Land Use and Land Cover Change Detection in Washington D.C. Using Sentinel-2 Data and Machine Learning Approaches

Seyed Mohammad Moein, Peyghambar Zadeh – 10921320

ABSTRACT
This study aims to examine land use and land cover (LULC) changes in Washington D.C. from 2017 to 2024, using satellite images from Sentinel-2. The aim is to classify and investigate these changes for three machine learning approaches; Random Forest (RF), Multi-Layer Perceptron (MLP) and Convolutional Neural Network(CNN). Some of the preprocessing steps that were taken involved masking clouds, mosaicking images, and histogram matching techniques used in normalizing images. Our findings reveal notable shifts from forested areas to urban expansions that demonstrate the impact of urbanization on green spaces. The outcomes show the importance of remote sensing data when combined with advanced machine learning techniques for environmental surveillance purposes. You can access all codes used in this research at our GitHub page [1].

1. INTRODUCTION

High-density urbanization, like capital city of countries, brings about a range of significant challenges. These include severe traffic congestion, increased air pollution, and heightened consumption of energy and resources. Additionally, densely populated areas can accelerate the spread of diseases, contribute to ozone layer depletion, and intensify heatwaves, all of which collectively diminish the quality of life for residents [1]. Urbanization has a profound impact on land cover and land use, resulting in significant environmental changes such as deforestation and increased built-up areas[2]. Monitoring these changes is crucial for sustainable urban planning and environmental conservation. Remote sensing data, particularly from Sentinel-2 satellites, provide high-resolution imagery that is ideal for observing and analyzing these changes over time.

Urban areas are expanding rapidly worldwide, leading to changes in land cover and land use that affect biodiversity, climate, and ecosystem services. As a result, understanding these changes is vital for developing strategies to mitigate negative impacts and promote sustainable development. Washington D.C., the capital of the United States (Figure 1), offers a unique setting for this study due to its mix of urban, suburban, and natural landscapes[3].

Satellite remote sensing provides a powerful tool for monitoring environmental changes over large areas and extended periods. Sentinel-2, part of the European Space Agency's Copernicus program, offers high-resolution optical imagery that is freely accessible and useful for various environmental monitoring applications. In this study, we use Sentinel-2 images to assess changes in land cover and land use in Washington D.C. over a seven-year period.

Figure 1 Washington DC. Area from the satellite

Several studies [4-6] have investigated the use of remote sensing and machine learning techniques for land cover and land use change detection.

One notable study by [5] employed Landsat data to map urban areas globally, demonstrating the utility of medium-resolution imagery in capturing urban expansion and its dynamics. They utilized a classification tree algorithm to achieve high accuracy in distinguishing urban from non-urban areas, emphasizing the importance of temporal consistency and spatial resolution in land cover mapping.

[6] and [7] some years ago analyzed the urban growth of the Washington D.C. metropolitan area using Landsat data from 1973 to 2000. They highlighted the significant increase in impervious surfaces due to urban expansion and its impact on the environment.

[1] https://github.com/moeinp70/LULC

The study employed a combination of spectral indices and change detection algorithms to monitor land cover transitions, providing a comprehensive assessment of urbanization trends over two decades.

The objective of this study is to detect and analyze changes in LULC between 2017 and 2024 using machine learning techniques. We employ three different classifiers: Random Forest (RF), Multi-Layer Perceptron (MLP), and Convolutional Neural Network (CNN)[8]. Each classifier has unique strengths, allowing us to compare their performance and effectiveness in LULC change detection. Additionally, we utilize histogram matching for image normalization to ensure consistent data quality across different time periods.

2. MATERIALS AND METHODS

2.1 Study Area

The study area is located in Washington D.C., the capital of the United States (see Figure 1). With a population exceeding 680 thousand as of the 2000 Census, it ranks among the most densely populated cities in the country. The capital city is characterized by a diverse mix of land covers, including urban areas, parks, water bodies like the Potomac River, and significant historical sites and designated landmarks like the Capitol Building and The White House, roads and bridges like The Arlington Memorial Bridge. Over the past two decades, significant urban and suburban expansion has transformed many forested areas into residential, commercial, and industrial zones. This urban growth is evident in the distinct geometric patterns of buildings and roads, as well as the widespread development of suburbs and housing subdivisions. According to [7] developed areas around Washington D.C. expanded by approximately 22 km² per year between 1973 and 1996. This rapid urbanization has intensified the Urban Heat Island (UHI) effect, highlighting the need for strategies to mitigate its adverse impacts [7]. The study area encompasses the entirety of Washington D.C., characterized by diverse land covers including water bodies, forests, parks, built-up areas, and road infrastructures. Washington D.C. provides a representative example of urbanization and its impact on various land cover types, making it an ideal location for this study.

