Detailed Project Report: Water Information Extraction Based on Multi-Model RF Algorithm and Sentinel-2 Image Data for GHMC Region

Manogna Darisa

May 10, 2025

Abstract

Accurate water body extraction is essential for sustainable resource management, especially in urban environments. This study uses a Random Forest-based multi-model approach to classify water and non-water areas over the GHMC (Greater Hyderabad Municipal Corporation) region using Sentinel-2 satellite imagery and terrain data. Multiple spectral indices were computed to enhance feature space, and a combination of manually collected water and non-water samples were used for training. Several Random Forest-based models, including feature selection techniques such as RF_16 and ReliefF_16, were evaluated. Results demonstrate that integrating spectral and terrain features significantly improves classification performance.

1 Introduction

Surface water detection using remote sensing techniques has become increasingly important for hydrological monitoring, disaster management, and urban planning. Traditional spectral methods are often insufficient due to challenges like shadows, urban noise, and mixed pixels. Machine Learning (ML), especially ensemble-based models like Random Forest (RF), offers robust classification even in complex, high-dimensional feature spaces.

In this project, we applied a multi-model RF-based method, inspired by previous research, adapted to the urban GHMC area using Sentinel-2 imagery. The goal was to extract water information accurately using different feature combinations and feature selection methods.

2 Study Area

The Greater Hyderabad Municipal Corporation (GHMC) region, located in Telangana, India, is characterized by a mixture of urban areas, vegetation, water bodies, and built-up land. The region presents a challenging environment for remote sensing classification due to complex land cover patterns, making it an ideal case study for evaluating machine learning-based water extraction methods.

3 Data Sources and Preprocessing

3.1 Satellite Imagery

Sentinel-2 Surface Reflectance Level-2A images were used. Images were filtered between 2023-10-01 and 2023-11-30, with less than 10% cloud cover. A median composite was generated to minimize cloud contamination.

3.2 Study Area

The Area of Interest (AOI) was defined using a shapefile of the GHMC boundary, imported from Google Earth Engine (GEE) assets.

