



A Data Engineering Framework for Multi-Modal Satellite Data Fusion and Land-Cover Classification Using Sentinel-1 and Sentinel-2 Imagery

¹Saima Anjum, ²Pooja M V

¹saimaanjum.fet.scst.cse@gmu.ac.in, Department of Data Engineering, GM UNIVERSITY, Davanagere – 577006, Karnataka, India.

²poojamv@gmit.ac.in, Assistant professor, Department of information Science and Engineering, School of Computer Science and Technology, Faculty of engineering, GM UNIVERSITY, Davanagere

ABSTRACT :

Land-cover classification using satellite imagery supports environmental monitoring, agricultural assessment, urban planning, and natural-resource management. Optical imagery provides spectral information about surface materials, while synthetic aperture radar (SAR) imagery provides complementary information related to surface structure and moisture and can be useful under cloudy conditions. This paper presents a data-engineering workflow for preparing and combining Sentinel-1 SAR and Sentinel-2 multispectral data in Google Earth Engine (GEE), with Davanagere District, Karnataka, India, as the study area. The workflow includes study-area selection, date and spatial filtering, image compositing, spectral feature selection, index calculation, radar-feature preparation, multi-modal feature fusion, supervised classification using Random Forest, and accuracy assessment. Sentinel-2 features include six selected spectral bands (B2, B3, B4, B8, B11, and B12) and three indices (NDVI, NDWI, and NDBI). Sentinel-1 features are planned to include VV, VH, and their difference in decibels. The evaluation design compares Sentinel-2-only, Sentinel-1-only, and fused-feature classifications using a separate validation sample and metrics such as overall accuracy, class-wise accuracy, confusion matrix, and Kappa coefficient. Initial GEE processing has confirmed 137 Sentinel-2 and 56 Sentinel-1 images meeting the configured collection filters for 2025, and the nine-band Sentinel-2 feature stack has been generated. Classification and accuracy results must be added after the remaining implementation and validation are completed.

Keywords: Google Earth Engine, Sentinel-1, Sentinel-2, Multi-Modal Data Fusion, Land-Cover Classification, Random Forest, Data Engineering.

1. Introduction

Satellite remote sensing provides repeated observations of the Earth's surface and is widely used to study vegetation, agriculture, built-up areas, water bodies, and bare land. Reliable land-cover maps can support land-use planning, ecosystem monitoring, and evidence-based regional decision-making. However, a single satellite sensor may not provide enough information to distinguish all land-cover types under different environmental and atmospheric conditions.

Sentinel-2 is an optical multispectral mission that records reflected solar radiation across visible, near-infrared, and short-wave infrared bands. These bands support the calculation of spectral indices such as the Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), and Normalized Difference Built-up Index (NDBI). Optical observations can be affected by cloud cover and atmospheric conditions. Sentinel-1 uses synthetic aperture radar and provides measurements such as VV and VH backscatter, offering complementary surface information and observations that are less dependent on daylight and cloud-free skies.

Google Earth Engine is a cloud-based geospatial processing platform that provides access to satellite collections and scalable analysis tools. In this project, GEE is used to filter and composite Sentinel-1 and Sentinel-2 imagery, derive features, and prepare a combined feature stack. The proposed experiments use Random Forest to compare land-cover classification based on each sensor independently and on the fused feature set.

The study area is Davanagere District in Karnataka, India. The target classes are forest/vegetation, agriculture, urban/built-up areas, water bodies, and bare land. The work emphasizes a reproducible data-engineering pipeline as well as a controlled comparison of the information contributed by optical and radar features.

2. Problem Statement

Land-cover classification can be challenging when different surface types have similar spectral responses, when land conditions vary across the study area, or when optical observations are affected by clouds. Relying on a single sensor may limit the information available to a classifier. In addition,

satellite data must be filtered, composited, transformed into useful features, and aligned consistently before it can be used for supervised classification.

Therefore, a structured and reproducible workflow is required to integrate Sentinel-1 radar and Sentinel-2 optical data, prepare suitable features, and evaluate whether multi-modal fusion provides a measurable benefit over single-source inputs. The performance must be assessed using independent validation samples rather than assuming that fusion automatically improves accuracy.

2.1 Objectives

- Acquire and filter Sentinel-1 and Sentinel-2 image collections for Davanagere District using Google Earth Engine.
- Prepare comparable image composites for the selected study period.
- Derive Sentinel-2 spectral features and NDVI, NDWI, and NDBI indices.
- Prepare Sentinel-1 VV and VH radar features and a VV–VH difference feature.
- Combine optical and radar features into a multi-modal feature stack.
- Train and compare Random Forest classifiers using Sentinel-2-only, Sentinel-1-only, and fused features.
- Evaluate classification using a confusion matrix, overall accuracy, class-wise accuracy, and Kappa coefficient.
- Document the data-engineering workflow so that the processing steps can be reproduced.

