

Estimating potato leaf chlorophyll content with limited samples using optimized models trained on mixed datasets

Zhenning Du

Inner Mongolia Agricultural University

Fei Li

lifei@imau.edu.cn

Inner Mongolia Agricultural University

Yuncai Hu

Technical University of Munich

Haibo Yang

Inner Mongolia Agricultural University

Kang Yu

Technical University of Munich

Article

Keywords: PROSPECT model, Optimized spectral indices, Machine learning, Potato leaf chlorophyll content, Small sample size

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11273437/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Estimating potato leaf chlorophyll content with limited samples using optimized models trained on mixed datasets

Zhenning Du; Fei Li; Yuncai Hu; Haibo Yang; Kang Yu

Affiliations: [Inner Mongolia Agricultural University, College of Resources and Environment, Hohhot 010011, Inner Mongolia Autonomous Region, China]

Corresponding author: [Fei Li lifei@imau.edu.cn]

Abstract

Accurate monitoring of potato leaf chlorophyll content (LCC) can support nitrogen management, but limited field samples often constrain spectral estimation models. Thus, this study combined PROSPECT-5B simulations with measured data to evaluate optimized spectral indices and six regression methods for potato LCC estimation under small-sample conditions. Leaf hyperspectral reflectance and LCC were measured in central Inner Mongolia in 2023 and 2025. Wavelengths selected by optimizing eight indices on simulated data were applied to measured spectra. The 2023 field samples were combined with different numbers of simulated samples for calibration, and the 2025 data were used for independent validation. The optimal simulated-to-measured sample ratio varied among methods, ranging from 2:1 to 5:1. The tabular prior-data fitted network (TabPFN) achieved the highest validation R^2 (0.73), with a root mean square error (RMSE) of 0.55 mg/g and a relative error of 17.21%. The four highest-ranked optimized indices accounted for 65% of total variable importance. Using these four indices retained comparable validation accuracy ($R^2 = 0.70$; RMSE = 0.56 mg/g) while improving computational efficiency. TabPFN trained solely on measured data yielded an R^2 of 0.52 and an RMSE of 0.67 mg/g. Combining measured and simulated data in an appropriate ratio improved LCC estimation for potatoes with limited field samples. The four-index TabPFN model offers a promising approach to non-destructive monitoring of potato chlorophyll across cultivars and environments.

Keywords PROSPECT model; Optimized spectral indices; Machine learning; Potato leaf chlorophyll content; Small sample size

Introduction

Potato is the fourth-largest food crop worldwide and plays an important role in China's food security. Nevertheless, overall productivity remains relatively low, and yields vary considerably across regions, leaving substantial scope to increase yields per unit area (Singh et al. 2018). Chlorophyll is a key photosynthetic pigment and an indicator of crop photosynthetic capacity, nutrient stress, and nitrogen status. Chlorophyll content is significantly and positively correlated with leaf nitrogen content and is widely used to assess crop growth and health (Qiao et al. 2022).

Traditional spectrophotometric methods for determining potato leaf chlorophyll content (LCC) are accurate but destructive and time-consuming (Ta et al. 2019). Handheld SPAD meters provide rapid

measurements, but numerous readings are required to characterize chlorophyll status at the crop population level (Wakahara et al. 2025). These limitations restrict the use of either method for efficient, non-destructive monitoring over large areas. Hyperspectral remote sensing provides rapid, non-destructive measurements and has been widely used for monitoring potato chlorophyll (Li et al. 2025). However, empirical models based on hyperspectral data generally require large numbers of field samples for calibration and validation (Liu et al. 2023). With limited samples, these models are prone to overfitting, which can substantially reduce their ability to generalize (Zarco-Tejada et al. 2001). Studies in maize (Verma et al. 2022; Zhao et al. 2018) and wheat (Zhang et al. 2022b) have also shown that conventional fixed-band combinations have limited capacity to suppress spectral interference from leaf structure and water content, restricting model applicability across environments and growth stages. Even after sensitive wavelengths have been optimized, substantial amounts of representative field data are still needed. Leaf radiative transfer models, which can generate large numbers of simulated spectra, offer a way to address this limitation under small-sample conditions.

PROSPECT simulates leaf reflectance as a function of combinations of physiological and biochemical parameters, including chlorophyll content, equivalent water thickness, and dry matter content. It can therefore expand the sample space by generating large numbers of spectra with defined physiological and biochemical properties (Zhang et al. 2022a). The model has been evaluated in several crops. Yang developed N-PROSPECT to retrieve leaf nitrogen density in wheat (Yang et al. 2022), while the RSPECT model incorporates the anatomical structure of rice leaves (Xiang et al. 2025). PROSPECT and its coupled models have also shown good applicability in peanut and barley (Kong et al. 2025). In potato, PROSPECT has been used successfully to retrieve LCC (Botha et al. 2006), although previous studies have largely focused on particular cultivars or growth stages and have lacked systematic simulations. Simulations spanning a range of LCC levels are therefore needed to characterize the spectral response of potato leaves to chlorophyll variation and provide suitable training data for hyperspectral estimation with limited field samples. Combining PROSPECT with optimized spectral indices offers the sample-expansion and parameter-decoupling capabilities of a physical model alongside the computational simplicity and direct spectral response of statistical indices. This combination may improve for potatoes under small-sample conditions.

Combining machine learning with spectral features has become an active area of research in crop chlorophyll estimation. Machine learning can improve estimation accuracy and address challenges such as canopy structural interference and background noise. In maize, a hybrid feature-extraction method combining competitive adaptive reweighted sampling (CARS) and long short-term memory (LSTM) was developed to account for growth-related variation and improve the stability of chlorophyll detection (Zhao et al. 2023). For wheat and rice, hybrid retrieval frameworks that combine PROSAIL with deep learning (Shen et al. 2026) and self-attention neural networks that use multimodal data (Chen et al. 2024) have been developed to improve generalization in the presence of canopy structural heterogeneity. Conventional methods, including random forest (RF), support vector machine (SVM), and backpropagation neural network (BPNN), have also demonstrated good predictive performance and stability in potato (Yang et al. 2025b) and soybean (Gao et al. 2024), supporting their use in

precision fertilization. Nevertheless, model performance depends strongly on the quantity and quality of training data. Destructive sampling and laboratory analyses are time-consuming and labor-intensive, making large, high-quality field datasets difficult to obtain. When sample size is limited, the high dimensionality of hyperspectral data increases the risk of overfitting and can substantially reduce generalization performance (Ahmad et al. 2025). Hybrid approaches that combine radiative transfer models and machine learning can expand the training sample space and increase the number and variability of measured samples (Yang et al. 2025a). However, most studies have focused on a single modeling framework, with few systematic comparisons of algorithms under small-sample conditions. Algorithms differ in their sample requirements, sensitivity to data distributions, and ability to generalize across cultivars and environments. Their performance on mixed datasets should therefore be compared systematically to identify methods suited to estimating LCC from limited field samples.

This study aimed to combine PROSPECT-5B simulations with field measurements collected in different years and locations in Inner Mongolia to construct mixed datasets of different sizes for potato LCC estimation. We evaluated eight optimized spectral indices: Opt-NDVI, Opt-RVI, Opt-DVI, Opt-Clred-edge, Opt-mSR705, Opt-mND705, Opt-NDDA, and Opt-NPDI. Six modeling methods were compared: partial least squares regression (PLSR), random forest (RF), support vector regression (SVR), a one-dimensional convolutional neural network (1D-CNN), Gaussian process regression (GPR), and a tabular prior-data fitted network (TabPFN). Their performance on independent validation data was compared across 10 levels of simulated data input (H1-H10) to determine the optimal mixing ratio and identify the best-performing model under small-sample conditions. The overall aim was to develop an accurate and broadly applicable approach to estimating potato LCC to support precision crop management.

