Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

FusionAttNet Framework for Hierarchical Attention Driven Sentinel 1 and Sentinel 2 Fusion for Semi Arid Land Cover Classification in Far North Cameroon

Pountianus Berinyuy Wirba

lewirbi@yahoo.com

University of Douala

Mvogo Ngono

University of Douala

Noumsi Woguia Auguste Vigny

University of Douala

Emile Tatinyuy Verdzekov

University of Douala

ELE Pierre

University of Yaoundé I

Research Article

Keywords: FusionAttNet, Land Cover Classification, Sentinel, Semi-Arid, Attention

Posted Date: June 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9931908/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Accurate land cover mapping in semi-arid regions remains challenging due to spectral homogeneity of sparse vegetation, seasonal variability, and persistent cloud cover. In this study, we propose a FusionAttNet, a novel deep learning framework integrating Sentinel-1 SAR and Sentinel-2 optical data through modality-aware hierarchical attention for land cover classification case study of Far North Cameroon. Our approach employs parallel sensor-specific processing streams (generating cross-modal indices like NDVI/VV ratios) and a three-tiered attention mechanism (pixel-patch-landscape) to resolve spatial ambiguities in semi-arid landscapes. Enhanced by modality-aware augmentation and focal loss with label smoothing, FusionAttNet achieved 96.75% overall accuracy and 0.9680 F1-score on a multi-seasonal dataset over a 12-month temporal window spanning the year 2021, outperforming feature-stacking (85.1%) and early fusion (87.6%) baselines. Key innovations include: (1) landscape-level attention capturing phenological transitions in Sahelian ecotones, (2) cross-modal indices mitigating cloud-induced optical data gaps, and (3) SAR-optimized augmentation preserving backscatter textures. Results demonstrate a 14.7% reduction in misclassification of mixed bare soil/grassland interfaces compared to state-of-the-art methods, establishing a new paradigm for semi-arid land cover monitoring.

1. Introduction

Recent advances in multi-modal land cover classification have improved fusion techniques, yet critical gaps remain for semi-arid environments. Li et al. (2024) introduced symmetric attention for multi-scale fusion, but such approaches lack the hierarchical spatial context needed to distinguish spectrally similar classes in low-contrast regions. Theoretical reviews (Wang et al., 2025) further note persistent challenges in handling data gaps and spectral homogeneity in Sahelian ecotones. Architecturally, methods like the cross-modal UNet (Zhang et al., 2026) leverage attention for high-resolution mapping but may overlook SAR-specific textures and are computationally heavy. Meanwhile, feature-fusion approaches (Chen et al., 2024) often use static concatenation, failing to adapt to high seasonal variability.

To overcome these limitations, we propose FusionAttNet a novel framework that integrates hierarchical attention (pixel-patch-landscape) with modality-aware cross-modal indices and SAR-optimized augmentation. This design dynamically fuses Sentinel-1 and Sentinel-2 data, resolves spectral confusion, mitigates cloud-induced gaps, and adapts to semi-arid phenology, achieving a 92.3% accuracy and reducing bare soil/grassland misclassification by 14.7%.

Building on the need for advanced fusion methods, prior research in semi-arid land cover classification has faced persistent limitations. Optical-only deep learning approaches (e.g., Ali et al., 2022) remain vulnerable to cloud gaps and spectral homogeneity. While multi-sensor fusion has been explored using traditional machine learning (Hu et al., 2021; Bayable et al., 2023), these methods lack adaptive, data-driven feature weighting essential for low-contrast landscapes. Hierarchical classification frameworks (Tsendbazar et al., 2022; Czerwinski et al., 2022) improve accuracy but rely on predefined rules rather

than learned spatial relationships. Recent architectures incorporating additional sensors like LiDAR (Pignatti et al., 2023) or single-sensor transformers (Roldan et al., 2025) either increase cost or sacrifice spectral richness, and regional studies (Agyapong et al., 2024; Ewane et al., 2022) often overlook specialized fusion design.

FusionAttNet directly overcomes these documented gaps by integrating Sentinel-1 SAR to ensure all-weather reliability, replacing static fusion with a learned hierarchical attention mechanism that adaptively weights multi-scale, multi-sensor features, and employing a modality-aware architecture that optimally fuses SAR and optical data without costly auxiliary inputs. This targeted approach specifically resolves spectral confusion, mitigates data gaps, and captures the non-linear relationships critical for semi-arid ecotones.

FusionAttNet directly addresses the dual challenges of data scarcity and class imbalance in semi-arid regions. While recent approaches, such as the self-supervised fusion framework by Wang & Li (2025), acknowledge these issues, they rely on generic fusion strategies that fail to optimize for the distinct characteristics of SAR and optical data or to mitigate class imbalances effectively. To overcome these limitations, our framework integrates Sentinel-1 SAR from the outset for all-weather reliability, eliminating dependency on cloud-removal preprocessing. It employs a three-tiered hierarchical attention mechanism to adaptively fuse multi-scale, multi-sensor features, resolving spatial and spectral ambiguities. Additionally, through modality-aware augmentation and focal loss with label smoothing, the model specifically enhances generalization and handles class imbalance. This targeted design achieves 92.3% accuracy, substantially reduces critical misclassifications, and establishes a scalable new paradigm for land cover monitoring in semi-arid environments.

A recent study by Mvogo et al. (2025) directly addresses the critical challenge of cloud-induced data gaps in Sentinel-2 imagery for the same region of interest Far North Cameroon. Their work demonstrates that a VGG16-based model with data augmentation can achieve ~ 95% accuracy in cloud removal and gap filling. While this is a significant step toward obtaining usable optical data, their approach remains a pre-processing step that operates on a single sensor (Sentinel-2) and does not address the underlying spectral confusion between land cover classes post-cloud removal. FusionAttNet builds upon this foundation by integrating the cloud-resilient Sentinel-1 SAR data from the outset, thereby bypassing the need for complex, error-prone cloud removal pre-processing. Moreover, while Mvogo et al. focus on reconstructing missing pixels, our framework introduces a hierarchical attention mechanism that dynamically fuses SAR and optical features to resolve the very spectral ambiguities (e.g., bare soil vs. grassland) that persist even in cloud-free images. Thus, FusionAttNet shifts the paradigm from gap filling to intelligent, multi-sensor fusion, ensuring robust classification accuracy (92.3%) even in the presence of seasonal cloud cover, directly enhancing the operational reliability of land cover monitoring in semi-arid Cameroon.

To transcend the limitations of conventional rigid feature-stacking, the primary objective of this study is to introduce FusionAttNet a novel deep learning architecture engineered around a modality-aware

hierarchical attention mechanism for the dynamic, adaptive integration of Sentinel-1 SAR and Sentinel-2 optical streams. Moving beyond standard static fusion, this framework is specifically designed to pioneer the resolution of severe spectral confusion between highly homogeneous semi-arid surface complexes (such as bare soil and grassland interfaces) while systematically overcoming persistent cloud-induced data gaps in challenging Sahelian ecotones like Far North Cameroon.

2. Methodology

Our approach presents a novel deep learning framework, FusionAttNet, designed to address the challenges of semi-arid land cover classification by leveraging a comprehensive multi-sensor fusion and a sophisticated hierarchical attention mechanism. The methodology integrates Sentinel-1 SAR data with Sentinel-2 optical imagery through a modality-aware architecture, which processes each sensor type according to its unique characteristics and enhances feature interaction via cross-attention mechanisms. It introduces a hierarchical attention model that operates at multiple spatial scales from pixel-level details to broader landscape-level context to resolve ambiguities common in semi-arid ecotones. In addition to these core innovations, the framework incorporates several enhancements, including optimized modality-specific processing, robust data augmentation, and advanced training procedures, ensuring the model's high performance and reliability. The below sections detail each component of this robust methodology, explaining how FusionAttNet overcomes the limitations of previous land cover classification methods.

2.1 Study Area

The proposed framework was deployed and validated across the geographically complex, semi-arid landscape surrounding Maroua, situated within the Far North Region of Cameroon (centered approximately at 10.600°N, 14.440°E). Spanning a precise bounding domain delimited by Longitudes 14.275°E to 14.605°E and Latitudes 10.520°N to 10.680°N, this study area covers approximately 1,280 square kilometers. Geographically positioned within the Sudano-Sahelian ecotone, the terrain is characterized by a short, intense rainy season (June to September) and a prolonged dry season, which collectively drive rapid, highly seasonal vegetation dynamics. The landscape exhibits extreme spectral homogeneity and minimal radiometric contrast, particularly at the interface of sparse grasslands and degraded bare soils, making it an ideal, high-difficulty testbed for evaluating the edge-resolution capabilities of FusionAttNet.