2.2 Scope

- Study area: Davanagere District, Karnataka, India.
- Study period used in the current GEE script: 1 January 2025 to 1 January 2026, with the end date exclusive.
- Optical input: Sentinel-2 Surface Reflectance Harmonized collection, filtered by study area, date, and CLOUDY_PIXEL_PERCENTAGE metadata.
- Radar input: Sentinel-1 GRD collection using IW mode and scenes containing both VV and VH polarizations.
- Feature engineering: six Sentinel-2 spectral bands, three spectral indices, two radar polarizations, and their dB difference.
- Classification: supervised Random Forest experiments for five land-cover classes, subject to creation and verification of labelled samples.
- Evaluation: independent validation samples and quantitative metrics; final values will be reported only after the experiments run.

2.3 Limitations

- Collection-level cloud-percentage filtering does not guarantee that every pixel in the composite is cloud-free; a pixel-level cloud mask may be needed.
- Classification quality depends on the representativeness and correctness of labelled training and validation samples.
- Results from one district and one selected year may not generalize to other regions or seasons.
- Sentinel-1 and Sentinel-2 have different measurement characteristics; preprocessing and feature definitions must be documented carefully.
- Random Forest performance depends on training data, class balance, feature selection, and parameter settings.
- Final classification accuracy, Kappa, and any claim of improvement from fusion remain unconfirmed until evaluation is completed.

3. Data Preprocessing and Feature Engineering

Data preprocessing prepares the satellite observations for consistent analysis. The current workflow selects the district boundary, filters satellite collections by spatial extent and date, applies collection-level quality filtering where configured, creates median composites, clips the composites to the study area, and derives features. The processing sequence is: Study-Area Selection → Sentinel-1/Sentinel-2 Collection Filtering → Composite Generation → Feature Engineering → Multi-Modal Fusion → Labelled Sample Preparation → Random Forest Classification → Accuracy Assessment → Map and Result Export.

Table 1 summarises the planned processing steps.

Sl. No.	Step	Purpose
1	Study-area selection	Select the Davanagere District boundary.
2	Sentinel-2 data collection	Filter optical images by district, date, and cloud-percentage metadata.
3	Sentinel-1 data collection	Filter radar images by district, date, IW mode, and VV/VH availability.
4	Composite generation	Create median composites and clip them to the district.
5	Optical feature selection	Select B2, B3, B4, B8, B11, and B12.
6	Spectral-index calculation	Calculate NDVI, NDWI, and NDBI.
7	Radar feature preparation	Select VV and VH and calculate VV minus VH in dB.
8	Feature fusion	Combine the optical feature stack and radar feature stack.
9	Sample preparation	Create verified labelled training and independent validation samples.
10	Classification and evaluation	Train Random Forest models, generate maps, and calculate accuracy metrics.

3.1 Study-Area and Data Collection

The study-area boundary is obtained from the FAO GAUL 2015 level-2 administrative boundary collection by filtering the district name for Davanagere. The Sentinel-2 collection used in the script is COPENICUS/S2_SR_HARMONIZED. It is filtered to the district and the date interval from 1 January 2025 to 1 January 2026, with CLOUDY_PIXEL_PERCENTAGE less than 20. The current console output reports 137 images meeting these collection

filters.

The Sentinel-1 collection used is COPERNICUS/S1_GRD. It is filtered by the same district and date interval, restricted to instrument mode IW, and filtered to scenes containing both VV and VH polarizations. The current console output reports 56 images meeting these filters. These counts describe collection items matching the configured filters; they do not by themselves indicate the number of cloud-free or usable pixels.

3.2 Composite Generation

Median composites are created from the filtered collections and clipped to the Davanagere boundary. The Sentinel-2 composite is displayed as a false-colour or natural-colour view using selected bands, while Sentinel-1 VV and VH layers are displayed separately for visual inspection. Compositing reduces the workflow to a representative image stack for the selected period. Additional quality masking can be applied if required.

3.3 Sentinel-2 Feature Engineering

Six spectral bands are selected from the Sentinel-2 composite: B2 (blue), B3 (green), B4 (red), B8 (near infrared), B11 (short-wave infrared), and B12 (short-wave infrared). Three normalized-difference indices are added to the feature stack.