Materials and Methods

Experimental Design

Field experiments were conducted in the main potato-producing region of central Inner Mongolia in 2023 and 2025. In 2023, cultivar 'Shepody' was grown at four nitrogen application rates (0, 165, 195, and 350 kg ha⁻¹), with four replications per treatment. In 2025, cultivar 'Xisen 6' was grown at five nitrogen application rates (0, 60, 120, 240, and 360 kg ha⁻¹), with three replications per treatment. Data were collected before and after flowering in both years. The experiment was conducted using a randomized complete block design. Drip irrigation was used, and other field management practices followed local recommendations.

Leaf spectral measurements

Leaf spectral reflectance was measured using a PSR+3500 portable spectroradiometer (SPECTRAL EVOLUTION, USA), which covers 350-2500 nm with a 1-nm output interval. Three representative potato plants were selected in each plot. Measurements were taken from the middle portion of each leaf using a leaf clip with an active light source, avoiding the veins. Leaves w

ere held horizontally and perpendicular to the leaf clip. Spectral data were recorded and stored on the accompanying personal digital assistant (PDA).

Determination of relative chlorophyll content

Fresh potato leaves were collected, the leaf surface was cleaned, the midrib was removed, and the leaves were cut into small pieces. A 0.20 g sample was placed into a brown volumetric flask containing 95% ethanol for extraction for 48 h, with thorough shaking every 12 h until the leaves turned white. Chlorophyll content was determined by ultraviolet spectrophotometry. The extract was poured into a 1 cm path-length cuvette, and absorbance was measured at 665 nm and 649 nm using 95% ethanol as the blank. The absorbance values were then substituted into the formulas to calculate the concentrations of chlorophyll a, chlorophyll b, and total chlorophyll, and subsequently the chlorophyll content in the plant tissue. The calculation formulas are as follows:

$$c_a = 13.95 \times A_{665} - 6.88 \times A_{649} \quad (1)$$

$$c_b = 24.96 \times A_{649} - 7.32 \times A_{665} \quad (2)$$

$$C_T = c_a + c_b \quad (3)$$

$$LCC (mg/g) = \frac{C \times V}{FW \times 1000} \quad (4)$$

where Ca, Cb, and CT are the concentrations of chlorophyll a, chlorophyll b, and total chlorophyll in the extract, respectively (mg/L); A649 and A665 are the absorbance at 649 and 665 nm, respectively; V is the extract volume; and FW is the leaf fresh weight.

PROSPECT simulations

PROSPECT is a physically based model of leaf directional – hemispherical reflectance and transmittance (Jacquemoud and Baret 1990). It was introduced by Jacquemoud and Baret in 1990 and further developed in 1995. Its efficiency and accuracy have supported its widespread use for simulating crop leaf spectra. In this study, we used PROSPECT-5B to simulate potato leaf reflectance. This version represents leaf structural and optical properties and simulates reflectance and transmittance across the visible and near-infrared regions. Its input parameters include leaf chlorophyll content (Cab, $\mu g cm^{-2}$), equivalent water thickness (Cw, $g cm^{-2}$), dry matter content (Cm, $g cm^{-2}$), carotenoid content (Car, $\mu g cm^{-2}$), leaf structure index (N), and brown pigment content (Cbrown). Parameter ranges are given in Table 1. We generated 1000 simulated leaf spectra over the 400-2500 nm range at 1 nm intervals. The simulated spectra were used to optimize spectral indices and identify sensitive wavelengths. These wavelengths were then applied to the measured spectra to calculate the corresponding indices.

Table 1 Parameter ranges used in PROSPECT-5B simulations

Variable	Abbreviation	Range
Leaf structure index	N (Dimensionless)	1-2
Leaf chlorophyll content	Cab ($\mu\text{g}/\text{cm}^2$)	20-100
Leaf carotenoid content	Car ($\mu\text{g}/\text{cm}^2$)	0-20
Leaf brown pigment content	Cbrown (Dimensionless)	0-1
Leaf equivalent water thickness	Cw (g/cm^2)	0-0.06
Leaf dry matter content	Cm (g/cm^2)	0-0.05

Machine learning algorithms

The modeling workflow was implemented in Python using Jupyter through Anaconda Navigator. The methods evaluated were PLSR (Berger et al. 2020), RF (Loozen et al. 2020), SVR (Miphokasap and Wannasiri 2018; Nigon et al. 2020; Vapnik et al. 1996), 1D-CNN (Lecun et al. 1998; Shi et al. 2022; Singh et al. 2018), GPR (Campos-Taberner et al. 2016), and TabPFN (Hollmann et al. 2025). Model inputs comprised eight optimized spectral indices: Opt-NDVI, Opt-RVI, Opt-DVI, Opt-CIred-edge, Opt-mSR705, Opt-mND705, Opt-NDDA, and Opt-NPDI. For the measured dataset, these indices were calculated using sensitive wavelengths selected from the PROSPECT-simulated dataset.

Selection and optimization of spectral indices

Eight spectral indices associated with potato LCC were selected for their differing wavelength combinations and mathematical formulations (Table 2). To identify the wavelength combinations most strongly correlated with LCC, we systematically evaluated all possible two-band and three-band combinations over 400-1000 nm in the PROSPECT-5B-simulated dataset. Optimized indices were constructed using the eight formulations listed in Table 2.

Table 2 Spectral indices associated with leaf chlorophyll content and their formulations

Spectral indices	Formula	Formula	Reference
Two-band spectral indices			
Normalized difference vegetation index (NDVI)	$(R_j - R_i)/(R_j + R_i)$	$(R_{800} - R_{670})/(R_{800} + R_{670})$	(Jordan 1969)
Ratio vegetation index (RVI)	R_j/R_i	R_{800}/R_{670}	(Verrelst et al. 2015)
Different vegetation index (DVI)	$R_j - R_i$	$R_{800} - R_{680}$	(Rouse et al. 1974)
Chlorophyll Index Red Edge (CIred edge)	$(R_j - R_k)/R_i$	$(R_{750} - R_{700})/R_{700}$	(Gitelson et al. 2003)
Three-band spectral indices			
Modified Red-edge Ratio index (mSR705)	$(R_j - R_k)/(R_i - R_k)$	$(R_{750} - R_{445})/(R_{705} - R_{445})$	(Sims and Gamon 2002)
Modified Red-edge Normalized Difference index (mND705)	$(R_i - R_j)/(R_i + R_j - 2 \times R_k)$	$(R_{750} - R_{705})/(R_{750} + R_{705} - 2 \times R_{445})$	(Sims and Gamon)

			2002)
Double peak N index (NDDA)	$(R_i + R_k - 2 \times R_j) / (R_i - R_k)$	$(R_{755} + R_{680} - 2 \times R_{705}) / (R_{755} - R_{680})$	(Feng et al. 2014)
Combined spectral indices			
Nitrogen planar domain index (NPDI)	$(CI_{red-edge} - CI_{red-edge}^{MIN}) / (CI_{red-edge}^{MAX} - CI_{red-edge}^{MIN})$	$(CI_{red-edge} - CI_{red-edge}^{MIN}) / (CI_{red-edge}^{MAX} - CI_{red-edge}^{MIN})$	(Li et al. 2012)

Data analysis

Microsoft Excel 2021 and MATLAB 2021b (The MathWorks, Inc., Natick, MA, USA) were used for spectral index calculations and machine learning modeling. Figures were prepared in Origin 2021. The calibration dataset combined field measurements collected in 2023 with PROSPECT-5B-simulated data. Field measurements collected in 2025 were used to validate the model.

