Multimodal Wheat Disease Diagnosis using RGB, Multispectral, and Hyperspectral UAV Imagery: From Baselines to Auxiliary-Head Late-Fusion Models

Kshitij Raj Sharma
Université Bretagne Sud
Vannes, France
skshitizraj@gmail.com

Sahar Mohamed
Université Bretagne Sud
Vannes, France
sahaarr208@gmail.com

Arunima Sen
Université Bretagne Sud
Vannes, France
arunimasen2001@gmail.com

***Abstract*—This paper summarizes methods and experiments developed for the Beyond Visible Spectrum: AI for Agriculture 2026 competition (Task 1), which targets automated wheat disease diagnosis from aligned RGB, multispectral (MS), and hyperspectral (HS) UAV patches. We detail the sequence of approaches explored: classical classifiers on hand-crafted features, late-fusion convolutional neural networks (CNNs) that jointly encode RGB/MS/HS, stacking with tree-based models, and a final improved late-fusion CNN with auxiliary per-modality heads, focal loss, and hand-crafted HS descriptors without PCA. We report out-of-fold, validation, and public leaderboard metrics, analyze what worked and what failed (including leakage in naïve stacking), and provide practical lessons for robust multimodal fusion. The final auxiliary-head late-fusion model achieves approximately 0.607 accuracy and 0.603 macro-F1 out-of-fold, and we outline promising directions to approach an accuracy target of 0.80.**

***Index Terms*—multimodal fusion, hyperspectral imaging, multispectral indices, late fusion, focal loss, auxiliary heads, crop disease classification.**

I. Introduction

Automatic detection of crop disease from remote sensing is crucial for precision agriculture, enabling early intervention and targeted treatment in large-scale wheat fields. Hyperspectral imagery provides rich spectral cues that capture subtle biochemical changes, while multispectral vegetation indices (e.g., NDVI, NDWI) encode targeted physiological information and RGB encodes spatial texture. However, HS data are high-dimensional, prone to noisy bands and overfitting, so careful fusion and regularization strategies are required.

This work, prepared as part of the Beyond Visible Spectrum: AI for Agriculture 2026 competition (Task 1), investigates strategies for combining RGB, MS, and HS information in a supervised classification setup with three classes: Health, Other, and Rust. We compare:

- classical machine learning on hand-crafted spectral and texture features;
- end-to-end late-fusion CNNs with separate modality encoders;
- stacking and ensembling with tree models on top of CNN features;
- architectural and training modifications—auxiliary heads, focal loss, and HS-specific design—to improve robustness and modality balance.

II. Dataset and Preprocessing

A. Dataset description

Each sample consists of three spatially aligned modalities:

- RGB: PNG images with 3 channels;
- MS: GeoTIFFs with 5 bands (Blue, Green, Red, Red-Edge, NIR);
- HS: GeoTIFFs with nominally 125 bands spanning approximately 450–950 nm.

The training set contains 600 labelled samples, while the competition validation/test split consists of 300 IDs for which predictions must be submitted. A small subset of totally black images was discovered in the training set and removed for training only; however, the full list of 300 validation/test IDs is preserved for submission.

B. Preprocessing steps

RGB patches are loaded, converted from BGR to RGB if needed, normalized per channel using dataset statistics, and resized to a common spatial resolution. MS patches are padded/cropped to 5 bands and used to compute spectral indices such as NDVI, GNDVI, NDRE, NDWI, and SAVI, which are concatenated with the raw bands to form a 10-channel MS tensor. HS patches are trimmed to remove noisy edge bands, the spectral axis is padded or cropped to a common channel count across files, and the spatial size is resized to match the network input; in the final pipeline, we explicitly avoid PCA and instead rely on a learnable spectral projection.

Data cleaning removes obviously corrupted or blank samples from the training set, but the corresponding IDs remain in the evaluation list so that predictions can still be produced for all required cases.

III. Problem Formulation

This task is formulated as a supervised multi-class classification problem.

Given:

- A dataset of multimodal UAV images per wheat patch (RGB, MS, HS),
- A corresponding label $y \in \{Health, Rust, Other\}$,

the goal is to learn a function:

$$f(X_{\text{rgb}}, X_{\text{ms}}, X_{\text{hs}}) \rightarrow y$$

that maximizes classification performance on unseen data. The primary evaluation metric is the macro F1-score, defined as the unweighted average of per-class F1 scores:

$$F1_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} F1_c$$

where C is the number of classes.

This metric ensures that each class contributes equally, preventing the model from favoring majority classes. Because the dataset is relatively small and high-dimensional (approximately 600 features versus approximately 577 samples), the core challenge is balancing model expressiveness with overfitting control. The combination of per-fold feature selection, regularized gradient boosting, and ensemble averaging was designed to address this trade-off.