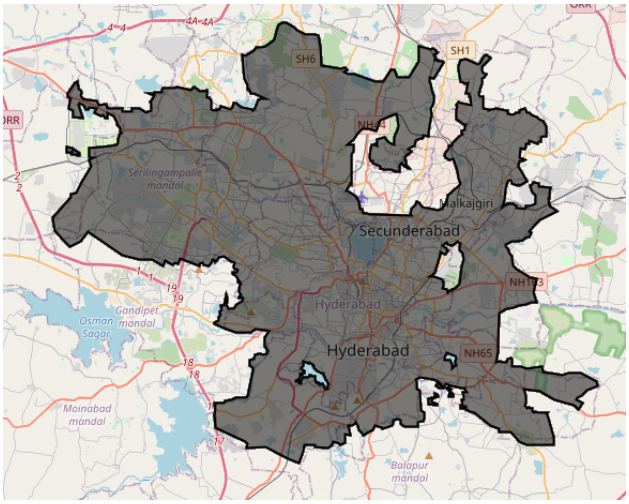

Figure 1: shapefile of AOI

3.3 False color composite of AOI

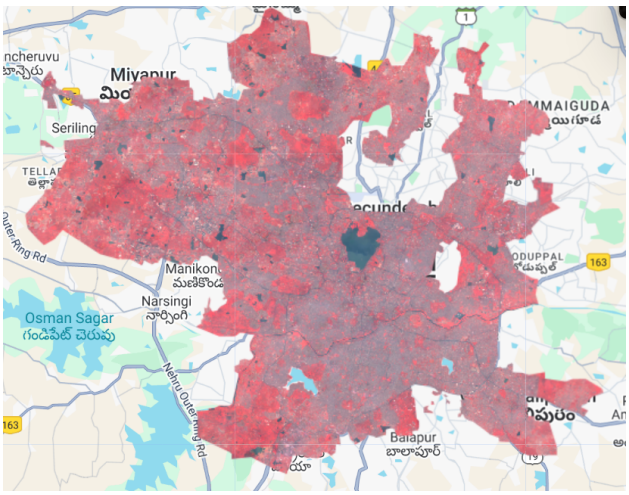

Figure 2: FCC1

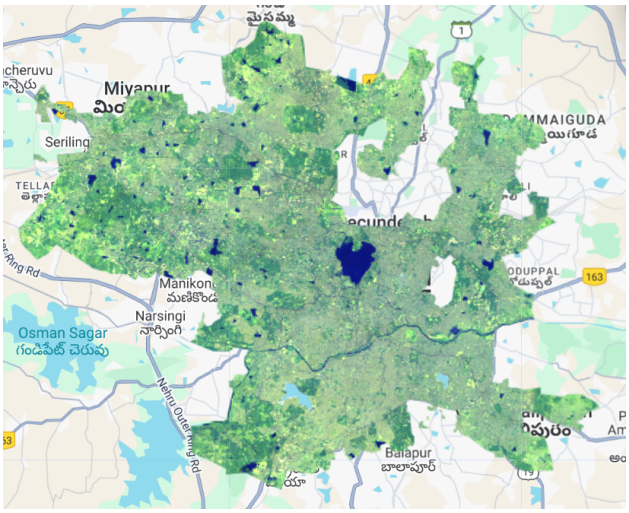

Figure 3: FCC2

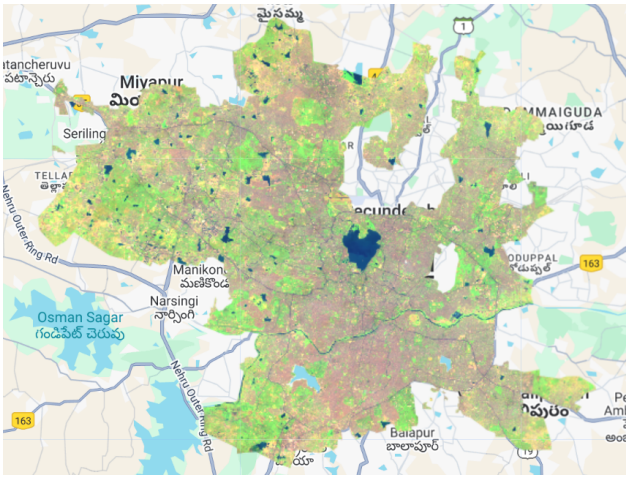

Figure 4: FCC3

3.4 Spectral Bands Used

The following six Sentinel-2 bands were selected:

- B2 (Blue)
- B3 (Green)
- B4 (Red)
- B8 (Near Infrared, NIR)
- B11 (Short Wave Infrared 1, SWIR1)
- B12 (Short Wave Infrared 2, SWIR2)

3.5 Feature Construction

In total, 24 features were constructed, divided into four categories:

- **Traditional Spectral Features**: B2, B3, B4, B8, B11, B12, NDVI, MSAVI
- **Red-Edge and Other Spectral Features**: B5, B6, B7, B8A, NDI45, MCARI, REIP, S2REP, IRECI, PSSRa
- **Water Indices**: NDWI, MNDWI, LSWI
- **Terrain Features**: DEM, SLOPE, ASPECT

Additional vegetation and red-edge indices were computed to improve land-water discrimination.

3.6 Spectral Indices and Their Meaning

Multiple spectral indices were computed to enhance water body discrimination. Below are the indices with their full forms, meanings, and formulas:

- **NDVI (Normalized Difference Vegetation Index)**: Measures vegetation health.
$$NDVI = \frac{B8 - B4}{B8 + B4}$$
- **NDWI (Normalized Difference Water Index)**: Enhances water features while suppressing vegetation.
$$NDWI = \frac{B3 - B8}{B3 + B8}$$
- **MNDWI (Modified Normalized Difference Water Index)**: Improves water feature extraction using SWIR.
$$MNDWI = \frac{B3 - B11}{B3 + B11}$$
- **BSI (Bare Soil Index)**: Identifies bare soil areas.
$$BSI = \frac{(B11 + B4) - (B8 + B2)}{(B11 + B4) + (B8 + B2)}$$
- **NDMI (Normalized Difference Moisture Index)**: Detects moisture content in vegetation.
$$NDMI = \frac{B8 - B11}{B8 + B11}$$
- **LSWI (Land Surface Water Index)**: Sensitive to leaf and soil moisture.
$$LSWI = \frac{B8 - B11}{B8 + B11}$$
- **MSAVI (Modified Soil Adjusted Vegetation Index)**: Reduces soil brightness influence on vegetation indices.
$$MSAVI = 0.5 \times \left(2B8 + 1 - \sqrt{(2B8 + 1)^2 - 8(B8 - B4)}\right)$$

- **NDI45 (Normalized Difference Index between B5 and B4)**: Indicates chlorophyll content.