Results

Identification of chlorophyll-sensitive wavelengths using simulated spectra and optimized indices

Leaf spectra over 400 - 1000 nm were simulated using PROSPECT-5B and compared with spectra measured in 2023 and 2025 (Fig. 1). The simulated and measured mean spectra exhibited similar patterns, with reflectance ranging from 0.05 to 0.70. Agreement was particularly evident around the green peak (550 nm), red edge (680 - 760 nm), and near-infrared plateau (760 - 1000 nm). These results indicate that the simulated spectra adequately represented the variation in field-measured potato leaf spectra and reproduced their principal spectral features.

Conventional spectral indices used to estimate potato chlorophyll content can be affected by soil background during early growth and by complex canopy structure at later stages. To address these influences, eight indices-Opt-NDVI, Opt-RVI, Opt-DVI, Opt- $CI_{red-edge}$, Opt-mSR705, Opt-mND705, Opt-NDDA, and Opt-NPD I- were optimized using the simulated dataset to identify wavelengths sensitive to potato LCC (Fig. 2; Table 3)

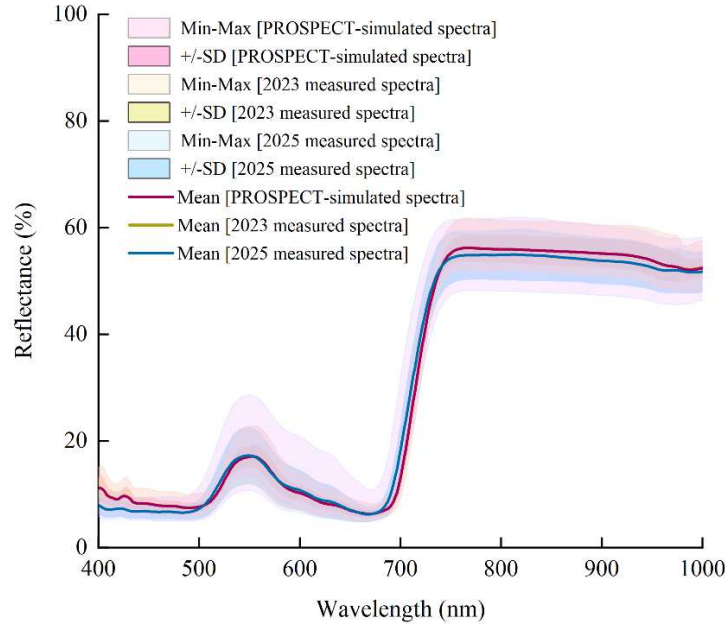


Fig.1 Comparison of PROSPECT-simulated leaf reflectance with spectra measured in 2023 and 2025. Lines show mean reflectance; shaded areas indicate the minimum – maximum range and ± 1 standard deviation (SD) for each dataset

The wavelengths selected for the eight optimized indices spanned the red, green, and blue regions, which are sensitive to potato LCC, providing a mechanistic basis for their strong correlations with chlorophyll content. The coefficient of determination (R^2) was calculated to assess the relationship between each index and simulated LCC (Table 3). Opt-DVI, Opt-mND705, and Opt-NPDI performed best, each yielding an R^2 of 0.65. All eight optimized indices were significantly correlated with simulated LCC, with R^2 values ranging from 0.60 to 0.65. By comparison, the conventional indices yielded R^2 values of only 0.02-0.15, indicating greater sensitivity to simulated LCC after wavelength optimization.

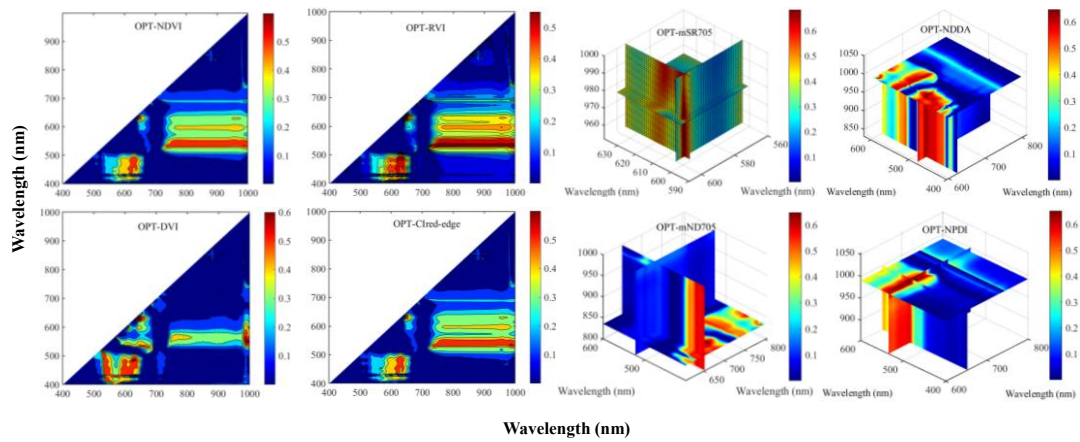


Fig.2 Sensitivity of spectral indices to wavelength combinations in the PROSPECT-simulated dataset

Table 3 Optimized wavelengths and coefficients of determination for spectral indices related to simulated potato leaf chlorophyll content

Spectral indices	Optimizing band	Coefficient
Two-band spectral indices		
Optimization-Normalized difference vegetation index (Opt-NDVI)	534, 987	0.60
Optimization-Ratio vegetation index (Opt-RVI)	534, 988	0.61
Optimization-Different vegetation index (Opt-DVI)	635, 648	0.65
Optimization-Chlorophyll Index Red Edge (Opt-CIred edge)	534, 988	0.61
Three-band spectral indices		
Optimization-Modified Red-edge Ratio index (Opt-mSR705)	595, 604, 975	0.63
Optimization-Modified Red-edge Normalized Difference index (Opt-mND705)	483, 625, 995	0.65
Optimization-Double peak N index (Opt-NDDA)	483, 625, 995	0.64
Combined spectral indices		
Optimization-Nitrogen planar domain index (Opt-NPDI)	535, 653, 990	0.65

Estimation of potato leaf chlorophyll content using mixed calibration datasets

Adding field samples to simulated training data, often termed spiking, can improve the predictive performance of models trained on simulated datasets (Guerrero et al. 2010). This approach has also improved LCC estimation (Li et al. 2023). Here, field measurements collected in 2023 were used as spike samples and combined with different numbers of simulated samples to construct mixed datasets. These datasets were used to train six models. It should be noted that the LCC data measured in 2025 were not used for simulated data generation, mixed dataset construction, determination of mixing ratios, selection of the number of features, algorithm selection, or model hyperparameter optimization, nor were they included in any model training process. They served only as external independent validation data to evaluate the generalization ability of the trained models on real observations. Calibration and validation results are presented in Tables 4 and 5, respectively. The proportion of simulated samples influenced prediction accuracy, with both R^2 and RMSE varying among mixing ratios. In validation, PLSR achieved its highest accuracy with H2; RF, 1D-CNN, and GPR performed best with H3; and SVR and TabPFN performed best with H5 (Table 5). Calibration and validation predictions for each model at its optimal mixing ratio are shown in Fig. 3.

The best-performing configurations of all six methods improved LCC estimation to varying degrees (Fig. 3). TabPFN achieved the highest validation accuracy, with predicted and measured values distributed close to the 1:1 line. Its validation R^2 was 0.73, with an RMSE of 0.55 mg/g and an RE of 17.21%. RF and GPR showed intermediate performance, retaining predictive ability but yielding lower accuracy than TabPFN. PLSR, SVR, and 1D-CNN performed less well. The 1D-CNN model had a validation R^2 of only 0.39, along with substantially higher RMSE and RE values. Considering all three metrics, TabPFN provided the best validation performance and showed good transferability across sites. It was therefore selected for further optimization of potato LCC estimation.