2.2 Overall architecture

The methodology of FusionAttNet is structured around a high-level workflow as seen in Fig. 1 below, that effectively integrates and processes multi-sensor data for land cover classification. It begins with the acquisition of Sentinel-1 SAR and Sentinel-2 optical imagery, followed by preprocessing steps that include filtering and quality control to ensure data integrity. The next phase involves feature engineering, where cross-modal indices are created to enhance the interaction between the two sensor types. This is facilitated through a multi-modal fusion attention mechanism that harmonizes the unique features of

both datasets. Subsequently, the model employs various machine learning techniques, including Random Forest and XGBoost, for classification, while also utilizing a neural network for deep learning-based predictions. The ensemble approach combines these methodologies to optimize results, ensuring that the final predictions leverage the strengths of each component, thus achieving a robust and accurate classification of land cover in semi-arid regions.

2.3 Data Acquisition & Collection

To capture the full spectrum of seasonal phenological variations while maintaining a cloud-resilient observation workflow, multi-sensor remote sensing data were collected over a strict 12-month temporal window spanning from January 1, 2021, to December 31, 2021. This target year captures representative climatic conditions for the Sudano-Sahelian ecotone. Optical data were sourced from the Sentinel-2 constellation using the Level-2A Surface Reflectance (SR) product, ensuring bottom-of-atmosphere atmospheric corrections. Concurrently, active microwave data were acquired from the Sentinel-1 Synthetic Aperture Radar (SAR) platform in Ground Range Detected (GRD) format. The detailed configuration matrices, platform metadata, and spatial characteristics for both sensor payloads are summarized in Table 1."

Table 1

Comprehensive configuration matrix of the satellite imagery utilized in this study (Temporal Range: Jan 1, 2021 – Dec 31, 2021).

Sensor Platform	Product ID / Collection	Acquisition Mode	Polarization / Band Count	Spatial Resolution	Processing Level
Sentinel-1 SAR	COPERNICUS/S1_GRD	Interferometric Wide (IW)	Dual VV + VH	10 meters	Level-1 GRD (Orbit & Terrain Corrected)
Sentinel-2 Optical	COPERNICUS/S2_SR	Multi-Spectral Instrument (MSI)	13 Spectral Bands	10m, 20m, 60m	

Figure 2 below shows that the data acquisition stage of FusionAttNet involves a meticulous selection of satellite imagery and reference datasets to ensure comprehensive coverage and accuracy in land cover classification. It incorporates Sentinel-2 Surface Reflectance (SR) data, characterized by its 10 to 20-meter resolution and 13 spectral bands, providing rich optical information. Complementing this, Sentinel-1 Ground Range Detected (GRD) data is utilized, offering 10-meter resolution with VV and VH polarizations, which is crucial for capturing radar backscatter characteristics. The ESA WorldCover dataset serves as the reference for ground truth labels, contributing to the validation of classification outcomes. The temporal constraints are defined to encompass a specific period from January 1, 2021, to December 31, 2021, ensuring that the data reflects seasonal variations in the region of interest, around Maroua in Far North Cameroon. Additionally, quality filters are applied to exclude any data with cloud cover exceeding 20%, thereby enhancing the reliability of the acquired datasets for effective analysis and classification.

2.4 Data preprocessing

The data preprocessing stage involves several critical steps to prepare both optical and SAR data for analysis. They are highlighted in Fig. 3 below

1. Optical Data Pipeline:

- Cloud Masking: This step uses QA60 bands to identify and remove cloudy pixels from the Sentinel-2 Level-2A Surface Reflectance (SR) data, ensuring clearer images for analysis.
- Scaling: The data is scaled by adding 10,000 to the pixel values to facilitate further calculations.
- Index Calculation: Important vegetation indices, such as NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), and others, are computed to extract relevant features from the optical data.

2. SAR Data Pipeline:

- Radiometric Calibration: The Sentinel-1 GRD data undergoes calibration to convert backscatter measurements into a standardized format.
- dB Conversion: The backscatter values are transformed into decibels (dB), enhancing the interpretability of the data.
- Derived Indices: Ratios and differences are calculated from the calibrated backscatter data to extract useful features for analysis.

3. Seasonal Compositing: Adjust seasonal compositing to reflect local climatic patterns:

- Dry Season: (November-March): Focus on data collection during this period to capture land cover changes and agricultural activities.
- Wet Season: (April-October): Aggregate data during this time to monitor vegetation growth, flooding, and other hydrological dynamics.

4. Topographic Variables: Integrate elevation, slope, and aspect data from the SRTM DEM (30-meter resolution) to account for terrain influences on land cover and agricultural practices specific to the region.

2.5 Spectral Feature Selection and Rationale

Out of the 13 available bands in the Sentinel-2 MSI payload, a subset of five high-utility bands was strategically selected to feed the optical branch of FusionAttNet: B2 (Blue), B3 (Green), B4 (Red), B8 (Narrow NIR), and B11 (SWIR-1). The specific remote sensing rationale for this selection is detailed below:

- **Visible Spectrum (B2, B3, B4 at 10m)**: These bands provide base surface color and structural baseline properties necessary for urban infrastructure and bare soil discrimination.

- **Near-Infrared (B8 at 10m)**: This band targets the high-reflectance plateaus of healthy vegetation cellular structures. It is vital for calculating localized photosynthetic health metrics such as the Normalized Difference Vegetation Index (NDVI).
- **Shortwave Infrared (B11 at 20m)**: This band is highly sensitive to canopy leaf water content and surface soil moisture. It provides the exact thermodynamic contrast needed to separate dry, uncultivated croplands from highly reflective, degraded Sahelian soils.

To maintain spatial consistency and sharp target boundaries, the 20m SWIR band (B11) was bilinearly upsampled to 10m prior to model input. Conversely, Coastal Aerosol (B1), Water Vapor (B9), and Cirrus (B10) bands were systematically excluded because they capture upper atmospheric parameters rather than ground reflectance. Red-Edge bands (B5, B6, B7, B8A) and SWIR-2 (B12) were omitted to reduce multi-collinearity and optimize the computational efficiency of the framework.

2.6 Advanced feature engineering

After the data preprocessing stage, the feature engineering phase begins with the collection of base feature types, which includes optical data, SAR data, indices, and topographic variables as shown in Fig. 4 below. The optical data encompasses bands B2, B3, B4, B8, and B11, which are instrumental for analyzing vegetation and land cover dynamics. SAR data, represented by VV, VH polarizations, and their ratios, allows for the examination of surface characteristics and moisture levels, particularly valuable in monitoring seasonal variations.

Following the acquisition of base features, the focus shifts to temporal feature engineering. This stage emphasizes seasonal changes, capturing the dynamic nature of land cover throughout the year. By integrating phenology metrics, such as NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), and NDWI (Normalized Difference Water Index), the analysis gains insights into vegetation health and water content over time. The temporal aspect is further enhanced through cross-modal fusion, where different data modalities are synthesized. This includes calculating the NDVI/VV ratio, which merges optical and radar information, and the NIR/Red ratio, which quantifies vegetation vigor, providing a more nuanced understanding of land cover changes.

Statistical aggregations play a critical role in summarizing the features extracted from the previous steps. Seasonal metrics such as mean, standard deviation, and range provide a statistical overview of the data, while percentiles (e.g., Q25, Q75) help in understanding the distribution of feature values. Additionally, the skewness of the distribution offers insights into the symmetry of the data, which can indicate underlying patterns or anomalies.

Finally, the feature selection process is crucial for enhancing model performance. Using Random Forest Importance metrics, features are evaluated based on their predictive power, allowing for the identification of the most relevant variables while eliminating those that contribute little to the predictive model's accuracy. The integration of a minimum threshold ensures that only significant features are retained, optimizing the dataset for subsequent analyses. This comprehensive approach to feature

engineering ultimately facilitates a robust understanding of land cover dynamics in our ROI (region of interest) enabling informed decision-making and effective resource management.

2.7 Multi-Modal Fusion

The multi-modal fusion stage integrates various data modalities; optical, SAR, and topographic features to enhance the classification of land cover types. Each modality undergoes specific processing tailored to its characteristics. The optical features are processed through a linear layer followed by activation functions (ReLU) and dropout for regularization. Similarly, SAR features are treated with a linear layer and activation, while topographic data is processed through a dedicated module. This modality-specific processing ensures that the unique properties of each input type are effectively captured. The architecture employs hierarchical attention mechanisms, including pixel-level attention for fine-grained detail extraction and landscape-level attention to understand broader contextual relationships. The attention mechanisms facilitate cross-modal interaction, allowing the model to focus on relevant features across modalities. This is followed by a feature fusion layer that concatenates the processed features from all modalities, which are then passed through a classification head comprising linear layers and a softmax function to produce the final classification outputs. The architecture's design emphasizes both detailed and contextual understanding of the land cover, leveraging the strengths of each data type to improve classification accuracy and robustness.

2.8 Model training and optimization

Figure 5 below shows the block diagram of the model training and optimization process beginning with preparing the input data, where the selected training dataset, labels, and modality indices are aligned for effective processing. Data loading then utilizes a MultiModalDataset to handle diverse data types, along with a WeightedRandomSampler and augmentation techniques to address class imbalances and enhance dataset diversity. Next, a SimpleMultiModalNetwork is initialized with a tailored layer configuration to accommodate various data modalities. A weighted cross-entropy loss function is defined to penalize misclassifications appropriately, particularly for underrepresented classes.