$$NDI45 = \frac{B5 - B4}{B5 + B4}$$

- **MCARI (Modified Chlorophyll Absorption Ratio Index)**: Measures vegetation chlorophyll absorption.

$$MCARI = (B5 - B4 - 0.2(B5 - B3)) \times (B5 - B4)$$

- **REIP (Red Edge Inflection Point)**: Locates the red-edge position, important for vegetation stress.

$$REIP = 700 + 40 \times \frac{(B4 + B7)/2 - B5}{B6 - B5}$$

- **S2REP (Sentinel-2 Red Edge Position)**: Similar to REIP but adapted to Sentinel-2.

$$S2REP = 705 + 35 \times \frac{(B4 + B7)/2 - B5}{B6 - B5}$$

- **IRECI (Inverted Red Edge Chlorophyll Index)**: An inverted index related to chlorophyll.

$$IRECI = \frac{B7 - B4}{B5/B6}$$

- **PSSRa (Pigment Specific Simple Ratio a)**: Another chlorophyll estimation index.

$$PSSRa = \frac{B7}{B4}$$

3.7 Terrain Features

Terrain parameters such as Digital Elevation Model (DEM), slope, and aspect were computed and stacked with spectral features.

4 Sample Collection

A total of 1110 manually selected water points and 1110 non-water points were collected by visual inspection using Sentinel-2 imagery and Google Earth reference images. The data was split into 70% training and 30% testing subsets while maintaining class balance.

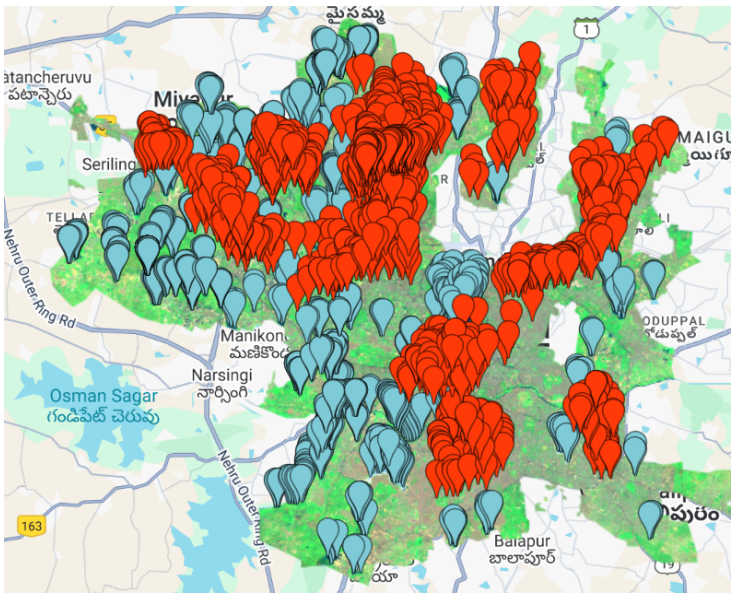

Figure 5: Manually Collected Training Samples for Water and Non-Water Classes. The Blue colored samples represents water and the Red colored samples represents nonwater

5 Model Construction

5.1 Four Models

- **Model A**: Basic spectral features and indices (NDVI, MSAVI).
- **Model B**: Adds red-edge and vegetation indices.
- **Model C**: Further includes water-related indices.
- **Model D**: Incorporates terrain features.

5.2 Optimized feature models

- **RF_16:** Top 16 features selected based on Random Forest feature importance.
- **Relief_16:** Top 16 features selected using ReliefF algorithm.

6 Random Forest Algorithm

6.1 What is Random Forest?

Random Forest is an ensemble machine learning algorithm that builds multiple decision trees and merges their results to achieve more accurate and stable predictions.

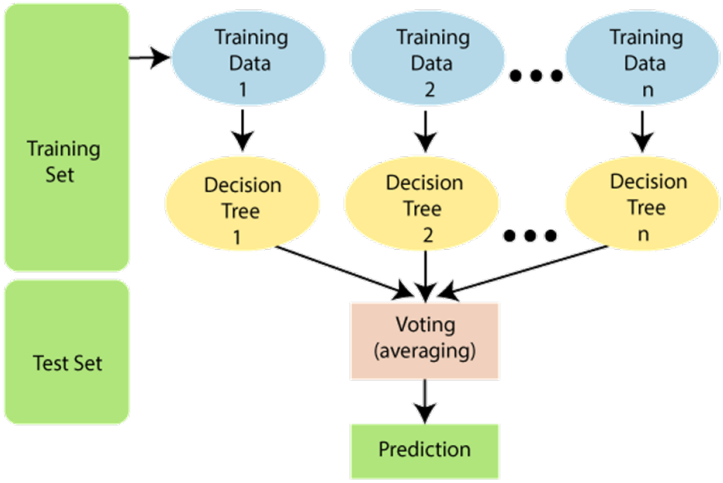

Figure 6: Working process of Random Forest: training multiple decision trees on different subsets and combining results through voting.