Table 4 Calibration performance of six modeling methods trained on mixed datasets comprising 2023 field measurements and simulated data

Simulated datasets sizes	PLSR		RF		SVR		1D-CNN		GPR		TabPFN	
	R ²	RMSE (mg/g)	R ²	RMSE (mg/g)	R ²	RMSE (mg/g)	R ²	RMSE (mg/g)	R ²	RMSE (mg/g)	R ²	RMSE (mg/g)
H1	0.66	0.29	0.93	0.21	0.21	0.45	0.36	0.40	0.97	0.16	0.69	0.27
H2	0.70	0.26	0.92	0.19	0.36	0.46	0.37	0.39	0.98	0.16	0.73	0.27
H3	0.72	0.26	0.93	0.21	0.34	0.42	0.46	0.39	0.97	0.17	0.83	0.30
H4	0.75	0.26	0.94	0.19	0.47	0.47	0.36	0.40	0.97	0.16	0.86	0.26
H5	0.71	0.26	0.94	0.19	0.67	0.42	0.31	0.45	0.97	0.17	0.89	0.21
H6	0.72	0.27	0.94	0.19	0.62	0.42	0.39	0.37	0.97	0.17	0.86	0.23
H7	0.72	0.26	0.94	0.19	0.67	0.46	0.43	0.39	0.97	0.17	0.89	0.27
H8	0.72	0.27	0.93	0.19	0.76	0.42	0.45	0.39	0.97	0.17	0.82	0.19
H9	0.71	0.27	0.94	0.19	0.73	0.42	0.46	0.37	0.97	0.17	0.90	0.29
H10	0.70	0.27	0.93	0.19	0.71	0.42	0.45	0.39	0.97	0.17	0.86	0.29

Table 5 Validation performance of six modeling methods using field measurements collected in 2025

Simulated datasets sizes	PLSR		RF		SVR		1D-CNN		GPR		TabPFN	
	R ²	RMSE (mg/g)	R ²	RMSE (mg/g)	R ²	RMSE (mg/g)	R ²	RMSE (mg/g)	R ²	RMSE (mg/g)	R ²	RMSE (mg/g)
H1	0.42	0.62	0.57	0.63	0.29	0.65	0.31	0.76	0.43	0.67	0.53	0.54
H2	0.59	0.63	0.56	0.63	0.23	0.65	0.36	0.76	0.57	0.67	0.53	0.54
H3	0.37	0.71	0.62	0.64	0.37	0.65	0.39	0.76	0.67	0.69	0.37	0.55
H4	0.31	0.62	0.50	0.62	0.32	0.66	0.35	0.75	0.49	0.69	0.52	0.57
H5	0.42	0.62	0.52	0.71	0.44	0.70	0.37	0.76	0.61	0.67	0.73	0.55
H6	0.36	0.63	0.36	0.72	0.41	0.67	0.32	0.75	0.51	0.69	0.61	0.59
H7	0.33	0.60	0.43	0.71	0.36	0.72	0.37	0.76	0.52	0.69	0.52	0.53
H8	0.39	0.61	0.41	0.72	0.37	0.75	0.31	0.76	0.52	0.69	0.57	0.53
H9	0.39	0.61	0.33	0.75	0.41	0.72	0.37	0.76	0.43	0.70	0.41	0.53
H10	0.31	0.62	0.37	0.74	0.42	0.72	0.33	0.76	0.51	0.69	0.55	0.53

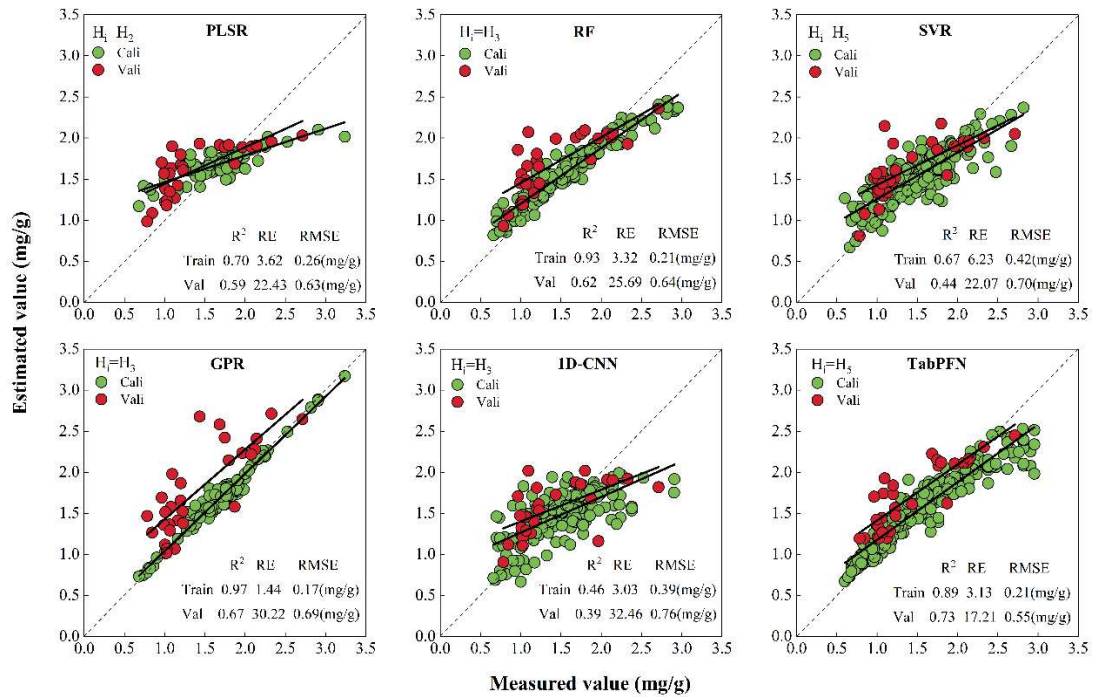


Fig.3 Calibration and validation results for six modeling methods using optimized spectral indices as inputs. Green and red points represent calibration and validation samples, respectively. Dashed lines indicate the 1:1 relationship. RMSE, root mean square error; RE, relative error

Optimization of the TabPFN model for potato leaf chlorophyll estimation

TabPFN achieved the highest accuracy in predicting potato LCC when all eight optimized indices were used as inputs. However, an excessive number of input variables may reduce accuracy and increase computational complexity. We therefore evaluated input importance using Gini importance scores. Gini importance is a built-in random forest metric that quantifies each variable's contribution to the model, with larger values indicating greater importance. Scores ranged from 0.06 to 0.19 (Fig. 4a). Optimized mSR705, NPDI, mND705, and NDDA were the most influential variables, together accounting for 65% of total importance. The optimized two-band indices contributed less than the three-band indices. To determine the appropriate number of inputs for TabPFN, indices were added sequentially in descending order of importance, and the resulting models were evaluated using RMSE and runtime. RMSE declined markedly as inputs were added and stabilized at approximately 0.56 mg/g after the fourth index was included (Fig. 4b). Runtime, by contrast, increased exponentially with the number of inputs. Using all eight indices reduced computational efficiency without significantly improving accuracy relative to using four. The four highest-ranked indices were therefore retained as the final inputs. This model achieved a validation R^2 of 0.70, an RMSE of 0.56 mg/g, and an RE of 17.02% (Fig. 5b). TabPFN trained solely on measured data showed substantially poorer predictive performance (Fig. 5a). These results further indicate that simulated data can expand the distribution of training samples and compensate for the limited variation represented in field measurements, improving the generalizability and robustness of potato LCC estimation.

