A PROJECT REPORT ON

CROP DISCRIMINATION USING MACHINE LEARNING AND GEE: A STUDY OF UDHAM SINGH NAGAR

M.Sc. Agriculture Analytics

Submitted By:

Koneenika Mallick – 202419008

Nidhi Chaudhary - 202419004

Guided By:

Mr. Abhishek Danodia
Scientist/Engineer – SF
(Agriculture and Soil Department)

Dr. Kamal Pandey
Scientist/Engineer - SF
(Geoweb Services, IT and Distance Learning)

Indian Institute of Remote Sensing

Indian Space Research Organization

Department of Space, Govt. of India Dehradun – 248001

Uttarakhand, India

April 2025

TABLE OF CONTENTS:

TABLE OF FIGURES

TABLE OF TABLES:

1.ACKNOWLEDGEMENT

I extend heartfelt gratitude to all those who contributed to the successful completion of my M.Sc. research under the Joint Education Program by DAIICT, IIRS, and AAU. Their unwavering support and invaluable guidance made this journey possible.

I am deeply thankful to my supervisors, Mr. Abhishek Danodia and Dr. Kamal Pandey, whose expert mentorship played a pivotal role in shaping my research and navigating me towards its completion. Their consistent support and insightful guidance were instrumental at every stage.

Finally, I am profoundly grateful to my family for their unwavering support and encouragement. Their dedication and encouraging words served as a constant source of motivation, inspiring me to persevere and accomplish my academic goals.

***

1. ABSTRACT:

Crop classification information is crucial for monitoring environmental changes, agricultural planning and sustainable development. This study evaluates the performance of optical and microwave remote sensing datasets for crop classification over Udham Singh Nagar district, Uttarakhand, India, during the cropping season from June to October 2023. Sentinel-2 optical imagery provided high resolution spectral details, while sentinel-1 microwave imagery contributed structural information unaffected by cloud cover. Ground truth data for rice and sugarcane crops were collected through field survey, while water bodies, settlements, forests and bare land were manually delineated. Five machine learning algorithms- Random Forest (RF), Support vector machine (SVM), K-Nearest Neighbors (KNN), Classification and regression (CART), and Gradient Boosted machine (GBM) – were employed. Results demonstration that combing optical and microwave data highest accuracy of 89.21% during the combined June-October period. This underscores the advantage of multi- sensor data fusion and ensemble learning for optical LULC mapping.

Keywords:

Crop Discrimination, Machine Learning, GEE, GBM, SVM, KNN , Random Forest

***

3 .INTRODUCTION:

Monitoring crop classification and discrimination along with land use and land cover dynamics is essential for agricultural management, urban planning, biodiversity conservation and disaster response. Remote sensing provides frequent, wide-area observations, enabling timely crop mapping.

Optical sensors like sentinel-2 capture surface reflectance across multiple bands but are limited by cloud cover. Microwave sensors like Sentinal-1 penetrate clouds and provide surface structure information independent of weather conditions. Combining these datasets offer a more complete and reliable view.

Machine learning classifiers, particularly ensemble methods like RF and GBM have proven effective in handling complex, high – dimensional remote sensing data. This study integrates sentinel -1 and sentinel -2 data with multiple machine learning models to produce detailed crop discrimination and LULC maps for the agricultural season of Utham Singh Nagar district.

***

4. Objective:

The objective of this project is to develop an accurate and robust crop discrimination framework by integrating multi-temporal optical (Sentinel-2) and microwave (Sentinel-1) remote sensing datasets using machine learning algorithms within the Google Earth Engine (GEE) platform. Specifically, the project aims to:

- Assess the effectiveness of optical and microwave data individually and in combination for crop classification over the Udham Singh Nagar district during the 2023 agricultural season.
- Implement and compare the performance of five machine learning classifiers — Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Classification and Regression Tree (CART), and Gradient Boosted Machine (GBM).

- Enhance classification accuracy through ensemble learning techniques based on majority voting.
- Analyze seasonal variations using vegetation and urban indices (NDVI, EVI, NDBI) for improved land use and land cover (LULC) discrimination.
- Demonstrate the advantages of multi-sensor data fusion and machine learning for operational crop mapping, contributing to sustainable agricultural monitoring and management practices.

***

5. Literature Review:

Crop classification using remote sensing data has gained significant momentum over the past decade, driven by advances in satellite technologies, machine learning algorithms, and cloud-based platforms like Google Earth Engine (GEE). Numerous studies have demonstrated the importance of multi-temporal and multi-sensor datasets for accurate land use and land cover (LULC) mapping, especially in complex agricultural landscapes.

Abdali et al. (2023) proposed a parallel-cascaded ensemble of machine learning models using multi-temporal Sentinel-1, Sentinel-2, and Landsat-8/9 data for crop type classification. Their work highlighted the advantages of combining optical and microwave imagery, achieving high classification accuracies, and underlined the power of ensemble learning to improve model robustness.

Similarly, Achahboun et al. (2022) utilized machine learning techniques on GEE for crop classification and emphasized the platform's ability to process large-scale remote sensing data efficiently. Their study validated that optical data alone can achieve good results, but integration with additional features significantly enhances performance, especially under cloud-prone conditions.

Shelestov et al. (2017) demonstrated the potential of GEE for large-scale crop classification by leveraging cloud computing capabilities. Their work paved the way for operational mapping by simplifying access to preprocessed datasets and implementing scalable machine learning models across vast agricultural regions.

In the context of semi-arid environments, Abubakar et al. (2023) focused on mapping maize cropland in Northern Nigeria. They emphasized the critical role of microwave data (Sentinel-1) in improving classification during periods of heavy cloud cover, validating the need for multi-sensor fusion for reliable crop monitoring.

Further, Yao et al. (2022) compared deep learning and machine learning approaches for crop classification, finding that while deep learning models offer superior performance with large annotated datasets, machine learning algorithms remain highly effective when ground truth data is limited, especially when coupled with time-series remote sensing indices.

Collectively, these studies underline several key insights:

Multi-sensor fusion (optical + microwave) consistently outperforms single-source approaches, particularly in regions affected by seasonal cloud cover.

Ensemble machine learning models, such as Random Forest and Gradient Boosted Machine, provide robust classification results across varied agricultural settings.

Vegetation and urbanization indices (NDVI, EVI, NDBI) significantly enhance the separability of land cover classes when used alongside spectral features.

Temporal analysis capturing seasonal growth dynamics improves crop type discrimination, making time-series datasets a critical asset.

Google Earth Engine provides an accessible and scalable platform for processing large datasets, enabling timely agricultural monitoring and decision support.

Building on these findings, the present study integrates optical and microwave datasets within GEE, employing multiple machine learning models and ensemble approaches to generate high-accuracy crop discrimination maps for the Udham Singh Nagar district.

***

6. STUDY AREA:

Udham Singh Nagar located in the state of Uttarakhand, India covering approximately **2536 km^{2}.** The region is agriculturally significant with major crops including rice and sugarcane. The district experiences a humid subtropical climate with heavy monsoon rains (June-September) influencing land cover dynamics. Fertile alluvial soils and seasonal waterlogging create complex land cover patterns, making it an ideal location for evaluating multi-sensor LULC mapping approaches.

