Manitoba Agronomists' Conference (MAC) 2025

Overcoming hyperspectral data gaps with generative models to improve machine-learning crop stress detection for crop management

Nooshin Noshiri[1], Christopher Henry[1], Christopher Bidinosti[2]

Computer Science Department, University of Manitoba, MB, Canada[1] | Physics Department, University of Winnipeg, MB, Canada[2]

Introduction

Hyperspectral imaging (HSI) is a powerful tool for detecting early signs of crop stress because it captures subtle spectral changes related to nutrients, disease, and environmental conditions. By recording hundreds of narrow wavelengths, HSI produces a three-dimensional hyperspectral cube (Figure 1), enabling detailed assessment of plant health beyond what is possible with standard RGB imaging. However, collecting large, high-quality HSI datasets is difficult and costly. Field and greenhouse imaging require specialized equipment, controlled conditions, and significant time, resulting in small or unbalanced datasets that limit the development of reliable machine-learning tools for crop management. To help overcome these data gaps, this study investigates the use of deep generative models to create realistic synthetic hyperspectral data. By augmenting limited datasets, we aim to strengthen machine-learning performance in agricultural analysis and support more effective, data-driven crop management decisions.

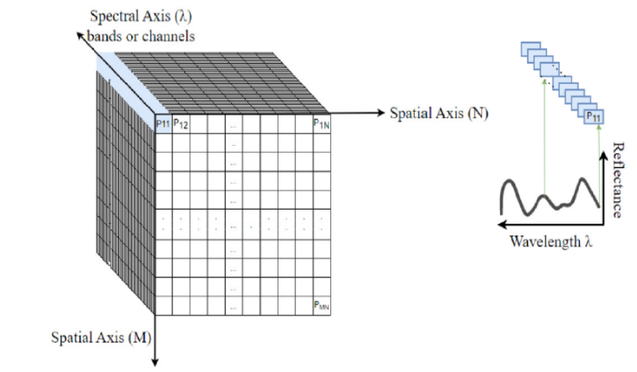


Figure 1. The hyperspectral cube

Objectives and Methodology

This study examines whether deep generative models can help address the challenges associated with the scarcity of HSI datasets by exploring the use of deep generative models to augment spectral data. The objectives are:

- ❑ Evaluate different deep generative models including a customized variational autoencoder (VAE) [2], deep convolutional generative adversarial networks (DCGAN) [3], and Wasserstein GAN with gradient penalty (WGAN-GP) [4], in generating synthetic HSI data.
- ❑ To compare the quality, variability, and representativeness of the augmented data from these generative models to a conventional noise addition approach.

We used hyperspectral images of Buttercrunch lettuce grown under varying nitrogen stress levels, captured with a SPECIM FX10 push-broom camera (spatial resolution 1024 pixels, 224 spectral bands, 400–1000 nm sensitivity with ~2.7 nm spectral resolution; Figure 2). From the 404 captured images, 104 samples were chosen for data augmentation, representing plants grown with a full standard dose of nitrogen. Algorithm 1 was employed for preprocessing, including radiometric calibration and removal of negative values, while Algorithm 2 (adapted from [6]) was used for segmentation to identify key regions of interest. Spectral signatures were then generated by averaging reflectance over all pixels within each segmented area across the 224 bands (Figure 4). A sample segmentation result is shown in Figure 3, and the normalized spectral dataset is visualized in Figure 5.

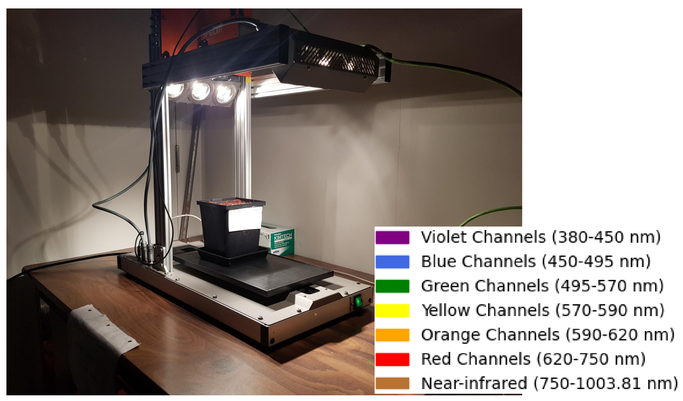


Figure 2: Data acquisition using SPECIM FX10

Algorithm 1. Preprocess hyperspectral data

1: Start
2: Read .raw data
3: Convert .raw data to 3D array
4: Calibrate the data using black and white calibration formula:

