HyperLeaf2024 – A Hyperspectral Imaging Dataset for Classification and Regression of Wheat Leaves

William Michael Laprade[1] Pawel Pieta[1] Svetlana Kutuzova[2] Jesper Cairo Westergaard[3]
Mads Nielsen[2] Svend Christensen[3] Anders Bjorholm Dahl[1]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark
[2]Department of Computer Science, University of Copenhagen
[3]Department of Plant and Environmental Sciences, University of Copenhagen

Abstract

Hyperspectral imaging is a widely used method in remote sensing, particularly for use in airborne and satellite-based land surveillance. Its versatility is, however, much larger and has also seen usage in everything ranging from food processing and surveillance to astronomy and waste sorting. It is also gaining inroads with agricultural research. With most available datasets focusing on per-pixel classification, there is, however, a potential for hyperspectral whole-image analysis, but there is a severe lack of datasets for whole-image analysis. To help fill this gap and facilitate methodological development in whole-image hyperspectral image analysis, we introduce the HyperLeaf2024 dataset. The dataset consists of 2410 hyperspectral images of wheat leaves, along with associated classification and regression targets at both the leaf level and the plot level. In addition to the dataset, we also provide experiments showing the importance of pretraining and highlighting the future research direction in whole-image hyperspectral image analysis.

1. Introduction

There is a need for large, high-quality, curated datasets to facilitate the development of deep learning-based algorithms for hyper-spectral imaging data. Existing publicly available datasets [6, 19, 28] are mainly for per-pixel classification tasks and are acquired from satellite [14, 38] or airborne [16, 23] imaging systems. There is great potential for using hyperspectral imaging in whole-image analysis settings where each imaged object has a set of derived properties rather than having properties defined on a per-pixel basis. To allow for developing deep learning-based methods to model such systems, we have created a comprehensive dataset of wheat leaves with associated meta-data.

Hyperspectral images contain reflectance information in each pixel across a portion of the light spectrum, typically in hundreds of bands across the visible or near-infrared spectrum, but also in mid- and long-wavelength infrared. Resolving the light spectrum into spectral bands reveals properties of the imaged material that may not be obtained from RGB images. Hereby, hyperspectral imaging provides insights into the physical and chemical properties of various materials, which has shown to be an effective tool for visual inspection. It has seen use in everything ranging from counterfeit detection [21], astronomy [7] and surveillance [44] to food safety [10], cancer detection [27] and waste sorting [24]. Hyperspectral imaging is also quite often used in agricultural research with investigations into drought stress [36] and disease detection [45], yield prediction [35] nutrient and water stress detection [36] among others.

Several challenges arise in computer vision method development in the field of hyperspectral image analysis compared to analyzing RGB images. Due to the difficulty in acquiring hyperspectral images (and labels) at scale, many of the existing hyperspectral imaging datasets suffer from a number of issues. Most datasets are either too small, not publicaly available and accessible, or lack whole-image labels. Method development for whole-image tasks is difficult without a large benchmark dataset for this use.

To facilitate method development of whole-image hyperspectral image analysis we introduce the HyperLeaf2024 dataset. The HyperLeaf2024 is one of the largest available hyperspectral imaging datasets for whole-image hyperspectral classification and regression. It contains 2410 images of wheat flag leaves with associated targets for cultivar, fertilizer content, yield, stomatal conductance, and chlorophyll fluorescence. It is easy to use and accessible via Kaggle.

1.1. Related work

Hyperspectral imaging datasets provide valuable spectral information for a range of applications. There are a number of existing datasets already available, however, most are severely limited in their applications for whole-image analysis Tab. 1. Among the available datasets, HySpecNet-11k [11] is one of the largest satellite-based hyperspectral imaging datasets currently available. However, it lacks any labels for benchmarking classification and regression models. There is also the LIB-HSI [15] dataset for building facade segmentation, but the labels are per-pixel and not whole-image. Similarly, the small HS-SOD [22] dataset for salient object detection includes only segmentation masks. The ICVL-HSI [1] dataset contains 200 natural outdoor hyperspectral images but lacks labels. Its newer version, Arad1K [2], is also unlabeled and the authors only made the spectrally downsampled versions of the images publicly available, even though they used a camera with a high spectral resolution. The low spectral resolution limits the use of this dataset e.g., for unsupervised pretraining.

A few larger datasets are also publicly available. One of them, the DeepHS [42] fruit dataset, contains over 4600 hyperspectral images, but only a subset, 1018 of these images have associated whole-image labels relating to fruit ripeness. HSIFoodIngr-64 [43] is one the largest datasets we could find that had whole-image labels. It contains nearly 3400 hyperspectral images of various cooked dishes with associated dish names, ingredients, and segmentation masks. The authors showed that the additional spectral information improves segmentation (per-pixel classification). However, they omit experiments on whole-image classification performance even though they mention it could be used for this task. Dish [32, 34] and ingredient [4, 8] classification from RGB images can already be done effectively. This limits the effectiveness of this dataset for whole-image hyperspectral analysis as the spectral information is not essential for the task.

2. Data

Our dataset contains hyperspectral images and physiological measurements of wheat flag leaves as well as variety classification and yield, from a 2023 winter wheat trial in Taastrup, Denmark (55.671060 N, 12.303205 E). Not used in this investigation, but available as additional data, is climate data for the trial location [41].