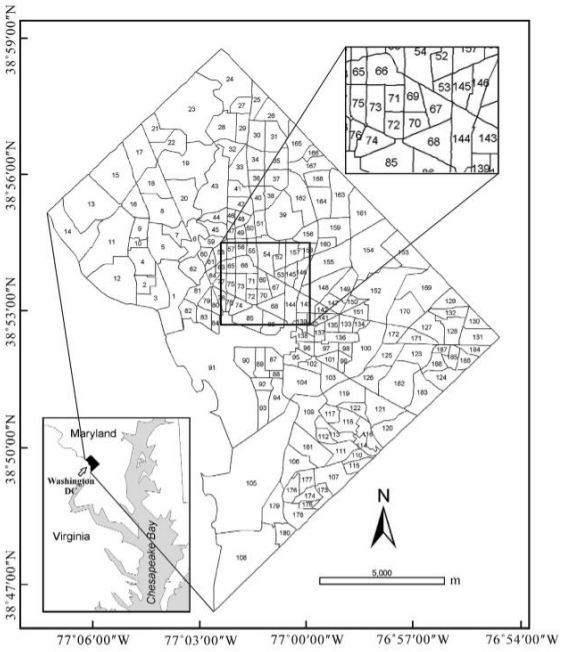

Figure 2 ***Location of study area and 2000 Census tracts of Washington DC where numbers are the tract Identity Documents (IDs).***

2.2 Data Acquisition

Sentinel-2 L2A level images from 2017 and 2024 were acquired using the Google Earth Engine (GEE). The GEE platform facilitates access to a vast archive of satellite data and offers powerful processing capabilities. The region of interest (ROI) was defined using a shapefile containing the boundaries of Washington D.C. We used 10 bands from Sentinel-2, excluding bands 1, 9, and 10 (Table 1) , to focus on bands that provide the most relevant information for land cover classification[9].

SENTINEL-2 BANDS	RESOLUTION [METERS]
Band 2 - Blue	10
Band 3 - Green	10
Band 4 - Red	10
Band 5 - Vegetation Red Edge	20
Band 6 - Vegetation Red Edge	20
Band 7 - Vegetation Red Edge	20
Band 8 - NIR	10
Band 8A - Vegetation Red Edge	20
Band 11 - SWIR	20
Band 12 - SWIR	20