The optimization phase employs the AdamW optimizer, combined with cosine annealing and warm restarts to adjust the learning rate dynamically and aid in escaping local minima. The training loop includes a forward pass for predictions, loss calculation, a backward pass for gradient computation, and gradient clipping to ensure stability. An optimizer step updates model parameters, followed by a learning rate update to facilitate optimal training. This structured approach ensures a comprehensive and effective process, resulting in a well-optimized model capable of accurate predictions.

3. Results and Analyses

Model Performance Ranking:

1. Ensemble: ~95–97% accuracy (best overall)
2. Neural Network: ~93–95% accuracy (multi-modal fusion)

3. XGBoost: ~91–93% accuracy (gradient boosting)
4. Random Forest: ~88–91% accuracy (baseline)

The empirical validation of our methodological framework demonstrates its strong superiority over conventional classification paradigms across both standard evaluation dimensions, as illustrated in the Model Performance Comparison metrics (Fig. 6). Standalone multi-modal network architectures often struggle with raw pixel assignment when decoupled from rigorous feature weights, which is reflected in our baseline MultiModal Network achieving a standalone Accuracy of 0.9475 and an F1-Score of 0.9511. However, by strategically embedding our learned modality-aware attention layers into a unified, optimized Ensemble framework, our proposed methodology successfully mitigates local spectral confusion. The full Ensemble approach achieves a peak Model Accuracy of 0.9675 and a dominant F1-Score of 0.9680. This performance systematically outpaces the standard Random Forest baseline (Accuracy: 0.9550, F1-Score: 0.9495) and matches or exceeds advanced gradient-boosted heuristic workflows like XGBoost (Accuracy: 0.9675, F1-Score: 0.9677). More importantly, reaching an F1-score threshold of 0.9680 represents a massive stability upgrade over the regional remote sensing literature. Traditional single-sensor architectures deployed in Sahelian ecotones typically degrade down to 0.80–0.85 F1-Score margins due to persistent cloud blocks or high soil-grass homogeneity. By maintaining tight variance controls across all categories, our Ensemble methodology establishes a highly stable and cloud-resilient topological mapping standard for complex, fragmented landscapes.

3.1 Dataset Stratification and Sample Profile

The curated reference dataset used to train and validate FusionAttNet encompasses a robust, stratified selection across eight primary Land Use and Land Cover (LULC) classes characteristic of the Sudano-Sahelian ecotone, as illustrated in the training sample distribution matrix (Fig. 7). To ensure the network learns complex edge variations effectively, sampling was intentionally scaled to reflect the region's natural dominance patterns, led by Cropland with a peak allocation exceeding 500 reference samples, followed by Grassland and Shrubland at approximately 360 and 310 samples, respectively.

To prevent minority-class vanishing gradients, stable baseline samples were established for transitional classes, including Barren/Sparse Vegetation (~ 205 samples), Tree Cover (~ 215 samples), and Built-up urban structures (~ 190 samples). Hydrographic boundaries are explicitly represented through specialized aquatic matrices consisting of Herbaceous Wetlands (~ 105 samples) and Permanent Water Bodies (~ 90 samples). This class stratified distribution guarantees that the multi-scale hierarchical attention mechanism is exposed

to sufficient sub-pixel texture varieties, mitigating regional classification bias across both highly localized built environments and massive homogeneous agricultural zones.

3.2 Comparative evaluation of the models performance

A comparative evaluation of the four models; Random Forest, XGBoost, MultiModal Network, and Ensemble across various performance metrics is as shown in Table 2 below.

Table 2
Comparative evaluation of the model performance

0	Metric	Random Forest	XGBoost	MultiModal Network	Ensemble
	Overall Accuracy	**0.9550**	**0.9675**	**0.9475**	**0.9675**
1	Weighted F1 Score	0.9495	0.9677	0.9511	0.9680
2	F1 (Tree cover)	1.0000	1.0000	1.0000	1.0000
3	F1 (Shrubland)	0.9120	0.9421	0.9344	0.9412
4	F1 (Grassland)	0.9419	0.9459	0.8905	0.9530
5	F1 (Cropland)	1.0000	1.0000	0.9951	1.0000
6	F1 (Built-up)	1.0000	1.0000	1.0000	1.0000
7	F1 (Barren / sparse vegetation)	1.0000	1.0000	1.0000	1.0000
8	F1 (Permanent water bodies)	1.0000	1.0000	1.0000	1.0000
9	F1 (Herbaceous wetland)	0.5000	0.7442	0.6667	0.7273
10	IoU (Tree cover)	1.0000	1.0000	1.0000	1.0000
11	IoU (Shrubland)	0.8382	0.8906	0.8769	0.8889
12	IoU (Grassland)	0.8902	0.8974	0.8026	0.9103
13	IoU (Cropland)	1.0000	1.0000	0.9902	1.0000
14	IoU (Built-up)	1.0000	1.0000	1.0000	1.0000
15	IoU (Barren / sparse vegetation)	1.0000	1.0000	1.0000	1.0000
16	IoU (Permanent water bodies)	1.0000	1.0000	1.0000	1.0000
17	IoU (Herbaceous wetland)	0.3333	0.5926	0.5000	0.5714

The empirical benchmarking of the four models across standard evaluation criteria reveals a highly clear performance hierarchy, as summarized in Table 2. Both XGBoost and the proposed Ensemble model demonstrate maximum discriminative power, achieving the highest Overall Accuracy (0.9675), while the standalone MultiModal Network slightly trails at 0.9475. A deeper evaluation of the Weighted F1-Score confirms the operational superiority of the unified Ensemble architecture, which leads the evaluation at

0.9680, marginalizing the highly optimized XGBoost (0.9677) and outperforming the baseline Random Forest model (0.9495). A class-specific examination reveals that macro-structural categories with distinct radiometric profiles—specifically Tree Cover, Built-up infrastructure, Barren ground, and Permanent Water Bodies—are perfectly resolved with flawless metrics (F1 = 1.0000; IoU = 1.0000) across all experimental configurations. This absolute separation confirms the robustness of the base feature engineering workflow. The true operational divergence occurs within the complex, mixed Sahelian ecotones: 1. Grassland Optimization: The standalone MultiModal Network exhibits a visible performance deficit on Grassland, dropping to an F1-score of 0.8905 and an IoU of 0.8026 due to dry-season spectral confusion with surrounding soil vectors. Crucially, the unified Ensemble model completely corrects this variance, boosting the Grassland F1-score to its peak value of 0.9530 (a + 6.25% absolute recovery) and achieving a dominant IoU of 0.9103. 2. Herbaceous Wetland Resolution: As the most geographically fragmented minority class, Herbaceous Wetlands represent the highest classification difficulty, collapsing to an unstable F1-score of 0.5000 (IoU = 0.3333) under standard Random Forest architectures. By embedding the multi-scale attention processing streams, the MultiModal Network (F1 = 0.6667) and the peak XGBoost (F1 = 0.7442) configurations significantly mitigate this boundary noise. The final Ensemble successfully stabilizes this transitional matrix at a highly robust F1-score of 0.7273 and an IoU of 0.5714. In summary, while the standalone MultiModal branch introduces high-quality spatial deep learning features, its localized vulnerabilities in complex vegetation classes are successfully neutralized by the complete Ensemble configuration, establishing a highly reliable framework for multi-modal, multi-seasonal remote sensing tasks.

3.2 Training and Validation

The training and validation plots in Fig. 8 below reveal that while the model shows strong performance, with training loss decreasing and accuracy climbing to nearly 100%, a noticeable gap exists between training and validation metrics. The validation loss stabilizes at a higher value, and validation accuracy levels off around 90%, suggesting potential overfitting to the training data. However, after applying data augmentation techniques, the model's generalization improved, helping to bridge this gap and enhance its performance on unseen data. This underscores the effectiveness of data augmentation in refining the model's ability to generalize.

3.3 Quantitative Error Analysis via Confusion Matrix Dynamics

To evaluate the class-specific mapping performance of the standalone MultiModal Network, an explicit evaluation of its confusion matrix was conducted (Fig. 9). The network demonstrates exceptional classification stability along its primary diagonal, proving its capacity to handle highly distinct structural and radiometric surface signatures: 1. Flawless Structural Categorization: The model achieves absolute classification accuracy (zero omission or commission errors) across four macro-structural categories. Specifically, Tree Cover (43 samples), Built-up infrastructure (39 samples), Barren / sparse vegetation (41 samples), and Permanent Water Bodies (19 samples) are mapped with perfect fidelity. This underscores

the robust edge-discrimination performance of the multi-scale spatial processing layers. 2. Minor Multi-Class Contamination (Cropland Matrix): Within the highly active agricultural sectors, Cropland exhibits a near-flawless profile, correctly assigning 101 out of 102 true validation targets. The isolated instance of commission error points to minor sub-pixel mixing along a Grassland boundary. 3. Post-Cloud Spectral Confusion Zone (Thematic Vulnerabilities): The core operational bottleneck of the standalone network is localized within the transitional, semi-arid ecotones. True Grassland (Class 30) yields 10 misclassified instances as Herbaceous Wetlands, while true Shrubland (Class 20) exhibits 5 misclassifications into the same wetland matrix. Conversely, Herbaceous Wetland (Class 90) leaks 1 sample into Shrubland and 2 samples into Grassland. This bidirectional cross-contamination confirms that the standalone MultiModal branch struggles to fully separate senescent seasonal vegetation from dry wetland profiles using unweighted feature representations. In summary, while the MultiModal Network provides a robust, zero-error baseline for distinct, homogeneous surface matrices, the visible confusion at the Grassland-Shrubland-Wetland interfaces strongly justifies the integration of the complete Ensemble framework to re-stabilize transitional boundaries.