2.1. Field setup

The trial field was sown in the fall of 2022 and consisted of a total of 144 plots, placed in four block repetitions, in a randomized split plot design, each block (39 m by 20.4 m) with 36 trial plots (each 10 m by 1.25 m) of wheat. Within a block, three columns of 12 plots had differing nutrient applications. The dataset we are releasing here includes imaging acquired from only the northwestern block of the trial Fig. 1.

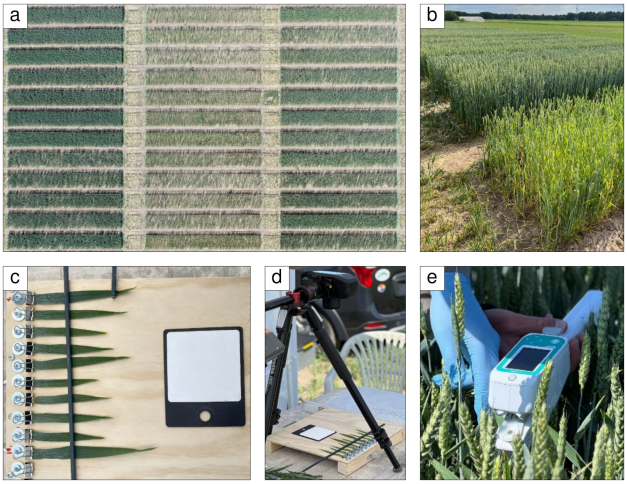

Figure 1. Data collection. a.) The 36 plots in the northwestern quadrant with visible differences in the columns due to varying fertilizer content. b.) The current state of growth during imaging. c.) The imaging platform with ten leaves, a bar, and a white reference. d.) Imaging setup with the camera mounted approximately 40 cm above the leaves. e.) LI-600 measurement device during measurement of a sunlit leaf.

Each plot is defined by three different properties: the wheat cultivar (variety), the amount of nutrition given, and the sowing density. The four cultivars were Heerup, Kvium, Rembrandt, and Sheriff. The nutrient application was given in normal amounts (for the area), half, and none. Sowing densities were chosen to give 150, 300, and 450 plants/m^2.

Fig. 1 shows an example of the crop during the data collection. Data was collected over two days beginning in the morning around 10:00 and completing late afternoon around 16:30. There was no precipitation immediately before or during the acquisition.

2.2. Hyperspectral imaging

The hyperspectral imaging of the leaves was performed with the leaves detached from the plant. Each leaf was clipped onto a wooden surface (see Fig. 1) for a total of ten leaves at a time. To minimize the effects on the reflectance from the leaves drying out before being imaged, a plastic tarp was kept over the leaves until all ten were ready to be taken to the imaging station. During imaging, the leaves were held flat against the surface by a metal rod placed across them (see Fig. 1). Leaves were not cleaned before imaging, so some images contain a visible insect presence.

Imaging was performed in a white tent to homogenize the light. A white reference panel is used to radiometrically calibrate the image signal to the true reflectance. It

Dataset	# of Images	# of Bands	Labels	Whole-image Labels	Task
ICVL-HSI	200	519	No	No	Spectral reconstruction from RGB
ARAD-1K	1000	31	No	No	Spectral reconstruction from RGB
HySpecNet-11k	11483	224	No	No	Image compression
HS-SOD	60	151	Yes	No	Salient object detection
LIB-HSI	513	204	Yes	No	Segmentation
DeepHS Fruit	4671 (1018)	224	Partial	Yes	Regression
HSIFoodIngr-64	3389	204	Yes	Yes	Segmentation and classification
HyperLeaf2024 (ours)	2410	204	Yes	Yes	Classification and regression

Table 1. Existing hyperspectral datasets.

was placed flat on the wooden mounting surface along with the leaves. The camera sensor was placed approximately 40 cm above the imaging area. The camera is the SPECIM IQ [3] (SPECIM, Specim Oulu, Finland) which delivers a hyperspectral image of spatial dimensions 512×512 pixels and 204 spectral bands from 400 to 1000 nm. Each group of ten leaves was imaged at two different exposure settings depending on lighting conditions.

2.3. Physiological sensing

Data on stomatal conductance and chlorophyll fluorescence was acquired using the portable LI-600 porometer/fluorometer (LI-COR Environmental GmbH, Bad Homburg, Germany).

A leaf was chosen, and the LI-600 was placed with the incident-light sensor pointing perpendicular to how the leaf was angled (before being placed under the aperture clamp), and the measurement was performed (see Fig. 1).

Stomatal conductance (g_{sw}) is a measure of the plant's "breathing". It is a measure of gas exchange (CO_2 going in, H_2O going out) through the numerous openings on a leaf where guard cells can control the aperture's closing and opening. The measure is of passage of molecules per volume per time, here in mol m^{-2} s^{-1}. Changes in lighting conditions [12], humidity, and temperature [5] can influence g_{sw}.

A leaf can either use light/photons for photosynthesis and thus growth, let it shine through the leaf (transmittance), reflect it (what we humans see) or change it into heat or fluorescence. Chlorophyll fluorescence measured as quantum yield of PSII (Φ_{PSII}) for light-adapted leaves, is thus a measure of how much light is absorbed by photosystem II chlorophyll [33] and used for photochemistry, and can give not just a real-time indicator of physiology but also depict a plant's flexibility with regards to environmental change [12].

Both g_{sw} and Φ_{PSII} can give insight into how well a plant handles different stresses [37].

