Application of Machine Learning Models for Rice Crop Mapping and Area Estimation Using Satellite Data of Fatehabad District (Haryana)

Dr. Rina[1], Dr. Rohtas Godara[2]
[1]Assistant Professor, Department of IT, GGDSD College, Chandigarh
[2]Associate Professor, Department of Geography, Govt. PG College, Panchkula

Abstract: *Accurate crop classification in large agricultural landscapes requires computationally efficient algorithms, scalable cloud infrastructure, and high-dimensional feature processing. This study presents a cloud-based geospatial machine learning framework implemented on the Google Earth Engine (GEE) platform. The proposed workflow leverages distributed computing architecture, parallelized pixel-wise processing, server-side JavaScript APIs, and optimized memory management for handling multi-temporal satellite big data. A high-dimensional feature space was constructed using multi-source remote sensing inputs, and supervised ensemble learning was applied with hyperparameter tuning (number of Crop signature = 300, bootstrap aggregation enabled, random feature selection per split). The study area, Fatehabad district (29.23°N–29.80°N and 75.13°E–75.98°E), covers 252,573.5 hectares within the Indo-Gangetic alluvial plains. The region is characterized by semi-arid climatic conditions and experienced a severe meteorological drought in 2017, with monsoon rainfall (137.8 mm) showing a 51% deficit from the long-period average, thereby intensifying dependence on groundwater irrigation for rice cultivation.*

The methodology integrates multi-temporal Sentinel-2 optical imagery (spectral bands and indices such as NDVI and LSWI)) with Sentinel-1 SAR data (VV and VH polarizations) through an image stacking and data fusion framework. A supervised Machine Learning Classification was performed using the Random Crop (RC) algorithm, a non-parametric ensemble learning technique capable of handling multi-modal datasets without assuming normal data distribution. The classifier was parameterized with 300 decision Crop signature point and trained using 80% of the sampled dataset, while 20% was reserved for independent validation.

Spatial area estimation results indicate that rice occupied 142,464.6 hectares (56.4%) and cotton covered 65,419.3 hectares (25.9%), together accounting for over 82% of the total geographical area. The study demonstrates that multi-sensor data fusion combined with machine learning provides a robust, scalable, and high-precision framework for crop mapping and acreage estimation in drought-prone agricultural landscapes.

Keyword: *Rice Crop, Random Crop (RC), Machine Learning, Google Earth Engine (GEE), Multi-Sensor Data, Remote sensing.*

I. Introduction:

Agriculture remains the cornerstone of the Indian economy, particularly in states like Haryana, where accurate monitoring of crop distribution is vital for food security and resource management. However, traditional methods of mapping crop types and land cover over large, heterogeneous landscapes are often labor-intensive and limited by temporal frequency. In recent years, the paradigm of remote sensing has shifted from local, desktop-based processing to planetary-scale cloud computing, enabling the analysis of massive datasets with unprecedented speed (Gorelick et al., 2017). This shift is particularly relevant for 2017, a pivotal year where the convergence of open-access satellite data and advanced computing platforms began to redefine agricultural monitoring.

The emergence of Google Earth Engine (GEE) has democratized access to high-performance geospatial analysis. As noted by Gorelick et al. (2017), GEE provides a parallel processing architecture that allows researchers to bypass the constraints of local storage and computing power. This capability has been instrumental in processing dense time-series data for land cover change detection, allowing for the derivation of phenological metrics crucial for distinguishing vegetation types (Azzari & Lobell, 2017). For a district like Fatehabad, which experiences dynamic seasonal cropping patterns, utilizing such cloud-based platforms facilitates the rapid integration of multi-temporal imagery necessary to capture phenological stages.

A significant challenge in optical remote sensing is data discontinuity caused by cloud cover, particularly during the *Kharif* season. To address this, contemporary research has increasingly focused on integrating multi-sensor data, such as Landsat-8 and Sentinel-2, to improve classification accuracy. Xiong et al. (2017) demonstrated the efficacy of combining these datasets on the GEE platform, successfully generating nominal 30-m cropland extent maps by integrating pixel-based and object-based algorithms. Similarly, Huang et al. (2017) highlighted that utilizing the full archive of available Landsat images within GEE significantly improves the mapping of major land cover dynamics by filling temporal gaps and providing a more robust spectral history for each pixel.