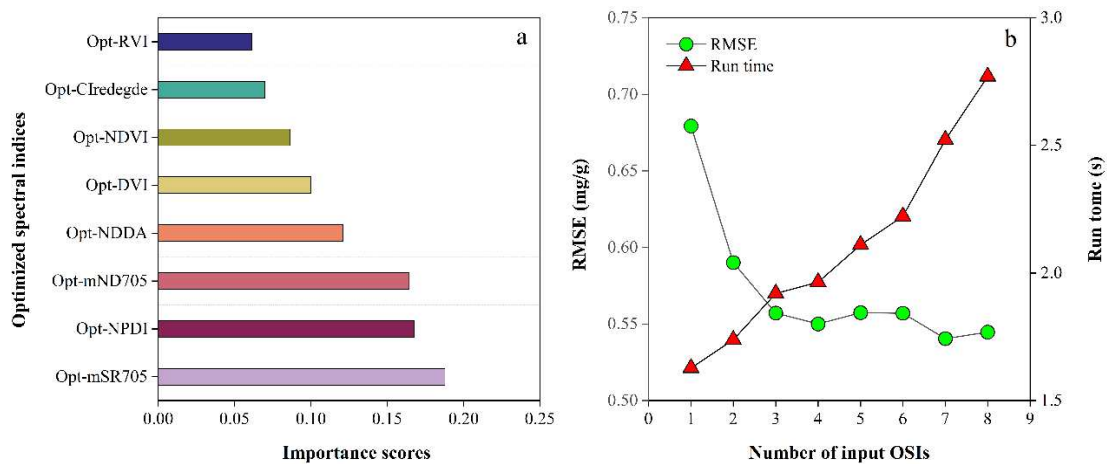


Fig.4 Optimization of the number of spectral indices used as inputs to TabPFN: (a) variable importance ranking and (b) root mean square error (RMSE) and runtime as a function of the number of optimized spectral indices (OSIs)

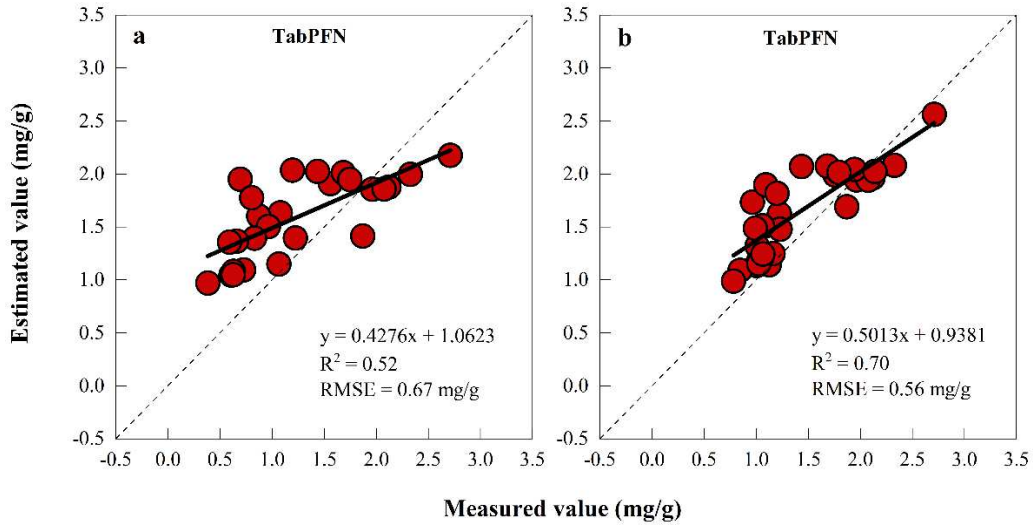


Fig.5 Validation results for TabPFN trained on measured data alone (a) and a mixed dataset (b)

Discussion

Accurate and efficient estimation of leaf chlorophyll content is important for monitoring crop health and supporting field management. Spectral sensing provides a means of estimating LCC as an indicator of crop growth. In this study, PROSPECT simulations and optimized spectral indices were used to develop models to estimate potato LCC from limited field samples. All six methods achieved their highest validation accuracy when the number of simulated samples was 2 to 5 times the number of measured samples, whereas further increases in the number of simulated samples reduced accuracy. TabPFN performed best (validation $R^2 = 0.73$). Reducing its inputs to four key indices retained accuracy close to that obtained with all eight indices ($R^2 = 0.70$). These findings highlight the importance of selecting an appropriate mixing ratio and learning algorithm under small-sample conditions. The following discussion considers the balance between measured and simulated data, differences among modeling methods, and the selection of input variables.

Combining measured and simulated data is an established approach to improving the estimation of vegetation physiological parameters. Field datasets often contain too few samples to represent the full range of relevant variation, whereas radiative transfer simulations can supplement spectral gradients and extend sample coverage (De Angeli et al. 2022). Our results show that mixed datasets can improve machine learning predictions of potato LCC even when calibration and validation datasets are completely independent. Similar benefits have been reported previously. Houborg and McCabe combined 87 measured samples with 1626 radiative transfer model (RTM) simulations to improve RF estimation of leaf area index (Houborg and McCabe 2018). The Spiking-Hybrid approach also combined measured data with simulated datasets of varying sizes to improve estimates of chlorophyll and water content compared with conventional machine learning (Wan et al. 2022). Other studies have shown that simulated data can support transfer learning for field LCC prediction (Yue et al. 2024). Such approaches incorporate the diversity of measured samples to improve calibration under target

conditions (Berger et al. 2018) and increase the diversity and coverage of the training data (Verrelst et al. 2021; Yang et al. 2021).

In the present study, the optimal mixed datasets contained 54 measured samples and 2 to 5 times as many simulated samples; adding more simulations was unnecessary to achieve satisfactory accuracy. Performance degraded beyond a 5:1 simulated-to-measured ratio, attributable to the increasing influence of domain shift. PROSPECT-5B simulates idealized leaf optics and does not account for micro-symptoms, surface wax, or canopy geometry. As the proportion of simulated samples increases, the synthetic distribution bias inherent in these idealized assumptions progressively overrides the real leaf structural signals contained in the measured spectra, so that additional simulations introduce more noise than useful information and reduce model performance. However, the optimal ratio varies among studies. Using RF and PLSR to estimate potato LCC, Yang reported an optimal number of simulated samples 11 - 14 times the number of measured samples, considerably higher than the ratio observed here (Yang et al. 2025a). Three factors may explain this difference. First, both our PROSPECT-5B simulations and measurements were at the leaf scale, ensuring consistency in physical scale. By comparison, Yang et al. used the PROSAIL canopy model, which includes leaf area index (LAI) and soil background. The resulting discrepancy between simulation and measurement scales requires more simulations to represent the additional variation associated with canopy structure. Second, we used low-dimensional optimized indices, whereas Yang et al. used high-dimensional, full-spectrum features, which inherently require more samples (Tong and Zhang 2024). Third, we used simulations to select sensitive wavelengths before constructing the mixed datasets. Simulations therefore, informed feature selection as well as increasing the sample size. In contrast, Yang et al. directly combined simulated and measured spectra, so simulations primarily supplemented sample numbers and a larger dataset was needed to improve accuracy.

These comparisons support the value of combining limited field measurements with simulated data to improve prediction and generalization. They also indicate that the optimal mixing ratio is not universal: it depends on the consistency between simulation and measurement scales, the information represented by the input features, and the representativeness of measured samples. The ratio should therefore be determined for each study's conditions.

Machine learning methods differ in their responses to sample size and data distribution when applied to vegetation spectral estimation. Deep learning is generally considered effective at extracting high-dimensional features (Song et al. 2025), yet SVR and 1D-CNN performed relatively poorly in this study. The 1D-CNN model yielded R^2 values below 0.5 in both the calibration and validation sets, indicating limited suitability for this small-sample estimation task. 1D-CNN suffers from severe over-parameterization when fed 8 scalar features that lack a 1D spatial/sequential topology. In contrast, TabPFN uses pre-trained Transformer priors designed for tabular contexts, enabling Bayesian-like posterior inference without gradient updates under small sample sizes. Yang noted that strong deep learning performance generally requires large datasets and that small samples increase the risk of overfitting, which is consistent with our interpretation of the results (Yang et al. 2022).