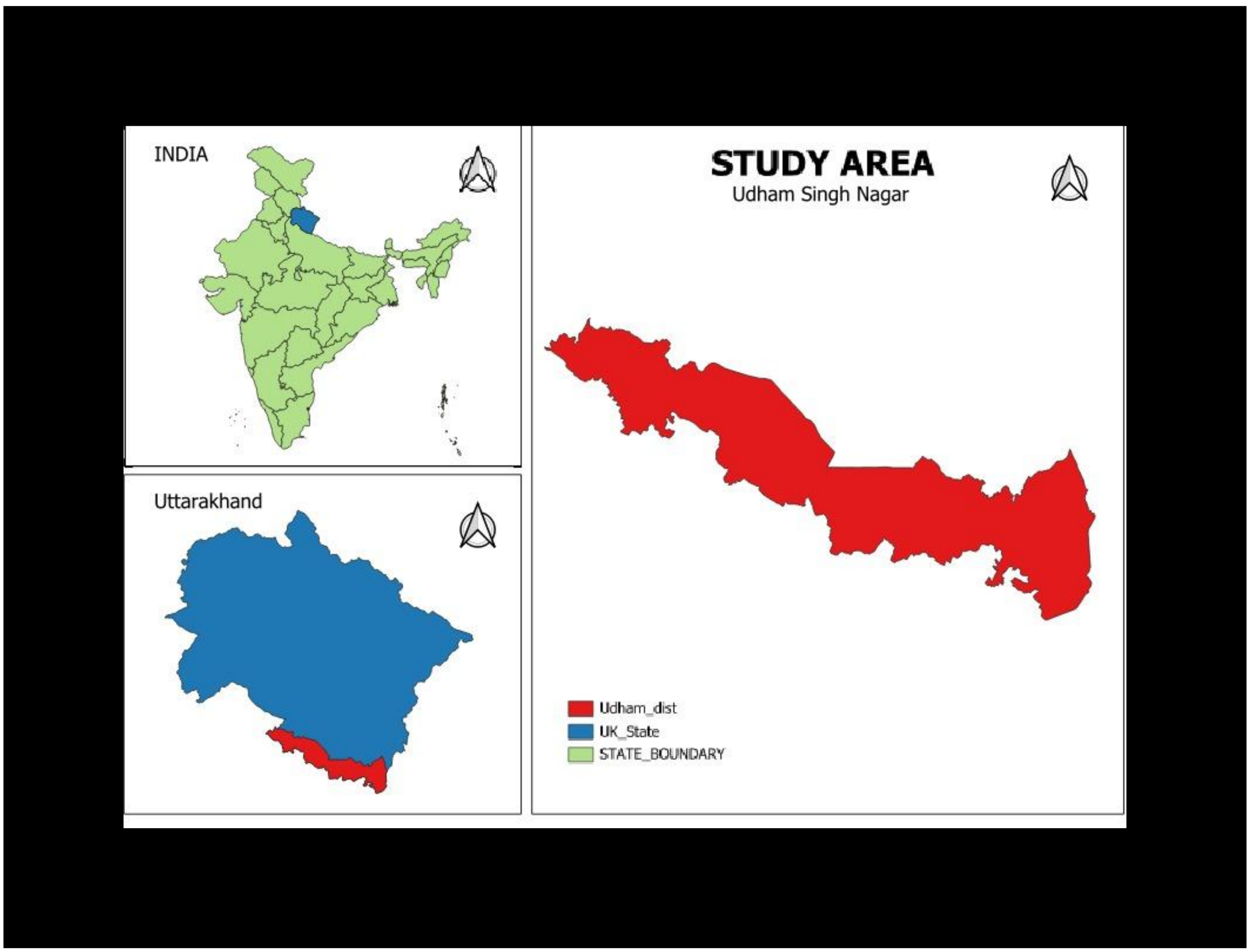

Fig.1: Study area map

***

7. DATASETS AND SOFTWARE USED:

a. Optical data(sentinal-2):

- Bands selected: B2(blue), B3(Green), B4(Near Infrared), B11(Shortwave Infrared 1), B12(Shortwave Infrared 2)
- Timeframe: June-1 to October-31, 2023

b. . Microwave data (sentinel-1):

- Polarization: VH (Vertical transmit, Horizontal receive)
- Time periods: June-1 to October-31, 2023

c. Ground Truth Data:

- Ground truth data: Obtained for rice and sugarcane crops
- Manual annotation: water bodies, settlements, forests and bareland classes delineated using high-resolution satellite imagery.

D. Software used:

- Google Earth Engine (GEE)
- Arc GIS

***

8. METHODOLOGY:

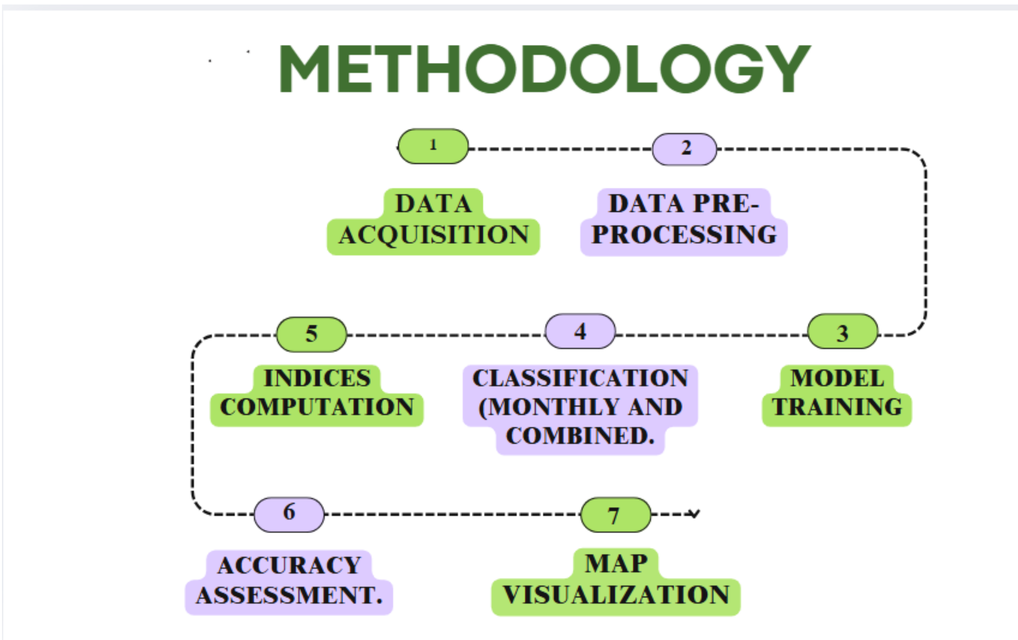

Fig.2: Methodology flowchart

8.1 Data preprocessing:

Accurate class cation requires high- quality and noise -free datasets. Therefore, a robust preprocessing workflow was adopted:

8.1.1 Optical data preprocessing:

Sentinel-2 images were filtered to retain only those with less than 30% cloud cover based on the **CLOUDY_PIXEL_PERCENTAGE** metadata. This ensured minimal atmospheric noise while maintaining temporal continuity. For each month, this dataset was generated from the filtered images to mitigate the effects of residual clouds, atmospheric variation, and outlier reflectance values.