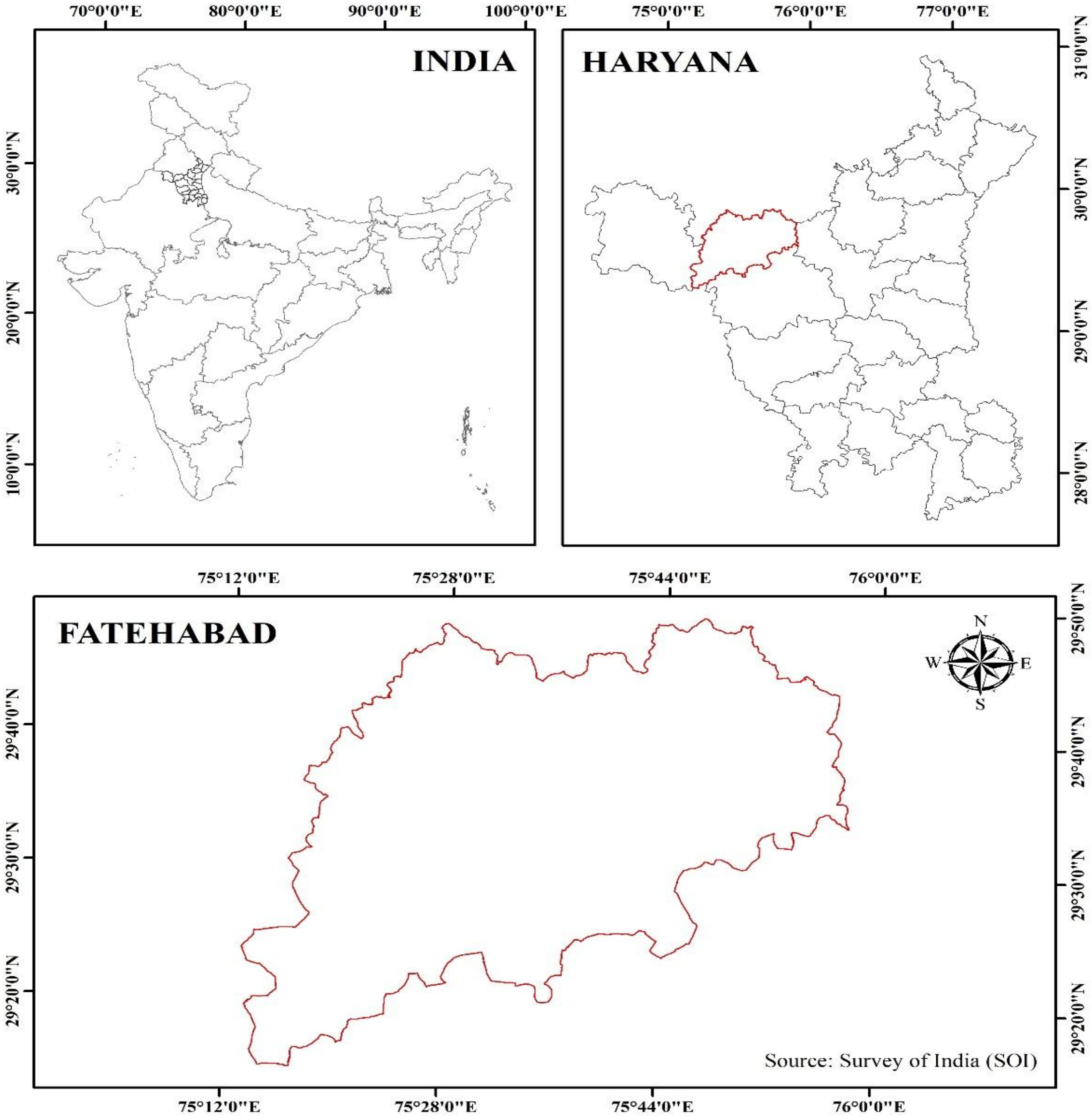
Parallel to advancements in data availability is the evolution of classification algorithms. Machine learning (ML) and Deep Learning (DL) approaches have shown superior performance over parametric classifiers like Maximum Likelihood. Kussul, Lavreniuk, Skakun, and Shelestov (2017) demonstrated that deep learning models, specifically Multi-Level Deep Learning (MLDL) architectures, outperform standard Random Forest and Support Vector Machine classifiers in distinguishing complex crop types. Furthermore, in a comparison of pixel-based approaches using GEE, Kussul et al. (2017) emphasized that advanced ML classifiers are essential for handling the high-dimensional data resulting from multi-sensor integration. Hird et al. (2017) further corroborated the utility of probabilistic machine learning in complex environments, noting its effectiveness in large-area mapping where spectral confusion between classes (such as wetlands or inundated crops) is common.