$$\text{Calibrated HSI Data} = \frac{\text{Raw HSI Data} - \text{Dark Reference}}{\text{WhiteReference} - \text{DarkReference}}$$

5: Replace negative voxel values with the average of neighboring voxels with values ≥ 0
6: Stop

Algorithm 2. Leaf segmentation from hyperspectral data using NDVI and EGI

1: Input: Hyperspectral image data
2: Output: Segmented leaf areas along all bands
3: Calculate mean across spectral bands for Blue (bands 1-38)
4: Calculate mean across spectral bands for Green (bands 39-76)
5: Calculate mean across spectral bands for Red (bands 77-114)
6: Calculate mean across NIR spectral bands (bands 115+)
7: Compute NDVI for each pixel: $\text{NDVI} = \frac{NIR-Red}{NIR+Red}$
8: Compute EGI for each pixel: EGI = 2 × Green – Blue – Red
9: Create binary masks based on thresholding NDVI and EGI
10: Element-wise multiplication of NDVI and EGI masks using thresholding to form the primary mask
Morphological Operations
11: Apply morphological closing using a 3×3 structuring element
12: Apply morphological opening using a 3×3 structuring element
13:Extract and output the final segmented leaf areas

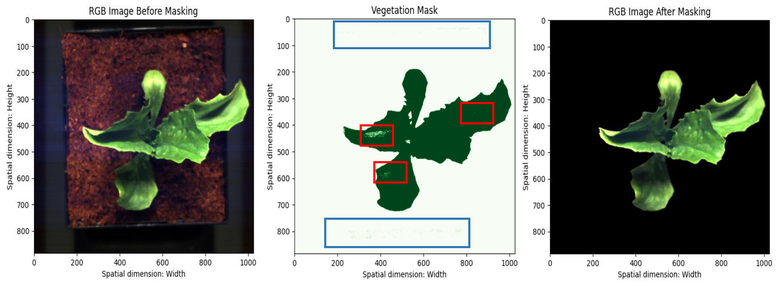


Figure 3. Segmentation process.

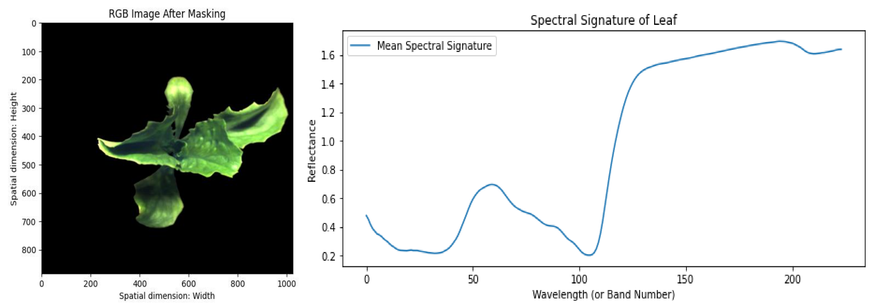


Figure 4. Mean reflectance of segmented regions across 224 bands.

