Image 2 biomass: Handcraft feature with Sentinel2 data into XGBoost

Guorun Huang
Khoury College of Computer Science
Northeastern University
Seattle, USA
huang.guor@northeastern.edu

Abstract*—This study explores pasture biomass prediction by integrating ground-level imagery with Sentinel-2 remote sensing data. We engineered a comprehensive set of features from the CSIRO dataset, including color indices, GLCM textures, and metadata, to capture vegetation variability. Five distinct modeling strategies were developed and compared, ranging from deep learning (ResNet) to structured approaches (XGBoost, MLP) and hybrid fusion models. Feature importance analysis was conducted to optimize input selection. The results demonstrate that combining proximal imagery with satellite spectral data creates a robust pipeline, providing a richer understanding of pasture conditions than single-source methods.

***Keywords*—feature engineering, Sentinel-2, XGBoost, ResNet, MLP, multi-modal data fusion, biomass prediction**

I. Introduction

Accurate pasture biomass estimation is fundamental to sustainable grazing management, directly impacting animal welfare, production consistency, and soil health. However, traditional assessment methods — such as manual clip and weigh or plate metering — are often labor-intensive, unscalable, and prone to errors under variable conditions. To address these limitations, this study utilizes the CSIRO Image2Biomass dataset to develop predictive models that estimate biomass components from ground-level imagery, ground-truth measurements, and remote sensing data (NDVI). By leveraging diverse data from Australian pastures, we aim to provide a scalable, automated solution for multi-target biomass assessment, evaluated through a globally weighted coefficient of determination (R^2) to ensure robust performance across key indicators like Green Dry Matter and Total Dry Weight.

II. Construct Features From Image

A. Raw data

The dataset consists entirely of images showing grass growing on the ground with a gray- brown soil background. The grass varies in color, density, and condition — some patches are dry, while others are fresh(Fig.1).

The original dataset includes the following features:

- sample_id — Unique identifier for each training sample (image).
- image_path — Relative path to the training image (e.g., images/ID1098771283.jpg).
- Sampling_Date — Date of sample collection.
- State — Australian state where sample was collected.
- Species — Pasture species present, ordered by biomass (underscore-separated).
- Pre_GSHH_NDVI — Normalized Difference Vegetation Index (GreenSeeker) reading.
- Height_Ave_cm — Average pasture height measured by falling plate (cm).
- target_name — Biomass component name for this row (Dry_Green_g, Dry_Dead_g, Dry_Clover_g, GDM_g, or Dry_Total_g).
- target — Ground-truth biomass value (grams) corresponding to target_name for this image.

B. Feature Engineering

To predict pasture biomass, I engineered a comprehensive set of features focusing on color indices, statistical distributions, and texture analysis. First, I utilized specific color indices to differentiate between vegetation states. Excess Green (ExG = 2G − R − B) and Excess Red (ExR = 1.4R − G) were calculated to highlight fresh green grass and dry or brown plant material, respectively. To ensure robustness against varying lighting conditions, I included the Visible Atmospherically Resistant Index (VARI = (G − R) / (G + R − B)) and the Color Index of Vegetation Extraction (CIVE = 0.441R − 0.811G + 0.385B + 18.787), which are effective at distinguishing green vegetation from bare soil and straw. Additionally, an RGB-based Normalized Difference Index (NDI = (G − R) / (G + R)) was used to separate green and dry grass similar to standard NDVI methods.

To capture the global color properties of the images, I analyzed histograms across both RGB and HSV color spaces. While RGB histograms provided basic insights into brightness and general color tendencies, HSV histograms and Hue distributions were incorporated to align more closely with human perception. The HSV features—specifically Hue for color category and Saturation for purity—proved superior for distinguishing between green grass, dry straw, and soil, particularly in complex lighting. While Hue distributions alone are sufficient for separating vegetation types, the full HSV histogram was retained to capture nuances in freshness (vivid vs. dull) and brightness.

Finally, I integrated texture and coverage metrics to characterize vegetation structure. GLCM (Haralick) features—such as contrast, homogeneity, and entropy—were extracted to distinguish the fine, uniform texture of dense grass from the rougher, high-contrast texture of dry straw or the smooth texture of soil. Complementing these was Green Coverage, calculated by thresholding ExG values to determine the percentage of green pixels. Although ExG statistical distributions (mean, variance) describe the intensity of greenness, Green Coverage serves as a lightweight, intuitive global indicator. By using both, the