This study focuses on Fatehabad, Haryana, for the year 2017, leveraging these advancements to address local agricultural challenges. By adopting the frameworks established by Gorelick et al. (2017) and Azzari and Lobell (2017), this research utilizes the GEE platform to process multi-sensor data. Drawing on the methodologies of Xiong et al. (2017) and Kussul, Lavreniuk, Skakun, and Shelestov (2017), the study aims to implement robust machine learning classification to accurately inventory crop acreage. This approach not only overcomes the limitations of single-date imagery but also establishes a reproducible workflow for monitoring agricultural dynamics in the region.

II. Study Area:

The study was conducted in the Fatehabad district of Haryana, India, geographically situated between 29.23°N and 29.80°N latitudes and 75.13°E and 75.98°E longitudes. The district covers a total geographical area of 252,573.5 hectares and forms part of the fertile Indo-Gangetic alluvial plains. Physiographically, the region is predominantly flat, interspersed with sand dunes in the southern and western peripheries, with soils ranging from sandy to clayey loam. The climate is semi-arid and sub-tropical with distinct seasonal variations. While the district typically receives an annual average rainfall of approximately 400 mm, the study period in 2017 was marked by significant meteorological drought conditions; the cumulative monsoon rainfall (June–September) was recorded at 137.8 mm, representing a deficit of 51% compared to the long-period average (LPA). This substantial rainfall deficiency underscores the region's heavy reliance on groundwater irrigation for sustaining water-intensive Kharif crops such as rice.

Map-1: *Locational map of the Study area (Fatehabad).*

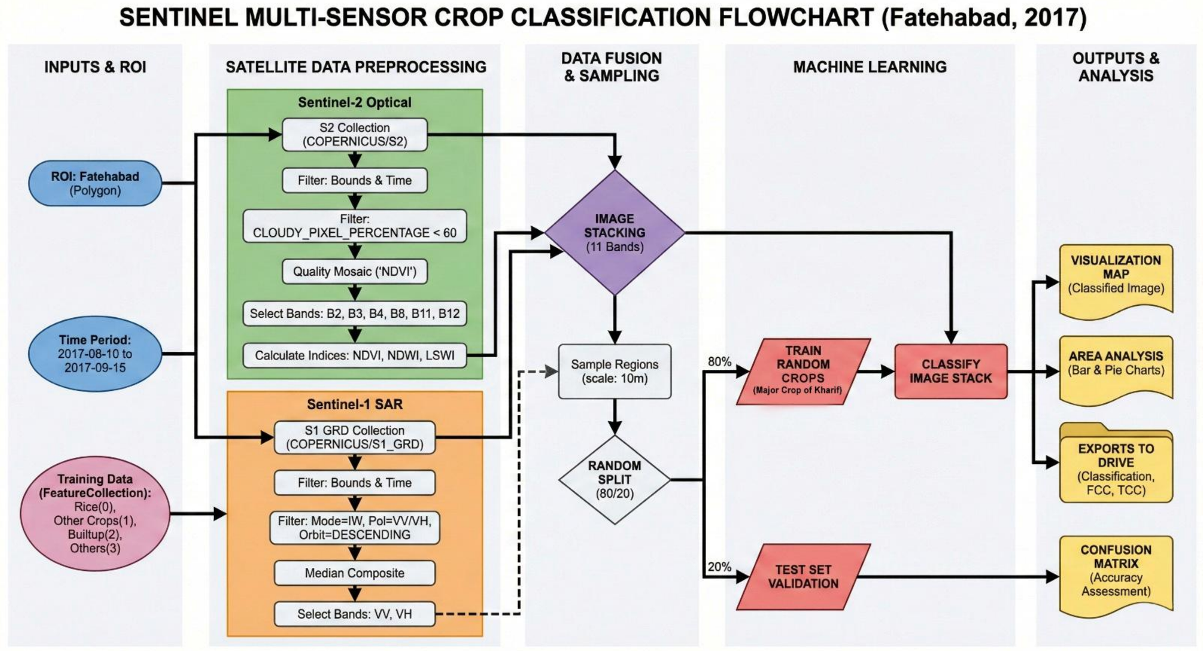


Source: Administrative boundary from SoI (Survey of India) and map composed in Q-GIS.