2.4. Crop yield

Before harvesting, the plots were cut to 6 m length, to omit the eastern part of each plot, where substantial destructive leaf sampling had taken place months prior. The yield is thus only representative of a 7.5 m^2 crop area in each plot. Grain weight yield was measured in real-time on the combine harvester.

2.5. Leaf segmentation and processing

We train a multi-class U-Net model to segment the leaves, metal bar, and white reference frame. The images are then standardized according to the pixels in the white reference frame to compute the true reflectance values. We then apply connected components to separate them into individual leaf images with spatial dimensions of 48×352 and 204 spectral channels. All non-leaf pixels are filled with zeros to ensure models learn only the information inside the leaf. Finally, the hyperspectral leaf images are then clipped to the 0-1 range and converted to 16-bit tiff files.

2.6. Data split

We define the train, validation, and test split at the plot level. The training set contains 24 plots, the validation contains 4 plots and the testing set contains 8 plots. This is done to avoid data leakage associated with dividing the leaves from the same image among the different sets. We also ensure that each split has, individually, at least one of every cultivar, seeding density, and fertilizer content.

3. Experiments

Several baseline experiments were run using a selection of high-performing deep learning-based methods to get an understanding of how well existing methods perform on this dataset. Each architecture is trained as classification task to predict the cultivar (Heerup, Kvium, Rembrandt, or Sheriff) and again as a regression task to predict the four regression targets (fertilizer content, yield, stomatal conductance, and chlorophyll fluorescence).

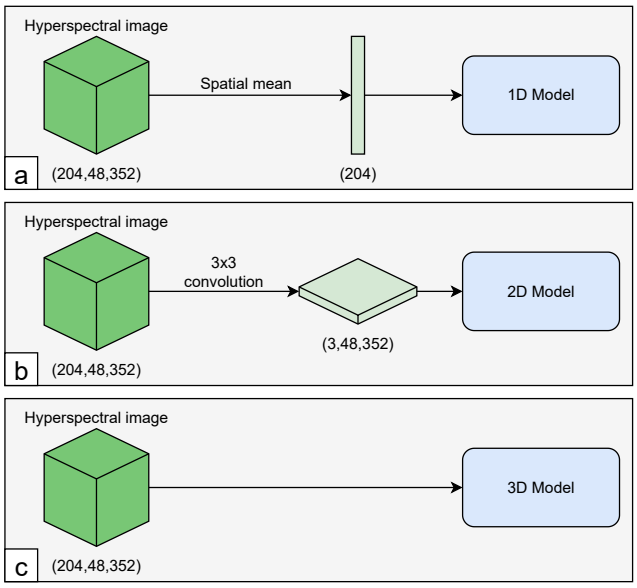

Figure 2. Differences between 1D, 2D, and 3D variants of deep learning architectures. The inputs are hyperspectral images with dimensions (spectral, height, width).

We divided the experiments into three categories: convolutional neural networks, transformers, and unsupervised methods. 2D and 3D variants of most architectures are trained (see Fig. 2 b-c). The 2D variants use a single 3×3 convolution to compress the 204 channels down to 3 channels to facilitate usage with the predefined architectures. In some of the architectures, a 1D variant was trained as well, where a mean was calculated with a reduction along the spatial dimensions of the input data (Fig. 2a). Unless otherwise stated, all described models were trained with a learning rate of 10^{-4} using AdamW [31] optimizer, with an effective batch size of 32.

For data augmentations, we used random horizontal and vertical flipping, random scaling (0.75, 1.25), random shift (0.25,0.25), Gaussian blurring, and Gaussian noise. All experiments use the same data augmentation setup. Per-channel standardization is applied to all images using a mean and standard deviation derived from the training set images.

We report the classification accuracy and mean squared error (MSE) for most models. The MSE is computed after targets have been standardized by the training set mean and standard deviation. This ensures an MSE of 1 is equivalent to simply guessing the mean and MSE of 0 is perfect accuracy.

3.1. Convolutional neural networks

We evaluate the performance across three different architectures. ResNet [17] was chosen as it is a widely used standard in deep learning image analysis. For a more recent, state-of-the-art architecture we chose ConvNeXt [30]. We also looked at MobileNetV3 [20] to get insight into what kind of performance we could achieve with architectures designed for limited hardware resources. We investigate these architectures in 1D, 2D, and 3D variants. The 2D variants are tested both with and without pre-trained ImageNet [39] weights. Each model is trained once as a classification task for predicting cultivar, and again as a regression task for predicting the remaining targets. Each model was trained with an early stopping criterion of validation loss not decreasing for 25 epochs.

Model	1D	2D	3D
ResNet18	4.4	11.7	33.7
ResNet50	18	25.6	48.2
ResNet152	40.4	60.2	119
ConvNeXt-S	48.9	50.2	55.8
ConvNeXt-B	86.5	88.6	96.7
ConvNeXt-L	193	197	211
MobileNetV3-S	2.5	2.5	2.8
MobileNetV3-L	5.4	5.5	5.8
ViT-B	—-	85.1	85.1
ViT-L	—-	302	302
ViT-H	—-	629	629

Table 2. Parameter counts (in millions) for the various architectures.