Table 1 Different Band set

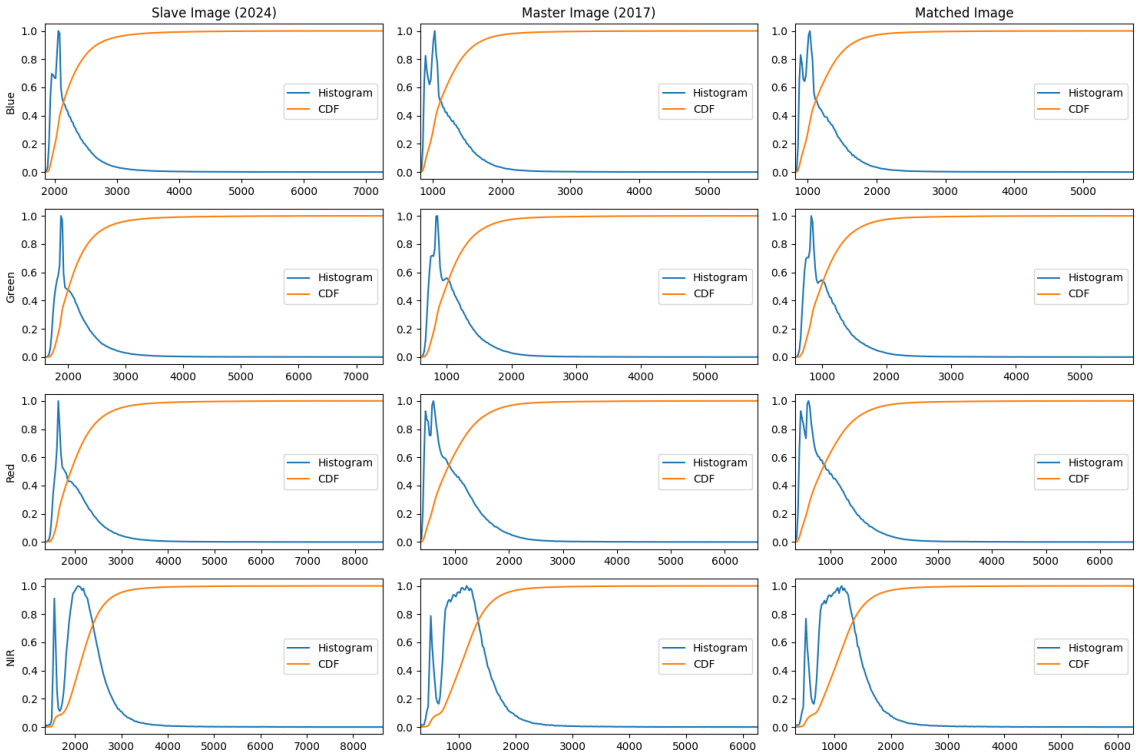

Figure 3 Histogram matching for 2017 and 2024

2.3 Preprocessing

Preprocessing steps included cloud masking, mosaicking, and histogram matching to normalize the images. Cloud masking is essential to remove cloud-covered pixels that can obscure land cover features. Mosaicking combines multiple images to create a complete and seamless view of the study area. These two are performed by defualt on Sentinel2 L2A[10] data that we used in our project. Histogram matching was performed using Scikit-image to align the distributions of pixel values between the two datasets, ensuring that the images are comparable despite being captured at different times. Figure 3 illustrates the histogram and cumulative distribution function (CDF) plots for four spectral bands (Blue, Green, Red, and Near Infrared (NIR)) of the slave image (2024), the master image (2017), and the matched image. Histogram matching is a crucial preprocessing step to ensure that the images from different time periods are comparable by aligning their intensity distributions.

2.4 Training Data

Training data was extracted from a GeoPackage file containing labeled polygons for five classes: Water, Forest, Field and Park, Built-up, and Road and Bridge. Pixel values within these polygons were extracted and used to train the classifiers. Each pixel was assigned a label corresponding to its land cover class, providing the necessary data for supervised classification.

- **Class 1: Water** - Areas covered by water bodies such as rivers, lakes, and ponds.
- **Class 2: Forest** - Regions with dense tree cover, including natural forests and managed woodlands.
- **Class 3: Field and Park** - Open areas primarily used for agriculture, recreation, or as urban green spaces.
- **Class 4: Built-up** - Urban areas with high-density buildings, including residential, commercial, and industrial zones.
- **Class 5: Road and Bridge** - Infrastructure dedicated to transportation, including roads, highways, and bridges.

2.5 Classification Approaches

2.5.1 Random Forest (RF)

The Random Forest classifier was implemented using Scikit-learn. RF is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification. It is robust to overfitting and can handle large datasets with higher accuracy.

2.5.2 Multi-Layer Perceptron (MLP)

The MLP was used to capture non-linear relationships between features. It consists of multiple layers of neurons, with each layer fully connected to the next one. MLPs are effective for a wide range of classification tasks and were implemented using Scikit-learn's MLPClassifier[11].The Multilayer Perceptron (MLP) model employed in this study features a two-layer architecture with hidden layers containing 128 and 64 neurons, respectively. It uses the Adam optimizer for training, with a maximum of 200 iterations, a learning rate of 0.001, and an alpha value for regularization set to 1e-4, providing a robust framework for effective land cover classification.

2.5.3 Convolutional Neural Network (CNN)

A CNN was designed to leverage spatial correlations between neighboring pixels. CNNs are particularly effective for image classification tasks due to their ability to capture spatial hierarchies[8]. The CNN was implemented using TensorFlow[12], with layers designed to extract features from the input images and classify them into one of the five land cover classes. The Convolutional Neural Network (CNN) architecture used in this study consists of three convolutional layers with ReLU activation functions and varying filter sizes, followed by a flattening layer and two dense layers, ultimately employing a softmax activation function for multiclass classification.