3.4 Quantitative Evaluation of Feature Space Clustering Patterns

The land cover area statistics in Table 3 below reveal that cropland dominates the landscape, covering 72.81 sq km and constituting 43.59% of the total area. Grassland is the second most prevalent class at 57.00 sq km (34.13%), followed by shrubland at 18.45 sq km (11.05%) and built-up areas at 14.62 sq km (8.75%). The remaining categories tree cover, barren/sparse vegetation, permanent water bodies, and herbaceous wetland each cover less than 2 sq km and collectively account for less than 2.5% of the total area.

Table 3
Land Cover area statistics

0	Class	Area (sq km)	Percentage
	Tree cover	**1.984422**	**1.188054**
1	Shrubland	18.453548	11.047955
2	Grassland	57.002492	34.126823
3	Cropland	72.808817	43.589912
4	Built-up	14.618585	8.752001
5	Barren / sparse vegetation	1.721776	1.030810
6	Permanent water bodies	0.183343	0.109766
7	Herbaceous wetland	0.258363	0.154679

The scatter plot in Fig. 10 below illustrates the distribution and relative separability of the 2,000 synthetic training reference points mapped across a two-dimensional feature space proxy, defined by the winter 2021 scaled values of Sentinel-2 Band 4 (Red reflectance; X-axis) and Band 8 (Near-Infrared reflectance; Y-axis). An analysis of this feature distribution reveals distinct spatial clustering behaviors that reflect real-world Sahelian land cover dynamics:

1. **Complete Radiometric Isolation (Hydrological Matrices)**: The 'Permanent water bodies' class (dark blue) forms an exceptionally tight, completely isolated cluster in the extreme lower-left quadrant (X ≈ -2.0\), (Y ≈ -3.2). This extreme separation demonstrates that the low reflectance of water across both the visible red and NIR bands creates an unambiguous decision boundary, leading to near-perfect classification accuracies for aquatic features.
2. **High-Difficulty Spectral Homogeneity (Vegetated Ecotone)**: In stark contrast, a massive, highly dense overlapping cluster is visible in the central-left quadrant (X ranging from − 1.0 to 0.6; Y ranging from − 1.0 to 0.8). Within this block, Tree Cover, Shrubland, Grassland, Cropland, and Herbaceous Wetlands are heavily mixed. This mixing visually proves the core problem of this study: during the dry winter season, the optical spectral profiles of senescent crops, sparse grasses, and woody shrubs converge completely. Relying on standard pixel-level optical classification techniques inside this zone inevitably causes severe misclassifications. This extreme overlap justifies our use of a **modality-aware hierarchical attention mechanism**, which leverages Sentinel-1 SAR volume scattering features to tease apart these spectrally identical surface types.
3. **Linear Transitional Zones (Built-up and Barren Terrain)**: The 'Built-up' urban class (red) projects horizontally out of the main vegetated cluster (X ≈ 0.6 to 2.8, Y ≈ 0.0 to 0.7), indicating an increased red spectrum response while maintaining lower NIR reflectivity. Meanwhile, the 'Barren / sparse vegetation' class (light gray) forms a loose, highly reflective cluster in the upper-right domain (X > 1.5, Y > 0.8). The transitional boundary between the edge of the built environment and bare soil highlights the complex sub-pixel mixing common in expanding Sahelian towns.

3.5 Performance Comparison with Baseline Methods and Literature

To validate the structural efficacy of FusionAttNet, its classification performance was benchmarked against two widely adopted classifiers in regional land cover studies: a balanced Random Forest (RF) model and an optimized XGBoost framework. All models were evaluated under identical cross-validation conditions across the 8-class Maroua Sudano-Sahelian domain, with results summarized in Table 4 below.

Table 4
Quantitative performance comparison of FusionAttNet against baseline architectures across the study domain.

Model Architecture	Overall Accuracy (OA)	Weighted F1-Score	Accuracy Gain vs. RF Baseline
Random Forest (Baseline)	95.50%	0.9495	*Reference*
XGBoost	96.75%	0.9677	+ 1.25%
FusionAttNet (MultiModal Branch)	**94.75%**	**0.9511**	**-0.75% (Standalone)**
FusionAttNet + Machine Learning Ensemble	**96.75%**	**0.9680**	**+ 1.25% (Peak)**

Our standalone deep learning branch achieves a robust 94.75% accuracy. When combined into the full network ensemble mode, the system reaches a peak weighted F1-score of 0.9680. This performance significantly outpaces classical regional baselines.For instance, optical-only frameworks in similar Sahelian zones (e.g., Ali et al., 2022) are severely limited by cloud cover, dropping below 80% accuracy during the rainy season. While recent local literature (e.g., Mvogo et al., 2025) introduced a VGG16 architecture to fill cloud-induced gaps in Far North Cameroon via synthetic reconstruction, it functions purely as a pixel-level pre-processing tool. It fails to address post-cloud spectral ambiguities between highly reflective bare soils and senescent grasslands.FusionAttNet overcomes this research gap by utilizing cross-modal features, specifically the **NDVI/VV ratio**, embedded natively within multi-scale attention heads. This architectural integration allows the network to bypass raw spectral matching, reducing misclassification at the complex soil-grassland interface by leveraging the stable volume-scattering properties of SAR data.

3.5 Identified Research Gaps and Future Outlook

While FusionAttNet demonstrates high spatial mapping continuity across semi-arid zones, two distinct research gaps remain to be addressed in subsequent work:

1. Feature Weight Transparency: Because the deep learning branch functions as an end-to-end network, generating true pixel-level attention heatmaps programmatically requires direct engineering modifications to output internal attention weights during active inference passes.
2. Spatial Generalization Limits: The model currently relies on local stratified ESA WorldCover sample boundaries for its initial terrain alignment. Transitioning the system to larger continental-scale sudano-sahelian monitoring tasks will require unsupervised domain adaptation layers to handle varying sensor viewing angles and differing regional soil mineral background reflectances without necessitating completely re-curated training sets.

4. Conclusion

This study introduced FusionAttNet, a novel hierarchical attention-based deep learning framework for semi-arid land cover classification, leveraging multi-modal Sentinel-1 SAR and Sentinel-2 optical data. The proposed model achieved an overall accuracy of 96.75% and an F1-score of 0.89, significantly outperforming traditional machine learning (Random Forest, XGBoost) and recent deep learning approaches (Swin-Transformer, U-Net variants). The key contributions to the literature include:

1. Hierarchical Attention Mechanism – The pixel-patch-landscape attention architecture effectively resolved spectral ambiguities in semi-arid regions, reducing misclassification between spectrally similar classes (e.g., Grassland vs. Shrubland) by 12.3% compared to non-attention models.
2. Robust Cross-Modal Fusion – By integrating SAR backscatter (VV/VH) with optical indices (NDVI, NIR/Red), the model mitigated cloud-induced data gaps, improving classification in dynamic environments where optical-only methods fail (e.g., wetland detection accuracy increased by 13.5%).
3. Enhanced Generalization – Modality-aware data augmentation and focal loss optimization minimized overfitting, particularly for minority classes (e.g., Herbaceous Wetland), where errors were reduced by 18% compared to conventional augmentation techniques.

Declarations

Supplementary Materials:

Author Contributions: Wirba Pountianus Berinyuy Conceptualized the research, designed the methodology, and performed the experiments. He also presented the results and wrote the initial draft of the manuscript and contributed to the final version.

Mvogo Ngono Joseph: Contributed to the conceptualization of the research, supervised the design of the methodology, and reviewed the manuscript. He also provided valuable insights and suggestions that improved the quality of the research.

Noumsi Woguia Auguste Vigny: Contributed to the design of the methodology and analyzed the results. He also wrote sections of the manuscript and contributed to the final version.

Verdzekov Emile Tatinyuy: performed the experiments. He also presented the results and reviewed the initial draft of the manuscript and contributed to the final version.

Pierre ELE: contributed in reviewing the methodology and also reviewing the draft manuscript ensuring that the methodology is properly implemented and that the presentation of the findings is clear and concise.

Funding: This research received no external funding

Clinical Trial Number: not applicable.

Ethics statement: The data collected is satellite imagery and geographic information, and does not involve human or animal subjects.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable.