3.2. Transformers

We also evaluate on the standard ViT [9] architectures in both 2D and 3D variants. The 3D variants use cube patching and take into account all 204 spectral channels. Zero padding is applied to images before patching to ensure divisibility by the patch size. ViT-B and ViT-L use patches of $3 \times 16 \times 16$ in the 2D variant or $1 \times 16 \times 16 \times 16$ in the 3D variants, while ViT-H uses $3 \times 14 \times 14$ in 2D and $1 \times 14 \times 14 \times 14$ in 3D. Positional encodings are learned during training and we opt for average pooling over a `cls` token to obtain a final set of features. We also tried all three variants of SwinV2 [29] in both 2D and 3D but they failed to converge, so we omit their results. Due to a higher volatility in training the early stopping criterion was set to 50 epochs for all transformer-based models.

3.3. Unsupervised methods

We also look into unsupervised pre-training methods with experiments using variational autoencoders (VAE) [25] and masked autoencoders (MAE) [18]. We train the VAEs using 3 different encoder architectures (ResNet50, ConvNeXt-Base, and MobileNetV3-Large) in both 2D and 3D variants. After pre-training, the encoders are then fine-tuned and evaluated on the test data. All VAEs were trained for

300 epochs with a latent space of 1056 dimensions. Binary cross-entropy loss was used as a reconstruction loss. For that reason, we dropped the per-channel standardization step of image pre-processing for VAE training. We used the KL-divergence annealing schedule with $\beta = 0$ for the first 50 epochs and β linearly increasing until reaching 10^{-4} at epoch 300. The pre-trained encoders were then finetuned to two downstream tasks: classification and regression. Both downstream networks consisted of a fully connected layer (1056×512) followed by batch normalization and ReLU activation function, followed by another linear layer (512×4). From the initial pre-trained encoder weights, each downstream model was finetuned end-to-end for the classification and regression tasks for 100 epochs.

Similar experiments are done with the MAEs. The MAEs use the 3 ViT variants in both 2D and 3D as encoders. They all use the same decoder architecture as defined in the MAE paper. For comparison with VAE, the per-channel standardization step is also dropped for these experiments.

4. Results

4.1. Convolutional neural networks

The majority of CNN-based models converged and reached the early stopping criterion, except a 3D version of ConvNeXt. For this ConvNeXt model, there was close to no decrease in training and validation loss both for classification and regression. The final test metrics for all the models can be found in Tab. 3. The scores for models that did not converge are excluded from the average calculation to ensure a fair comparison between the input data variants.

The average metric values in both prediction tasks demonstrate that pre-trained models generally provided better results than their non-pre-trained versions. They were also faster to converge during training. In some cases, the pre-trained versions converged to a higher training loss, which wasn't then reflected in validation or test sets, suggesting stronger generalization power of these models.

In classification, the average values show a clear increase in accuracy after increasing the dimensionality of the input data. This is especially visible for the 1D variant, where the main focus was placed on the spectral information. Although much weaker, an opposite trend can be found in regression results with the best average MSE reached for the 1D variant.

On the level of individual models, the best result in both tasks was reached using ConvNeXt, which is considered to be the state-of-the-art convolutional model for classification. Importantly, it is the smallest version of this model that provided the best result. Similar behavior is also present in the ResNet models, suggesting that small models are a better fit for these tasks. This behavior is no longer true for

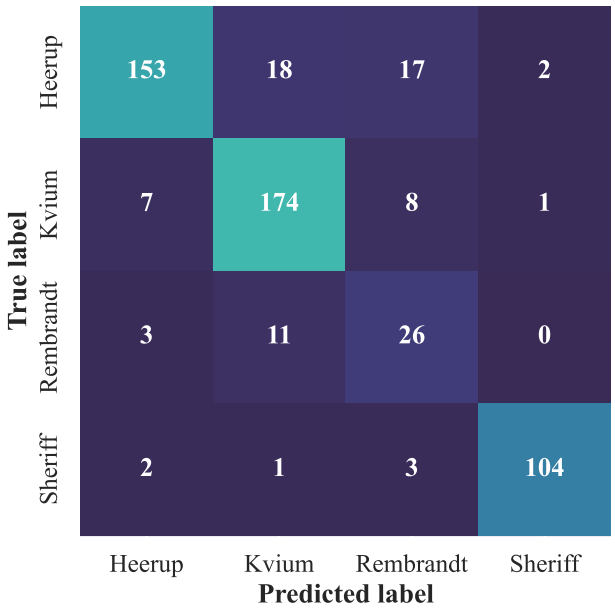

Figure 3. Classification confusion matrix for individual targets, using 2D pre-trained ResNet18.

MobileNetV3 models, but their results are still comparable to the other architectures, despite their design constraints.

We use 2D pre-trained ResNet18 model for a fine-grained analysis of results, as it has achieved high scores in both classification and regression. It is also simple and small enough to provide a representation of general trends found in the data. In classification (confusion matrix on Fig. 3), Kvium and Sheriff have the highest accuracy. It is also rare that other cultivars are incorrectly classified as Sheriff. In regression (Tab. 4), the fertilizer level scores the best, while the stomatal conductance (g_{sw}) is clearly the hardest to predict.

4.2. Transformers

As with convolutional architectures, the vision transformers show improved cultivar classification performance and reduced regression performance when using the 3D variant over the 2D variants, see Tab. 3. We can see that the ViT-H variant with 629M parameters struggles with overfitting during the classification task. Overall, however, the results of the transformer models are worse than the convolutional models. Transformers are known to need a lot of training data, which might explain this result.