8.1.2 Microwave data preprocessing:

Sentinel-1 Ground range detected (GRD) Products were processed focusing exclusively on the VH-polarizations. The VH band images were mosaiced monthly using a median reducer to suppress speckle noise and ensure surface consistency.

8.2 Indices Computation:

There widely used indices were computed to enhance the feature space for classification:

1. **Normalize Difference Vegetation Index (NDVI)**: Indicates the density and vigor of green vegetation, critical for crop mapping.
2. **Enhanced vegetation Index (EVI)**: Improves vegetation detection under high biomass conditions and reduces soil and atmosphere- induced noise.
3. **Normalized difference built- up Index (NDVI)**: Useful for identifying urban and bare land classes, especially as the cropping season progresses into harvest.

All derived indices were stacked along with the spectral bands to form the final feature space for classification.

8.3 Sampling splitting:

A success of any supervised classification heavily relies on the quality and representativeness of the training data.

- **70%** of samples were randomly selected for training the machine learning models.
- **30%** of samples were reserved for independent testing and validation.

This approach ensured that each classification was proportionally and adequately represented in both the training and testing datasets.

8.4 Accuracy Assessment:

The performance of the classifiers was evaluated using a confusion matrix and standard accuracy metrics, including Overall Accuracy, User's Accuracy, Producer's Accuracy, and the Kappa Coefficient. These measures provided a comprehensive assessment of classification reliability, completeness, and agreement with reference data.

8.5 Classification workflow:

Five ML classifiers were implemented separately to evaluate and compare classification performances:

- **Random Forest (RF):** An ensemble method using multiple decision trees with bootstrap aggregation, robust against overfitting and effective with high dimensional remote sensing data.
- **Support Vector Machine (SVM):** A non-parametric classifier that constructs hyperplanes in a high dimensional space useful for handling complex class boundaries.
- **K-Nearest Neighbors (KNN):** A simple yet effective algorithm relying on the majority vote of nearest neighbors for classification.
- **Classification and regression Trees (CART):** Decision tree-based algorithm using a hierarchical structure to model class distribution.
- **Gradient Boosted Machine (GBM):** A learning technique that builds models sequentially where each subsequent model attempts model attempts to correct the errors of the previous ones.

Classification procedure:

Each classifier was trained and tested separately for:

- Optical – only dataset
- Optical+ microwave - combined dataset

The models were applied for:

  - Each month individually (June, July, august, September, October)

- Combined periods (June-July, June- august, June-September, June-October)

This allowed for both temporal trend analysis and understanding the impact of data integration on classification accuracy.

8.6 Ensemble Classification:

In addition to individual classifier, an ensemble model was generated by performing a majority voting across the outputs of RF, SVM, CART, and GBM classifiers. Each pixel was assigned the most frequent class label among the five classifiers. This ensemble approach aimed to leverage the strengths of multiple algorithms to produce a more robust and accurate land cover map.

***

9. RESULT AND DISCUSSION:

9.1 Monthly Analysis of Indices:

Vegetation indices (NDVI, EVI) and urbanization index (NDBI) were analyzed for each month from June to October 2023 to assess land surface changes:

- **NDVI (Normalized Difference Vegetation Index):**

  NDVI values showed a steady increase from June to a peak in August, highlighting the vigorous vegetation growth during the monsoon season. In September and October, NDVI slightly declined, consistent with the harvesting period and exposure of bare soil.

- **EVI (Enhanced Vegetation Index)**:

  EVI followed a similar seasonal trend, increasing through July and august and reaching its maximum in September. EVI provided a sharper detection of dense vegetation compared to NDVI, particularly during peak biomass periods.

- **NDBI (Normalized Difference Built-up Index):**

NDBI remained low during the initial months, reflecting dominant vegetation cover. However, it increased sharply in September and October, indicating the exposure of built-up or bare surface post-harvest.

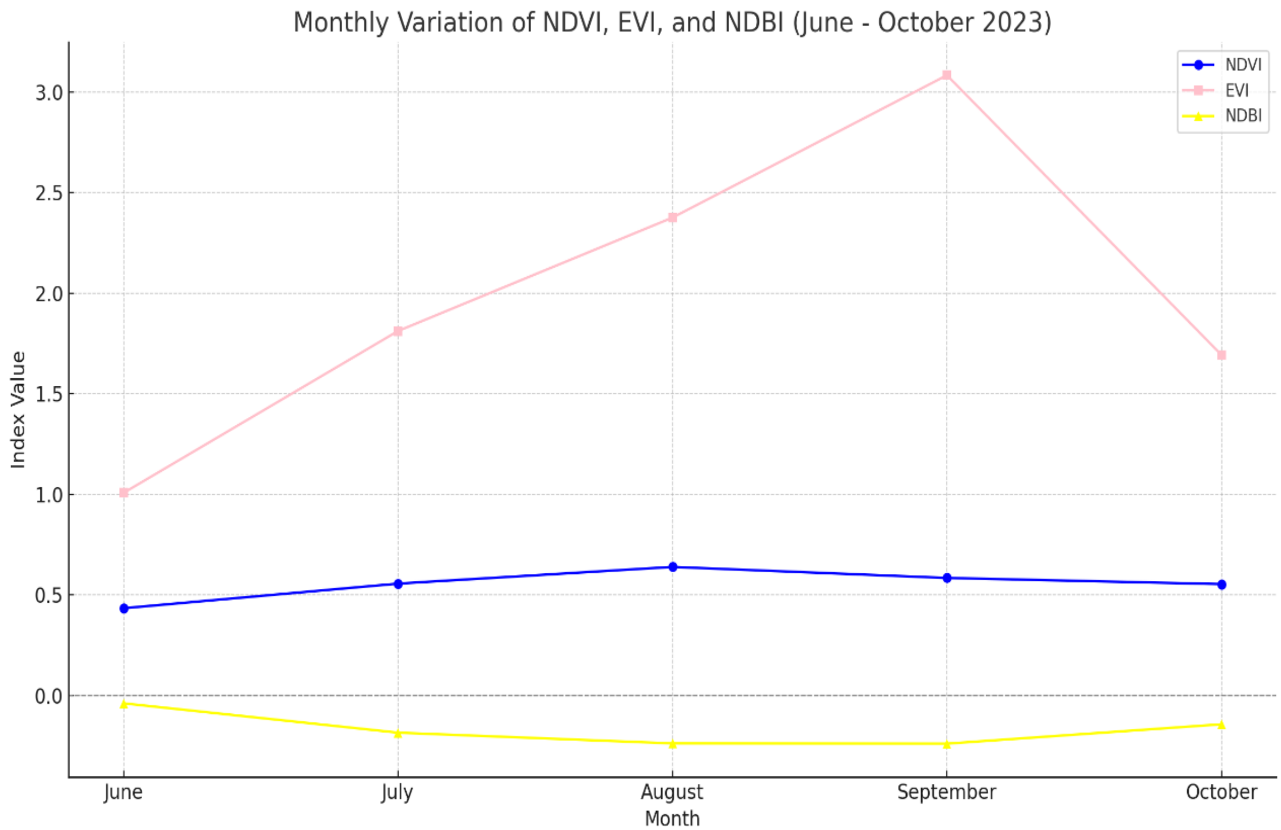

Fig.3: Monthly variation graph of indices

INFERENCE:

The seasonal trends of NDVI, EVI and NDBI clearly aligned with the agricultural cycles in Udham Singh Nagar. July and August represented peak vegetation vigor, whereas September and October marked transitional phases with increased bare land, aiding discrimination between different land cover classes in subsequent classification tasks.