$NDVI = (B8 - B4) / (B8 + B4)$. NDVI highlights vegetation-related spectral contrast.

$NDWI = (B3 - B8) / (B3 + B8)$. This formulation is used to highlight water-related spectral contrast.

$NDBI = (B11 - B8) / (B11 + B8)$. NDBI can help distinguish built-up surfaces from vegetation, although it is not a definitive built-up detector.

The resulting Sentinel-2 feature stack contains nine bands: B2, B3, B4, B8, B11, B12, NDVI, NDWI, and NDBI. The current GEE console confirms a list of nine Sentinel-2 feature bands.

3.4 Sentinel-2 Feature Engineering

Sentinel-1 GRD provides VV and VH backscatter measurements in decibels in the selected collection. The planned radar feature stack includes the VV band, the VH band, and the difference $VV - VH$, named $VV_minus_VH_dB$. The difference is a subtraction of values represented in dB; it should not be described as a linear-power ratio. The three-band radar stack is to be verified in the GEE console before fusion.

3.5 Multi-Modal Feature Fusion

The fusion step combines the nine-band Sentinel-2 feature stack with the three-band Sentinel-1 feature stack using the GEE `addBands` operation. If all planned features are available, the fused stack will contain 12 bands: six optical bands, three optical indices, VV, VH, and $VV_minus_VH_dB$. The fused feature stack is the input for the multi-modal classification experiment.

3.6 Training and Validation Data

Supervised classification requires labelled samples for the five target classes: forest/vegetation, agriculture, urban/built-up, water bodies, and bare land. Samples should be drawn from reliable reference information and checked against imagery or other appropriate evidence. Training and validation samples must be separated to avoid evaluating a model on the same observations used for training. The class labels, sampling method, sample counts, and split strategy should be documented in the final manuscript.

4. System Design and Implementation

The proposed framework uses Google Earth Engine as the data acquisition and geospatial processing environment. It is organized as a sequence of modular operations: boundary selection, satellite-data ingestion, filtering, compositing, feature engineering, feature fusion, labelled-sample preparation, classification, accuracy assessment, visualization, and export.

4.1 Hardware and Software Requirements

Table 2. Hardware and Software Requirements

Category	Technology or tool	Purpose
Geospatial platform	Google Earth Engine	Accesses satellite collections and performs cloud-based geospatial processing.
Optical data	Sentinel-2 SR Harmonized	Provides multispectral reflectance features.
Radar data	Sentinel-1 GRD	Provides VV and VH backscatter features.
Programming environment	GEE Code Editor JavaScript	Implements filtering, compositing, feature engineering, classification, and evaluation.
Administrative boundary	FAO GAUL 2015 level 2	Defines the Davanagere District study area.
Classifier	Random Forest in GEE	Trains supervised land-cover classification models.
Evaluation	Confusion matrix and accuracy metrics	Measures performance on independent validation samples.
Visualization and export	GEE map layers and export functions	Displays composites and classification results and exports outputs when configured.

4.2 Functional Modules

1. **Study-Area Module** — selects the Davanagere District boundary and defines the spatial extent for processing.
2. **Optical Data Module** — filters Sentinel-2 imagery and creates a clipped median composite.
3. **Radar Data Module** — filters Sentinel-1 IW scenes containing VV and VH polarization and creates a clipped median composite.
4. **Feature Engineering Module** — selects optical bands, calculates NDVI/NDWI/NDBI, and prepares radar features.
5. **Fusion Module** — combines optical and radar feature bands into a common multi-modal input image.
6. **Sample Preparation Module** — organizes labelled samples for the five land-cover classes and separates training from validation data.
7. **Classification Module** — trains Random Forest models for Sentinel-2-only, Sentinel-1-only, and fused feature inputs.
8. **Accuracy Assessment Module** — calculates confusion matrices, overall accuracy, class-wise accuracy, and Kappa using held-out validation samples.
9. **Visualization and Export Module** — displays composites and classification maps and exports the required outputs if configured.

4.3 Data Organization

The current implementation is based on Earth Engine image collections, images, and feature collections rather than a conventional relational database. The following logical data structures describe the main objects used in the workflow. They are conceptual structures for documenting the pipeline, not claims that database tables have been created.