4.3. Unsupervised methods

The results of the VAE models, pre-trained in an unsupervised manner, can be found in Tab. 5a. Both classification and regression tasks are better in the 3D versions than the 2D versions. Individually, the ResNet50 model outperforms

Task/Metric	Classification - Accuracy			
Model	1D	2D	2D-P	3D
ResNet18	0.485	**0.840**	0.862	0.845
ResNet50	0.374	0.832	0.815	0.845
ResNet152	0.455	0.826	0.804	**0.874**
ConvNeXt-S	0.534	0.747	**0.893**	0.398*
ConvNeXt-B	0.466	0.721	0.779	0.362*
ConvNeXt-L	**0.572**	0.817	0.840	0.317*
MobileNetV3-S	0.434	0.796	0.832	0.774
MobileNetV3-L	0.389	0.762	0.864	0.740
CNN Average	0.463	0.793	0.836	0.815
ViT-B	—	0.679	—	0.617
ViT-L	—	0.715	—	0.619
ViT-H	—	0.474	—	0.711
ViT Average	—	0.623	—	0.649

(a) Classification results

Task/Metric	Regression - MSE			
Model	1D	2D	2D-P	3D
ResNet18	0.420	0.434	0.388	0.413
ResNet50	0.464	0.421	0.451	**0.388**
ResNet152	0.449	0.481	0.398	0.443
ConvNeXt-S	**0.383**	0.431	0.406	0.645*
ConvNeXt-B	0.421	0.433	0.395	0.626*
ConvNeXt-L	0.401	0.445	0.402	0.592*
MobileNetV3-S	0.471	**0.393**	0.442	0.524
MobileNetV3-L	0.424	0.419	**0.387**	0.443
CNN Average	0.429	0.432	0.409	0.442
ViT-B	—	0.437	—	0.441
ViT-L	—	0.406	—	0.433
ViT-H	—	0.398	—	0.565
ViT Average	—	0.414	—	0.480

(b) Regression results

Table 3. Cultivar classification accuracy and the average mean squared error of the four regression targets. `P` indicates the usage of pre-trained ImageNet weights and * indicates that the model failed to converge and is not included in the average.

Target	MSE	R^2
Grain weight	0.333	0.557
g_{sw}	0.619	0.188
Φ_{PSII}	0.446	0.546
Fertilizer	0.154	0.837

Table 4. Regression MSE and R^2 scores for individual targets, using 2D pre-trained ResNet18.

the others for classification and MobileNetV3-L is the best for regression, while the ConvNeXt models fail to converge. Overall, however, the VAE's show worse results given the same encoder network than the CNN models.

We see even worse performance in the fine-tuned MAE models. Classification performance improves when moving from 2D to 3D and while regression performance stays fairly consistent between 2D and 3D. The ViT-H classification models overfit quickly, while the smaller ViT-B and ViT-L achieve a higher accuracy. All variants perform similarly for the regression task.

5. Discussion

5.1. Data quality

In this study, we have introduced a novel hyperspectral imaging dataset aimed at facilitating method development in hyperspectral image analysis with a focus on whole-image analysis. One of the critical aspects of making a dataset is minimizing sources of error as best as possible during data collection. Most notably, after removing a leaf from the plant it quickly begins to degrade altering its spectral signature. To minimize this effect we did two things: Firstly, we used a plastic tarp to cover the leaves during collection, prior to imaging, to hold in some of the moisture and prevent them from drying out too quickly. Secondly, we made the collection as efficient as possible and thus limited the maximum time a leaf has been detached prior to imaging to approximately 5 minutes. An investigation of the differences in spectral reflectance between the longest detached leaf and the most recently detached leaf in each image suggests this impact is minimal.

The LI-600 measurements of stomatal conductance and chlorophyll fluorescence are sensitive to both the time of day and lighting conditions. This introduces a possible source of noise in the dataset as well. We attempted to minimize the effect of varying sunlight conditions by homogenizing the light using a white tent covering the imaging area. For the full and zero fertilizer plots, we completed multiple passes across the field each day, one in the morning and one in the afternoon to ensure that we have images taken in varying lighting conditions. Further investigation may also benefit from including climate data such as solar radiation, temperature, and humidity.

The accuracy of the true reflectance values is also sensitive to the accuracy of the white reference frame segmentation. To minimize this possible source of error, the mask of the white reference frame includes a binary erosion step as a safety measure to ensure that only pixels inside the white reference are used for white balancing. Similarly, the masks