9.2 Optical Data Accuracy:

Month	RF Accuracy (%)	SVM Accuracy (%)	KNN Accuracy (%)	CART Accuracy (%)	GBM Accuracy (%)
June	75.91	69.68	75.69	69.68	76.77
July	86.86	81.51	85.08	83.07	88.42
August	85.53	73.24	82.01	79.61	84.87
September	82.99	66.21	77.55	75.28	82.08
October	83.04	72.39	80.65	78.04	85.21
June+July	76.11	68.51	76.95	69.75	78.19
June+July+August	80.65	73.89	77.62	75.52	82.98
June+July+August+September	84.54	72.18	81.23	75.28	85.43
June to October	85.99	82.54	85.13	78.66	85.78

Table.1. optical data accuracy

INFERENCE:

- Across all months, the GBM classifier consistently achieved the highest accuracy, peaking in July at **88.42%.** RF closely followed particularly strong during the peak growing season (June- August). SVM underperformed relative to other models, particularly during transitional months like September where spectral likely increased due to harvesting actives.
- Aggregating multiple months improved classification performance significantly. The longer temporal composites captured both phenological development and harvesting offering richer spectral variation for the classifiers. GBM remained the top performer overall, while RF achieved excellent performance for the full season (June-October).

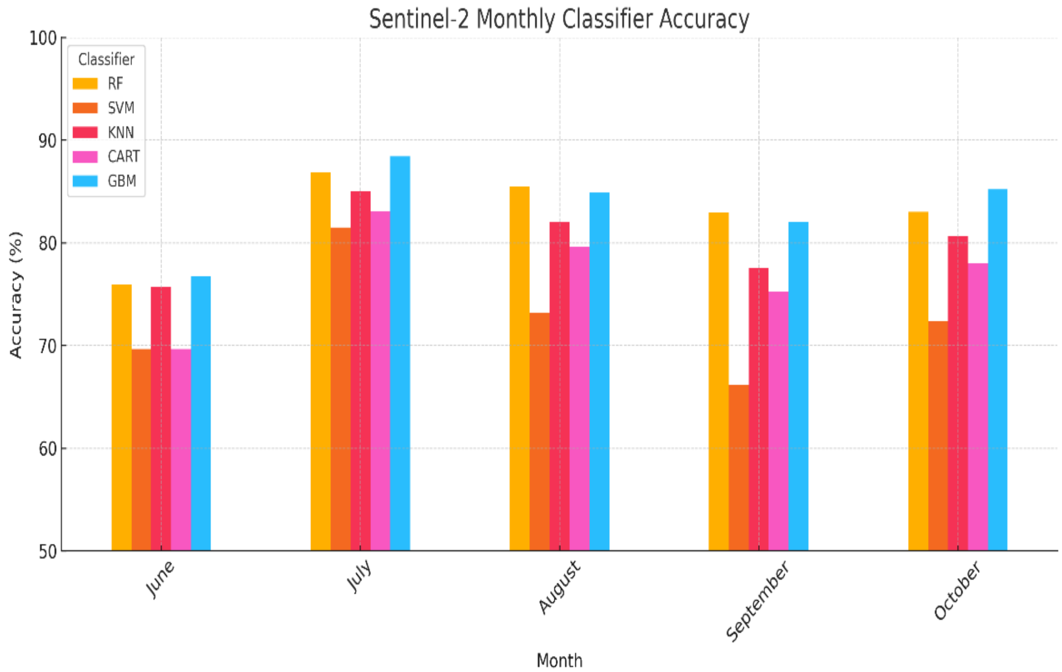

Fig.4. sentinel-2 dataset monthly classifier accuracy

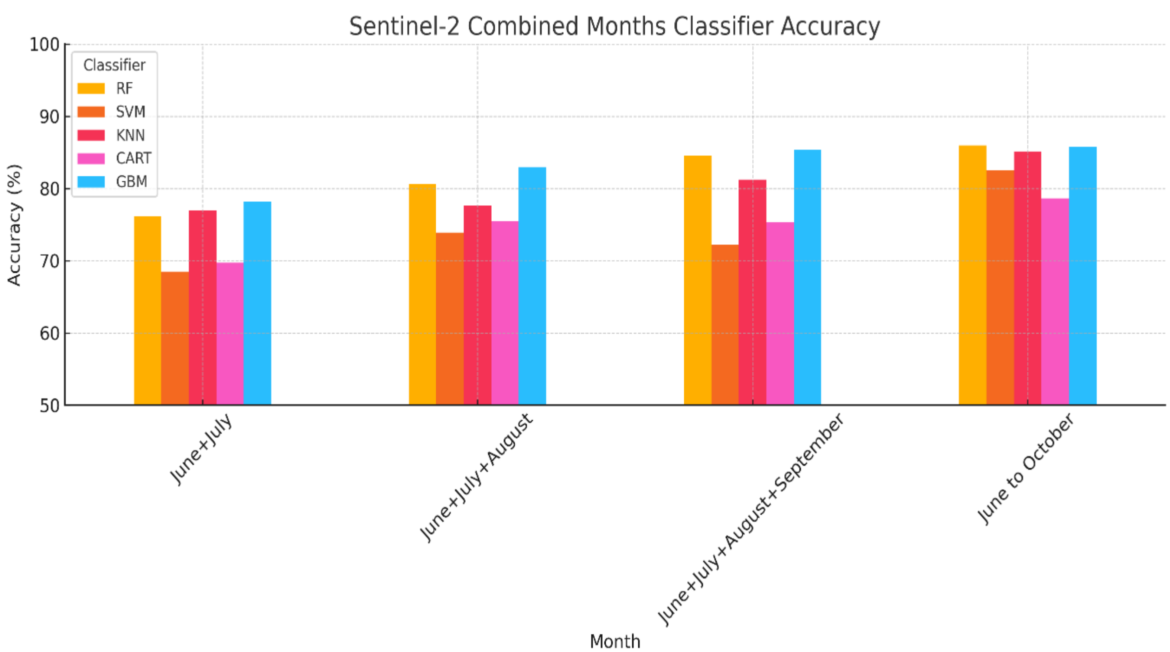

Fig.5: Optical monthly and combined periods accuracy graph

9.3 Optical + Microwave Accuracy:

Month	RF Accuracy (%)	SVM Accuracy (%)	KNN Accuracy (%)	CART Accuracy (%)	GBM Accuracy (%)
June	77.97	73.65	71.27	70.41	77.54
July	88.27	78.89	85.71	85.5	88.49
August	87.1	56.77	80.65	81.94	88.25
September	82.98	74.58	78.36	79.2	84.03
October	80.96	74.9	78.45	77.4	83.47
June+July	82.7	79.11	76.37	80.8	83.12
June+July+August	87.85	85.93	77.18	84.43	88.27
June+July+August+September	89.18	82.78	81.01	83.22	88.14
June to October	89.21	80.95	83.45	86.17	88.66

Table.2. Optical and Microwave accuracy

INFERENCE:

- The Fusion of optical and microwave data improved monthly accuracies by **2-5%** compared to optical only datasets, particularly under challenging information effectively complemented spectral data, enabling better discrimination mature crops, harvested fields and settlements.
- GBM again outperformed all other classifiers, peaking at 88.49% in July, followed closely by RF.
- The highest overall accuracy **(89.21%)** was achieved by Random Forest (RF) when using the full June-October optical + microwave information significantly enhances model robustness, especially over longer periods encompassing complex land cover transitions.