III. Methodology:

The study employed a cloud-based geo-computational approach using the Google Earth Engine (GEE) platform to map crop distribution in the Fatehabad district. The workflow integrated multi-sensor satellite data, spectral indices, and machine learning techniques. The overall methodological framework, illustrating the processing chain from data acquisition to accuracy assessment, is presented in Figure 1.

Fig-1: *Flow diagram of Methodology.*


SENTINEL MULTI-SENSOR CROP CLASSIFICATION FLOWCHART (Fatehabad, 2017)
INPUTS & ROI
SATELLITE DATA PREPROCESSING
DATA FUSION & SAMPLING
MACHINE LEARNING
OUTPUTS & ANALYSIS
ROI: Fatehabad (Polygon)
Time Period: 2017-08-10 to 2017-09-15
Training Data (FeatureCollection): Rice(0), Other Crops(1), Builtup(2), Others(3)
Sentinel-2 Optical
S2 Collection (COPERNICUS/S2)
Filter: Bounds & Time
Filter: CLOUDY_PIXEL_PERCENTAGE < 60
Quality Mosaic ('NDVI')
Select Bands: B2, B3, B4, B8, B11, B12
Calculate Indices: NDVI, NDWI, LSWI
Sentinel-1 SAR
S1 GRD Collection (COPERNICUS/S1_GRD)
Filter: Bounds & Time
Filter: Mode=IW, Pol=VV/VH, Orbit=DESCENDING
Median Composite
Select Bands: VV, VH
IMAGE STACKING (11 Bands)
Sample Regions (scale: 10m)
RANDOM SPLIT (80/20)
80%
20%
TRAIN RANDOM CROPS (Major Crop of Kharif)
CLASSIFY IMAGE STACK
TEST SET VALIDATION
VISUALIZATION MAP (Classified Image)
AREA ANALYSIS (Bar & Pie Charts)
EXPORTS TO DRIVE (Classification, FCC, TCC)
CONFUSION MATRIX (Accuracy Assessment)

3.1. Satellite Data Acquisition and Pre-processing

To characterize the heterogeneous agricultural landscape during the *Kharif* season (August–September), the study utilized a data fusion approach combining optical and Synthetic Aperture Radar (SAR) imagery.

3.1.1. Optical Data Processing (Sentinel-2 MSI)

Optical data were acquired from the Sentinel-2 MultiSpectral Instrument (MSI) Level-1C collection (COPERNICUS/S2). The temporal window was defined from August 10, 2017, to September 15, 2017, to capture the peak vegetative growth stages of the target crops.

Given the prevalence of cloud cover during the monsoon season, a strict pre-processing protocol was implemented:

1. **Cloud Filtering:** The image collection was initially filtered to exclude scenes with a CLOUDY_PIXEL_PERCENTAGE exceeding 60%.
2. **Quality Mosaicking:** To generate a seamless, cloud-free composite, a qualityMosaic function was applied using the Normalized Difference Vegetation Index (NDVI) as the quality band. This algorithm iterates through the image stack and retains the pixel with the highest NDVI value, effectively removing cloud and shadow artifacts while preserving vegetation information.
3. **Band Selection:** Six spectral bands were retained for the analysis: Blue (B2), Green (B3), Red (B4), Near-Infrared (B8), and Shortwave Infrared (B11, B12).

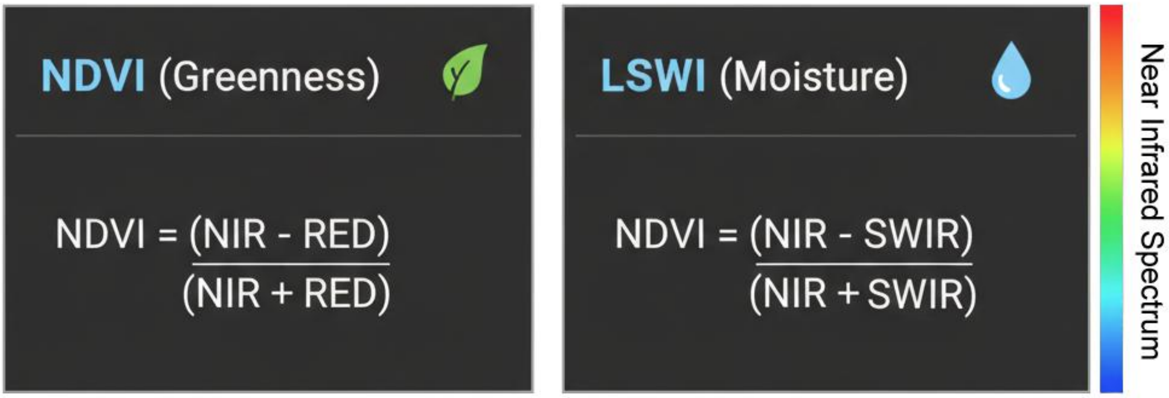
3.1.2. SAR Data Processing (Sentinel-1)