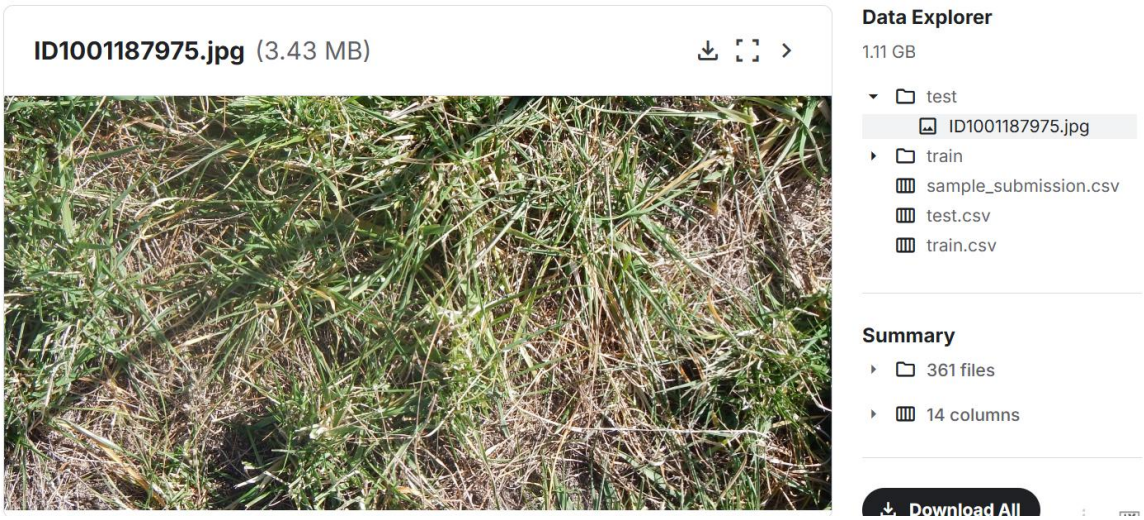

Fig. 1. Raw image data example.

model benefits from detailed statistical data alongside a direct measure of overall vegetation extent without redundancy.

III. Chose of Features

After computing all the features, I combine them into a single vector so each image corresponds to one feature vector. XGBoost can learn the weights by itself and decide which features matter more. The advantages are: we don't need to manually pick "the best" features — the model optimizes this on its own. Combining different types of features makes the model more robust and prevents a single feature from failing under certain conditions.

However, not all features help the prediction. Some may be redundant or even add noise. By checking how much each feature contributes to the model, we can keep the important ones and remove the useless or repetitive ones, improving both efficiency and generalization.

Feature importance analysis Idea: during training, the model measures how much each feature reduces error or increases information gain at split nodes.Outcome: the model produces a feature-importance score that shows which features are most useful for prediction.In XGBoost, this can be accessed through model.feature_importances_.

IV. Sentinel2 Data

There is no single pasture remote-sensing image dataset for Australia in 2015, but we can obtain the data from public satellite sources such as Landsat and Sentinel.

A. Choose data source

Landsat (NASA) Landsat provides 30-meter multispectral imagery covering the entire world, including Australia. It is widely used for vegetation indices such as NDVI and EVI, and for monitoring pasture cover and biomass. For 2015, both Landsat 7 and Landsat 8 are available, but Landsat 7 suffers from striping issues, so Landsat 8 is generally preferred.

Sentinel-2 (ESA Copernicus) Sentinel-2 offers 10–20 m multispectral imagery. Sentinel-2A began operating in July 2015, providing coverage over Australian pasture regions.

Key differences between Sentinel-2 (S2) and Landsat 8 (L8):

- Spatial resolution: S2: 10 m, 20 m, 60 m L8: 15 m, 30 m, 100 m → S2's 10 m bands are better for smaller fields.
- Revisit time: S2 (two satellites): 5 days L8: 16 days → S2 provides much more frequent observations.
- Spectral bands: S2: 13 bands L8: 11 bands → S2 includes multiple red-edge bands, which are very useful for vegetation monitoring.

Because of these advantages, I chose Sentinel-2 data. The data is not raw Level, it is Level-2A Analysis Ready Data (ARD). We can think of it as a four-dimensional data cube:

- X: longitude
- Y: latitude
- Time: a time series from 2015 to 2019 (roughly every 5 days, excluding cloudy scenes)
- Bands: each pixel contains values for 10 spectral bands

All 10 m bands are resampled to 20 m so that every band has the same resolution. DEA provides a Python environment (the Sandbox) where we load data using dc.load(). Product: ga_s2am_ard_3 (S2A). Measurements: the 10 bands used in the paper (nbart_blue, nbart_green, nbart_red, nbart_nir_1, etc.). Location: the latitude–longitude bounding box of the study area.The output is an xarray. Dataset, essentially a large array containing reflectance values for all times and all bands.

B. Date Formatting

After loading the data, several steps are required:

- Cloud masking The paper states that a cloud-detection algorithm removed more than 75% of cloudy pixels. DEA includes an oa_fmask band. Pixels marked as cloud or cloud-shadow must be set to NaN.

- Resampling If using raw data, bands B2, B3, B4, and B8 are 10 m. They must be resampled to 20 m (nearest-neighbour) to align with the other bands.
- The paper uses the median reflectance of all pixels inside each paddock. This requires a paddock boundary file. We overlay the polygon on the satellite image, extract all pixels inside it, and compute the median for each band.

