PlantHyper: An EnMAP-Based Hyperspectral Dataset for Crop Trait Retrieval and Agricultural Monitoring

Asherl Bwatiramba[1,2], **Irina Zlotnikova**[1], **Rajalakshmi Selvaraj**[1], **Dimane Mpoeleng**[1]
[1]Botswana International University of Science and Technology, Palapye, Botswana
[2]Botho University, Gaborone, Botswana

Study Area

PlantHyper is based on EnMAP imagery acquired over Pandamatenga in northern Botswana (18°32S, 25°38E). The broader study area covers 280,380 ha, including a commercial farming zone of 25,074 ha. The site is a major rainfed farming hub on vertisols (black cotton soils), where maize, sorghum, and sunflower are cultivated under annual rainfall of about 600 mm. The dataset includes extensive cropland and surrounding natural vegetation

Problem and Motivation

- Use of Deep Learning (DL) and Machine Learning (ML) in hyperspectral remote sensing is limited by by the scarcity of labelled ground-truth data
- Field campaigns are costly, time-consuming, and difficult to conduct at large scale
- Existing datasets are often often small, geographically limited, and class-imbalanced and do not generalise well

Aim and Contribution

- **Aim:** To develop PlantHyper, a large-scale hyperspectral dataset derived from EnMAP imagery and synthetic labels generated using a hybrid Radiative Transfer Model (RTM) and ML approach
- **Contribution:** A hybrid pipeline based on PROSAIL simulations, Artificial Neural Network (ANN) inversion, and retrieval of Leaf Area Index (LAI) and Leaf Chlorophyll Content (LCC) to generate a labelled dataset

Dataset Snapshot

Table 1. Comparison of PlantHyper with commonly used public hyperspectral benchmark datasets (platform, spatial resolution, spectral range, number of bands, classes, and scene size).

Dataset	Platform / Sensor	Spatial res.	Spectral range (nm)	Bands	Classes	Scene size (px)
Indian Pines	Airborne / AVIRIS	20 m	400–2500	220	16	145×145
Salinas Valley	Airborne / AVIRIS	3.7 m	400–2500	224	16	512×217
Pavia University	Airborne / ROSIS	1.3 m	430–860	115	9	610×340
Heilongjiang Benchmark	Spaceborne / AHSI (ZY1-02D)	30 m	400–2500	149	8	897×483; 843×719
PlantHyper (this work)	**Spaceborne / EnMAP**	**30 m**	**420–2450**	**224–246**	**4**	**128×128**

Methods and Pipeline

Method (Hybrid RTM–ML Workflow)

- **Input data:** EnMAP Level-2A hyperspectral image (30 m; 420–2450 nm, Visible and Near-Infrared (VNIR)/ Short-Wave Infrared (SWIR))
- **Crop masking:** Random Forest (RF) classification used to remove non-crop pixels and reduce uncertainty
- **Physics-based simulation:** PROSAIL is parameterised (PROSPECT leaf + SAIL canopy) and run in forward mode to generate a large Look-Up Table (LUT) of reflectance spectra paired with vegetation traits (LAI, LCC)
- **Sensor alignment:** Simulated spectra are resampled using EnMAP spectral response functions (SRFs) and noise is added to improve realism
- **Learning-based inversion:** An ANN is trained on simulated spectra (with Principal Component Analysis (PCA) for dimensionality reduction) and applied to EnMAP imagery to retrieve LAI and LCC traits
- **Output:** Trait maps + labelled PlantHyper dataset; performance evaluated using independent in situ measurements of coefficient of determination R^2 and Root Mean Square Error (RMSE)

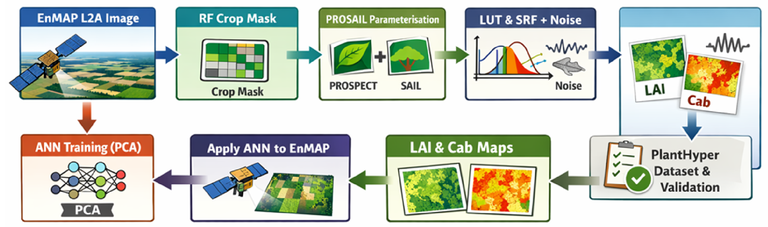

Figure 1. Hybrid RTM–ML pipeline used to construct PlantHyper from EnMAP hyperspectral imagery. EnMAP Level-2A input is first crop-masked using Random Forest classification. PROSAIL is then parameterised for forward LUT simulations, followed by SRF resampling and noise addition. An ANN with PCA is used to retrieve LAI and chlorophyll traits, producing the labelled PlantHyper dataset and validation outputs

Contacts

For more information, contact Mr. Asherl Bwatiramba: BA24020039@BIUST.AC.BW

Study Results

- **High trait-retrieval accuracy:** The hybrid PROSAIL–ANN approach achieved strong agreement with independent in situ measurements
- **LAI:** R^2 = 0.81, RMSE = 0.58 (robust LAI estimation)
- **LCC:** R^2 = 0.79, RMSE = 1.19 (reliable LCC retrieval)
- **Operational output:** The trained ANN produces spatial LAI and LCC maps from EnMAP imagery, supporting creation of a labelled PlantHyper dataset
- **Validation note:** Accuracy was assessed using independent field data (Pandamatenga in situ data unavailable; comparable region used)

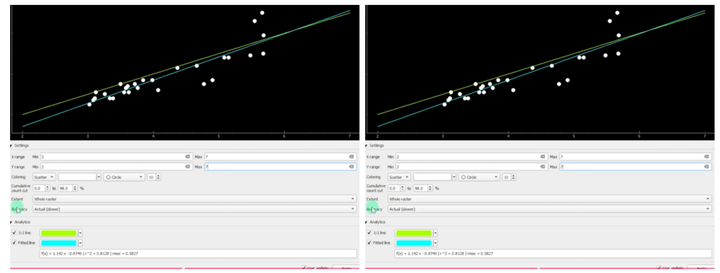

Figure 2. Validation of the hybrid RTM and ANN approach against independent in situ measurements: (a) LAI and (b) LCC, shown as extracts from EnMAP-Box outputs

Conclusions and Next Steps

- The PlantHyper dataset is created from EnMAP using a hybrid PROSAIL–ANN workflow to generate physically consistent labels
- Strong validation performance: LAI R^2=0.81 (RMSE=0.58); LCC R^2=0.79 (RMSE=1.19)
- Next steps: collect Pandamatenga in situ data and expand to more scenes/crops/seasons to improve generalisation