To mitigate the limitations of optical sensors in persistent cloud cover, Sentinel-1 C-band SAR Ground Range Detected (GRD) data (COPERNICUS/S1_GRD) were integrated. The SAR collection was filtered for the Interferometric Wide (IW) swath mode and dual-polarization (VV and VH) in the DESCENDING orbit pass. A temporal median composite was generated from the filtered collection. This compositing method serves as a speckle reduction technique, smoothing out random noise while retaining the stable backscattering characteristics of the underlying land cover.

3.2. Feature Engineering and Image Stacking

To enhance the spectral separability between spectrally similar classes (e.g., Rice and Other Crops), three spectral indices were computed derived from the Sentinel-2 composite:

1. **Normalized Difference Vegetation Index (NDVI):** Calculated to assess vegetation vigor and biomass.


NDVI (Greenness)
NDVI = (NIR - RED) / (NIR + RED)
LSWI (Moisture)
NDVI = (NIR - SWIR) / (NIR + SWIR)
Near Infrared Spectrum

Reference: Xiao, X et al (2006)

2. **Land Surface Water Index (LSWI):** Derived using NIR and SWIR bands to monitor leaf water content and soil moisture, which is particularly effective for identifying irrigated paddy fields.

These indices were stacked with the six optical bands and two SAR bands (VV, VH) to create a high-dimensional 11-band composite image for classification.

3.3. Training Data and Sampling Strategy

The classification scheme defined four primary land cover classes: Rice (Class 0), Cotton(Class 1), Non-Agriculture (Class 2), and Others (Class 3). Reference data were compiled into a FeatureCollection and merged to create a unified training dataset.

The sampleRegions function was employed to extract spectral and backscatter signatures from the 11-band composite at a spatial resolution of 10 meters. To ensure robust model evaluation, the sampled dataset was subjected to a stratified random split:

- **Training Set:** 80% of the samples were used to train the classifier.
- **Validation Set:** 20% of the samples were reserved for independent accuracy assessment.

3.4. Machine Learning Classification (Random Forest)

The study utilized the Random Crop (RC) algorithm, a non-parametric ensemble learning method known for its high accuracy and ability to handle multi-modal data (optical and SAR) without assuming a normal distribution.

The classifier was parameterized with 300 decisions Signature. The model was trained using the 80% training subset, mapping the relationship between the 11 input features and the target land cover classes. The trained model was subsequently applied to the entire composite image to generate a categorical land use/land cover map of the ROI.

3.5. Accuracy Assessment and Area Estimation

The reliability of the classified map was evaluated using the 20% independent validation subset. A confusion matrix was computed to derive standard accuracy metrics, including

$$PA = \frac{\text{Correctly classified samples of a class}}{\text{Total reference samples of that class}} \times 100$$

$$UA = \frac{\text{Correctly classified samples}}{\text{Total predicted samples of that class}} \times 100$$

IV. Results and Discussion

This section details the findings derived from the multi-sensor crop classification of the Fatehabad district during the *Kharif* season of 2017. The study employed a Random Crop (RC) classifier (configured with 300 decision signature) to integrate active microwave data from Sentinel-1 SAR (VV and VH polarizations) with passive optical data from Sentinel-2. This fusion approach, augmented by spectral indices was specifically designed to overcome atmospheric challenges and enhance class separability.