6.2 How Random Forest Works

1. Draw multiple bootstrap samples from the dataset.
2. Train a decision tree on each sample.
3. At each tree node, a random subset of features is considered for splitting.
4. Final prediction is made by majority voting across all trees.

7 Feature Selection Methods

7.1 RF_16 (Random Forest Top 16 Features)

Random Forest assigns importance scores to features based on Gini impurity decrease.

Gini Coefficient measures the impurity of a dataset. It is used to decide the best feature for splitting the data at each node of a decision tree.

The Gini formula is given by:

$$Gini = \sum \left(\frac{f(C_i, T)}{T} \times \left(1 - \frac{f(C_i, T)}{T} \right) \right)$$

where C_i is a class, $f(C_i, T)$ is the frequency of class i in dataset T, and T is the total number of samples.

Using the Gini Coefficient, Random Forest splits data at every node and forms multiple decision trees to improve classification performance.

The top 16 features with highest importance were selected to form the RF_16 feature subset.

7.2 ReliefF Algorithm

ReliefF estimates feature importance based on how feature values differentiate between neighboring instances of different classes.

7.3 ReliefF_16

The top 16 features selected using the ReliefF algorithm were used to train a separate Random Forest model.

8 Results and Evaluation

To evaluate the performance of the land cover classification, a confusion matrix was generated. This matrix helps to understand how well the classifier is predicting each land cover class.

Confusion Matrix

From the confusion matrix:

- The diagonal values represent correctly classified pixels for each land cover class.

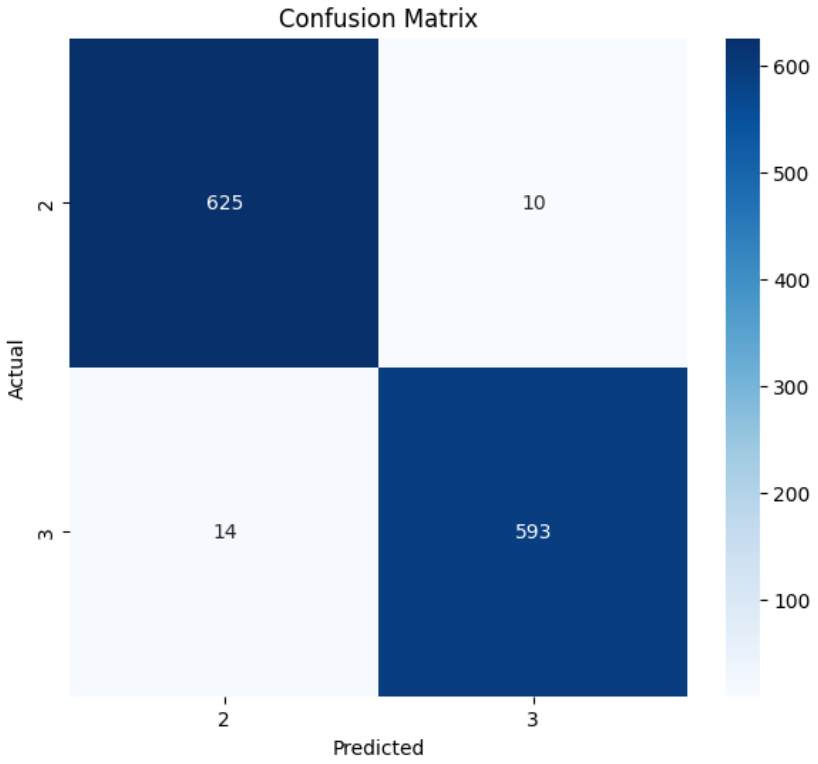

Figure 7: Confusion Matrix for the Classified Land Cover Map

- High values along the diagonal suggest that the model has performed well in distinguishing between classes.

Overall Accuracy and Kappa Coefficient

- **Overall Accuracy (OA)** is computed as the total number of correctly classified pixels divided by the total number of reference pixels.
- **Kappa Coefficient** is calculated to measure agreement between prediction and reference data, adjusting for chance agreement.