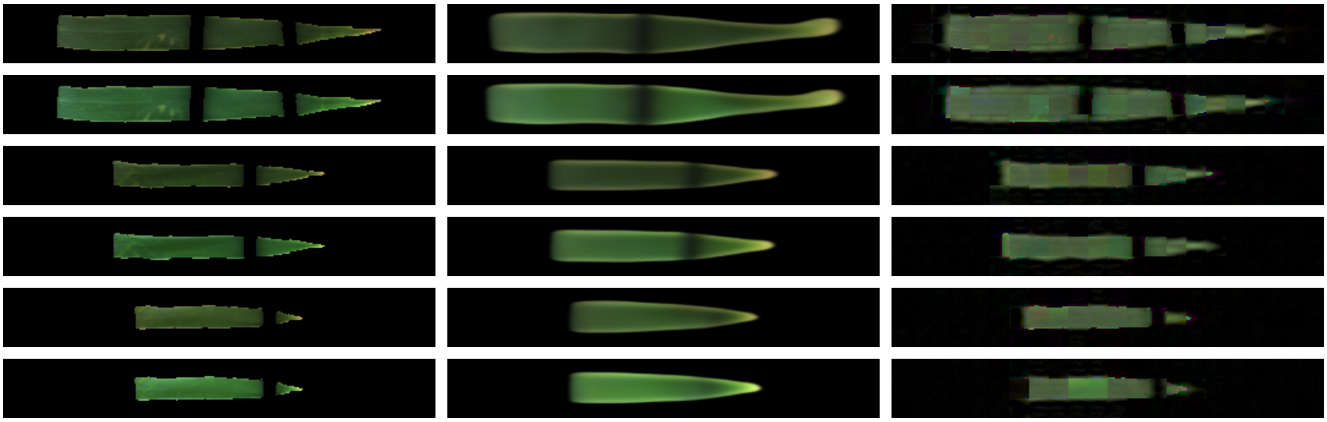

Figure 4. RGB reconstructions from the VAE and MAE models. Left is ground truth, middle is VAE and right is MAE. The vertical gaps in the images are caused by the removal of the pixels corresponding to the metal bars that hold down the leaves during imaging.

Task	Classification		Regression	
VAE Encoder	2D-P	3D	2D-P	3D
ResNet50	0.725	0.842	0.451	0.447
ConvNeXt-B	0.555*	0.574*	0.500*	0.587*
MobileNetV3-L	0.757	0.677	0.445	0.415
Average	0.710	0.759	0.448	0.431

(a) VAE fine-tuning results

Task	Classification		Regression	
MAE Encoder	2D	3D	2D	3D
ViT-B	0.498	0.653	0.432	0.441
ViT-L	0.708	0.551	0.421	0.413
ViT-H	0.309	0.432	0.445	0.435
Average	0.505	0.545	0.433	0.430

(b) MAE fine-tuning results

Table 5. Cultivar classification accuracy and average MSE regression scores for the fine-tuned VAEs and MAEs. P indicates the usage of pre-trained ImageNet weights and * indicates that the model failed to converge and is not included in the average

for the leaves also have binary erosion applied to ensure only leaf pixels are inside the mask. In this case, the erosion operation is less aggressive to avoid losing the tips of the leaves. Thus, there is a possibility for a few non-leaf pixels to be included in some of the segmentation masks.

5.2. Experiments

Model pretraining seems to be important for obtaining good results. In the CNN architectures, models pretrained using ImageNet outperformed the non-pretrained variants quite significantly. The highly generalizable weights from the pretrained models provide an improved starting point for fine-tuning using hyperspectral imaging. This suggests that the knowledge gained by pretraining on a wide variety of data can significantly help, even when we use a naive compression method to fit the hyperspectral images to the pretrained models. It also hints that a large-scale, all-encompassing, hyperspectral imaging dataset similar to ImageNet may provide an effective boost to many hyperspectral imaging models.

Ignoring the pretrained variants and with the exception of 3D ConvNeXt models, which failed to train properly, it seems the 3D CNN models outperform their lower dimensional counterparts for classification, while the regression models follow the opposite trend. It suggests that the classification task is highly dependent on the spatial information while the regression tasks are more dependent on the spectral information.

We found previously [26] with similar data that 3D autoencoders outperform their 2D counterparts on downstream tasks. The fine-tuned VAE results here follow that trend, despite performing worse than the classifcation and regression models. Additional tests showed that this is likely due to the lack of normalization. We dropped the normalization step to facilitate reconstruction with cross-entropy loss. The lack of normalization along with the inherent difficulty in training transformer-based models with limited data also explains the poor MAE results.

In the classification task, we see that Sheriff is the most distinctive of the four cultivars, while Rembrandt is the most difficult to identify. Given that you can visually see the differences between the amount of fertilizer (light to dark green) (Fig. 1), it is expected that the easiest regression task is predicting fertilizer content. Fertilizer usage correlates with yield [40], so it is not unexpected that the grain weight is the second easiest target to predict. The stomatal conductance and chlorophyll fluorescence however, are the most difficult to predict, with chlorophyll fluorescence

being slightly easier due to chlorophyll's association with reflectance [13]. A possible reason for the difficulty in predicting these two values is due to their high variability depending on the time of day and lighting conditions.

Models trained on this dataset will likely not perform well on other wheat experiments (e.g., different soil types, cultivars, climate, etc.). However, such use is not the purpose of this dataset. We provide this dataset for whole-image hyperspectral method development. Even though we have here a specialized agriculture research task, it functions well for method development because it provides tasks that require the integration of the spatial and spectral properties in different ways. Namely, the classification task is more dependent on the spatial information and the regression task depends more on the spectral information. Methods that are developed to perform well on both of these tasks will ideally perform well on a new dataset that may require the integration of the spatial and spectral information in a different manner.

6. Conclusion

The HyperLeaf2024 dataset presents a significant resource for advancing whole-image hyperspectral imaging methodologies. With over 2400 hyperspectral images accompanied by whole-image targets, this dataset will facilitate the development and evaluation of hyperspectral image analysis algorithms. Crucially, having distinct spatially-dependent and spectrally-dependent targets for each image allows the development of methods that excel in both aspects and that can adapt to the varying requirements in hyperspectral applications.

Despite the naive method of compression we employ to fit the hyperspectral images to the pretrained 2D models, they still outperform all non-pretrained variants. These experiments highlight the importance of large-scale, comprehensive datasets in pretraining hyperspectral imaging models. They not only demonstrate the value of pretraining in hyperspectral image analysis but also highlight a clear path forward for future research in hyperspectral imaging.

