(IJPS) 2023, Vol. No. 15, Jan-Jun **e-ISSN: 2455-6270 p-ISSN: 2455-7455**

Application Of The Machine Learning (ML) Approaches Based On Classification Tools In Predicting Of Urban Land Cover[1]

Tejas Thakral
Vivekananda Institute of Professional Studies
Delhi

ABSTRACT

Understanding the evolution of the urban environment and its implications requires an understanding of the classification of urban land cover. In this work, we present a comparative examination of machine learning techniques for urban land cover classification using remote sensing data. The research makes use of the UCI Machine Learning Repository's Urban Land Cover dataset, which is made up of high-resolution photos of cities. Together with their accuracy scores, a comparison of a number of well-known machine learning classification algorithms is also conducted, including the Decision Tree, Random Forest, Support Vector Machine, XGBoost, K-Nearest Neighbors, and Ridge classifiers. After eliminating the outliers and fine-tuning the hyper-parameters using Grid Search CV, the Random Forest algorithm beats the other machine learning algorithms, with an overall accuracy of 91.38%.

INTRODUCTION

Understanding the development of urban environments and its effects on the ecology, resource management, and urban planning depends heavily on the classification of urban land cover. For a thorough examination, observation, and detection, remote sensing data from satellites or other aerial platforms offers invaluable details about the earth's surface [1]. Accurate mapping and monitoring of urban land cover are aided by it.

Information on urban land cover is an essential tool for planning and managing urban areas. It can be created using high-resolution aerial or satellite photography and is useful for mapping impervious surfaces, green space, and building footprint updates, among other things. Data from Geographic Information Systems (GIS) have been extensively utilized in the examination of land-based sustainability [9]. But it's a difficult challenge to accurately extract land cover information from high-resolution images. Traditional pixel-based picture classification methods can be severely hindered by the high degree of spectral variability among land cover classes, which can be caused by sun angle, gaps in tree canopies, and shadows [5]. The modifiable areal unit issue (MAUP), which results in a mismatch between pixels and real-world items of interest, is the reason for this [4].

With an emphasis on the MAUP, this research attempts to investigate the methods and obstacles associated with obtaining urban land cover information from high-resolution data. The possibility of novel approaches like object-based image analysis (OBIA) and machine learning methods to raise the model's classification accuracy will also be examined in this work [4]. Information about land cover was extracted from Because the pixels in remote sensing photographs might have any size, it can be hard to match them up with actual things in the real world [2]. Numerous research have employed the method of geographic object-based image analysis to address this issue. Prior to classification, the OBIA technique divides the image into homogeneous sections. The characteristics of these segments are utilized for classification rather than the characteristics of individual pixels. This method decreases the susceptibility of classification to the modifiable areal unit problem (MAUP), integrates geographical and contextual information, and can aid in reducing within-class spectral variability.

The purpose of this study is to use the Urban Land Cover dataset to create a machine learning strategy for predicting urban land cover. We will also investigate the different machine learning methods that are applied to classification.

[1] *How to cite the article:*
Thakral T., Application Of The Machine Learning (ML) Approaches Based On Classification Tools In Predicting Of Urban Land Cover; *International Journal of Professional Studies*; Jan-Jun 2023, Vol 15, 70-76

These algorithms, which include the Random Forest classifier, XGBoost classifier, Ridge classifier, Decision Tree classifier, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) classifier, will assist us in determining the most precise and effective technique for classifying urban land cover [5].

Our research aims to create an accurate land cover classification model that may be used for real-world purposes like urban planning, environmental monitoring, and resource management. The findings of this study could be useful in sustainability, urban management, and decision-making. They will also contribute to the body of knowledge in the field of urban land cover classification utilizing remote sensing and machine learning.

A. Motivation

Urbanization is a worldwide phenomenon with profound effects on society, the environment, and the economy. As cities continue to grow and expand, analyzing and comprehending urban land cover patterns becomes more crucial for resource management, environmental monitoring, and urban planning. Data from remote sensing, such as satellite or aerial photography, offer a wealth of information for researching urban land cover. [1] The Urban Land Cover Dataset provides an extensive and varied set of urban land cover information.