Approximating paddock locations. Because the dataset does not provide exact paddock coordinates, I estimated locations using the state-level coordinates with a small random offset. This is a rough approximation because a single state contains many paddocks in different conditions—some resting, some newly planted—and with different grass species and growth stages. Even paddocks close to each other can look very different, which is visible in the images.

In this project, my goal is mainly to enrich the information sources and explore this type of remote-sensing data. Unlike close-range camera images, satellite imagery is captured from far above and includes infrared bands, making it a very interesting data source.

Also, DEA outputs time as an index in the format 2015-07-02 00:00:00. To match the image dataset, I convert it to the format 2015/7/2.

C. *Determining the Geographic Window Size*

To decide how large the latitude–longitude window should be, I looked at the average size of Australian farms. Agricultural land covers about 55–61% of Australia (roughly 393–426 million hectares). There are about 136,000 farms, most of them livestock or pasture farms. This averages to about 3,100 hectares per farm, or 31 km^2.

Sentinel-2 coverage: At around 30° S, each Sentinel-2 scene covers roughly 290 km × 290 km. Spatial resolution depends on the band:

- 10 m (visible + NIR)
- 20 m (red-edge + SWIR)
- 60 m (atmospheric bands)

If we load a 1° × 1° area (~100 km × 100 km on earth), then at 20 m resolution the image becomes a 5000 × 5000 pixel array. Loading a full month of data at this size can cause memory issues, so I reduced the window to 0.1° × 0.1°.

V. Model Overview

Model 1: Images + ResNet. Each image is fed directly into a ResNet model, and the resulting embedding is used to predict the five target values.

Model 2: Hand-crafted features + metadata + XGBoost. All manually engineered features, together with metadata, are used as input to an XGBoost model.

Model 3: Hand-crafted features + metadata + MLP. The same set of engineered features and metadata is fed into a multilayer perceptron.

Model 4: Sentinel-2 median band values + hand-crafted features + metadata + MLP. For each sample, the median value of all Sentinel-2 bands is combined with the engineered features and metadata, then passed into an MLP.

Model 5: Fusion model (ResNet + XGBoost). This model combines image embeddings from ResNet with predictions or features from XGBoost.

VI. Running Envoronment Setting

Kaggle Notebook is an online interactive coding environment provided by Kaggle, with free access to CPUs and GPUs. It also integrates datasets, so we can directly load public datasets or competition data by selecting them through the 'Add Data' button on the left side of the Notebook.

VII. Results

Best Model achieved the highest validation R^2 on both GDM_g and Dry_Total_g, demonstrating that structured features paired with XGBoost is the optimal approach for this data.

Value of Satellite Data: Comparing the MLP models, adding Sentinel-2 data raised the GDM_g score from 0.3956 to 0.5116 over the baseline, validating the utility of satellite inputs.

Image Data Issues: (ResNet) failed to converge, resulting in negative scores. Consequently, the fusion model underperformed, as the image data effectively acted as noise.

I focus here on the two highest-weighted targets (accounting for over 60% of the score) to demonstrate model performance. Detailed results will be submitted to Gradescope along with the code.

TABLE I.

Model Input + Algorithm	**Performance**		
	Target: GDM_g (R^2) (Train / Val)	***Target: Dry_Total_g (R^2) (Train / Val)***	***Training time (sec)***
Images + ResNet	-0.4491 / -1.2452	-0.6701 / -1.2970	267.15 (Slowest)
Hand-crafted + Meta + XGBoost	0.9997 / 0.7140	0.9998 / 0.6618	251.00
Hand-crafted + Meta + MLP	0.5519 / 0.3956	0.4625 / 0.1569	112.91 (Fastest)
Sentinel-2 + Features + MLP	0.3524 / 0.5116	0.3898 / 0.2550	126.72
Fusion (ResNet + XGBoost)	0.9511 / 0.6518	0.9999 / 0.5423	182.58

Fig. 2. Running results for 5 models.

Acknowledgment

Thanks to Professor Maxwell and the TA Sihe for your support. I also referred to the Kaggle, DEA, and various online tutorials throughout the project.

References

[1] Yun Chen, Juan Guerschman, Yuri Shendryk, Dave Henry, and Matthew Tom Harrison, "Estimating Pasture Biomass Using Sentinel-

2 Imagery and Machine Learning," Remote Sens. 2021, 13(4), 603, Feb 2021.

[2] Bernardo Cândido, Ushasree Mindala, Hamid Ebrahimy, Zhou Zhang, and Robert Kallenbachan, "Integrating Proximal and Remote Sensing with Machine Learning for Pasture Biomass Estimation," Sensors. 2025, 25(7), Mar 2025.

[3] Qiyu Liao, Dadong Wang, Rebecca Haling, Jiajun Liu, Xun Li, Martyna Plomecka, Andrew Robson, Matthew Pringle, Rhys Pirie, Megan Walker, and Joshua Whelan, "Estimating Pasture Biomass from Top-View Images: A Dataset for Precision Agriculture," arXiv:2510.22916