Although RF and PLSR have often outperformed other conventional methods in estimating plant nitrogen concentration (López-Calderón et al. 2020; Prado Osco et al. 2019) and the nitrogen nutrition index (Zha et al. 2020), their accuracy here remained limited. One possible explanation is that these methods did not fully capture the complex relationships between simulated and measured data. With limited observations, noise and meaningful response signals may be difficult to distinguish, preventing the models from learning relationships that adequately represent measured variation.

TabPFN, in contrast, achieved a validation R^2 of 0.73, an RMSE of 0.55 mg/g, and an RE of 17.21%. For a new small-sample classification task, TabPFN uses the training samples as context and predicts new observations in a single forward pass, without gradient-based training or hyperparameter tuning. This mechanism avoids overfitting when fitting models to small datasets (Hollmann et al. 2022). Its small-sample capabilities are grounded in prior-data fitted network theory (Hollmann et al. 2025). From a Bayesian perspective, a Transformer pretrained on synthetic data can approximate posterior inference under suitable priors, including Gaussian processes. This enables accurate, well-calibrated predictive distributions even when relatively few real observations are available.

Selecting appropriate inputs can improve both predictive accuracy and computational efficiency (Dormann et al. 2012). Gini importance ranking identified optimized mSR705, NPDI, mND705, and NDDA as the four most important indices, accounting for 65% of total importance. Using these four indices produced accuracy comparable to that obtained with all eight. The optical basis for the selected wavelengths helps explain their effectiveness. The 705 nm band lies on the steep slope of the red edge, where the chlorophyll absorption coefficient (K_{ab}) is moderate. The reflectance response to chlorophyll content remains approximately linear, so that this band avoids the saturation that occurs near 680 nm, where chlorophyll absorption is close to its maximum and reflectance changes little once chlorophyll content is high (Gitelson et al. 2003; Jordan 1969). This is why mSR705 and mND705, which incorporate the 705 nm red-edge band, can reduce the influence of leaf structure and provide a linear response to chlorophyll content. The 534 nm and 653 nm bands correspond to the green and red chlorophyll absorption regions, respectively, and are therefore sensitive to pigment variation; NPDI and NDDA exploit these bands, and their sensitivity to pigment variation has likewise been reported (Rouse et al. 1974). The 990 nm band, in contrast, is dominated by leaf internal structure scattering (N) and water absorption, so it carries information about leaf structure and water status rather than chlorophyll alone. Including this band allows the indices to capture structural and water-related variation that the visible bands cannot represent. The inclusion of auxiliary information, such as blue reflectance, can also make three-band indices less susceptible to saturation than two-band indices (Jordan 1969).

In this study, TabPFN runtime increased exponentially with the number of inputs, while accuracy stabilized after the fourth index was added, suggesting redundancy among the remaining variables. Removing redundant inputs using Gini importance can improve efficiency and reduce the risk of overfitting (Guyon and Elisseeff 2003). This strategy has been widely used for retrieving crop biochemical parameters (Yang et al. 2025b).

The four-index TabPFN model achieved a validation R^2 of 0.70, an RMSE of 0.56 mg/g, and an RE of 17.02%. This accuracy is comparable to that reported for potato chlorophyll estimation using RF

by Botha et al. R^2 approximately 0.70-0.72 (Botha et al. 2006), and it also demonstrates rapid adaptation to limited sample sizes. Previous comparisons of red-edge position algorithms for maize LCC estimation have emphasized that cross-validation alone cannot establish broad model applicability and stability; independent validation is also needed (Tang et al. 2024). Our validation data came from different sites, cultivars, and nitrogen treatments. These differences resulted in a slightly higher RE, although overall accuracy remained acceptable. Feature reduction based on Gini importance, therefore, improved computational efficiency without compromising accuracy. The selected indices also have estimation.

Limitations and Perspectives

However, several limitations of this study should be noted. First, PROSPECT-5B simulations and measurements were conducted at the leaf scale but canopy structure, soil background, and illumination geometry were not represented. If PROSPECT-5B were coupled with canopy models such as PROSAIL or RSPECT (Xiang et al. 2025), the transferability to UAV- and satellite-based observations would be improved. Such coupling would allow the four-index TabPFN framework to be extended from leaf-level estimation to canopy-scale mapping of potato chlorophyll, supporting field-scale N diagnosis and variable-rate fertilization. Future work should also verify the model based on UAV hyperspectral imagery across contrasting canopy cover stages to quantify the influence of canopy structure on prediction accuracy on the field-scales. Second, the PROSPECT-5B parameter ranges were defined a priori rather than constrained by measured leaf traits, which may lead to bias into the simulated feature space. Constraining simulation parameters with measured leaf traits, or adopting a Bayesian calibration of the radiative transfer model, would improve the realism of the simulated dataset (Berger et al. 2020). Future studies should further establish multi-cultivar, multi-site, and multi-year datasets to verify whether the identified optimal simulated-to-measured ratio remains stable, and to develop practical guidelines for determining the mixing ratio under varying conditions. Incorporating active learning strategies could also improve sampling efficiency by selecting the most informative field samples for spiking, thereby reducing reliance on simulated data and mitigating synthetic-distribution bias (Hollmann et al. 2025). Third, TabPFN may have limited interpretability, resulting in extending the framework to simultaneous retrieval of chlorophyll, water, and dry matter would provide a more complete basis for precision nitrogen management (Hollmann et al. 2025). Combining TabPFN with explainable machine learning techniques or physically based sensitivity analysis would help clarify how the selected red-edge and green bands drive the predictions. In the longer term, integrating the optimized spectral indices and TabPFN into a portable or UAV-mounted monitoring system could enable real-time, non-destructive diagnosis of potato growth status in commercial production.

Conclusion

Leaf chlorophyll content is an important indicator of crop nutritional status. This study integrated the PROSPECT-5B radiative transfer model with machine learning to develop a non-destructive method for estimating potato LCC. Sensitive wavelengths identified by optimizing indices on simulated leaf reflectance were consistent with the measured spectra and were successfully applied to estimation from field measurements. Optimized spectral indices reduced the number of model inputs, improved computational efficiency, and increased prediction accuracy. Their combination with machine learning enabled accurate retrieval of potato LCC. Among the methods evaluated, TabPFN performed particularly well, achieving a calibration R^2 of 0.89 and an RMSE of 0.21 mg/g, and a validation R^2 of 0.73 and an RMSE of 0.55 mg/g. These results support the potential of this approach for rapid and accurate estimation of potato leaf chlorophyll content.

Funding: This study was financially supported by the Development and Application of Precision Water–Fertilizer Management and Green Pest and Disease Control Technologies for Potato (2025YFHH0184).

Data Availability: The data used in this study are freely available. Anyone who wants to use the data can contact the corresponding author Fei Li.

Declarations

Conflict of Interest: The authors declare no competing interests.