IV. Methods

All deep-learning experiments are implemented in PyTorch, using `timm` backbones for the RGB branch. We employ stratified K-fold cross-validation with $K = 5$ and report out-of-fold (OOF) metrics where available.

A. Baseline classical classifiers

We first consider classical models to provide fast sanity checks. An XGBoost classifier is trained either on hand-crafted features or on embeddings/probabilities extracted from CNNs. A logistic regression model is also used as a simple linear stacker on top of CNN outputs. These baselines help identify obvious data or label issues before investing in more complex architectures.

B. Late-fusion CNN (baseline)

The baseline late-fusion CNN employs three modality-specific branches followed by a shared classifier. The RGB branch uses a pretrained `timm` backbone (e.g., ConvNeXt-Tiny or EfficientNet) followed by global pooling and a linear projection into a compact embedding. The MS branch uses a small convolutional encoder applied to the 10-channel MS tensor (raw bands plus indices) to produce a spatially pooled embedding. Early experiments on the HS branch employed PCA to reduce spectral dimensionality, but the final version replaces PCA with a 1×1 learnable spectral projection followed by a spatial CNN encoder. The three embeddings are concatenated and passed through a linear classification head.

Training uses class-balanced sampling and standard augmentations (horizontal/vertical flips, 90° rotations); later experiments add label smoothing or focal loss. At inference time, test-time augmentation (TTA) averages predictions over flips and rotations.

C. Stacking and XGBoost on CNN features

To exploit the representational power of CNNs and the flexibility of gradient-boosted trees, we perform second-stage learning on top of CNN outputs. For each fold, we extract OOF embeddings or probability vectors from the late-fusion CNN and train XGBoost or logistic regression models on these features. A key lesson is that stacking must be trained only on true OOF predictions; naïve stacking that fits on training predictions leads to severe over-optimism (train accuracy up to 0.96) and fails to generalize.

D. Feature Engineering with LightGBM

As an alternative to deep learning, we explored handcrafted feature engineering with gradient boosting.

1) Feature Extraction: We engineered features across modalities and spatial domains. RGB: texture descriptors (GLCM statistics, LBP histograms), color statistics (HSV moments), and per-channel statistics. MS: per-band statistics plus vegetation indices (NDVI, NDRE, GNDVI, SAVI, EVI); band order was corrected empirically (Band 1=NIR, Band 4=Red) after debugging negative NDVI values. HS: spectral statistics on sampled bands, first/second derivatives, and curve characteristics across 101 bands. Spatial: quadrant analysis (mean/variance), edge-vs-center contrast, and gradient magnitude features. This yields $\sim$150–185 features total, from which we select the top 120 by LightGBM feature importance.

2) Model Configuration: LightGBM uses $n_{\text{estimators}} = 2000$–$3000$, $\text{max\_depth} = 8$–10, and $\text{learning\_rate} = 0.008$–$0.015$. Class balancing applies a weight of 1.5–1.8$\times$ to Health. Pseudo-labeling incorporates 76–112 high-confidence test samples (threshold 0.75–0.80) with reduced weights (0.4–0.5$\times$). A 5-model ensemble uses different random seeds for diversity. Results achieve $\sim$55% OOF macro-F1; vegetation indices rank highest in importance, but performance trails CNNs, likely due to limited spatial resolution (64$\times$64 RGB, 32$\times$32 HS vs. typical 224$\times$224 inputs).

E. Improved late-fusion with auxiliary heads

After diagnosing performance issues and HS underutilization, we design an improved late-fusion model with several modifications.

1) HS treatment without PCA: We remove PCA on HS because it tended to compress discriminative spectral cues for rust vs. healthy crops. Instead, we apply a learnable 1×1 spectral projection that reduces the number of channels while allowing the network to discover task-relevant spectral combinations.

2) Hand-crafted HS descriptors: To provide robust low-dimensional cues, we compute per-pixel or per-patch HS descriptors (mean, standard deviation, minimum, maximum, median, spectral centroid, etc.) and append them to the raw HS channels. These features complement the learned spectral embedding and improve the resilience of the HS branch to noise.

3) Auxiliary modality heads: We attach small classifier heads to each modality branch (RGB, MS, HS) and define the total loss as

$$L = L_{\text{main}} + \alpha \, \frac{L_{\text{rgb}} + L_{\text{ms}} + L_{\text{hs}}}{3}, \tag{1}$$

where L_{main} is the loss of the fused classifier, L_{rgb}, L_{ms}, and L_{hs} are the auxiliary losses, and α controls their relative weight. This encourages each modality to remain predictive and reduces the tendency of the model to ignore the HS branch.