With HyperLeaf2024, we have a foundation for further advancing deep learning-based analysis models for whole-image hyperspectral imaging both for the specific use in agricultural research, but also as a basis for the general wider use of hyperspectral imaging.

References

[1] Boaz Arad and Ohad Ben-Shahar. *Sparse Recovery of Hyperspectral Signal from Natural RGB Images*, page 19–34. Springer International Publishing, 2016. 2

[2] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Luc Van Gool, Shuai Liu, Yongqiang Li, Chaoyu Feng, Lei Lei, Jiaojiao Li, Songcheng Du, Chaoxiong Wu, Yihong Leng, Rui Song, Mingwei Zhang, Chongxing Song, Shuyi Zhao, Zhiqiang Lang, Wei Wei, Lei Zhang, Renwei Dian, Tianci Shan, Anjing Guo, Chengguo Feng, Jinyang Liu, Mirko Agarla, Simone Bianco, Marco Buzzelli, Luigi Celona, Raimondo Schettini, Jiang He, Yi Xiao, Jiajun Xiao, Qiangqiang Yuan, Jie Li, Liangpei Zhang, Taesung Kwon, Dohoon Ryu, Hyokyoung Bae, Hao-Hsiang Yang, Hua-En Chang, Zhi-Kai Huang, Wei-Ting Chen, Sy-Yen Kuo, Junyu Chen, Haiwei Li, Song Liu, Sabarinathan Sabarinathan, K Uma, B Sathya Bama, and S. Mohamed Mansoor Roomi. Ntire 2022 spectral recovery challenge and data set. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2022. 2

[3] Jan Behmann, Kelvin Acebron, Dzhaner Emin, Simon Bennertz, Shizue Matsubara, Stefan Thomas, David Bohnenkamp, Matheus Kuska, Jouni Jussila, Harri Salo, Anne-Katrin Mahlein, and Uwe Rascher. Specim iq: Evaluation of a new, miniaturized handheld hyperspectral camera and its application for plant phenotyping and disease detection. *Sensors*, 18(2):441, 2018. 3

[4] Marc Bolaños, Aina Ferrà, and Petia Radeva. Food ingredients recognition through multi-label learning, 2017. 2

[5] James A. Bunce. Responses of stomatal conductance to light, humidity and temperature in winter wheat and barley grown at three concentrations of carbon dioxide in the field. *Global Change Biology*, 6(4):371–382, 2000. 3

[6] G. Camps-Valls and L. Bruzzone. Kernel-based methods for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 43(6):1351–1362, 2005. 1

[7] P. Chauhan, Prabhjot Kaur, N. Srivastava, Rishitosh Sinha, Nirmala Jain, and S. Murty. Hyperspectral remote sensing of planetary surfaces: An insight into composition of inner planets and small bodies in the solar system. *Current Science*, 108:915–924, 2015. 1

[8] Jingjing Chen, Bin Zhu, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. A study of multi-task and region-wise deep learning for food ingredient recognition. *IEEE Transactions on Image Processing*, 30:1514–1526, 2021. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 4

[10] Yao-Ze Feng and Da-Wen Sun. Application of hyperspectral imaging in food safety inspection and control: A review. *Critical Reviews in Food Science and Nutrition*, 52 (11):1039–1058, 2012. 1

[11] Martin Hermann Paul Fuchs and Begüm Demir. Hyspecnet-11k: A large-scale hyperspectral benchmark dataset, 2023. 2

[12] Théo Gerardin, Cyril Douthe, Jaume Flexas, and Oliver Brendel. Shade and drought growth conditions strongly impact dynamic responses of stomata to variations in irradiance in nicotiana tabacum. *Environmental and Experimental Botany*, 153:188–197, 2018. 3

[13] Anatoly A. Gitelson, Yuri Gritz †, and Mark N. Merzlyak. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *Journal of Plant Physiology*, 160(3):271–282, 2003. 8

[14] Luis Guanter, Hermann Kaufmann, Karl Segl, Saskia Foerster, Christian Rogass, Sabine Chabrillat, Theres Kuester, André Hollstein, Godela Rossner, Christian Chlebek, Christoph Straif, Sebastian Fischer, Stefanie Schrader, Tobias Storch, Uta Heiden, Andreas Mueller, Martin Bachmann, Helmut Mühle, Rupert Müller, Martin Habermeyer, Andreas Ohndorf, Joachim Hill, Henning Buddenbaum, Patrick Hostert, Sebastian van der Linden, Pedro Leitão, Andreas Rabe, Roland Doerffer, Hajo Krasemann, Hongyan Xi, Wolfram Mauser, Tobias Hank, Matthias Locherer, Michael Rast, Karl Staenz, and Bernhard Sang. The enmap spaceborne imaging spectroscopy mission for earth observation. *Remote Sensing*, 7(7):8830–8857, 2015. 1

[15] Nariman Habili, Ernest Kwan, Weihao Li, Christfried Webers, Jeremy Oorloff, Mohammad Ali Armin, and Lars Petersson. *A Hyperspectral and RGB Dataset for Building Façade Segmentation*, page 258–267. Springer Nature Switzerland, 2023. 2

[16] Christine Hbirkou, Stefan Pätzold, Anne-Katrin Mahlein, and Gerhard Welp. Airborne hyperspectral imaging of spatial soil organic carbon heterogeneity at the field-scale. *Geoderma*, 175–176:21–28, 2012. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 4