4.1. Acreage Estimation

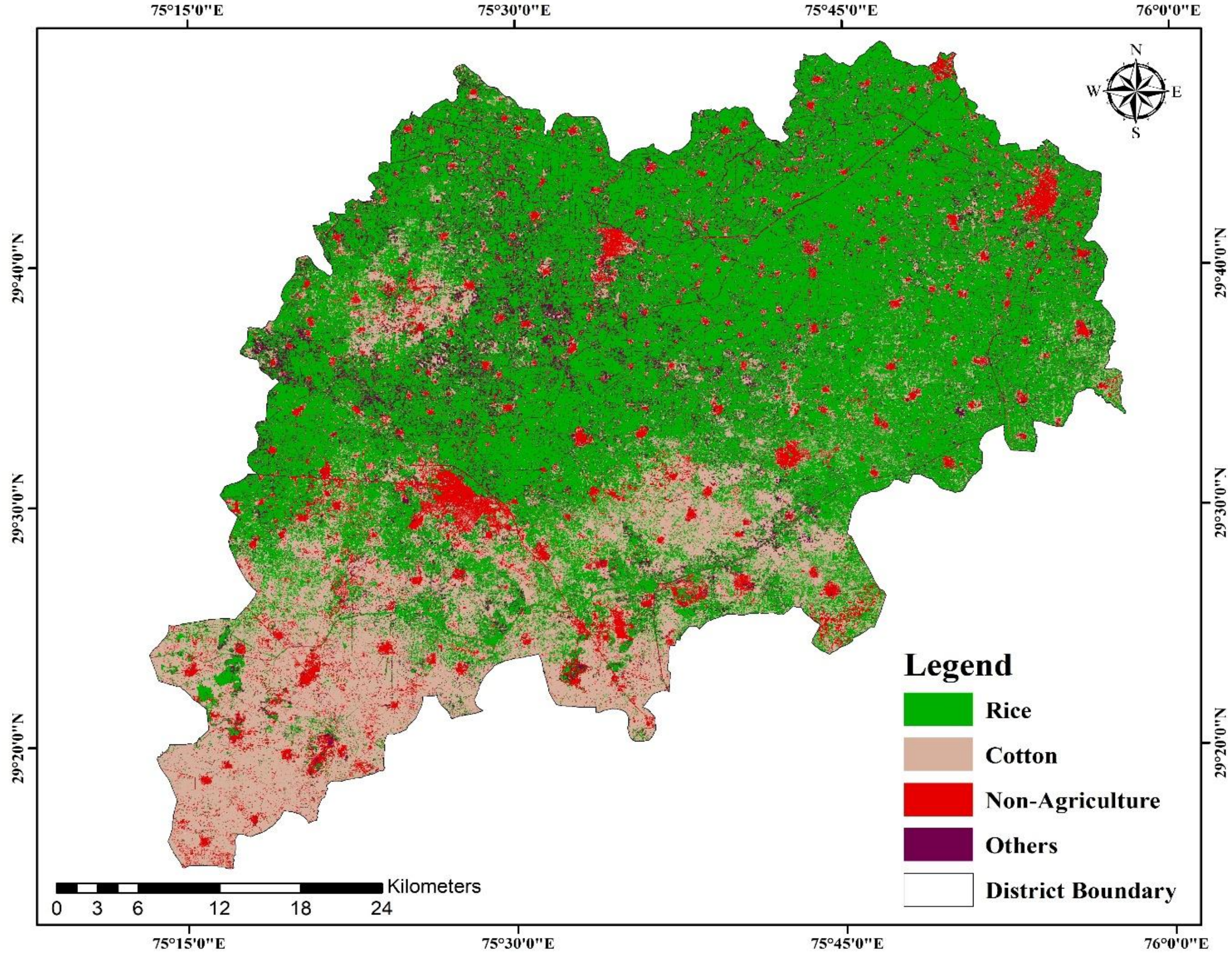
The pixel-based area statistics derived from the classification provide a quantitative breakdown of land use in the district. The total geographical area analysed was 252,573.5 hectares.

Table 1: *Estimated Area of Classified Land Cover Classes (Fatehabad, Kharif 2017).*

Sr. no.	Class	Area in Hac	Area in %
1	Rice	142464.6	56.4
2	Cotton	65419.3	25.9
3	Non-Agriculture	18440.3	7.3
4	Others	26249.3	10.4
5	Total Geographical Area	252573.5	100

The results highlight a clear monocultural dominance, with Rice occupying approximately 56.4% of the district's total area. Cotton follows as the second major crop, covering 25.9%. Together, these two crops constitute over 82% of the landscape, underscoring the intensive agricultural nature of Fatehabad during the monsoon season.