References

- Ahmad M, Distefano S, Khan AM, Mazzara M, Li C, Li H, Aryal J, Ding Y, Vivone G, & Hong D. (2025). A comprehensive survey for Hyperspectral Image Classification: The evolution from conventional to transformers and Mamba models. *Neurocomputing*, 644. <https://doi.org/10.1016/j.neucom.2025.130428>
- Berger K, Atzberger C, Danner M, D’Urso G, Mauser W, Vuolo F, & Hank T. (2018). Evaluation of the PROSAIL Model Capabilities for Future Hyperspectral Model Environments : A Review Study. *Remote Sensing*, 10(1). <https://doi.org/10.3390/rs10010085>
- Berger K, Verrelst J, Feret JB, Wang Z, Woche M, Strathmann M, Danner M, Mauser W, & Hank T. (2020). Crop nitrogen monitoring: Recent progress and principal developments in the context of imaging spectroscopy missions. *Remote Sens Environ*, 242, 111758. <https://doi.org/10.1016/j.rse.2020.111758>
- Botha EJ, Zebarth BJ, & Leblon B. (2006). Non-destructive estimation of potato leaf chlorophyll and protein contents from hyperspectral measurements using the PROSPECT radiative transfer model. *Canadian Journal of Plant Science*, 86(1), 279-291. <https://doi.org/10.4141/p05-017>
- Campos-Taberner M, García-Haro FJ, Camps-Valls G, Grau-Muedra G, Nutini F, Crema A, & Boschetti M. (2016). Multitemporal and multiresolution leaf area index retrieval for operational local rice crop monitoring. *Remote Sensing of Environment*, 187, 102-118. <https://doi.org/10.1016/j.rse.2016.07.018>

ps://doi.org/10.1016/j.rse.2016.10.009

- Chen Y, Wang X, Zhang X, Xu X, Huang X, Wang D, & Amin A. (2024). Spectral-based estimation of chlorophyll content and determination of background interference mechanisms in low-coverage rice. *Computers and Electronics in Agriculture*, 226. <https://doi.org/10.1016/j.compag.2024.109442>
- De Angeli K, Gao S, Danciu I, Durbin EB, Wu XC, Stroup A, Doherty J, Schwartz S, Wiggins C, Damesyn M, Coyle L, Penberthy L, Tourassi GD, & Yoon HJ. (2022). Class imbalance in out-of-distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types. *J Biomed Inform*, 125, 103957. <https://doi.org/10.1016/j.jbi.2021.103957>
- Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Laurcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, & Lautenbach S. (2012). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27-46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Feng W, Guo B-B, Wang Z-J, He L, Song X, Wang Y-H, & Guo T-C. (2014). Measuring leaf nitrogen concentration in winter wheat using double-peak spectral reflection remote sensing data. *Field Crops Research*, 159, 43-52. <https://doi.org/10.1016/j.fcr.2014.01.010>
- Gao C, Li H, Wang J, Zhang X, Huang K, Song X, Yang W, Feng M, Xiao L, Zhao Y, Shafiq F, Wang C, Qiao X, & Li F. (2024). Combined use of spectral resampling and machine learning algorithms to estimate soybean leaf chlorophyll. *Computers and Electronics in Agriculture*, 218. <https://doi.org/10.1016/j.compag.2024.108675>
- Gitelson AA, Gritz Y, & Merzlyak MN. (2003). Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J Plant Physiol*, 160(3), 271-282. <https://doi.org/10.1078/0176-1617-00887>
- Guerrero C, Zornoza R, Gómez I, & Mataix-Beneyto J. (2010). Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. *Geoderma*, 158(1-2), 66-77. <https://doi.org/10.1016/j.geoderma.2009.12.021>
- Guyon I, & Elisseeff A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 3, 1157-1182. <https://doi.org/10.1162/153244303322753616>
- Hollmann N, Müller S, & Eggensperger K. (2022). TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. *The Eleventh International Conference on Learning Representations (ICLR)*.
- Hollmann N, Muller S, Purucker L, Krishnakumar A, Korfer M, Hoo SB, Schirrmeister RT, & Hutter F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045), 319-326. <https://doi.org/10.1038/s41586-024-08328-6>
- Houborg R, & McCabe MF. (2018). A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135, 173-188. <https://doi.org/10.1016/j.isprsjprs.2017.10.004>
- Jacquemoud S, & Baret F. (1990). PROSPECT: A model of leaf optical properties spectra. *Remote Sensing of Environment*, 34(2), 75-91. [https://doi.org/10.1016/0034-4257\(90\)90100-z](https://doi.org/10.1016/0034-4257(90)90100-z)

- Jordan CF. (1969). Derivation of Leaf-Area Index from Quality of Light on the Forest Floor. *Ecology*, 50(4), 663-666. <https://doi.org/10.2307/1936256>
- Kong X, Zhao J, Liu Y, Bai B, Yang J, Fan Y, Li G, Li Z, & Wan S. (2025). The PROSPECT model in high-throughput phenotyping for peanut leaf parameter estimation: Comparative performance of hyperspectral inversion models. *Current Plant Biology*, 43. <https://doi.org/10.1016/j.cpb.2025.100498>
- Lecun Y, Bottou L, Bengio Y, & Haffner P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324. <https://doi.org/10.1109/5.726791>
- Li F, Mistele B, Hu Y, Yue X, Yue S, Miao Y, Chen X, Cui Z, Meng Q, & Schmidhalter U. (2012). Remotely estimating aerial N status of phenologically differing winter wheat cultivars grown in contrasting climatic and geographic zones in China and Germany. *Field Crops Research*, 138, 21-32. <https://doi.org/10.1016/j.fcr.2012.09.002>
- Li J, Wijewardane NK, Ge Y, & Shi Y. (2023). Improved chlorophyll and water content estimations at leaf level with a hybrid radiative transfer and machine learning model. *Computers and Electronics in Agriculture*, 206. <https://doi.org/10.1016/j.compag.2023.107669>
- Li Q, Li F, & Yang H. (2025). Optimizing wavelet functions to improve the accuracy of hyperspectral estimation of chlorophyll in potato leaves. *Journal of Plant Nutrition and Fertilizers* 31(11), 2252-2265.
- Liu N, Li F, Yang H, Yin H, Gao F, & Jia Y. (2023). Machine learning models fed with optimized spectral indices to improve inversion accuracy of potato chlorophyll content. *Journal of Plant Nutrition and Fertilizers*, 29(8), 1531–1542.
- Loozen Y, Rebel KT, de Jong SM, Lu M, Ollinger SV, Wassen MJ, & Karssen D. (2020). Mapping canopy nitrogen in European forests using remote sensing and environmental variables with the random forests method. *Remote Sensing of Environment*, 247. <https://doi.org/10.1016/j.rse.2020.111933>
- López-Calderón MJ, Estrada-Ávalos J, Rodríguez-Moreno VM, Mauricio-Ruvalcaba JE, Martínez-Sifuentes AR, Delgado-Ramírez G, & Miguel-Valle E. (2020). Estimation of Total Nitrogen Content in Forage Maize (*Zea mays* L.) Using Spectral Indices: Analysis by Random Forest. *Agriculture*, 10(10). <https://doi.org/10.3390/agriculture10100451>
- Miphokasap P, & Wannasiri W. (2018). Estimations of Nitrogen Concentration in Sugarcane Using Hyperspectral Imagery. *Sustainability*, 10(4). <https://doi.org/10.3390/su10041266>
- Nigon T, Yang C, Dias Paiao G, Mulla D, Knight J, & Fernández F. (2020). Prediction of Early Season Nitrogen Uptake in Maize Using High-Resolution Aerial Hyperspectral Imagery. *Remote Sensing*, 12(8). <https://doi.org/10.3390/rs12081234>
- Prado Osco L, Marques Ramos AP, Roberto Pereira D, Akemi Saito Moriya É, Nobuhiro Imai N, Takashi Matsubara E, Estrabis N, de Souza M, Marcato Junior J, Gonçalves WN, Li J, Liesenberg V, & Eduardo Creste J. (2019). Predicting Canopy Nitrogen Content in Citrus-Trees Using Random Forest Algorithm Associated to Spectral Vegetation Indices from UAV-Imagery. *Remote Sensing*, 11(24). <https://doi.org/10.3390/rs11242925>
- Qiao L, Tang W, Gao D, Zhao R, An L, Li M, Sun H, & Song D. (2022). UAV-based chlorophyll content estimation by evaluating vegetation index responses under different crop coverages. *Computers and Electronics in Agriculture*, 196. <https://doi.org/10.1016/j.compag.2022.106444>