3. RESULTS DISCUSSION

3.1 Classification Accuracy

Accuracy metrics for each classifier were computed and compared. In this study, the dataset was split into training and testing sets to evaluate the performance of the different classification approaches. Specifically, 70% of the data was used for training the models, and 30% was reserved for testing. This split ensures that the models are trained on a substantial portion of the data while retaining enough samples to rigorously evaluate their performance. The train_test_split function from Scikit-learn was employed, with a random_state of 42 to ensure reproducibility of the results across different runs. This consistent splitting strategy was applied to all classification approaches, including the Convolutional Neural Network (CNN), Multilayer Perceptron (MLP), and Random Forest models, providing a fair and comparable basis for performance evaluation. To visualize the result, Figure 4 Shows the LULC map of the study area using RF. The CNN approach demonstrated slightly inferior performance about 97% accuracy due to the lack of training data but its ability to capture spatial

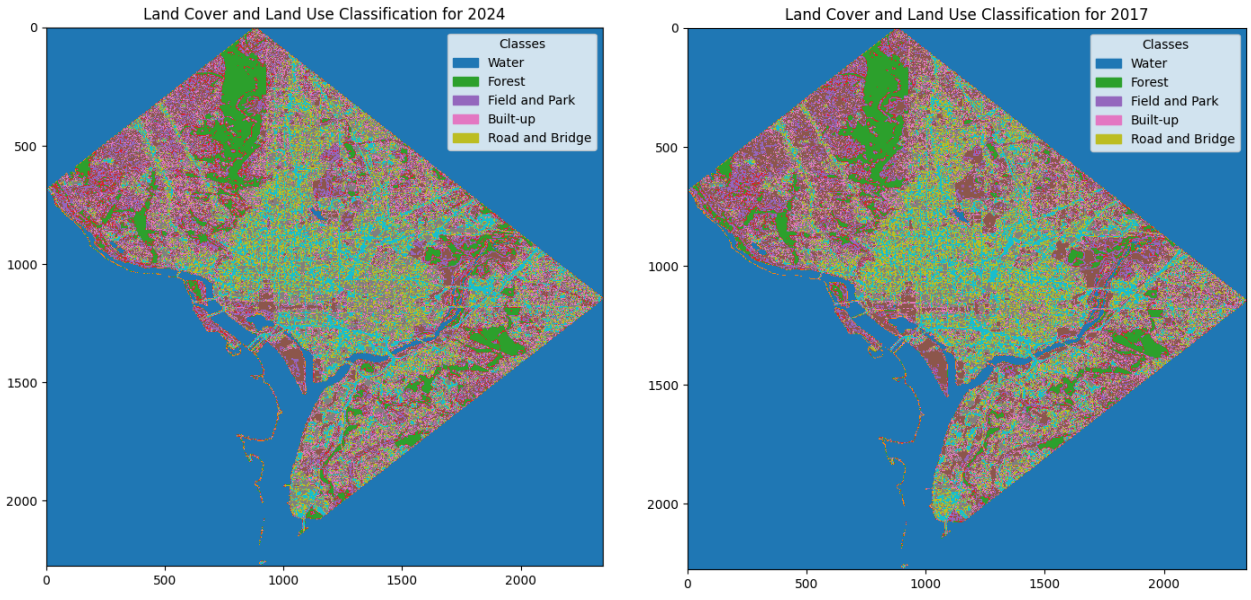

Figure 4 LCLU of Washington both for 2017 and 2024 using RF

patterns , while RF (Table 2) and MLP also provided satisfactory results about 99%. For each classifier, we evaluated the classification accuracy using standard metrics such as overall accuracy, precision (Eq 1), recall (Eq 2), and F1 score (Eq 3).