Map-2: *Spatial distribution of Kharif Crops in Fatehabad using Google Earth Engine (GEE) August 2017.*



Source: Sentinel-2A image, Processed by GEE.

4.1. Accuracy Assessment

The efficacy of the classification model was rigorously evaluated using a confusion matrix generated from an independent validation dataset. The integration of SAR backscatter with optical spectral signatures resulted in a robust Overall Accuracy (OA) of 98%, indicating a high degree of reliability in the resulting land use/land cover maps.

Table-2: Overall Accuracy, Producer's Accuracy (PA), and User's Accuracy (UA) used by Confusion Matrix.

Class	0	1	2	3
Built-up	0	1	89	0
Cotton	0	92	0	0
Others	0	0	1	14
Rice	97	0	0	1

Rows- Reference (Ground Truth) data and Columns- Predicted (Classified) data

Class-Specific Performance

- **Rice (Paddy):** The classifier exhibited exceptional performance for the Rice class, achieving a Producer's Accuracy (PA) of 100% and a User's Accuracy (UA) of 98%. This near-perfect detection is attributed to the distinct interaction between radar signals and paddy fields.

$$PA_{Rice} = \frac{97}{98} \times 100 = 98.97\% \qquad UA_{Rice} = \frac{97}{97} \times 100 = 100\%$$

  The transplanted stage of rice involves flooded fields, which significantly alters the dielectric constant of the surface. Sentinel-1's sensitivity to these structural properties—specifically the volume scattering from the crop canopy and double-bounce scattering from the interaction between the vertical stalks and the underlying water—allowed for a precise delineation of rice from other vegetation.
- **Cotton:** The Cotton class (encompassing cotton and guar) also showed strong separability, achieving a PA of 98%. While minor spectral confusion was initially observed between young cotton crops and mixed vegetation due to overlapping phenological stages early in the season, the inclusion of water-sensitive indices proved critical. The LSWI (Land Surface Water Index) and NDWI (Normalized Difference Water Index) successfully captured subtle variances in leaf water content, thereby refining the boundary between cotton and spectrally similar "Other Crops."

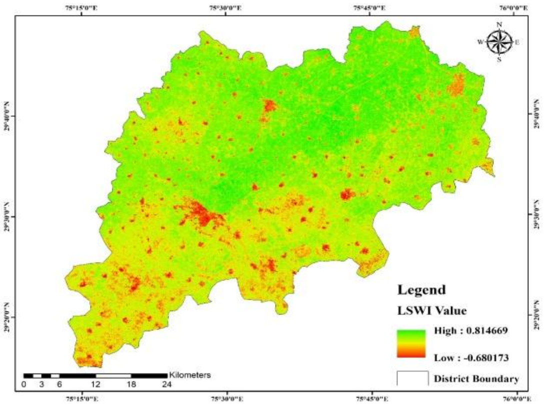
4.2. Spatial Distribution of Crops

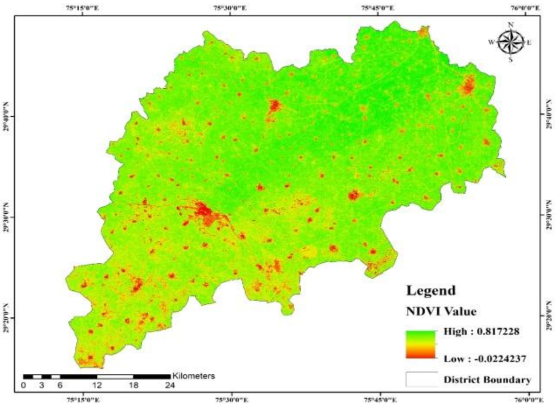
The classified map unveils a distinct spatial dichotomy in the agricultural landscape of the Fatehabad district, governed largely by hydrological and agro-climatic factors.

- **Rice Dominance (North & North-East):** Rice cultivation is heavily concentrated in the northern and north-eastern sectors of the district. This spatial clustering correlates strongly with the established canal irrigation networks and the availability of groundwater resources in these zones, which are essential for sustaining water-intensive paddy cultivation.
- **Cotton Dominance (West & South):** Conversely, the western and southern regions are dominated by the Cotton class. These areas are characterized by relatively drier conditions and a lower water table. The prevalence of cotton here reflects an adaptive agricultural strategy, leveraging the region's agro-climatic suitability for crops that are less water-demanding than rice.

