of the Classification Algorithms

Classification Algorithms Used	Before Dropping Outliers	Last 21 features without dropping outliers	After Dropping Outliers	Last 21 features after dropping outliers	Using Average Values
Decision Tree	74.95%	58.18%	68.63%	57.59%	71.40%
Random Forest	79.09%	61.14%	73.76%	59.76%	70.41%
XGBoost	80.27%	67.74%	74.75%	64.50%	73.77%
KNN	39.50%	29.98%	29.58%	25.44%	32.14%
SVM (SVC)	56.60%	61.14%	48.71%	52.47%	68.63%
Ridge	60.94%	62.91%	48.71%	52.27%	65.88%
Random Forest with Grid Search CV	85.71%	77.38%	91.38%	83.81%	85.61%

V. Result

The study produced observations as referred in table 1. Table 1 contains respective accuracy and comprehensive exploration of various classification algorithms, each algorithm bringing distinct attributes to the forefront. A meticulous evaluation of these classifiers was performed across differing feature selections, thereby unveiling unique accuracy scores associated with each of them.

Ultimately, the study concludes that the Random Forest classification algorithm emerges as the most effective choice for the model. Characterized by its ensemble learning technique and amalgamation of multiple decision trees, Random Forest leverages the majority vote of these trees for accurate class prediction. To optimize its performance, the study employs Grid Search CV [12].

Grid-Search CV, which stands for Grid Search Cross-Validation, is a crucial technique in machine learning for systematically fine-tuning hyper-parameters of a model to achieve optimal performance. It helps in finding the best hyper-parameters for the machine learning model and facilitates the enhancement of the model's fit to the dataset. Paper [8] gives an overview of using grid search cv for tuning the hyper parameters of random forest classifier on Sentinel-2 data which plays a pivotal role in attaining heightened accuracy of this research.

VI. Conclusion

In this paper, different machine learning algorithms have been used for urban land cover classification. A number of different machine learning algorithms have been evaluated and it is found that Random Forest has achieved the best performance. Also, the use of different feature selection methods have been explored and it was found that the best results are obtained by using all the features and by not dropping any since every feature in the dataset gives a very important insight into the spectral, textural, size and shape of the land cover.

The results clearly suggests that machine learning is a promising approach for urban land cover classification. Random Forests are able to achieve high accuracy and they are relatively insensitive to noise. The use of spectral and spatial features can further improve the accuracy of classification.

VII. Future Scope

The Future Scope of this project is to make functionality available to any and every domain which can benefit from spatial imagery, we have already set foot in that domain by deploying our current model and making it globally available. With the help of our model, we can optimize the urbanization process with ease, mitigate the problem of under/over population and allow better mapping of underdeveloped or newly discovered areas. However, this requires the image extraction algorithm used to extract numeric data from the satellite images. Once integrated with the image extraction method our model is only limited by pixel imperfections and processing power. The model can also be used on high-definition spatial images of distant land masses to help better understand the cosmos. We can also try and explore classification using neural networks such as multi-scale convolutional neural network as used in paper [7]. Integrating our model with environmental monitoring systems, such as weather stations and air quality sensors will also provide deep insights on changes our planet faces over both long and short periods of time.

The model further can also be used in different domains like rent prediction, drone imagery, anomaly detection and the domain of agriculture too.

As the future scope is very broad, we have started expanding the project by the following:

- Interactive Dashboard with graphs and visualization techniques for insights on the produced data.
- Incorporation of Deep learning algorithms
- Improving overall accuracy of the model.

References

[1] Ursula C. Benz, Peter Hofmann, Gregor Willhauck, Iris Lingenfelder, and Markus Heynen. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for gis-ready information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(3):239–258, 2004. Integration of Geodata and Imagery for Automated Refinement and Update of Spatial Databases.

[2] Bal Choudhary, Puneeta Pandey, R. Kohli, V.K. Garg, and Ashok Dhawan. *Applications of Remote Sensing and GIS in Land Resource Management*. 02 2018.

[3] Stefanos Georganos, Tais Grippa, Sabine Vanhuysse, Moritz Lennert, Michal Shimoni, and Eléonore Wolff. Very high resolution object-based land use–land cover urban classification using extreme gradient boosting. *IEEE Geoscience and Remote Sensing Letters*, 15(4):607–611, 2018.

[4] Brian Johnson and Zhixiao Xie. Classifying a high resolution image of an urban area using super-object information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 83:40–49, 2013.

[5] Brian A. Johnson. High-resolution urban land-cover classification using a competitive multi-scale object-based approach. *Remote Sensing Letters*, 4(2):131–140, 2013.

[6] T. Kavzoglu and I. Colkesen. A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 11(5):352–359, 2009.

[7] Chun Liu, Doudou Zeng, Hangbin Wu, Yin Wang, Shoujun Jia, and Liang Xin. Urban land cover classification of high-resolution aerial imagery using a relation-enhanced multiscale convolutional network. *Remote Sensing*, 12(2), 2020.

[8] Giandomenico De Luca, João M. N. Silva, Salvatore Di Fazio, and Giuseppe Modica. Integrated use of sentinel-1 and sentinel-2 data and open-source machine learning algorithms for land cover mapping in a mediterranean region. *European Journal of Remote Sensing*, 55(1):52–70, 2022.

[9] Jacek Malczewski. Gis-based land-use suitability analysis: a critical overview. *Progress in Planning*, 62(1):3–65, 2004.

[10] Chong Peng and Qiang Cheng. Discriminative ridge machine: A classifier for high-dimensional data or imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2595–2609, 2021.

[11] Darius Phiri, Matamyo Simwanda, Serajis Salekin, Vincent R. Nyirenda, Yuji Murayama, and Manjula Ranagalage. Sentinel-2 data for land cover/use mapping: A review. *Remote Sensing*, 12(14), 2020.

[12] Muhammad Murtadha Ramadhan, Imas Sukaesih Sitanggang, Fahrendi Rizky Nasution, and Abdullah Ghifari. Parameter tuning in random forest based on grid search method for gender classification based on voice frequency. *DEStech transactions on computer science and engineering*, 10(2017), 2017.

[13] Phan Thanh Noi and Martin Kappas. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery. *Sensors*, 18(1), 2018.

[14] Asma Trabelsi, Zied Elouedi, and Eric Lefevre. Decision tree classifiers for evidential attribute values and class labels. *Fuzzy Sets and Systems*, 366:46–62, 2019. Selected Papers from LFA 2016 Conference.

[15] Tianxiang Zhang, Jinya Su, Zhiyong Xu, Yulin Luo, and Jiangyun Li. Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier. *Applied Sciences*, 11(2), 2021.