

UAV-Based Nitrogen Assessment in Wheat: A Systematic Review of Target Traits, Validation Rigor, Growth Stage Evidence, and Machine Learning Approach

Muhammad Waqar Nasir^a, Mawra Ishaq^a, Muhammad Haseeb Javaid^b, Muhammad Daniyal Junaid^c, Changwei Tan^a

^aJiangsu Key Laboratory of Crop Genetics and Physiology/Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops, Agricultural College of Yangzhou University, Yangzhou 225009, China

^bInstitute of Vegetable Sciences, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China

^cFaculty of Life Sciences, Department of Biology, University of Okara, Pakistan

*Corresponding author: 009048@yzu.edu.cn

Supplementary Note S1. Decision Rules for Coding and Comparative review

Operational rules used to harmonize the extracted dataset and select study-level values for comparative analysis

S1. Purpose

This note documents the operational rules used to convert the raw extraction workbook into the final harmonized dataset and the analysis-ready subsets used for the manuscript. The purpose of these rules is to make the review transparent, reproducible, and internally consistent.

S2. Study inclusion and final dataset

- Only studies retained after title, abstract, and full-text screening were eligible for harmonization and comparative review.
- The final study pool comprised 88 articles. Harmonization and coding did not change the final study count.
- Studies removed after full-text review because the complete article was not accessible were not carried forward into the analytical dataset.

S3. Rules for coding target directness

- Direct N was used when the primary target was a direct nitrogen-status variable, such as plant nitrogen concentration (PNC), plant nitrogen accumulation (PNA), leaf nitrogen concentration (LNC), leaf nitrogen accumulation (LNA), canopy nitrogen content (CNC), or nitrogen nutrition index (NNI).
- Close proxy was used when the target was strongly related to nitrogen status but was not itself a direct nitrogen variable, for example SPAD, chlorophyll, or canopy chlorophyll content.
- Indirect proxy was used when the target was a broader agronomic trait influenced by nitrogen, such as biomass, yield, or leaf area index.
- Treatment class only was used when the study predicted nitrogen treatment classes or related categorical labels rather than a continuous nitrogen-status variable.
- When multiple outcomes were reported, the primary study-level coding was assigned to the outcome judged most central to the nitrogen-diagnosis purpose of the study, based on the abstract, methods, results emphasis, and reported best model.

S4. Rules for coding validation rigor

- Internal-only validation was used for random train–test splits, holdout validation within the same dataset, k-fold cross-validation, repeated internal resampling, and related within-dataset partitioning approaches.
- Independent-like validation was used when the validation design included temporal separation, spatial separation, site-to-site transfer, experiment-to-experiment transfer, or any explicitly external or cross-domain testing scenario.

S5. Rules for coding growth stage

- Growth-stage terms reported in the articles were standardized to the following labels: regreening, tillering, jointing, booting, heading, anthesis, grain filling, and maturity.
- Stage assessment type was coded as single-stage, multi-stage, or unclear depending on the overall sampling design.
- Best-stage comparison was used only when the article explicitly identified one practically optimal stage or when the evidence clearly supported one selected stage for comparison.
- Studies that sampled multiple stages but did not explicitly isolate one best stage were retained for stage-frequency counting but were excluded from best-stage comparison.

S6. Rules for sensor and model harmonization

- Sensor type was harmonized into RGB, multispectral, hyperspectral, or multisensor/fused systems.
- Data fusion was coded as Yes when the study explicitly combined multiple sensor streams, feature domains, or data sources; otherwise it was coded as No.
- Model family labels reported in the articles were collapsed into broader standardized classes such as traditional machine learning, ensemble machine learning, deep learning, and hybrid or physics-informed approaches.
- Tier labels were used only to support high-level comparison of sensing complexity and modeling complexity across the literature.

S7. Rules for selecting the comparative R^2 value

- R^2 was selected as the common comparative metric because it was the most frequently reported and broadly comparable performance index across studies.
- For Q1 and Q2, a study was included in R^2 comparison only if a usable study-level R^2 could be identified from the article or extraction sheet.
- When multiple models were reported, the selected R^2 corresponded to the model judged as the study's best or most emphasized result for the coded target used in the comparative review.
- When a study reported outcomes outside the main coded target category, the selected value was aligned as closely as possible to the target used for the study-level coding decision.
- Studies without a usable R^2 were retained for descriptive counting when appropriate but excluded from R^2 -based comparison.

S8. Rules for question-specific analysis subsets

- Q1 compared Direct N, Close proxy, and Indirect proxy studies using the selected R^2 value where available.
- Q2 compared Internal-only and Independent-like validation groups using the selected R^2 value where available.
- Q3 was treated primarily as descriptive and exploratory because the evidence base for explicit best-stage comparison was small and uneven.
- Sensitivity-oriented core subsets were created to support interpretation, including a Direct N core dataset for Q1 and a True N diagnosis core dataset for Q2.

S9. Quality control and manual checks

- Quality assurance flags in the analysis-ready sheets were retained to preserve transparency about studies requiring caution in interpretation.
- The harmonized dataset should be treated as the final analytical dataset for the current manuscript unless a future re-check of the original article justifies revision.

S10. Interpretation notes

The resulting review should be interpreted as a structured comparative quantitative review rather than a formal meta-analysis. This reflects the heterogeneity of study targets, validation designs, and reporting structures, as well as the absence of fully standardized effect-size information across the included literature.