Table 3. Data Organization

Logical object	Representative fields or bands	Purpose
Study-area feature collection	ADM2_NAME, geometry	Stores the selected district boundary.
Sentinel-2 image collection	system:time_start, CLOUDY_PIXEL_PERCENTAGE, B2/B3/B4/B8/B11/B12	Holds filtered optical scenes and their metadata.
Sentinel-1 image collection	system:time_start, instrumentMode, VV, VH	Holds filtered radar scenes and their metadata.
Sentinel-2 feature image	B2, B3, B4, B8, B11, B12, NDVI, NDWI, NDBI	Provides the nine optical features.
Sentinel-1 feature image	VV, VH, VV minus VH dB	Provides the three planned radar features.
Fused feature image	All nine optical and three radar features	Provides the combined input for fusion classification.
Labelled sample collection	geometry, land-cover class, sample identifier	Provides reference labels for model training and validation.
Evaluation output	confusion matrix, overall accuracy, class-wise accuracy, Kappa	Stores or reports model evaluation results.

4.4 Algorithm and Process Flow

Algorithm: Multi-Modal Satellite Land-Cover Classification

Input: Davanagere District boundary, filtered Sentinel-1 and Sentinel-2 collections, and verified labelled samples.

Output: Sentinel-2-only, Sentinel-1-only, and fused land-cover maps with evaluation metrics.

Start the GEE script and select the Davanagere District boundary.

Filter Sentinel-2 imagery by region, date, and configured cloud-percentage threshold.

Filter Sentinel-1 imagery by region, date, IW mode, and VV/VH polarization availability.

Create median composites and clip them to the district boundary.

Select six Sentinel-2 bands and calculate NDVI, NDWI, and NDBI.

Prepare Sentinel-1 VV and VH features and calculate VV minus VH in dB.

Combine the optical and radar features into a fused image.

Prepare verified labelled samples and separate training and validation samples.

Train a Random Forest model using Sentinel-2 features and classify the study area.

Train a Random Forest model using Sentinel-1 features and classify the study area.

Train a Random Forest model using fused features and classify the study area.

Evaluate all three models on an independent validation sample using the same evaluation protocol.

Compare the confusion matrices and metrics, display the maps, and export outputs as required.

4.5 Implementation Details

At the current implementation stage, the GEE script successfully selects the Davanagere District boundary, filters the 2025 Sentinel-2 and Sentinel-1 collections, creates composites, and generates the nine-band Sentinel-2 feature stack. The console reports 137 Sentinel-2 collection items, 56 Sentinel-1 collection items, and nine Sentinel-2 feature bands. The next implementation tasks are to verify the three Sentinel-1 features, create the 12-band fused stack, prepare reference-labelled samples, train the three Random Forest models, and compute independent accuracy results. These remaining steps should be completed before the manuscript is presented as a final empirical study.

5. Results and Discussion

This section reports the outputs available from the current implementation and defines the evaluation needed to compare the three feature configurations. Classification performance values are not included here because the classification and independent accuracy assessment have not yet been completed in the current workflow.

5.1 Experimental Setup

The configured study period is 1 January 2025 through 31 December 2025. The Sentinel-2 collection is filtered using CLOUDY_PIXEL_PERCENTAGE less than 20, and the Sentinel-1 collection is filtered for IW mode with both VV and VH polarizations. The classifier planned for all experiments is Random Forest. For a fair comparison, the same study area, class definitions, training/validation strategy, and comparable classifier settings should be used for Sentinel-2-only, Sentinel-1-only, and fused-feature models.

Item	Current value or plan
Study area	Davanagere District, Karnataka, India
Study period	2025 calendar year
Sentinel-2 collection count	137 matching image-collection items reported by GEE
Sentinel-1 collection count	56 matching image-collection items reported by GEE
Sentinel-2 feature stack	9 bands confirmed by GEE console
Sentinel-1 feature stack	3 planned bands; verify in GEE
Fused feature stack	12 planned bands after both stacks are verified
Land-cover classes	Forest/vegetation, agriculture, urban/built-up, water bodies, bare land
Classifier	Random Forest
Validation metrics	Confusion matrix, overall accuracy, class-wise accuracy, Kappa; F1 may also be reported

5.2 Feature Preparation Results

The current console output confirms that the filtered collections contain 137 Sentinel-2 image items and 56 Sentinel-1 image items for the configured region and period. The Sentinel-2 feature band list contains nine elements, corresponding to six selected spectral bands and three normalized-difference indices. These outputs verify the initial collection filtering and optical feature-engineering stages, but they do not establish land-cover classification accuracy.

5.3 Classification Results

To complete this subsection, insert the three classification maps and report the actual results produced by GEE. The final manuscript should include a table with the measured overall accuracy and Kappa for Sentinel-2-only, Sentinel-1-only, and fused features, along with class-wise metrics where possible. Do not state that the fused model is superior unless the independent validation results support that conclusion.