- **Precision**: The ratio of correctly predicted positive observations to the total predicted positives. High precision indicates a low false positive rate.
- **Recall**: The ratio of correctly predicted positive observations to all observations in the actual class. High recall indicates a low false negative rate.
- **F1-Score**: The weighted average of precision and recall. It takes both false positives and false negatives into account.
- **Support:** The number of actual occurrences of the class in the dataset.
- **Overall Accuracy:** The overall accuracy of the classifier is high, as indicated by the dominance of the diagonal elements (true positive rates) in the confusion matrix. The few off-diagonal elements represent the misclassifications, which are relatively minor.

- $\text{Precision} = \frac{TP}{(TP + FP)}$ (Eq 4)
- $\text{Recall} = \frac{TP}{(TP + FN)}$ (Eq 5)
- $\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ (Eq 6)

Where True Positive (TP) refers to the number of predictions where the classifier correctly predicts the positive class as positive. True Negative (TN) refers to the number of predictions where the classifier correctly predicts the negative class as negative. False Positive (FP) refers to the number of predictions where the classifier incorrectly predicts the negative class as positive. False Negative (FN) refers to the number of predictions where the classifier incorrectly predicts the positive class as negative.

Class	Precision	Recall	F1-Score	Support
1 (Water)	1.00	1.00	1.00	2824
2 (Forest)	0.99	1.00	0.99	3438
3 (Field and Park)	0.99	0.98	0.98	1876
4 (Built-up)	0.98	0.96	0.97	1828
5 (Road and Bridge)	0.94	0.98	0.96	1093
Accuracy			0.99	11059

Table 2 Random Forest metrics

Also the confusion matrix in Figure 5 for each model provides detailed insights into the performance across different land cover classes.

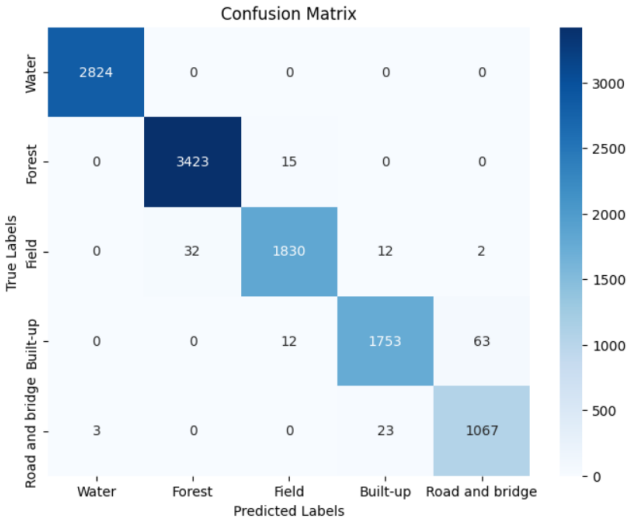

Figure 5 Confusion matrix for Random forest method

It visually represents the accuracy of the classifier by comparing the true labels with the predicted labels. The matrix includes the following classes:

- **Water**: The classifier correctly identified all 2824 instances of Water, with no misclassifications. This demonstrates the model's high accuracy in detecting Water regions.
- **Forest**: Out of 3438 true instances of Forest, the classifier correctly predicted 3423, with only 15 instances being misclassified as Field and Park. This indicates a high precision and recall for the Forest class.
- **Field and Park**: For this class, the classifier correctly identified 1830 out of 1876 instances. There were minor misclassifications, with 32 instances labeled as Forest and 12 as Built-up.
- **Built-up**: The model correctly classified 1753 out of 1828 instances as Built-up. A small number of instances were misclassified into other categories, including 12 as Field and Park and 63 as Road and Bridge.
- **Road and Bridge**: The classifier identified 1067 out of 1093 instances correctly, with a few misclassifications, primarily into the Built-up class.

The highest misclassification rates are observed between classes that are visually or contextually