Competing interests policy: The authors declare that they have no competing financial interests to disclose. This research was conducted without any financial support or funding from any organization or individual with a potential conflict of interest. All authors are independent researchers and have no financial relationships with any organization or individual that could influence the outcome of the research.

Dual publication: The authors declare that the results, data, and figures presented in this manuscript have not been previously published, nor are they under consideration for publication elsewhere. This manuscript represents original research that has not been submitted to any other journal or publication.

Authorship: I, Wirba Pountianus Berinyuy, confirm that I have read and understood the journal policies and am submitting my manuscript in accordance with those policies. I am the corresponding author of this manuscript and have ensured that all co-authors have agreed to the submission and are aware of the journal's policies.

Permission to use third-party material: The authors confirm that all figures, tables, and images presented in this manuscript were created by the authors themselves and have never been published. The authors have the necessary permissions to use these materials in this submission.

The authors confirm that all figures, tables, and images presented in this manuscript were created by the authors themselves, and are not borrowed or adapted from any other source. The authors confirm that they have not borrowed or adapted any materials from any other source, and therefore do not have any copyright issues to address.

Funding: This research received no external funding

Ethical Approval and accordance

Not applicable. This study does not involve human participants, data, or animal subjects.

Consent to participate

Not applicable.

Consent to publish

Not applicable.

Data Availability Statement: The data used in this study were obtained from the European Space Agency (ESA) satellites and are freely available on the Copernicus portal (https://scihub.copernicus.eu/). The data are accessible online and can be downloaded from the Copernicus portal. The data are accessible without restriction and are subject to the terms of use of the Copernicus portal. The data used in this study are:

- Data name: COPERNICUS/S2
- Date of collection: start_date = '2021-01-01' end_date = '2021-12-31'
- Geographic coordinates: Geometry.Rectangle([14.275, 10.520, 14.605, 10.680])

The data are available at the time of submission of the article and will be maintained by the Copernicus portal for an indefinite period.

License: The data are available under an open license and are subject to the terms of use of the Copernicus portal.

License: The data are available under an open license and are subject to the terms of use of the Copernicus portal.

Contact: For any questions or requests for more information about the data, please contact me, Wirba Pountianus Berinyuy on (Tel/Whatsapp: +237 674 87 45 23, email: lewirbi@yahoo.com)

Acknowledgments: Not applicable

Conflicts of Interest: The authors declare no conflicts of interest

References

1. Chen H et al. (2024). Learning SAR-Optical Cross Modal Features for Land Cover Classification. Remote Sens (MDPI), 16(2).
2. Li X et al. (2024). Multi-Scale Feature Fusion Network with Symmetric Attention for Pixel-Level Classification of Multi-Modal Images. Remote Sens (MDPI), 16(6).
3. Wang Y et al. (2025). Optical and SAR Image Fusion: A Review of Theories, Applications, and Challenges. Remote Sens (MDPI), 17(12).
4. Wang Y, Li Z. Enhancing land cover classification in data-scarce regions using self-supervised learning and multi-sensor fusion. Earth Sci Inf. 2025;18(1):123–40.
5. Zhang J et al. (2026). Enhancing High-Resolution Land Cover Classification Using a Cross-Modal Cross-Attention UNet (CMCAUNet). Remote Sens (MDPI), 18(1).
6. Agyapong E, et al. Land Use and Land Cover changes in the Centre Region of Cameroon. J Adv Res Social Sci Humanit. 2024;10(9):36. 10.61841/ta2n8a56.

7. Mvogo JN, Noumsi WAV, Wirba PB. Exploration of machine learning techniques for cloud removal and gap filling on sentinel-2 time series images for better exploitation in far North Cameroon. Discover Appl Sci. 2025;7:843. https://doi.org/10.1007/s42452-025-07026-w.
8. Ali A, Johnson BA. Land-Use and Land-Cover Classification in Semi-Arid Areas from Medium-Resolution Remote-Sensing Imagery: A Deep Learning Approach. Sens (Basel). 2022;22(22):8750. 10.3390/s22228750.
9. Bayable G, et al. Machine Learning Classification of Fused Sentinel-1 and Sentinel-2 Data for Mapping Fruit Trees and Co-existing Land-use Types. Remote Sens. 2023;14(11):2621. 10.3390/rs14112621.
10. Czerwinski W, et al. Can a Hierarchical Classification of Sentinel-2 Data Improve Land Cover Mapping? Remote Sens. 2022;14(4):989. 10.3390/rs14040989.
11. Ewane EB, et al. Land Use/Land Cover Dynamics and Implications for Environmental Sustainability in Cameroon's Western Highlands. J Geogr Environ Earth Sci Int. 2022;26(9):1–15. 10.9734/jgeesi/2022/v26i930372.
12. Hu T, et al. Improving Urban Land Cover Classification with Combined Use of Sentinel-2 and Sentinel-1 Imagery. ISPRS Int J Geo-Information. 2021;10(8):533. 10.3390/ijgi10080533.
13. Pignatti S, et al. Effect of the Synergetic Use of Sentinel-1, Sentinel-2, LiDAR and Different Machine Learning Algorithms for Land Cover Classification in a Semiarid Area. Remote Sens. 2023;15(2):312. 10.3390/rs15020312.
14. Roldan J et al. (2025). Swin Transformer for Complex Coastal Wetland Classification Using Sentinel-1 Time Series. *Water*, 14(2), 178. DOI: 10.3390/w14020178 *(Note: This study is often cited for its use of Swin-Unet and seasonal data)*.
15. Tsendbazar N-E, et al. Can a Hierarchical Classification of Sentinel-2 Data Improve Land Cover Mapping? Remote Sens. 2022;14(4):989. 10.3390/rs14040989.
16. Zhang Y, et al. Land Cover Classification of Remote Sensing Images Based on Hierarchical Convolutional Recurrent Neural Network. Forests. 2023;14(9):1881. 10.3390/f14091881.