4) Loss, regularization, and ensembling: To address class imbalance (Health being the hardest class), we employ focal loss with focusing parameter $\gamma = 2.0$ and class weights, combined with weighted sampling. Regularization includes mixup, flips and rotations, coarse dropout, and other spatial augmentations. At evaluation, we average predictions across folds and TTA transforms to obtain final probabilities.

V. Experiments and Results

We report metrics with explicit labels: OOF (out-of-fold CV), VAL (held-out validation during training), PUBLIC (Kaggle public leaderboard), and TRAIN (training set).

A. Representative performance

Table I summarizes representative accuracy and macro-F1 scores for key methods.

Some models, particularly naïve stacked XGBoost on training predictions, report very high training accuracy (e.g., 0.96), but these are not reliable indicators of generalization due to leakage.

B. Classical hand-crafted + LightGBM pipeline

In addition to the deep pipeline, we develop a strong classical approach based on engineered features and LightGBM. Aligned HS (125 bands), MS (5 bands), and RGB (3 channels) patches are converted into approximately 600 features spanning spectral, spatial, texture, and cross-modal descriptors. We then apply minimum Redundancy Maximum Relevance (mRMR) feature selection per fold, filter low-variance features, and standardize the selected features.

Optionally, we perform hyperparameter optimization with Optuna over:

- the number of selected features and variance threshold;
- LightGBM parameters (number of trees, depth, learning rate, subsampling, regularization, etc.);
- a class weight for the Health class.

We train an ensemble of three LightGBM models with different seeds and cap the number of boosting rounds at the median of the best iterations across folds to reduce overfitting.

A representative 5-fold CV shows mean train macro-F1 of about 0.812 and validation macro-F1 of about 0.708, with a train–val gap near 0.10. A typical OOF classification report yields macro-F1 around 0.71 with Health (precision 0.61, recall 0.55), Other (0.84, 0.75), and Rust (0.69, 0.82).

VI. Discussion

Our experiments highlight several practical lessons for multimodal wheat disease classification. First, HS is powerful but fragile: blindly applying PCA can discard discriminative spectral structure, so learnable spectral embedding and domain-specific descriptors are preferable. Second, auxiliary heads encourage each modality to remain predictive and mitigate the risk that the network relies solely on RGB or MS. Third, stacking requires honest OOF predictions; otherwise, leakage produces misleadingly high training scores and poor leaderboard performance.

We also observe that maintaining the full validation/test ID list is essential in competition settings, even if some inputs are degenerate, since submissions must provide predictions for all IDs. Class imbalance, with Health being the hardest class, benefits from focal loss, class weighting, and weighted sampling.

VII. Conclusions and Future Work

We evaluated a range of methods for multimodal crop disease classification from aligned RGB, MS, and HS UAV imagery in the context of the AI for Agriculture 2026 challenge. Our recommended pipeline centers on a late-fusion CNN with a learned HS spectral projector (no PCA), appended HS descriptors, auxiliary per-modality heads, focal loss, weighted sampling, K-fold ensembling, and TTA. The final robust OOF performance is approximately 0.607 accuracy and 0.603 macro-F1.

To approach an ambitious 0.80 accuracy target, promising directions include stronger RGB backbones (e.g., EfficientNet-B3, ConvNeXt-B), per-sample modality attention or gating to exploit HS selectively, self-supervised pretraining on unlabeled HS/MS stacks, careful pseudo-labeling, and more aggressive domain-specific augmentations (spectral shifts, simulated illumination, and lesion patterns).

Acknowledgements

We thank the Kaggle competition organizers and the dataset providers and our teacher, Professor Sébastien Lefèvre, for encouraging us to participate in this.

TABLE I
REPRESENTATIVE PERFORMANCE OF SELECTED METHODS.

Method	**Notes**	**Accuracy**	**Macro-F1**
XGBoost on CNN features	initial public submission	0.6421	—
Stacked XGBoost (train-only)	overfit, not honest	0.9600	0.9598
CNN late-fusion (single)	baseline CNN, no aux	0.6667	0.6652
XGBoost on CNN probabilities	XGB on OOF probs	0.6133	0.6094
Honest stacked XGBoost	proper OOF stacking	0.4367	0.4318
Logistic regression stack	linear stack	0.3633	0.3388
PCA + XGBoost	HS PCA then XGB	0.4167	0.3944
LightGBM, hand-crafted	texture + spectral feats	0.5500	0.5539
LightGBM + indices	veg. indices + derivatives	0.5500	0.5490
LightGBM ensemble + PL	5 models, pseudo-labels	0.4800	0.4782
Improved late-fusion	aux heads, no PCA	0.6066	0.6026