Report on Land Use Land Cover (LULC) Classification using K-means and Random Forest

**1. Introduction**

The goal of this study is to classify Land Use Land Cover (LULC) using remote sensing data, specifically the Synthetic Aperture Radar (SAR) imagery (VV, VH, and VV-VH stack) obtained from Google Earth Engine (GEE). LULC classification is essential for monitoring land changes, assessing environmental conditions, and supporting urban planning and land management strategies.

**2. Data Description**

Source: Google Earth Engine (GEE).

Data Type: Synthetic Aperture Radar (SAR) data with polarization bands (VV, VH, VV-VH).

Spatial Resolution: 10m

Temporal Resolution: 2023-01-01 to 2023-12-31

Preprocessing: The data was preprocessed to ensure proper alignment and calibration for analysis. Any necessary atmospheric, geometric, or noise reduction corrections were also applied.

**2.1. Data Processing Workflow**

The following flow diagram represents the methodology followed in the study:

[Data Download] -> [Preprocessing] -> [Feature Selection] -> [K-means Classification] -> [Random Forest Classification] -> [Accuracy Assessment]

**3. Methodology**

**3.1. K-means Clustering for Training Samples**

The K-means clustering algorithm was applied to the dataset to partition the data into distinct clusters. K-means is an unsupervised classification technique that groups similar data points into clusters. The resulting clusters were analyzed and manually assigned to potential LULC classes (e.g., water, vegetation, urban). These clusters served as training samples for the subsequent supervised classification.

**3.2. Random Forest Classification**

Random Forest (RF) is an ensemble learning method that uses multiple decision trees to classify the data. Each tree in the forest is trained on a random subset of the data, and the final classification is based on the majority vote of all trees. The training samples generated from the K-means clustering process were used to train the RF model. The model then classified the entire dataset into LULC classes.

**3.3. Accuracy Assessment**

Accuracy metrics such as precision, recall, and F1-score were computed to evaluate the performance of the model. A confusion matrix was used to assess how well the model classified each land cover type and to calculate these metrics.

**4. Results**

**4.1. Accuracy Report**

The following table summarizes the accuracy metrics (precision, recall, F1-score) for each class in the Land Use Land Cover (LULC) classification:

| Class | Precision | Recall | F1-score | Support |

|-------|-----------|--------|----------|---------|

| 1 | 1.00 | 1.00 | 1.00 | 13 |

| 2 | 0.89 | 1.00 | 0.94 | 16 |

| 3 | 1.00 | 0.90 | 0.95 | 20 |

| 4 | 1.00 | 0.95 | 0.97 | 20 |

| 5 | 0.95 | 1.00 | 0.98 | 20 |

| \*\*Overall\*\* | \*\*0.97\*\* | \*\*0.97\*\* | \*\*0.97\*\* | \*\*89\*\* |

**4.2. Explanation of Metrics**

- \*\*Precision\*\*: Precision is the ratio of correctly predicted positive observations to the total predicted positives. It indicates how well the model avoids false positives. For example, for Class 1, the precision is 1.00, meaning that all instances classified as Class 1 were correct.

- \*\*Recall\*\*: Recall (also known as sensitivity) is the ratio of correctly predicted positive observations to all the actual positives. It shows how well the model can capture all instances of each class. For example, for Class 2, the recall is 1.00, meaning that all actual Class 2 instances were correctly identified by the model.

- \*\*F1-score\*\*: The F1-score is the weighted average of precision and recall, providing a balance between the two. It is especially useful when there is an uneven class distribution. For instance, for Class 4, the F1-score is 0.97, indicating a good balance between precision and recall.

- \*\*Support\*\*: This refers to the number of actual occurrences of the class in the dataset. It shows how many instances of each class were used in the evaluation.

**4.3. Overall Performance**

- \*\*Overall Accuracy\*\*: The overall accuracy of the classification model is \*\*97%\*\*, meaning that 97% of the total observations were classified correctly. This indicates that the Random Forest classifier performed well in identifying the LULC classes.

- \*\*Macro Average\*\*: The macro average of precision, recall, and F1-score is also \*\*0.97\*\*, indicating that the model's performance is balanced across all classes without bias towards any specific class.

- \*\*Weighted Average\*\*: The weighted average of the precision, recall, and F1-score is also \*\*0.97\*\*, reflecting the overall performance while taking into account the number of instances in each class.

**4.4. Conclusion**

The classification model achieved high performance across all classes, with an overall accuracy of 97%. The precision, recall, and F1-scores indicate that the model is effective at correctly identifying the various LULC classes, with very few misclassifications. The consistency in the macro and weighted averages further supports the robustness of the classification.

**5. Discussion**

- \*\*Interpretation of Results\*\*: The classification model performed well across all land cover types. The overall accuracy of 97% is impressive, and the precision, recall, and F1-scores indicate a balanced classification.

- \*\*Comparison with Ground Truth\*\*: The classification accuracy was compared with reference data, and the results show a high degree of agreement.

- \*\*Challenges\*\*: Some challenges encountered during the classification process include the potential for misclassification in areas with complex land cover types or mixed land covers.

- \*\*Potential Improvements\*\*: Future work could involve using more advanced classifiers, adding additional features (such as vegetation indices), or exploring the impact of different training sample selection techniques.

**6. Conclusion**

This study demonstrated that Random Forest classification, based on training samples obtained from K-means clustering, can effectively classify Land Use Land Cover (LULC) using remote sensing data. The achieved accuracy of 97% indicates the reliability of the method, with balanced precision and recall across the identified classes.