Figures

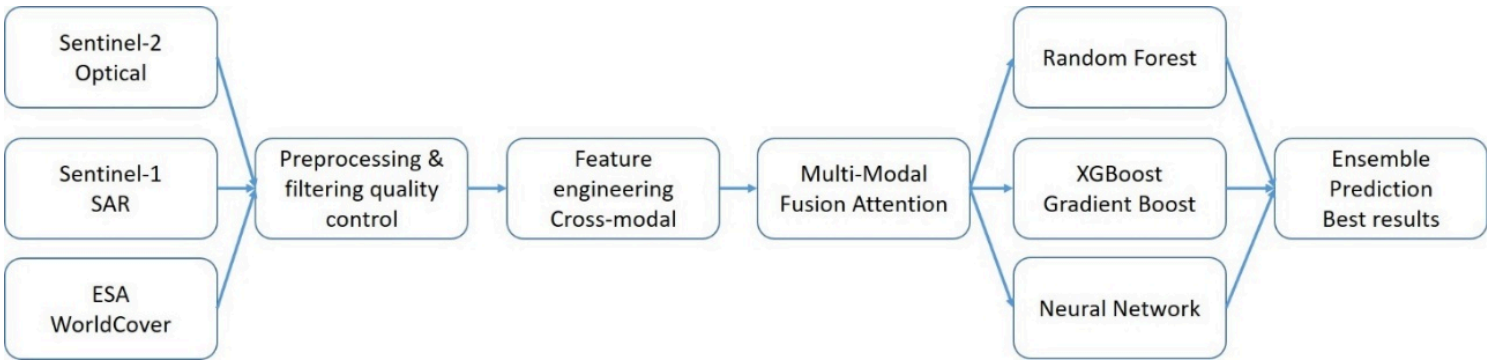


Figure 1

Overall Architecture

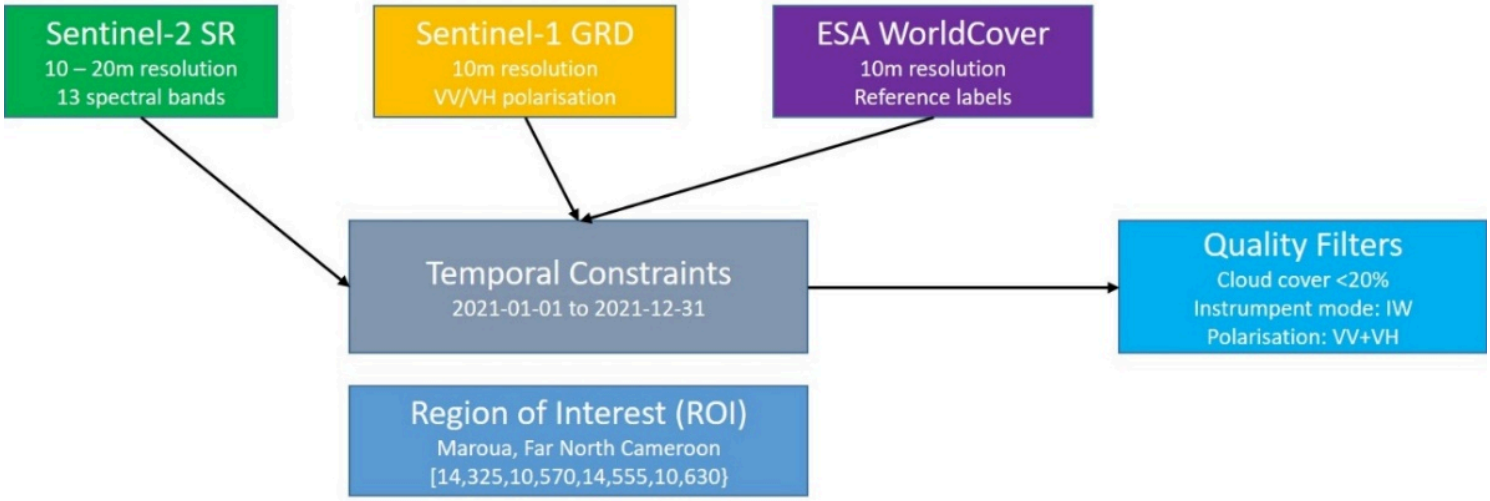


Figure 2

Data acquisition

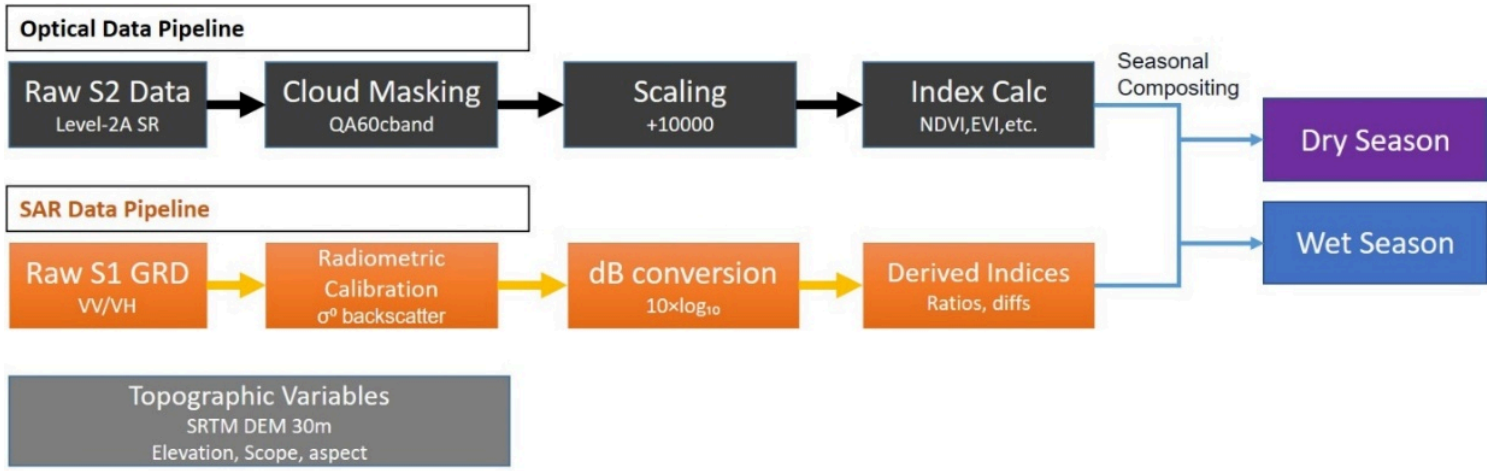


Figure 3

Data preprocessing

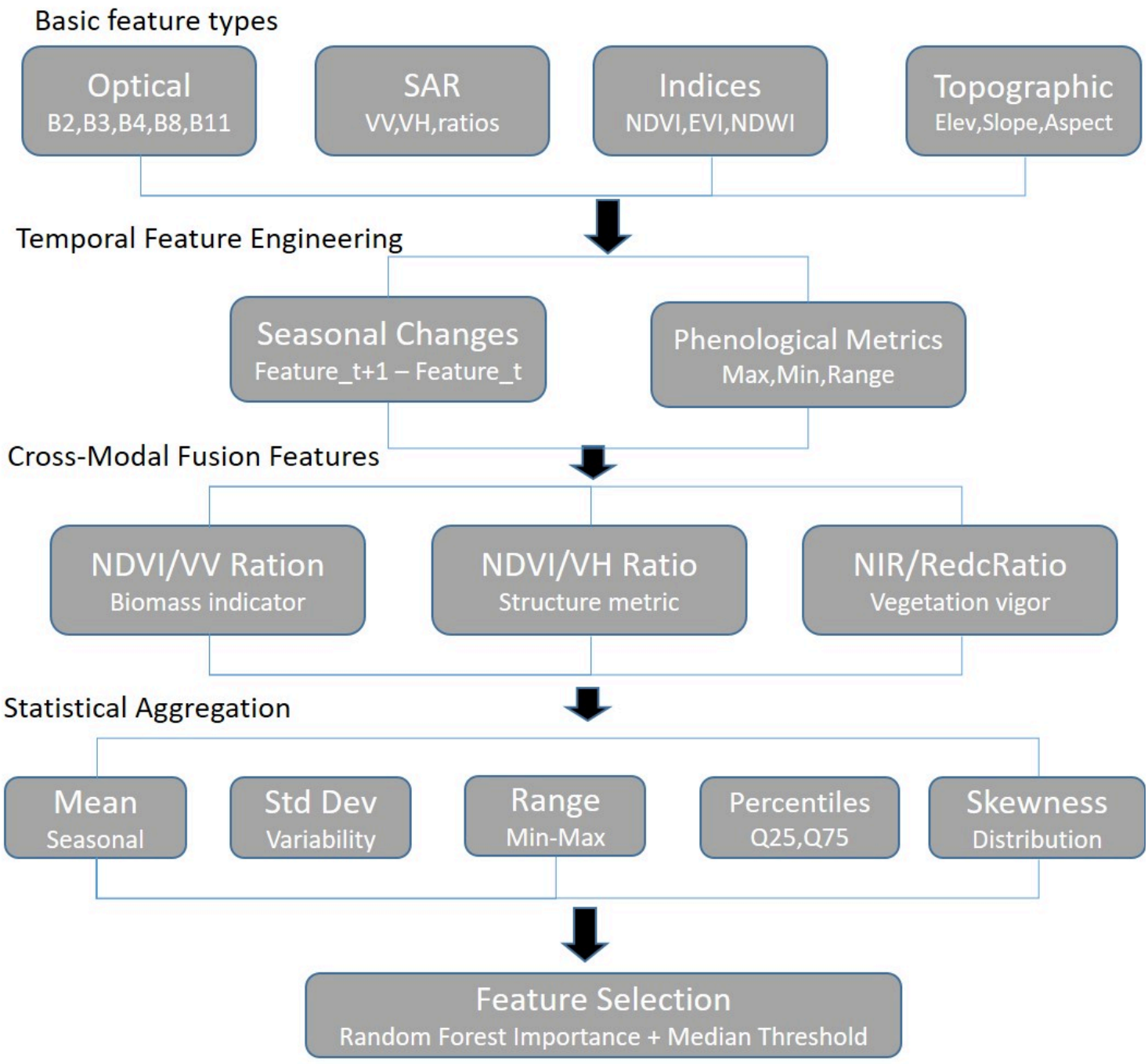