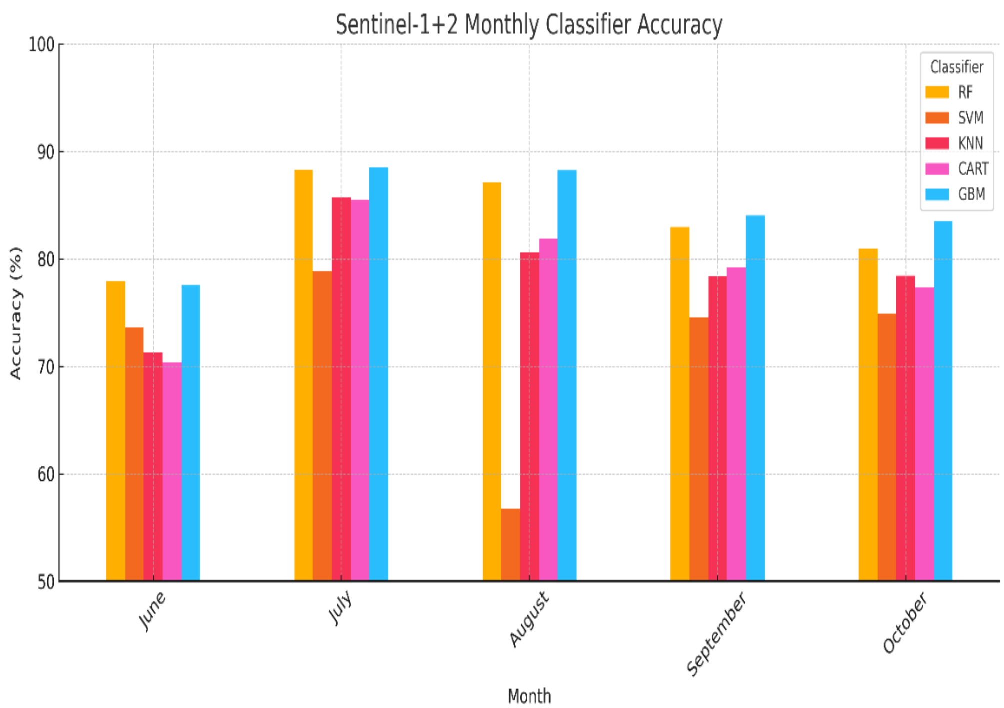

Fig.6. sentinel-1 & sentinel-2 monthly classifier accuracy

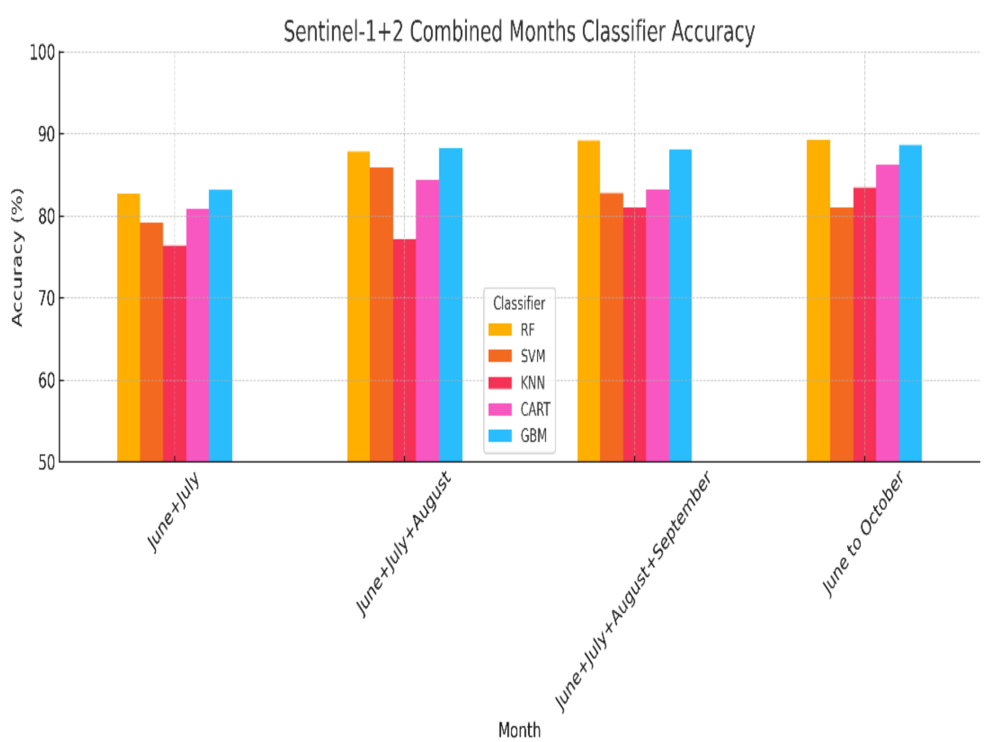

Fig.7. Optical + microwave monthly and combined periods accuracy graph

9.4 Area estimation:

- Rice dominated the agricultural landscape, covering approximately **1,77,749 hectares ,** accounting for nearly **98%** of the classified crop area.
- Sugarcane occupied around 2838 hectares, representing about **2%** of the total area.

9.5 Ensemble classification:

- The ensemble classification, created through majority voting among RF, SVM, KNN, CART, and GBM outputs achieved an overall accuracy of **90.25%** demonstrating the strength of integration multiple machine learning models for LULC mapping.