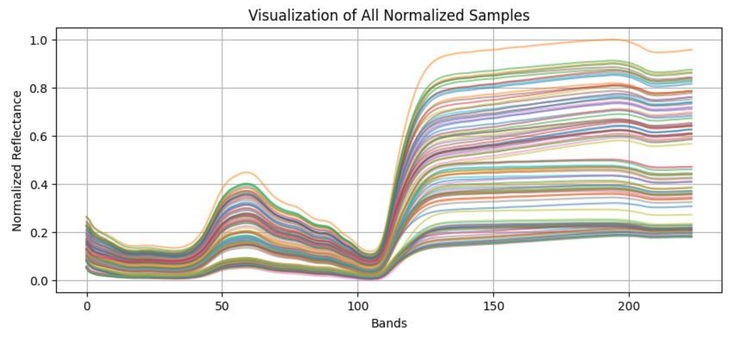
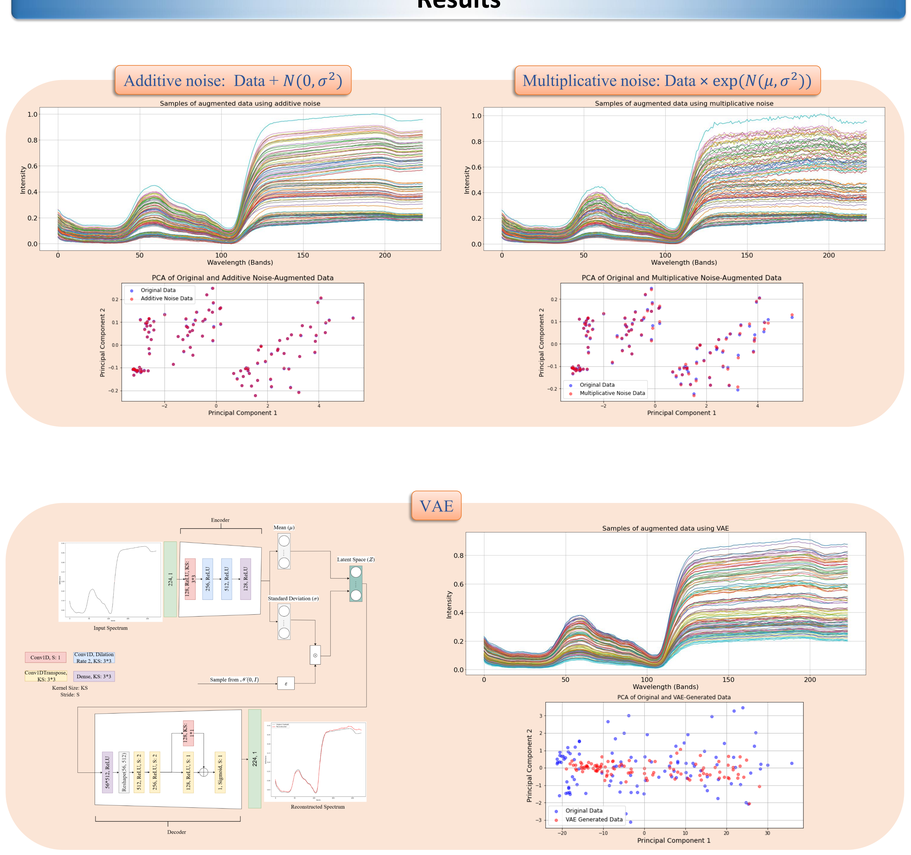