Experiment	Input features	Overall accuracy	Kappa	Notes
Model 1	Sentinel-2: 9 features	To be measured	To be measured	Run and evaluate on held-out samples.
Model 2	Sentinel-1: 3 features	To be measured	To be measured	Run and evaluate on the same validation protocol.
Model 3	Fused: 12 features	To be measured	To be measured	Compare with both single-source models.

5.4 Accuracy Assessment

A confusion matrix summarizes the agreement between reference labels and predicted classes. Overall accuracy measures the proportion of validation samples classified correctly. Class-wise producer's and user's accuracy can help identify classes that are omitted or confused with other classes. The Kappa coefficient can be reported as an additional agreement statistic. If F1-score is included, specify whether it is calculated per class or averaged across classes. The validation data must not be reused as training data.

5.5 Discussion

The proposed experimental design can show whether radar information contributes complementary value to optical spectral features for the selected district and period. Sentinel-2 indices provide useful feature representations for vegetation, water-related contrast, and built-up surfaces, while Sentinel-1 VV and VH provide radar backscatter information. However, the direction and size of any improvement from fusion are empirical questions. They depend on sample quality, class separability, seasonal variation, preprocessing, and validation design. A defensible discussion should interpret the actual confusion matrices and metrics and report limitations such as cloud contamination, class imbalance, and limited geographic coverage.

6. Conclusions and Future Work

6.1 Conclusion

This paper draft describes a data-engineering framework for multi-modal land-cover classification using Sentinel-1 and Sentinel-2 imagery in Google Earth Engine. The workflow is designed for Davanagere District and combines optical-band selection, spectral-index calculation, radar-feature preparation, feature fusion, Random Forest classification, and accuracy assessment. Initial implementation outputs confirm that the satellite collections can be filtered and that a nine-band Sentinel-2 feature stack has been generated. Completion of the Sentinel-1 feature stack, fused classification, labelled-sample preparation, and independent evaluation is required before final conclusions can be drawn about the benefit of multi-modal fusion.

6.2 Future Work

- Complete Sentinel-1 feature engineering and verify the 12-band fused feature stack.
- Create and document reliable labelled samples for all five land-cover classes.
- Train and compare Sentinel-2-only, Sentinel-1-only, and fused Random Forest models.
- Use independent validation samples and report confusion matrices, overall accuracy, class-wise metrics, and Kappa.
- Improve cloud and shadow masking for Sentinel-2 where appropriate.
- Test alternative seasonal windows or multiple years to assess temporal robustness.
- Investigate class imbalance, feature importance, and additional classifiers only after the baseline comparison is complete.
- Export final classification maps and evaluation tables for reproducibility.

Acknowledgements

The author expresses sincere gratitude to the Department of Data Engineering, GM University, Davanagere, and the project guide for their guidance and support. This acknowledgement should be revised to reflect the actual contributions and institutional requirements before submission.

REFERENCES

- [1] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R., "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18–27, 2017. <https://doi.org/10.1016/j.rse.2017.06.031>
- [2] Drusch, M., Del Bello, U., Carlier, S., et al., "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services," *Remote Sensing of Environment*, vol. 120, pp. 25–36, 2012. <https://doi.org/10.1016/j.rse.2011.11.026>
- [3] Torres, R., Snoeij, P., Geudtner, D., et al., "GMES Sentinel-1 mission," *Remote Sensing of Environment*, vol. 120, pp. 9–24, 2012. <https://doi.org/10.1016/j.rse.2011.05.028>
- [4] Breiman, L., "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. <https://doi.org/10.1023/A:1010933404324>
- [5] ESA, "Sentinel-1 User Handbook," European Space Agency. Available: <https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-1-sar>
- [6] ESA, "Sentinel-2 User Handbook," European Space Agency. Available: <https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi>
- [7] Google Earth Engine, "Sentinel-2 SR Harmonized," Earth Engine Data Catalog. Available: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED
- [8] Google Earth Engine, "Sentinel-1 SAR GRD," Earth Engine Data Catalog. Available: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S1_GRD
- [9] Google Earth Engine, "ee.Classifier.smileRandomForest," Earth Engine API Documentation. Available: <https://developers.google.com/earth-engine/apidocs/ee-classifier-smilerandomforest>
- [10] Google Earth Engine, "Classification," Earth Engine Guides. Available: <https://developers.google.com/earth-engine/guides/classification>

[11] FAO, "Global Administrative Unit Layers (GAUL), 2015," Earth Engine Data Catalog. Available: https://developers.google.com/earth-engine/datasets/catalog/FAO_GAUL_2015_level2