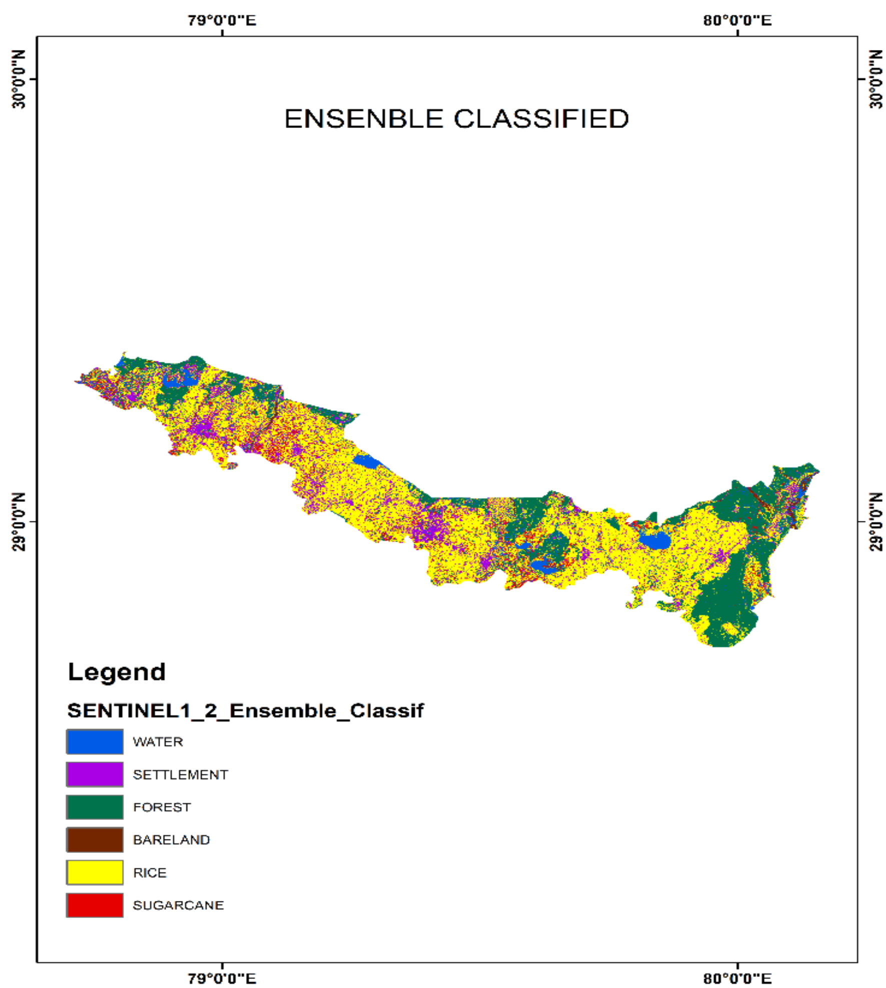

Fig.8: Ensemble map visualization

9.6 Highest-performing classifier maps visualization:

- The final maps generated using the highest- performing classifiers and the ensemble model clearly depict the distribution of major land covers including rice fields, sugarcane plantations, water bodies, forests, settlements, and bare lands across the Udham Singh Nagar district.