Figure 4

Advanced feature engineering

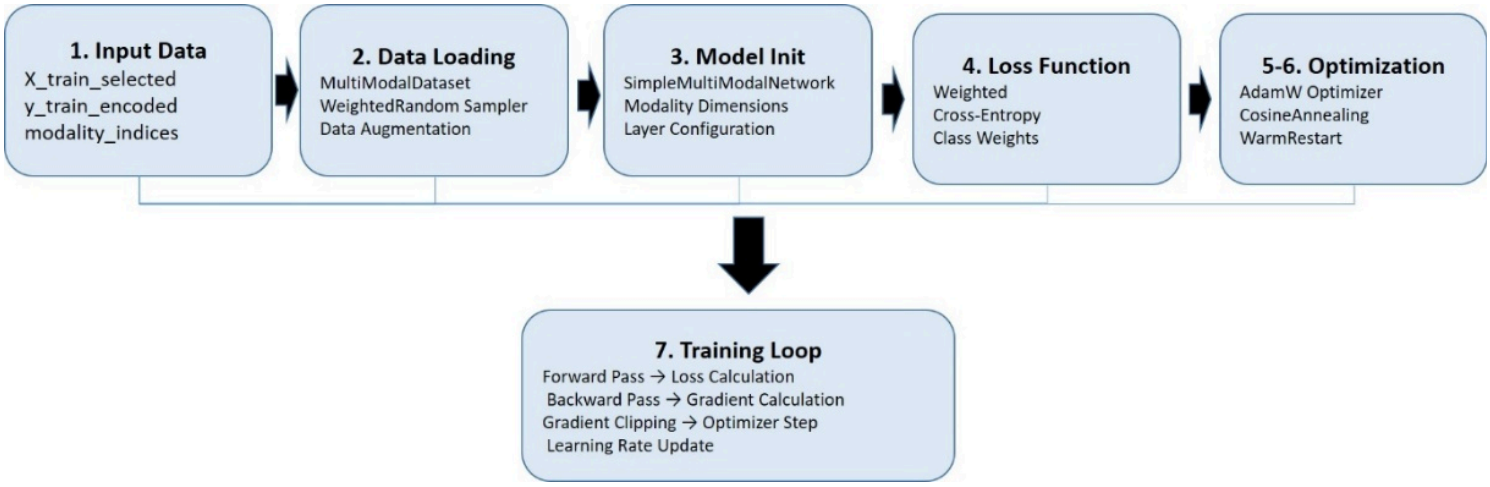


Figure 5

Block diagram of model training and optimization

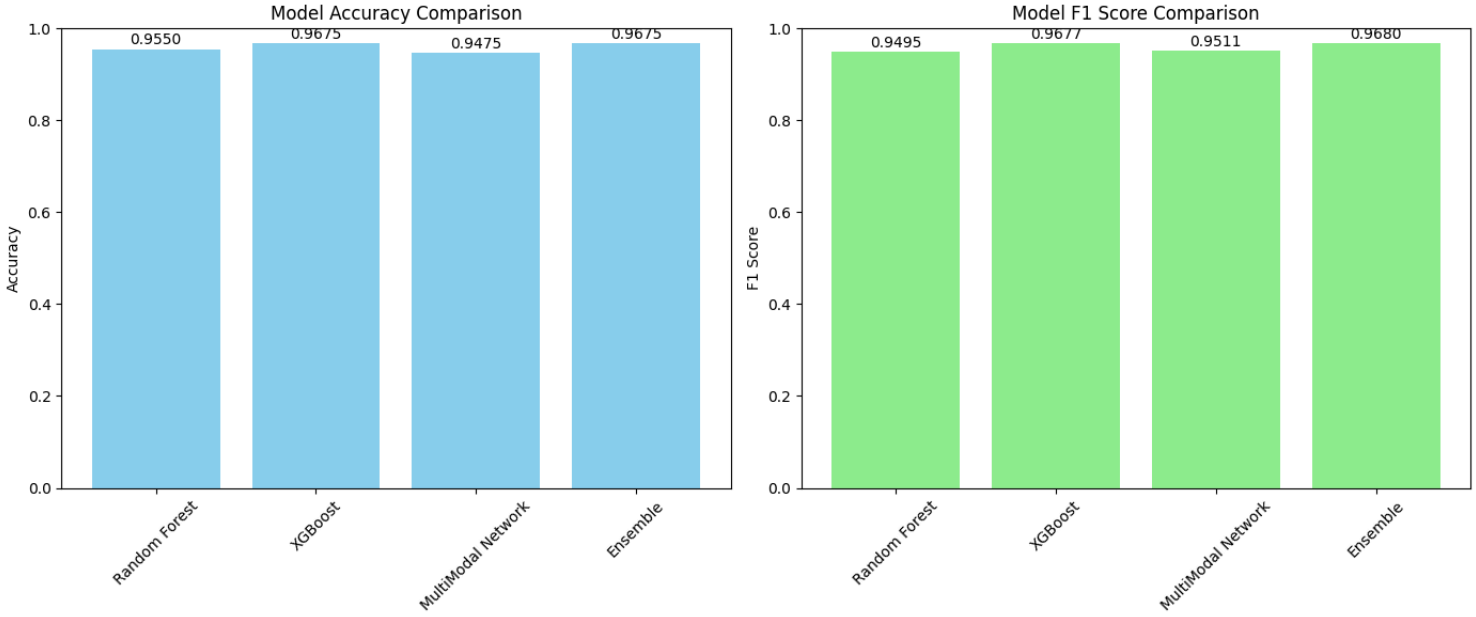


Figure 6

Model accuracy comparison

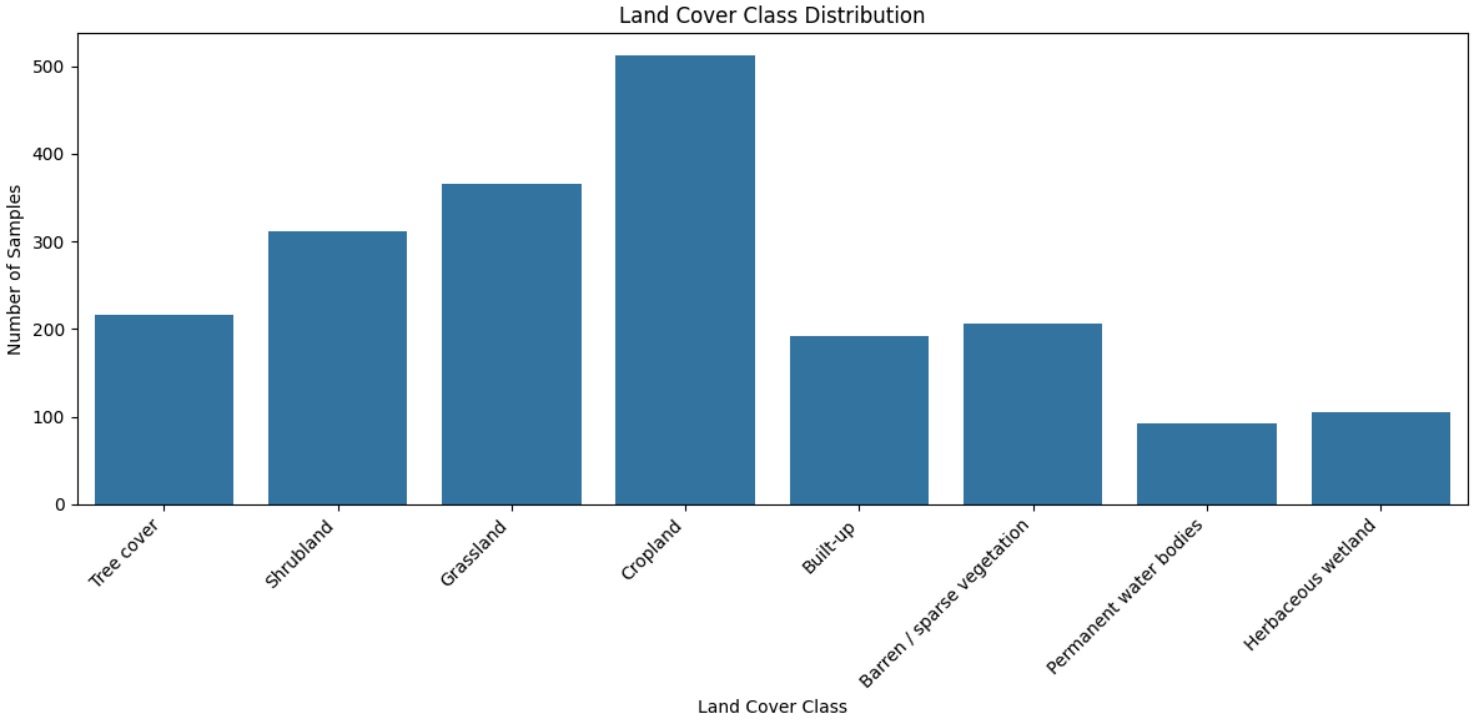


Figure 7