To gain a deeper understanding and analysis of the dynamic and intricate nature of urban landscapes, a research study on this topic was conducted. Unsupervised machine learning techniques such as clustering algorithms offer a viable method for pattern recognition in urban land cover data without the need for pre-established class labels.By using clustering techniques on the dataset, we may be able to find hidden patterns and structures, distinguish between similar land cover types, and learn more about the dynamics and spatial distribution of various land cover types present in metropolitan settings.

DETAILS OF THE DATASET

The UCI repository, which contains numerous publicly available datasets is where we got the urban land cover dataset that we used for our study. The Sentinel-2 satellite provided the multispectral images included in this collection, which covers an area of around 140 square kilometers. Since sentinel-2 data offers high spatial resolution, it can aid in reaching high accuracy, making it very helpful for these kinds of classifications [11]. With an image resolution of 10 meters, it is possible to analyze urban land cover patterns in great detail and at a fine grain level.

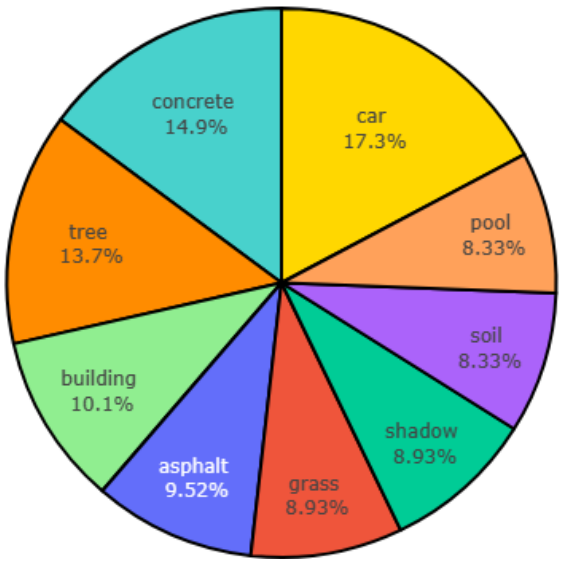

Fig. 1. Distribution of Classes

The collection includes land cover labels for classes including concrete, pool, soil, tree, shadows, buildings, cars, grass, and asphalt that reflect common urban elements that are essential for resource management, environmental monitoring, and urban planning. The collection has 148 features in total, or essentially 21 features collected over 7

different land areas. Among the features that offer a wealth of data for classifying land cover are spectral bands, indices, and other derived values. Every image in the collection is a 256x256 RGB image, with the red, green, and blue channels' color intensities represented by RGB numbers. Each image in this dataset has been manually annotated to identify the many forms of identifiable land cover that are present. Experts in the field have completed these annotations. Every image was annotated by multiple people to guarantee the correctness and uniformity of every label. This dataset provides a rich and diverse source of information for evaluating and developing land cover classification models, examining various machine learning algorithms, and researching pre-processing techniques for optimizing model performance. It is relevant for research studies that focus on machine learning, remote sensing, and urban planning. Planning for urban management, sustainability, and environmental effects can all be studied with this dataset. The dataset is widely used for activities related to mapping urban land cover, such as creating and assessing algorithms for classifying images. Furthermore, this dataset is particularly helpful for assessing how well algorithms work in difficult situations where several classes are present in a single image and the classes are visually similar to one another.