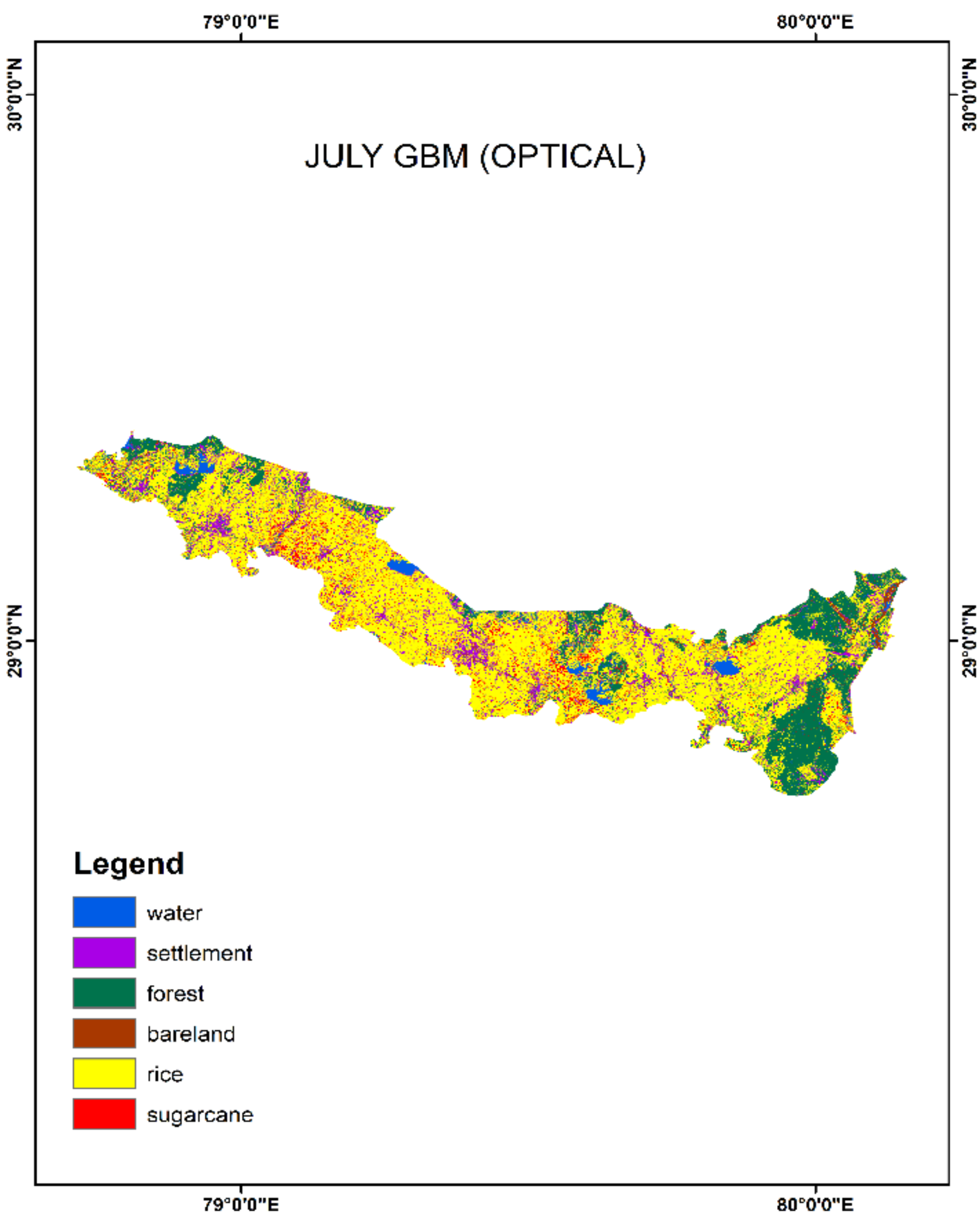

Fig.9.July GBM (optical) map visualization

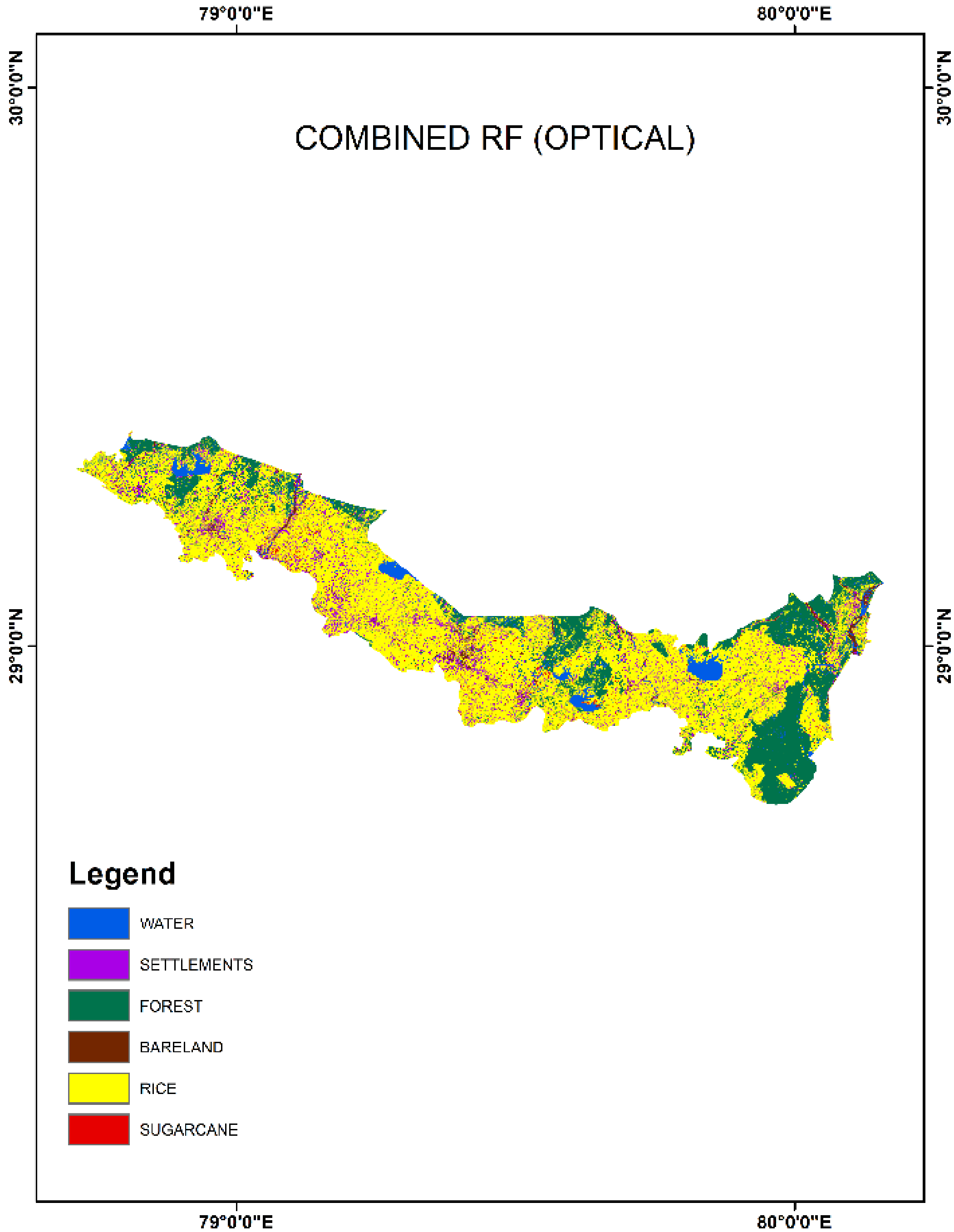

Fig.10.Combined (optical) map visualization (June-October)