Integrated Analysis of LSWI and NDVI

The LSWI and NDVI maps collectively reveal the agro-hydrological dynamics of Fatehabad district during the Kharif season of 2017. The LSWI map highlights zones of high surface moisture corresponding to irrigated rice fields, particularly in the northern and north-eastern regions. The NDVI maps demonstrate spatial variability in vegetation vigor, with high NDVI values indicating dense crop canopy and low NDVI values reflecting moisture stress or non-cropped areas. The combined interpretation of moisture-sensitive and vegetation indices provides strong evidence of irrigation-supported agricultural resilience under drought conditions and enhances crop discrimination accuracy in machine learning-based classification.





4.4. Discussion on Multi-Sensor Integration

The high accuracy achieved in this study validates the synergistic advantage of fusing active (SAR) and passive (Optical) remote sensing data.

1. **Complementary Information:** Sentinel-2 optical data provided critical information regarding crop health (via NDVI) and canopy water content (LSWI). However, optical sensors are often limited by cloud cover, a persistent issue during the Kharif (monsoon) season.
2. **The Role of SAR:** Sentinel-1 SAR data proved indispensable in mitigating the data gaps caused by cloud cover. The radar backscatter (VV and VH polarizations) is sensitive to physical structure rather than just surface color. This allowed the model to detect the unique structural evolution of crops—such as the vertical growth of rice stalks against a water background—regardless of atmospheric conditions.
3. **Impact of Indices:** The LSWI was particularly pivotal. By strictly defining the spectral signature of surface moisture, LSWI helped differentiate flooded rice paddies from non-flooded crops like cotton, acting as a critical filter that reduced misclassification errors significantly.

V. Conclusion

This study developed a reliable and high-precision framework for agricultural monitoring through the integration of active microwave Sentinel-1 SAR and passive optical Sentinel-2 satellite data, classified using the Random Forest algorithm. The multi-sensor fusion approach significantly enhanced classification performance, achieving an Overall Accuracy of 98%, and proved especially effective during the monsoon season when cloud cover limits optical imagery. SAR backscatter (VV and VH) played a crucial role in accurately identifying transplanted rice, resulting in near-perfect Producer's Accuracy, while spectral indices such as LSWI and NDWI helped reduce confusion between cotton and other vegetation types. Spatial analysis revealed that rice dominated the agricultural landscape, covering 56.4% (142,464 ha), mainly in canal-irrigated northern and north-eastern blocks, whereas cotton (25.9%, 65,419 ha) was concentrated in relatively drier western and southern regions. The findings underscore the heavy dependence on water-intensive rice cultivation and highlight the importance of precise crop mapping for supporting sustainable groundwater management and crop diversification policies in Haryana.

References:

[1]. Azzari, G., & Lobell, D. B. (2017). Land use and land cover change detection using Landsat data in Google Earth Engine. *Remote Sensing of Environment, 202*, 64–74.

[2]. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment, 202*, 18–27.

[3]. Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters, 14*(5), 778–782.

[4]. Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). *Comparison of pixel-based approaches to crop mapping in Ukraine using Google Earth Engine*. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. xxx–xxx). IEEE.

[5]. Hird, J. N., DeLancey, E. R., McDermid, G. J., & Kariyeva, J. (2017). *Machine learning in support of large-area probabilistic wetland mapping*. Remote Sensing, **9**(12), 1315.

[6]. Xiong, J., Thenkabail, P. S., Tilton, J. C., Gumma, M. K., Teluguntla, P., Oliphant, A., Congalton, R. G., Yadav, K., & Gorelick, N. (2017). *Nominal 30-m cropland extent map of continental Africa by integrating pixel-based and object-based algorithms using Sentinel-2 and Landsat-8 data on Google Earth Engine*. Remote Sensing, **9**(10), 1065.

[7]. Huang, H., Chen, Y., Clinton, N., Wang, J., Wang, X., Liu, C., Gong, P., Yang, J., Bai, Y., Zheng, Y., & Zhu, Z. (2017). *Mapping major land cover dynamics in Beijing using all Landsat images in Google Earth Engine*. Remote Sensing of Environment, 202, 166–176.

[8]. Xiao, X., Boles, S., Frolking, S., Li, C., Babu, J. Y., Salas, W., & Moore, B. (2006). Mapping paddy rice agriculture in South and Southeast Asia using multi-temporal MODIS images. *Remote Sensing of Environment, 100*(1), 95–113.