- Rouse J, Haas J, Schell J, Deering D, & JC H. (1974). Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. *NASA/GSFC Type III Final Report*, 144661. <https://doi.org/19740022555>
- Shen Y, Meng R, Li J, Zhao P, & Zhao F. (2026). A hybrid method for major food crops leaf chlorophyll content inversion driven by remote sensing mechanisms and deep learning. *National Remote Sensing Bulletin*, 5, 1392-1321.
- Shi S, Xu L, Gong W, Chen B, Chen B, Qu F, Tang X, Sun J, & Yang J. (2022). A convolution neural network for forest leaf chlorophyll and carotenoid estimation using hyperspectral reflectance. *International Journal of Applied Earth Observation and Geoinformation*, 108. <https://doi.org/10.1016/j.jag.2022.102719>
- Sims DA, & Gamon JA. (2002). Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sensing of Environment*, 81(2-3), 337-354. [https://doi.org/10.1016/s0034-4257\(02\)00010-x](https://doi.org/10.1016/s0034-4257(02)00010-x)
- Singh AK, Ganapathysubramanian B, Sarkar S, & Singh A. (2018). Deep Learning for Plant Stress Phenotyping: Trends and Future Perspectives. *Trends Plant Sci*, 23(10), 883-898. <https://doi.org/10.1016/j.tplants.2018.07.004>
- Song Y, Zhang J, Liu Z, Xu Y, Quan S, Sun L, Bi J, & Wang X. (2025). Deep learning for hyperspectral image classification: A comprehensive review and future predictions. *Information Fusion*, 123. <https://doi.org/10.1016/j.inffus.2025.103285>
- Ta X, An D, & Wei Y. (2019). Erratum to: ‘Dissolved oxygen prediction method for recirculating aquaculture system, based on a timing attenuation matrix and a convolutional neural network’ *Computers and Electronics in Agriculture* 145 (2018) 302–310. *Aquaculture*, 503. <https://doi.org/10.1016/j.aquaculture.2019.02.020>
- Tang Y, Li F, Jin Y, Wu L, & Yan D. (2024). Comparison of Methods for Estimating Chlorophyll Content in Maize Leaves by Red Edge Position. *Journal of Maize Sciences*, 32(4), 46-54. <https://doi.org/10.13597/j.cnki.maize.science.20240407>
- Tong F, & Zhang Y. (2024). Tree Species Classification on Hyperspectral Imagery Using Fewer Training Samples. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-4-2024, 71-76. <https://doi.org/10.5194/isprs-archives-XLVIII-M-4-2024-71-2024>
- Vapnik V, Golowich S, & Smola A. (1996). Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing *Advances in neural*
- Verma B, Prasad R, Srivastava PK, Yadav SA, Singh P, & Singh RK. (2022). Investigation of optimal vegetation indices for retrieval of leaf chlorophyll and leaf area index using enhanced learning algorithms. *Computers and Electronics in Agriculture*, 192. <https://doi.org/10.1016/j.compag.2021.106581>
- Verrelst J, Camps-Valls G, Muñoz-Marí J, Rivera JP, Veroustraete F, Clevers JGPW, & Moreno J. (2015). Optical remote sensing and the retrieval of terrestrial vegetation biogeophysical properties – A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 108, 273-290. <https://doi.org/10.1016/j.isprsjprs.2015.05.005>
- Verrelst J, Rivera-Caicedo JP, Reyes-Munoz P, Morata M, Amin E, Tagliabue G, Panigada C,

- Hank T, & Berger K. (2021). Mapping landscape canopy nitrogen content from space using PRISMA data. *ISPRS J Photogramm Remote Sens*, 178, 382-395. <https://doi.org/10.1016/j.isprsjprs.2021.06.017>
- Wakahara S, Miao Y, Li D, Zhang J, Gupta SK, & Rosen C. (2025). Evaluating the Potential of Improving In-Season Potato Nitrogen Status Diagnosis Using Leaf Fluorescence Sensor as Compared with SPAD Meter. *Remote Sensing*, 17(13). <https://doi.org/10.3390/rs17132311>
- Wan L, Zhou W, He Y, Wanger TC, & Cen H. (2022). Combining transfer learning and hyper spectral reflectance analysis to assess leaf nitrogen concentration across different plant species datasets. *Remote Sensing of Environment*, 269. <https://doi.org/10.1016/j.rse.2021.112826>
- Xiang S, Wang S, Jin Z, Xiao Y, Liu M, Yang H, Feng S, Feng Z, Liu T, Yu F, & Xu T. (2025). RSPECT: A PROSPECT-based model incorporating the real structure of rice leaves. *Remote Sensing of Environment*, 330. <https://doi.org/10.1016/j.rse.2025.114962>
- Yang H, Hu Y, Yin H, Jin Q, Li F, & Yu K. (2025a). Improving potato leaf chlorophyll content prediction using a machine learning model with a hybrid dataset. *International Journal of Remote Sensing*, 46(8), 3064-3088. <https://doi.org/10.1080/01431161.2025.2465916>
- Yang H, Li F, Hu Y, & Yu K. (2021). Hyperspectral indices optimization algorithms for estimating canopy nitrogen concentration in potato (*Solanum tuberosum* L.). *International Journal of Applied Earth Observation and Geoinformation*, 102. <https://doi.org/10.1016/j.jag.2021.102416>
- Yang H, Li F, Hu Y, & Yu K. (2025b). Improving the performance of plant nitrogen assessment in drip-irrigated potatoes using optimized spectral indices-based machine learning. *Precision Agriculture*, 26(3). <https://doi.org/10.1007/s11119-025-10248-y>
- Yang J, Guo X, Li Y, Marinello F, Ercisli S, & Zhang Z. (2022). A survey of few-shot learning in smart agriculture: developments, applications, and challenges. *Plant Methods*, 18(1), 28. <https://doi.org/10.1186/s13007-022-00866-2>
- Yue J, Leng M, & Tian Q. (2024). Estimation of Leaf Physical and Chemical Parameters Based on Hyperspectral Remote Sensing and Deep Learning Technologies. *Spectroscopy and Spectral Analysis*, 44(10), 2873-2883.
- Zarco-Tejada PJ, Miller JR, Noland TL, Mohammed GH, & Sampson PH. (2001). Scaling-up a band and model inversion methods with narrowband optical indices for chlorophyll content estimation in closed forest canopies with hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 39(7), 1491-1507. <https://doi.org/10.1109/36.934080>
- Zha H, Miao Y, & Wang T. (2020). Improving Unmanned Aerial Vehicle Remote Sensing-Based Rice Nitrogen Nutrition Index Prediction with Machine Learning. *Remote Sensing*, 12(2). <https://doi.org/10.3390/rs12020215>
- Zhang J, Gao D, & Song D. (2022a). Wavelengths Optimization and Chlorophyll Content Detection Based on PROSPECT Model. *Spectroscopy and Spectral Analysis*, 1514-1521.
- Zhang X, Sun H, Qiao X, Yan X, Feng M, Xiao L, Song X, Zhang M, Shafiq F, Yang W, & Wang C. (2022b). Hyperspectral estimation of canopy chlorophyll of winter wheat by using the optimized vegetation indices. *Computers and Electronics in Agriculture*, 193.

<https://doi.org/10.1016/j.compag.2021.106654>

- Zhao B, Duan A, Ata-Ul-Karim ST, Liu Z, Chen Z, Gong Z, Zhang J, Xiao J, Liu Z, Qin A, & Ning D. (2018). Exploring new spectral bands and vegetation indices for estimating nitrogen nutrition index of summer maize. *European Journal of Agronomy*, 93, 113-125. <https://doi.org/10.1016/j.eja.2017.12.006>
- Zhao R, An L, Tang W, Qiao L, Wang N, Li M, Sun H, & Liu G. (2023). Improving chlorophyll content detection to suit maize dynamic growth effects by deep features of hyperspectral data. *Field Crops Research*, 297. <https://doi.org/10.1016/j.fcr.2023.108929>