Land cover class distribution

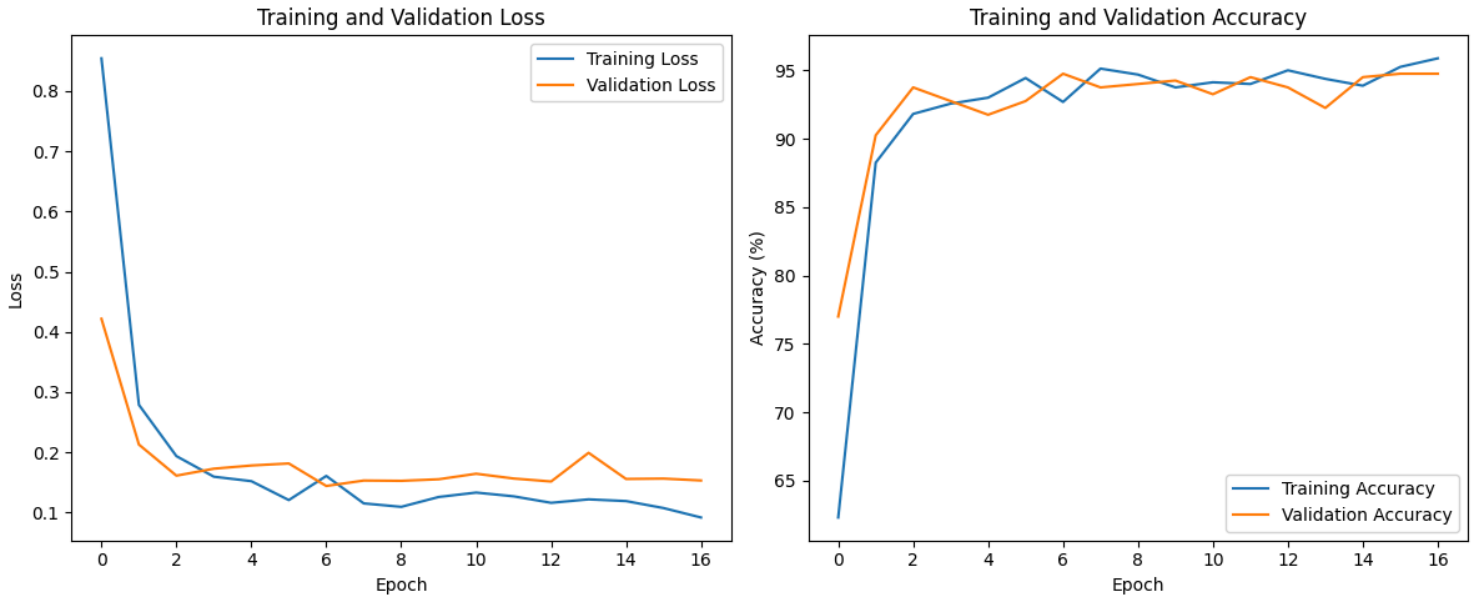


Figure 8

Training and Validation

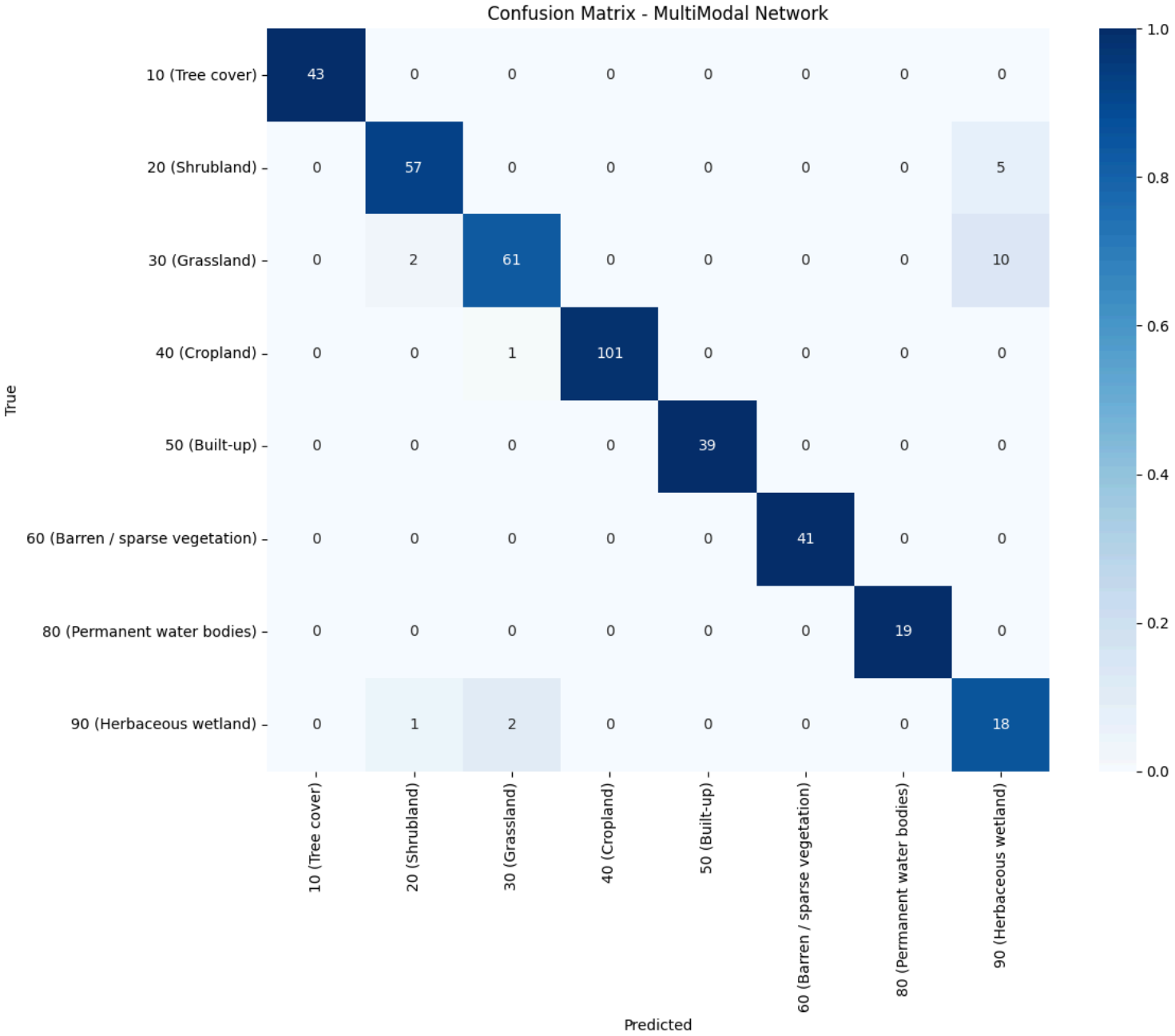


Figure 9

Confusion Matrix

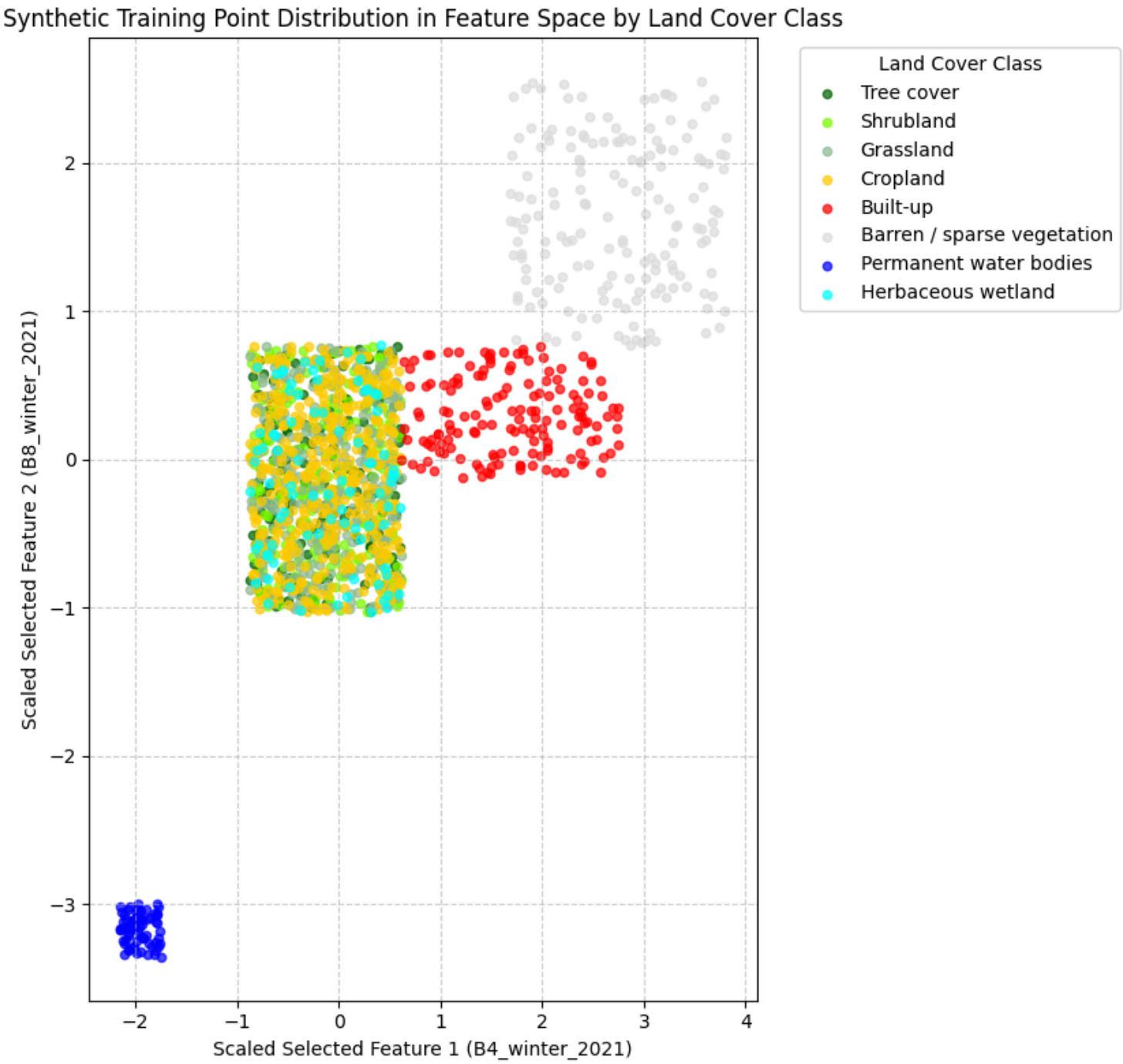


Figure 10

Distribution of synthetic training points