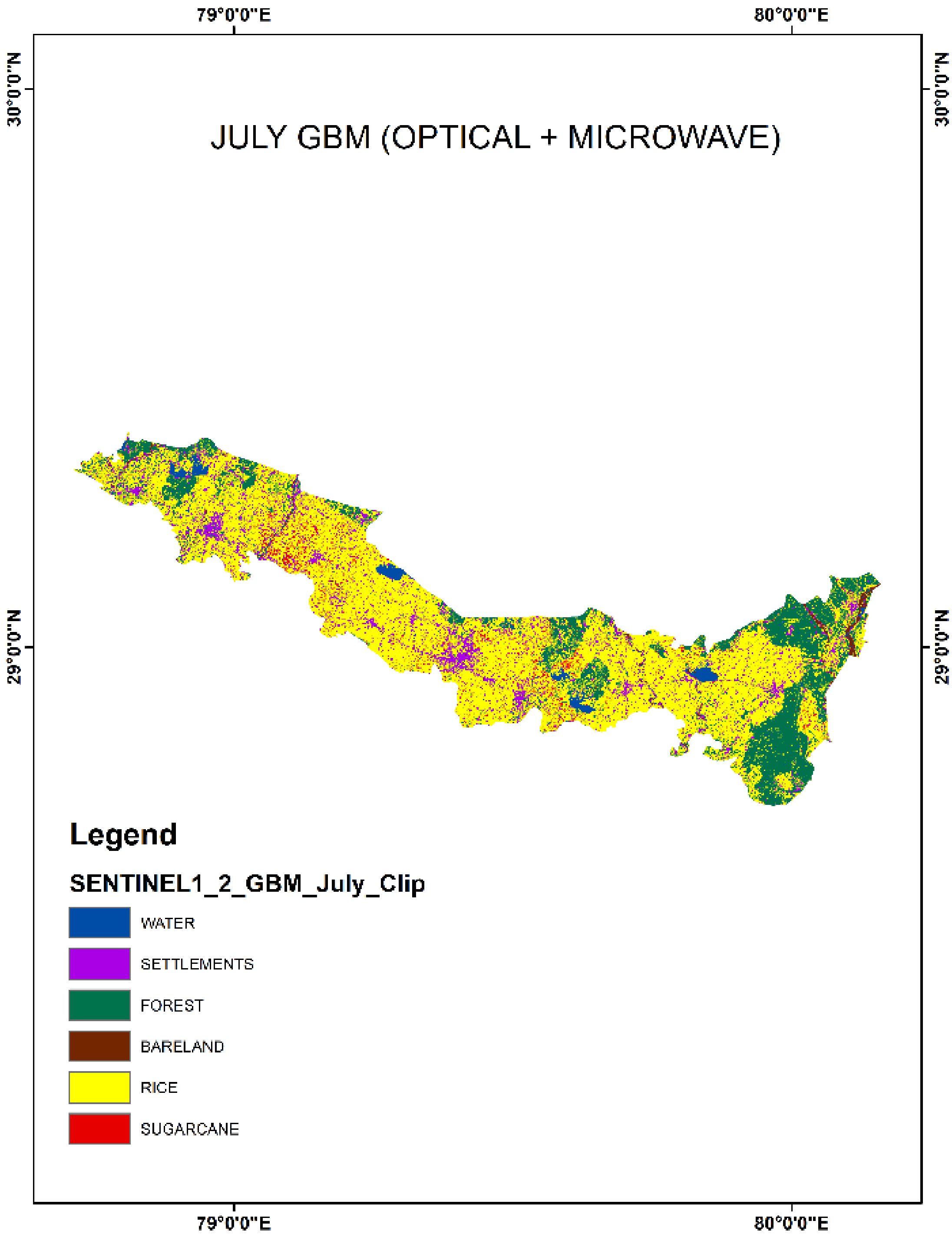

Fig.11.July GBM (optical+ Microwave) map visualization

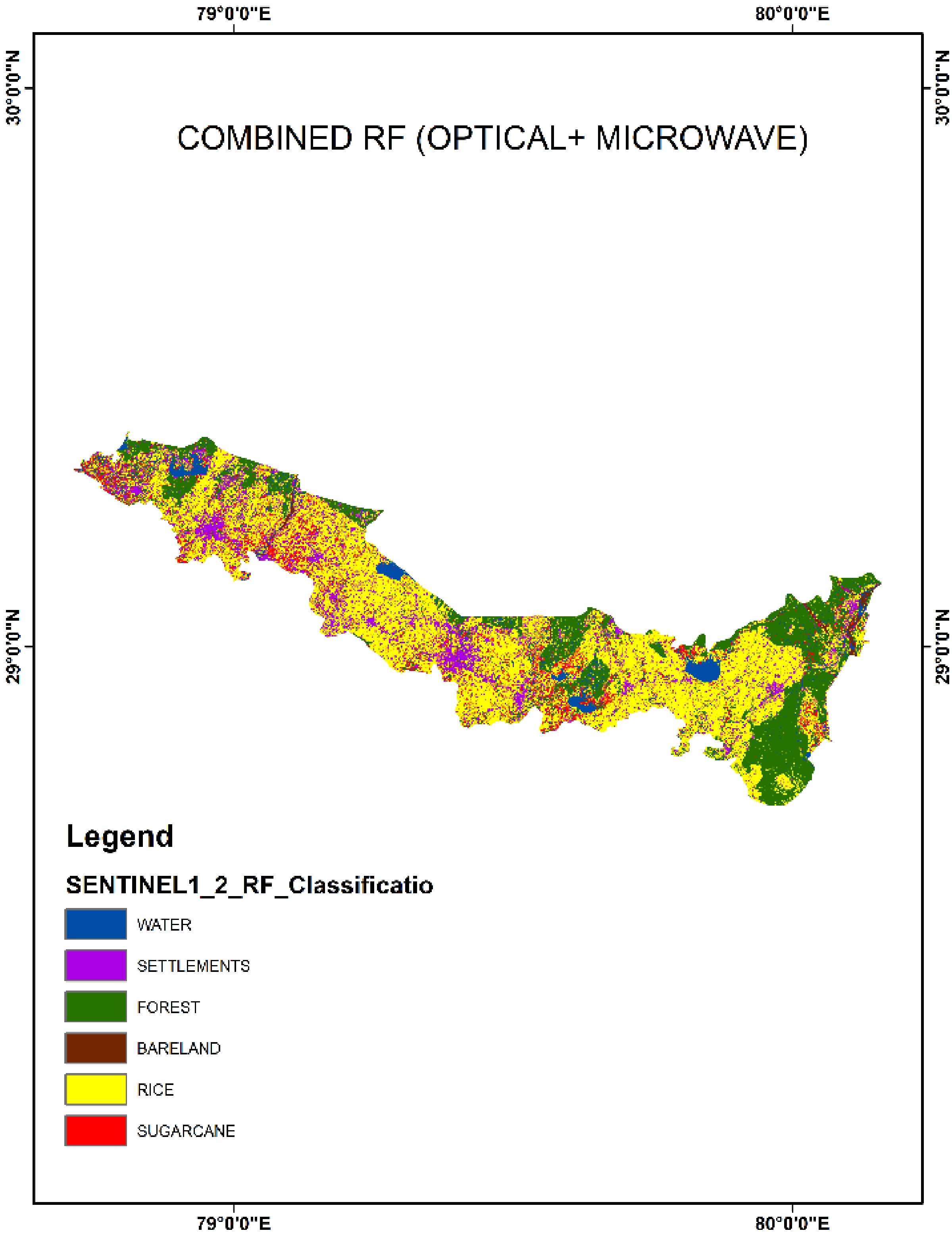

Fig.12. Visualized maps of highest performing classifiers

***

10.CONCLUSION:

This study comprehensively evaluated the effectiveness of integrating Optical and microwave remote sensing datasets for multi-temporal crop and land use/land cover(LULC) classification in Udham Singh Nagar district, Uttarakhand, India, during the 2023 agricultural season. By leveraging a combination of sentinel-2 optical data and sentinel-1 microwave backscatter information, classification accuracies improved notably compared to optical – only approaches.

Among the five machine learning classifiers tested- Random Forest(RF), Support Vector Machine(SVM), K-Nearest Neighbors(KNN), Classification and Regression Tree(CART), and Gradient boosted machine(GBM)- ensemble methods, particularly RF and GBM, Consistently and cumulative datasets. The highest overall accuracy achieved was 89.21% using the Random Forest classifier over the June- October period when combining optical and microwave datasets.

Seasonal vegetation dynamics, as captured through NDVI, EVI and NDBI closely followed known agricultural cycles, underscoring the value of time specific indices for improving classification separability. The integration of microwave data, less affected by atmospheric conditions was particularly valuable during cloud- prone monsoon months and in discriminating post-harvest bare fields and built-up areas.

Overall, the findings emphasize the critical role of multi- sensor data fusion and ensemble machine learning models in enhancing the reliability. Robustness and applicability of LULC mapping for operational and scientific purposes.

***

11.FUTURE WORK:

While the study yielded promising results, further advancements can be pursued to optimize classification performance and broader applicability.

- **Incorporate Deep Learning Models:**

  Future research can explore the use of deep learning architectures such as Convolution Neural Networks(CNNs), recurrent Neural Networks(RNNs) and transformance in handling complex spatial-temporal patterns in remote sensing imagery.

- **Apply time- series and phenology based classification:**

  Instead of using monthly median composites, implementing continuous time- series analysis could capture subtle intraseasonal phenological changes, improving crop specific discrimination.

- **Use Higher- Resolution Datasets:**

  Datasets with finer spatial resolution can be integrated to enhance classification granularity, particularly for smallholder agriculture and fragmented land cover classes.

- **Conduct Multi- year Studies:**

  Expanding the analysis to cover multiple years would enable the assessment of inter-annual variability, crop rotation patterns, and long -term land use changes, providing insights critical for sustainable land management.

- **Expand Ground Truth Data and semi- supervised learning:**

  Increasing the volume and diversity of ground truth samples through participatory mapping or citizen science can improve model generalization. Semi supervised or self-supervised learning techniques can further reduce the reliance on expensive field data collection.

***

12.REFERENCES:

- Abdali, E., Valadan Zoej, M. J., Taheri Dehkordi, A., & Ghaderpour, E. (2023). A Parallel-Cascaded Ensemble of Machine Learning Models for Crop Type Classification in Google Earth Engine Using Multi-Temporal Sentinel-1/2 and Landsat-8/9 Remote Sensing Data. Remote Sensing, 15(4), 1063.
- Achahboun, C., Chikhaoui, M., Naimi, M., & Bellafkih, M. (2022). Crops Classification Using Machine Learning and Google Earth Engine. Proceedings of the 2022 International Conference on Intelligent Systems and Computer Vision (ISCV), 1–6.
- Shelestov, A., Lavreniuk, M., Kussul, N., Novikov, A., & Skakun, S. (2017). Large-Scale Crop Classification Using Google Earth Engine Platform. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 777–780.
- Abubakar, G. A., Wang, K., Koko, A. F., Husseini, M. I., Shuka, K. A. M., Deng, J., & Gan, M. (2023). Mapping Maize Cropland and Land Cover in Semi-Arid Region in Northern Nigeria Using Machine Learning and Google Earth Engine. Remote Sensing, 15(11), 2835.
- Yao, J., Wu, J., Xiao, C., Zhang, Z., & Li, J. (2022). The Classification Method Study of Crops Remote Sensing with Deep Learning, Machine Learning, and Google Earth Engine. Remote Sensing, 14(12), 2758.

***