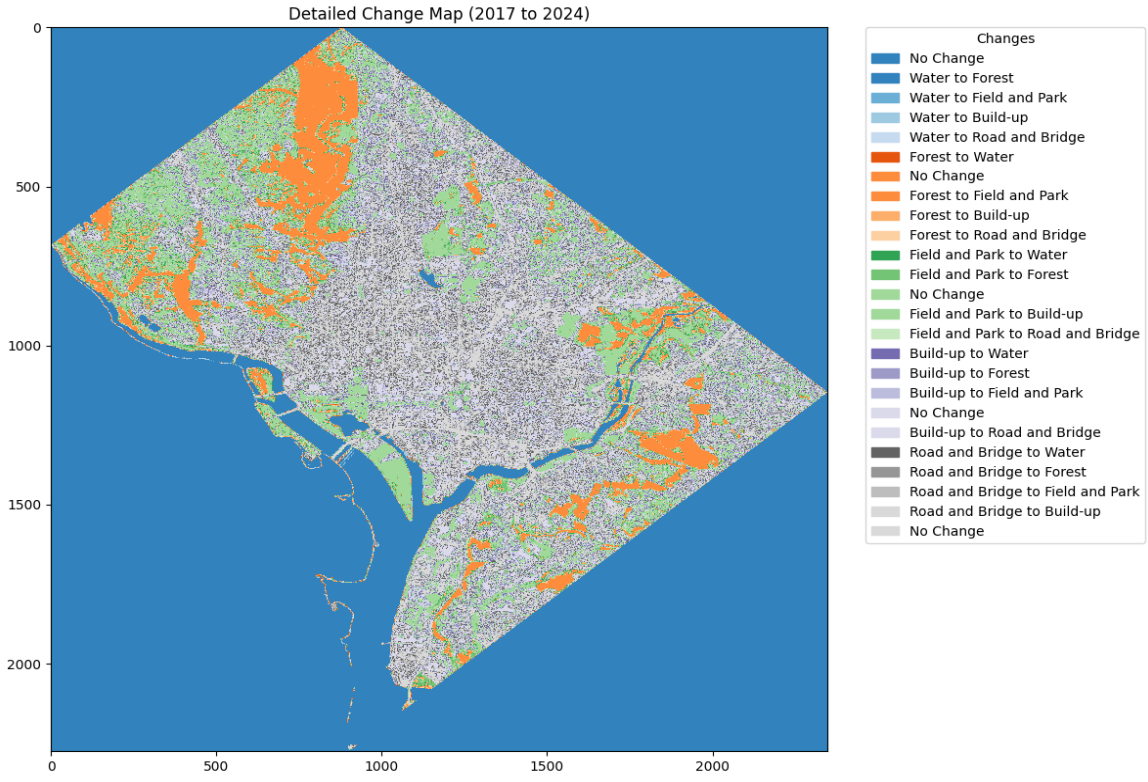

Figure 6 Detailed Change Map

similar, such as Forest and Field and Park, or Built-up and Road and Bridge. This indicates areas where further model refinement or additional features could enhance classification performance.

3.2 Change Detection

Detailed change maps were generated to identify specific transitions such as Forest to Built-up and Park to Road and Bridge. Figure 6 illustrates the detailed change map for Washington D.C., depicting land cover and land use changes between 2017 and 2024. The map categorizes changes into 25 distinct classes, each represented by a unique color. The legend, positioned outside the plot, provides a clear description for each class, facilitating easy interpretation of the changes. This detailed change map provides valuable insights into the dynamics of urban growth and its impact on natural and semi-natural landscapes in Washington D.C. The visualization aids in understanding the extent of urbanization and helps in identifying critical areas where conservation efforts may be necessary to mitigate adverse environmental impacts.

3.3 Analysis of Specific Land Cover and Land Use Changes

Figure 7 also highlights specific land cover and land use transitions in Washington D.C. between 2017 and 2024. This visualization focuses on critical changes from forest and park areas to built-up regions and transportation infrastructure (roads and bridges). The transitions are coded and color-mapped using the 'hot' colormap to emphasize areas of significant change.