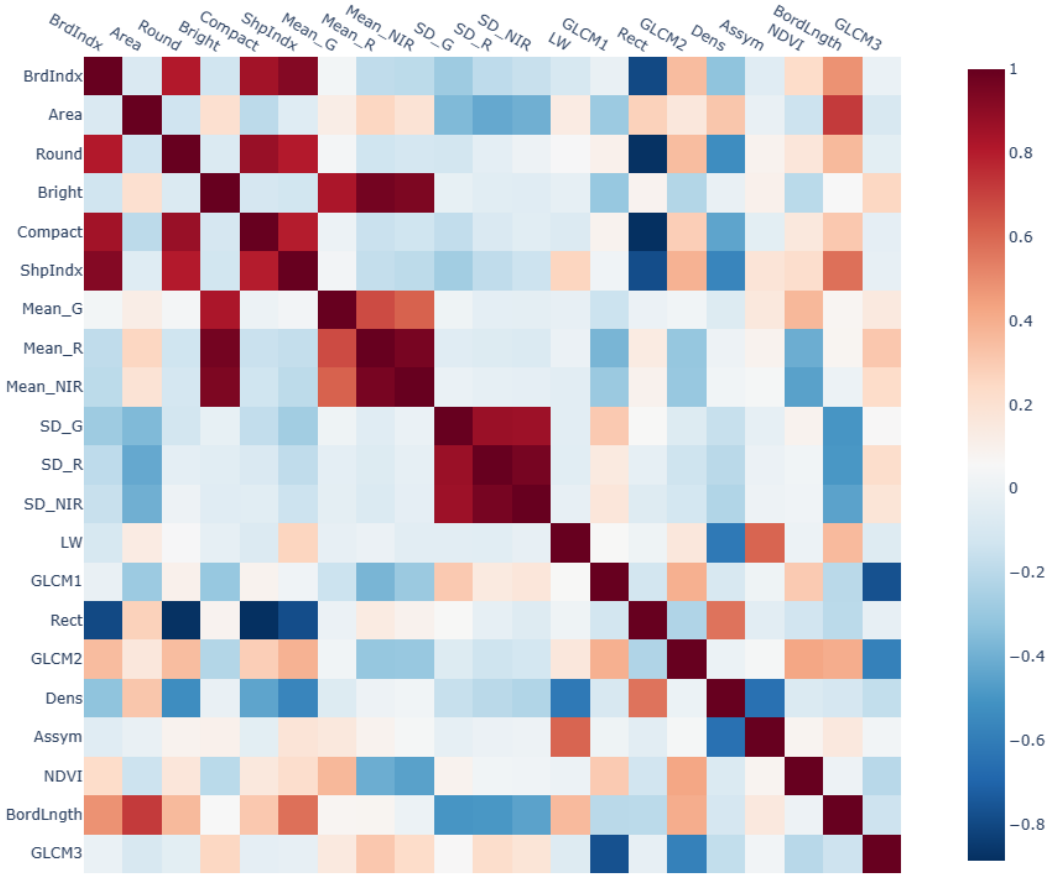

Fig 2: Features' Correlation

FORMULATION OF PROBLEMS

In this research article, the issue of accurately classifying urban land cover using machine learning approaches is being addressed. Complex and dynamic patterns of land cover, encompassing many natural and man-made elements such as concrete, pools, soil, trees, shadows, buildings, cars, grass, and asphalt, are characteristics of urban environments. Accurate and timely mapping of various land cover types is essential for resource management, urban planning, and environmental monitoring. However, conventional techniques for classifying land cover frequently have drawbacks such poor accuracy, labor-intensive manual analysis, and the inability to record changes over time.

METHODOLOGY

A. Preprocessing Data

The dataset was first examined to determine whether any null values were present, but none were. After that, the dataset's statistical summary and any appropriate visualizations were created so that it could be seen and understood. We started by displaying our dataset before identifying and eliminating any outliers. To calculate the IQR (InterQuartile Range), the first (Q1) and third (Q3) quartiles were calculated. In this instance, 1.5 times the obtained IQR was chosen as the threshold value. An outlier is defined as any result that is less than the sum of Q3 and IQR or the difference between Q1 and IQR.

$$Q1 = [(n+1)/4]$$

$$Q3 = [3(n+1)/4]$$

$$IQR = Q3–Q1$$

B. Choosing features

As stated in the dataset description, the dataset has 21 primary features that were captured at seven distinct image scales (20, 40, 60, 80, 100, 120, and 140 meters). In order to collect as much data as possible, various sets of 21 features have been chosen for the classification model. Given that a dataset can collect more information for a given class the larger its size, we have selected the last 21 features of the dataset, which correspond to the image scale of 140 meters.