Class-wise Accuracy

- **Producer's Accuracy (PA)** is the ratio of correctly classified pixels in a class to the total pixels of that class in the reference data.
- **User's Accuracy (UA)** is the ratio of correctly classified pixels in a class to the total pixels assigned to that class by the classifier.

8.1 Classification Accuracy

Model	Accuracy (%)	Kappa Score
Model A	83.24	0.8162
Model B	85.71	0.8283
Model C	87.23	0.8342
Model D	90.14	0.8940
ReliefF_16	96.50	0.9613
RF_16	**98.04**	**0.9677**

8.2 Model Comparison

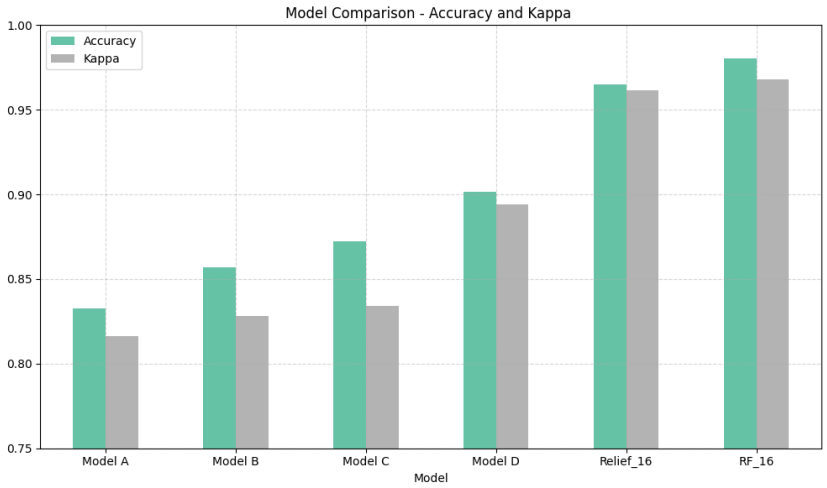

Figure 8: Classification Accuracy and Kappa Score Comparison among Different Models

9 Classification image of model RF_16

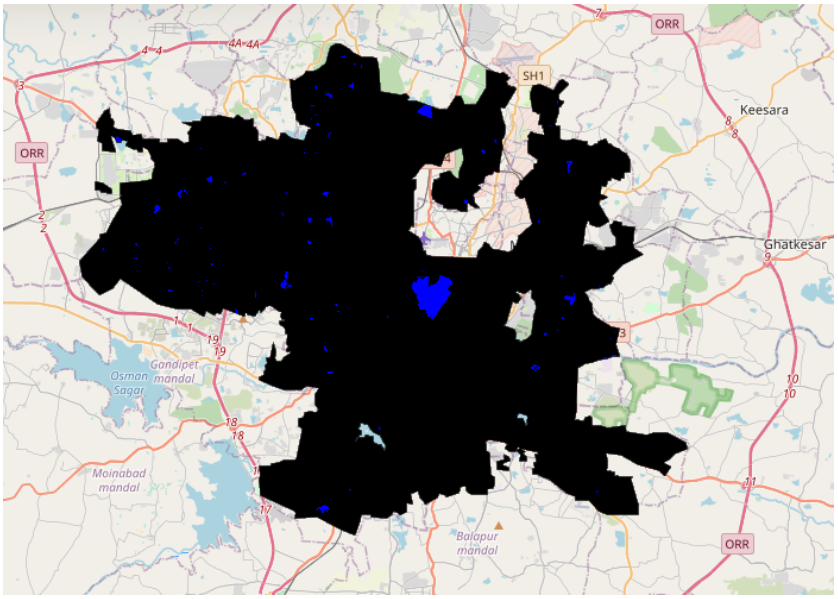

Figure 9: output Image

10 Conclusion

The Random Forest classifier with feature selection (RF_16) achieved the best performance for water body extraction over the GHMC region. Incorporating spectral indices, red-edge bands, water indices, and terrain parameters significantly improved classification accuracy. The methodology demonstrates strong potential for urban water resource monitoring.

References

- Jiang, Z., Wen, Y., Zhang, G., Wu, X. (2022). Water Information Extraction Based on Multi-Model RF Algorithm and Sentinel-2 Image Data. Sustainability, 14, 3797. `https://doi.org/10.3390/su14073797`