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 4

[19] Lin He, Jun Li, Chenying Liu, and Shutao Li. Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines. *IEEE Transactions on Geoscience and Remote Sensing*, 56(3):1579–1597, 2018. 1

[20] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3, 2019. 4

[21] Shuan-Yu Huang, Arvind Mukundan, Yu-Ming Tsao, Youngjo Kim, Fen-Chi Lin, and Hsiang-Chen Wang. Recent advances in counterfeit art, document, photo, hologram, and currency detection using hyperspectral imaging. *Sensors*, 22 (19):7308, 2022. 1

[22] Nevrez Imamoglu, Yu Oishi, Xiaoqiang Zhang, Guanqun Ding, Yuming Fang, Toru Kouyama, and Ryosuke Nakamura. Hyperspectral image dataset for benchmarking on salient object detection, 2018. 2

[23] Jianxin Jia, Yueming Wang, Jinsong Chen, Ran Guo, Rong Shu, and Jianyu Wang. Status and application of advanced airborne hyperspectral imaging technology: A review. *Infrared Physics & Technology*, 104:103115, 2020. 1

[24] Ali Can Karaca, Alp Erturk, M. Kemal Gullu, M. Elmas, and Sarp Erturk. Automatic waste sorting using shortwave infrared hyperspectral imaging system. In *2013 5th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2013. 1

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. 4

[26] William Michael Laprade, Jesper Cairo Westergaard, Jon Nielsen, Mads Nielsen, and Anders Bjorholm Dahl. *An Analysis of Spatial-Spectral Dependence in Hyperspectral Autoencoders*, page 191–202. Springer Nature Switzerland, 2023. 7

[27] Raquel Leon, Beatriz Martinez-Vega, Himar Fabelo, Samuel Ortega, Veronica Melian, Irene Castaño, Gregorio Carretero, Pablo Almeida, Aday Garcia, Eduardo Quevedo, Javier A. Hernandez, Bernardino Clavo, and Gustavo M. Callico. Non-invasive skin cancer diagnosis using hyperspectral imaging for in-situ clinical support. *Journal of Clinical Medicine*, 9(6):1662, 2020. 1

[28] Shutao Li, Weiwei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jon Atli Benediktsson. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9): 6690–6709, 2019. 1

[29] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. 2021. 4

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. 4

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. 4

[32] Runyu Mao, Jiangpeng He, Zeman Shao, Sri Kalyan Yarlagadda, and Fengqing Zhu. *Visual Aware Hierarchy Based Food Recognition*, page 571–598. Springer International Publishing, 2021. 2

[33] Kate Maxwell and Giles N. Johnson. Chlorophyll fluorescence—a practical guide. *Journal of Experimental Botany*, 51(345):659–668, 2000. 3

[34] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition, 2021. 2

[35] Ali Moghimi, Ce Yang, and James A. Anderson. Aerial hyperspectral imagery and deep neural networks for high-throughput yield phenotyping in wheat. *Computers and Electronics in Agriculture*, 172:105299, 2020. 1

[36] Mohd Shahrimie Mohd Asaari, Stien Mertens, Lennart Verbraeken, Stijn Dhondt, Dirk Inzé, Koirala Bikram, and Paul Scheunders. Non-destructive analysis of plant physiological traits using hyperspectral imaging: A case study on drought stress. *Computers and Electronics in Agriculture*, 195:106806, 2022. 1

[37] Hafeez Noor, Min Sun, Hussah I. M. Algwaiz, Alam Sher, Sajid Fiaz, KOTB A. Attia, Shabir Hussain Wani, Muneera D. F. AlKahtani, Latifa Al Husnain, Wen Lin, and Zhiqiang Gao. Chlorophyll fluorescence and grain filling characteristic of wheat (triticum aestivum l.) in response to nitrogen application level. *Molecular Biology Reports*, 49(7): 7157–7172, 2022. 3

[38] Shen-En Qian. Hyperspectral satellites, evolution, and development history. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:7032–7056, 2021. 1

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 4

[40] W. M. Stewart, D. W. Dibb, A. E. Johnston, and T. J. Smyth. The contribution of commercial fertilizer nutrients to food production. *Agronomy Journal*, 97(1):1–6, 2005. 7

[41] Simon Fiil Svane, Jesper Svensgaard, and Carsten Tilbæk Petersen. Meteorological data from the taastrup climate and water balance station 2014-2023, 2024. 2

[42] Leon Amadeus Varga, Jan Makowski, and Andreas Zell. Measuring the ripeness of fruit with hyperspectral imaging and deep learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021. 2

[43] Xiaojie Xia, Wei Liu, Liuan Wang, and Jun Sun. Hsifoodingr-64: A dataset for hyperspectral food-related studies and a benchmark method on food ingredient retrieval. *IEEE Access*, 11:13152–13162, 2023. 2

[44] P WT Yuen and M Richardson. An introduction to hyperspectral imaging and its application for security, surveillance and target acquisition. *The Imaging Science Journal*, 58(5): 241–253, 2010. 1

[45] Xiaohu Zhao, Jingcheng Zhang, Yanbo Huang, Yangyang Tian, and Lin Yuan. Detection and discrimination of disease and insect stress of tea plants using hyperspectral imaging combined with wavelet analysis. *Computers and Electronics in Agriculture*, 193:106717, 2022. 1