C. Choosing a classification algorithm

In Supervised Machine Learning, classification is a predictive modeling issue that entails labeling or classifying input data into predetermined groups. It matters for the reasons listed below:

• Pattern recognition;

• Feature selection;

• Data organization and summarization;

• Prediction and decision making

The classifiers that our model employed to identify the best-fitting classification algorithm are listed below:

• Classifier using Decision Trees (DT) [14] This method has a structure that resembles a tree. It starts with a root node and branches into decision nodes, or sub-nodes, which have leaf nodes that contain the tree's final result. It employs a divide and conquer strategy that is recursive and top-down.

To choose the ideal attributes for the root and decision nodes, there are two categories of attribute selection techniques available:

a. Information Gain (IG) is a measurement of how an attribute's entropy has changed.

b. The Gini Index: This gauges how pure a dataset is.

•The optimized distributed gradient boosting toolkit XGBoost classifier [3] is well-known for its scalable machine learning training capabilities. It improves forecasts by integrating inadequate models.It is well-liked for managing large datasets in applications like classification and regression because it effectively handles missing values without requiring extensive pre-processing.

• The classifier Random Forest (RF) [15]. One kind of bootstrap aggregating ensemble method is this algorithm.

• KNN classifier (K-Nearest Neighbors) [13] - One of the most straightforward machine learning classification methods is this one, and it can withstand data noise well. It's a classifier that uses distance. It is non-parametric and lazy, which means that it saves the training data without learning anything from it and begins classifying as soon as it receives the testing data. It also makes no assumptions on the data that is supplied. However, there are several drawbacks to utilizing this technique, namely the fact that it needs substantial processing power, it struggles with missing value data and cannot manage outliers.

• SVM, or Support Vector Classifier [6] - Strong supervised machine learning methods for regression and classification are support vector machines. Finding the best hyperplane in N-dimensional space to maximize the margin between different class data points is the aim of support vector machines (SVM). SVMs perform exceptionally well in applications such as text/image classification, anomaly detection, and the efficient use of high-dimensional and nonlinear data.

• Ridge Classifier [10]: This algorithm is applicable to binary and multi-class classification issues. To the cost function, a penalty term is added to make it function. In order to ensure that the coefficients are minimal and prevent over-fitting, the penalty term is often the sum of the squares of the feature coefficients. The L2 penalty term plus the mean square of the loss between the actual and predicted values make up the loss function for this kind of classification.

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%

RESULT

The observations from the study are listed in Table 1.

Table 1 presents the accuracy and thorough investigation of several categorization algorithms, with each methodology emphasizing unique qualities. These classifiers underwent a thorough test across a range of feature selections, revealing distinct accuracy scores for each of them.

In the end, the research finds that the Random Forest classification method is the best option for the model. Random Forest, which combines several decision trees and uses an ensemble learning technique, uses the majority vote of these trees to predict classes with high accuracy. The study uses Grid Search CV to maximize its performance [12].

Grid-Search Cross-Validation, or Grid-Search CV for short, is a critical machine learning methodology for methodically optimizing a model's hyper-parameters. It makes it easier to optimize the machine learning model's hyperparameters and makes it easier to improve the model's fit to the dataset. An overview of utilizing grid search CV to adjust the random forest classifier's hyperparameters using Sentinel-2 data is provided in Paper [8], which is essential to achieving the higher accuracy levels in this study.

CONCLUSION

Several machine learning techniques have been applied to the classification of urban land cover in this work. After a variety of machine learning algorithms are tested, it is discovered that Random Forest performs the best.

Additionally, various feature selection techniques have been investigated. It was discovered that the best outcomes come from utilizing every feature in the dataset and not removing any, as each one provides a crucial window into the spectral, textural, size, and shape of the land cover.

The findings unequivocally show that machine learning is a potential method for classifying urban land cover. Random Forests are comparatively noise-insensitive and capable of achieving great accuracy. The accuracy of categorization can be further increased by utilizing spectral and spatial characteristics.

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 El´eonore 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, Jo˜ao 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, knearest 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.