Figure 5. Normalized spectral signatures

Results



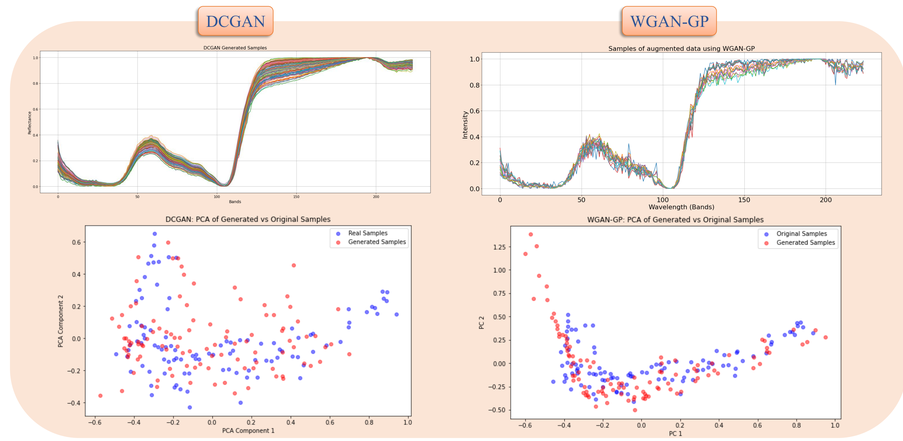


Table 1. Performance Comparison of Data Augmentation Techniques

Metric	Additive Noise	Multiplicative Noise	VAE	WGAN-GP	DCGAN
Cosine Similarity	0.9972	0.9972	0.9985	0.9974	0.9985
Jensen-Shannon Divergence	0.0052	0.0132	0.1908	0.0652	0.0359
KL Divergence	0.0001	0.0007	0.2194	0.0163	0.0053
Nearest Neighbor Distance	0.0149	0.0495	0.1488	0.3759	0.1539

To evaluate augmentation performance, we used four metrics commonly applied to spectral data analysis:

- ❑ Cosine Similarity measures how closely the shape of synthetic spectra aligns with the real FN signature (higher = more similar).
- ❑ Jensen–Shannon (JS) Divergence quantifies how different the synthetic and real data distributions are (lower = more similar).
- ❑ Kullback–Leibler (KL) Divergence evaluates distributional divergence between synthetic and real spectra, though it is asymmetric and more sensitive to mismatches in certain regions.
- ❑ Nearest Neighbor Distance (NND) measures the distance from each synthetic sample to the nearest real sample in feature space (lower = synthetic samples lie closer to the real data manifold).

Quantitative and visual evaluations show distinct differences among augmentation methods (Table 1). Additive noise produced spectra almost identical to the original FN signatures (cosine similarity 0.9972), confirming strong preservation of the data distribution but very limited variability. Multiplicative noise maintained high similarity but introduced more instability (JS = 0.0132, NND = 0.0495) compared to the additive noise. This matches the visual results where spectra appeared noisier and wavier. Among deep generative models, VAE and DCGAN produced the smoothest and least wavy synthetic spectra. VAE achieved the highest cosine similarity (0.9985), showing strong alignment with the average FN spectral shape. Its higher JS and KL values indicate increased sample-to-sample variability, but the spectra remain visually realistic and smooth. DCGAN combined high spectral integrity (0.9985) with low divergence (JS = 0.0359), generating synthetic spectra that are both smooth and highly aligned with the real distribution. In contrast, WGAN-GP produced the waviest and most fluctuating spectra, reflected in its larger NND (0.3759), indicating excessive variability and reduced realism.

Discussion and Future Work

Noise-based augmentation preserved the original spectral shape but contributed little to meaningful variability, limiting its usefulness for machine-learning models that require diverse training samples. Among the generative approaches, VAE and DCGAN performed best, producing smooth, biologically plausible spectra while still expanding dataset diversity. Their quantitative scores—high cosine similarity paired with moderate JS and KL divergence—show that they introduce controlled variation without distorting the underlying FN spectral structure. In contrast, WGAN-GP generated spectra with pronounced fluctuations, as reflected in its higher Nearest Neighbor Distance, which indicates that excessive variability can reduce the usefulness of synthetic samples for downstream modelling.

Future work will focus on developing morphology-aware hyperspectral data synthesis methods that capture both spectral patterns and the structural characteristics of crop leaves. By integrating spatial context, plant shape, and morphology into generative models, we aim to produce synthetic hyperspectral cubes that better reflect real crop physiology and field variability. This direction will involve transformer-based and diffusion architectures capable of learning joint spectral–spatial representations. Morphology-aware synthesis is expected to support more realistic data augmentation, strengthen downstream modeling, and enhance the overall utility of hyperspectral imaging in crop research and management.

Reference

[1] https://doi.org/10.1016/j.atech.2023.100316
[2] https://doi.org/10.48550/arXiv.1312.6114
[3] https://doi.org/10.48550/arXiv.1511.06434
[4] https://doi.org/10.48550/arXiv.1704.00028
[5] https://hdl.handle.net/10680/2127
[6] https://doi.org/10.3389/fpls.2018.01182