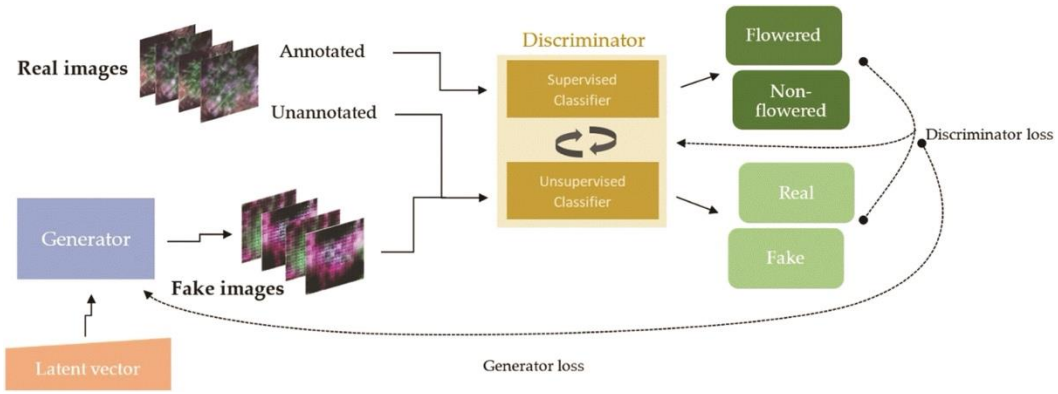
CABBI
CENTER FOR ADVANCED BIOENERGY AND BIOPRODUCTS INNOVATION

Breaking the Barrier of Human-annotated Training Data for Machine Learning-aided Plant Research using Aerial Imagery

Background/Objective

Machine learning (ML) can accelerate biological research. However, the adoption of such tools to facilitate phenotyping based on sensor data has been limited by (i) the need for a large amount of human-annotated training data for each context in which the tool is used and (ii) phenotypes varying across contexts defined in terms of genetics and environment. This is a major bottleneck because acquiring training data is generally costly and time-consuming. This study demonstrates how a ML approach can address these challenges by minimizing the amount of human supervision needed for tool building. The objective is to understand the trade-offs between predictive ability and the level of dependence on manual annotation for each of the algorithms.



ESGAN and data workflow including the generator (G) and discriminator (D) submodels utilized to assess flowering status.

Approach

A case study was performed to compare ML approaches that examine images collected by an uncrewed aerial vehicle to determine the presence/absence of panicles (i.e. "heading") across thousands of field plots containing genetically diverse breeding populations of two species of *Miscanthus*, a highly productive grass crop.

Results

Automated analysis of aerial imagery enabled the identification of heading approximately 9 times faster than in-field visual inspection by humans. Leveraging an Efficiently Supervised Generative Adversarial Network (ESGAN) learning strategy reduced the requirement for human-annotated data by 1 to 2 orders of magnitude compared to traditional, fully supervised learning approaches. The ESGAN model learned the salient features of the data set by using thousands of unlabeled images to inform the discriminative ability of a classifier so that it required minimal human-labeled training data.

Significance/Impacts

This method can accelerate the phenotyping of heading date as a measure of flowering time in *Miscanthus* across diverse contexts (e.g. in multistate trials) and opens avenues to promote the broad adoption of ML tools. More broadly, this work could address the need for advanced modeling techniques that can produce robust accuracy while reducing the operational cost of collecting time-consuming annotated data for many computer vision problems in plant science applications.

Varela et al. 2025. "Breaking the barrier of human-annotated training data for machine-learning-aided plant research using aerial imagery." *Plant Physiology*. DOI: 10.1093/plphys/kiaf132.