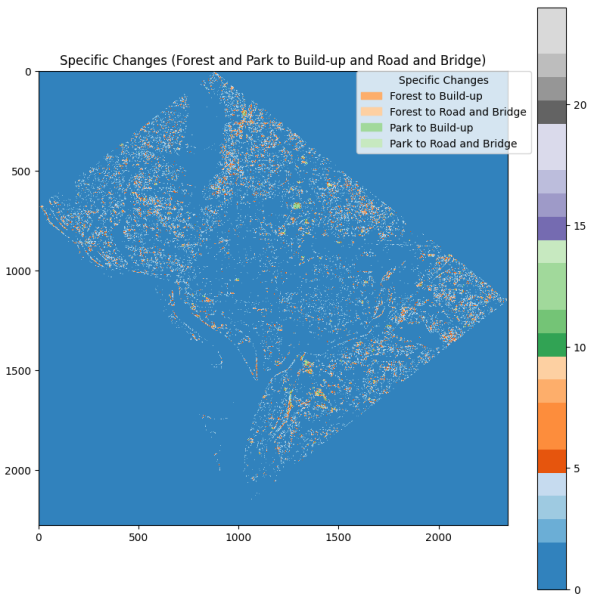

Figure 7 Specific Changes in urban development

- **Forest to Build-up (2 to 4)**: These transitions are represented in bright colors, indicating substantial deforestation and subsequent urban development. This pattern is predominantly visible in the northern and eastern parts of the city, where urban expansion has encroached upon previously forested regions.

- **Forest to Road and Bridge (2 to 5)**: Marked in contrasting shades, these changes highlight the construction of new roads and bridges within forested areas. These developments are crucial for understanding the impact of infrastructure projects on natural landscapes.

- **Field and Park to Build-up (3 to 4)**: This transition shows the conversion of green spaces, such as fields and parks, into urbanized areas. The visualization reveals scattered but significant patches of green space being developed into residential or commercial zones.

- **Field and Park to Road and Bridge (3 to 5)**: These transitions illustrate the development of new transportation infrastructure within existing green spaces, further emphasizing the expansion of the city's transport network into less developed areas.

These maps revealed significant urban expansion at the expense of green spaces. The analysis highlighted areas where urban development encroached on previously vegetated areas, providing insights into the dynamics of urban growth in Washington D.C.

4. CONCLUSIONS

This study successfully applied RF, MLP, and CNN classifiers to detect LULC changes in Washington D.C. from 2017 to 2024. The RF model outperformed the other approaches, highlighting the importance of spatial context in remote sensing classification. The study demonstrates the effectiveness of combining remote sensing data with advanced machine learning techniques for environmental monitoring. Future work could explore the integration of additional data sources, such as socioeconomic data, to enhance classification accuracy and provide more comprehensive insights into urbanization patterns.

REFERENCES

1. Yue, W., et al., *The relationship between land surface temperature and NDVI with remote sensing: application to Shanghai Landsat 7 ETM+ data.* International journal of remote sensing, 2007. **28**(15): p. 3205-3226.
2. Zhou, G., et al., *Impacts of Urban land surface temperature on tract landscape pattern, physical and social variables.* International Journal of Remote Sensing, 2020. **41**(2): p. 683-703.
3. Lakshmi, V., *Enhancing human resilience against climate change: Assessment of hydroclimatic extremes and sea level rise impacts on the eastern shore of Virginia, United States.* Science of The Total Environment, 2024: p. 174289.
4. Sexton, J.O., et al., *Urban growth of the Washington, DC–Baltimore, MD metropolitan region from 1984 to 2010 by annual, Landsat-based estimates of impervious cover.* Remote Sensing of Environment, 2013. **129**: p. 42-53.
5. Schneider, A., M.A. Friedl, and D. Potere, *A new map of global urban extent from MODIS satellite data.* Environmental research letters, 2009. **4**(4): p. 044003.
6. Yuan, D. and C. Elvidge, *NALC land cover change detection pilot study: Washington DC area experiments.* Remote sensing of environment, 1998. **66**(2): p. 166-178.
7. Masek, J., F. Lindsay, and S. Goward, *Dynamics of urban growth in the Washington DC metropolitan area, 1973-1996, from Landsat observations.* International journal of remote sensing, 2000. **21**(18): p. 3473-3486.
8. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.
9. Zhang, T.-X., et al., *Potential bands of sentinel-2A satellite for classification problems in precision agriculture.* International Journal of Automation and Computing, 2019. **16**: p. 16-26.
10. Gascon, F., et al., *Copernicus Sentinel-2A Calibration and Products Validation Status.* Remote Sensing, 2017. **9**(6): p. 584.
11. Taud, H. and J.-F. Mas, *Multilayer perceptron (MLP).* Geomatic approaches for modeling land change scenarios, 2018: p. 451-455.
12. Abadi, M., et al. *{TensorFlow}: a system for {Large-Scale} machine learning*. in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 2016.