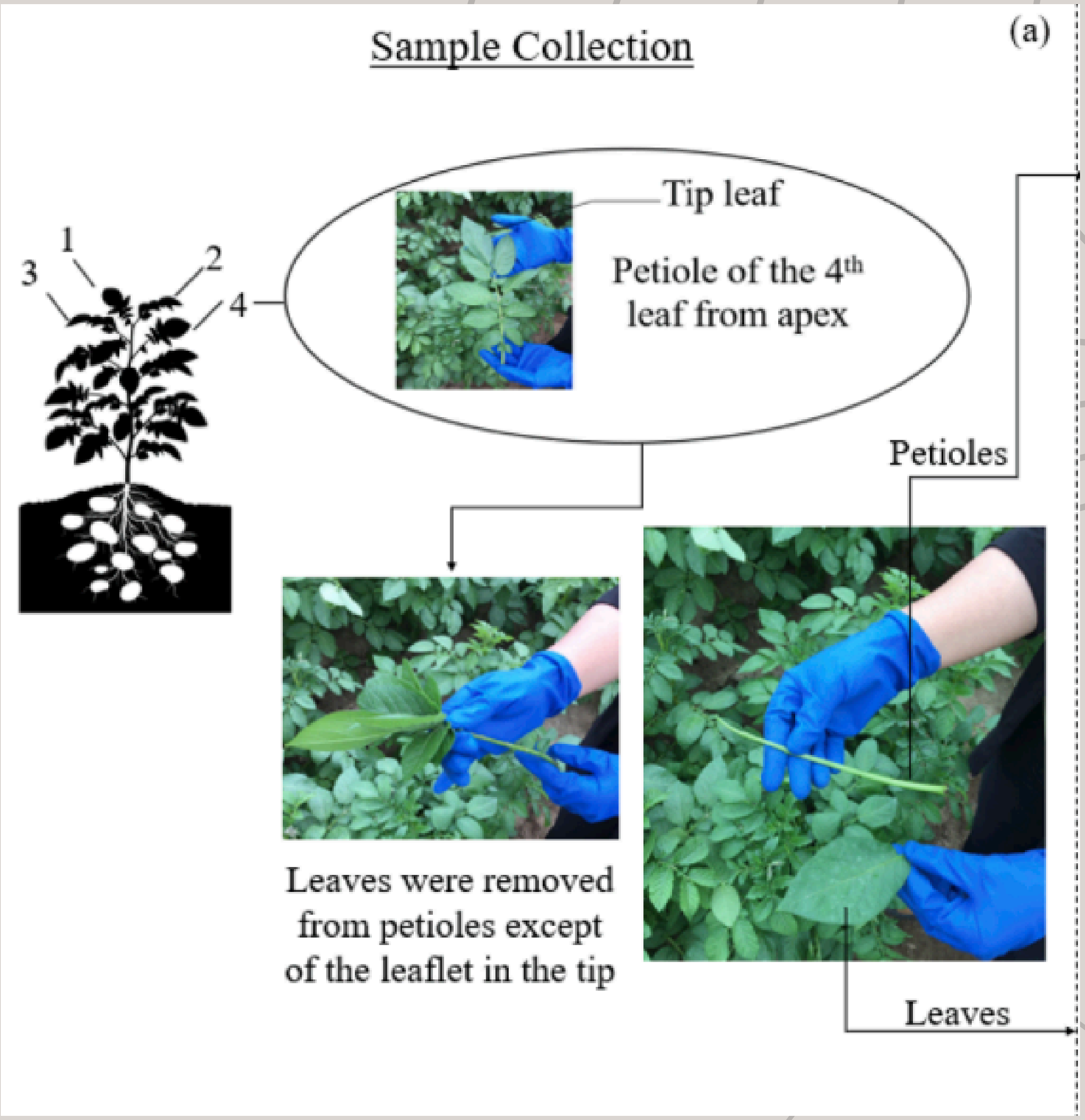
DATA SCIENCE MINI PROJECT
GROUP T14: PROBLEM D

Introduction

- Traditional nutrient assays (soil/plant chemical tests) are accurate but time-consuming, costly and destructive.
- Vis–NIR reflectance (400–2500 nm) offers a fast, non-invasive alternative for assessing plant nutrients.
- Previous studies show leaf spectra can estimate multiple nutrients; notably, dried leaf samples often yield higher accuracy than fresh.


Sample Collection
(a)
1
2
3
4
Tip leaf
Petiole of the 4th leaf from apex
Petioles
Leaves were removed from petioles except of the leaflet in the tip
Leaves

Sample Data Collection

Picture taken from: Abukmeil, R., Al-Mallahi, A., Campelo, F. (2022). https://doi.org/10.1016/j.compag.2022.107074

Data Description and Preparation

01

Data from Dalhousie University (Canada): 7 datasets of potato leaf and petiole samples, in two modes (fresh vs dried)

02

Spectra collected over 400–2500 nm range; samples equally divided into fresh and dried leaves

03

Nutrient features converted to uniform units (ppm to %) and all features were standardised (zero mean, unit variance)

Data Description and Preparation

04

Missing values handled per season: in one set, Boron and Chloride (3 values each) imputed with XGBoost regressor; Aluminum (3 values) imputed by column mean due to poor model fit

05

Other seasons: isolated missing N values (imputed with XGBoost) and rows without spectral data were removed

06

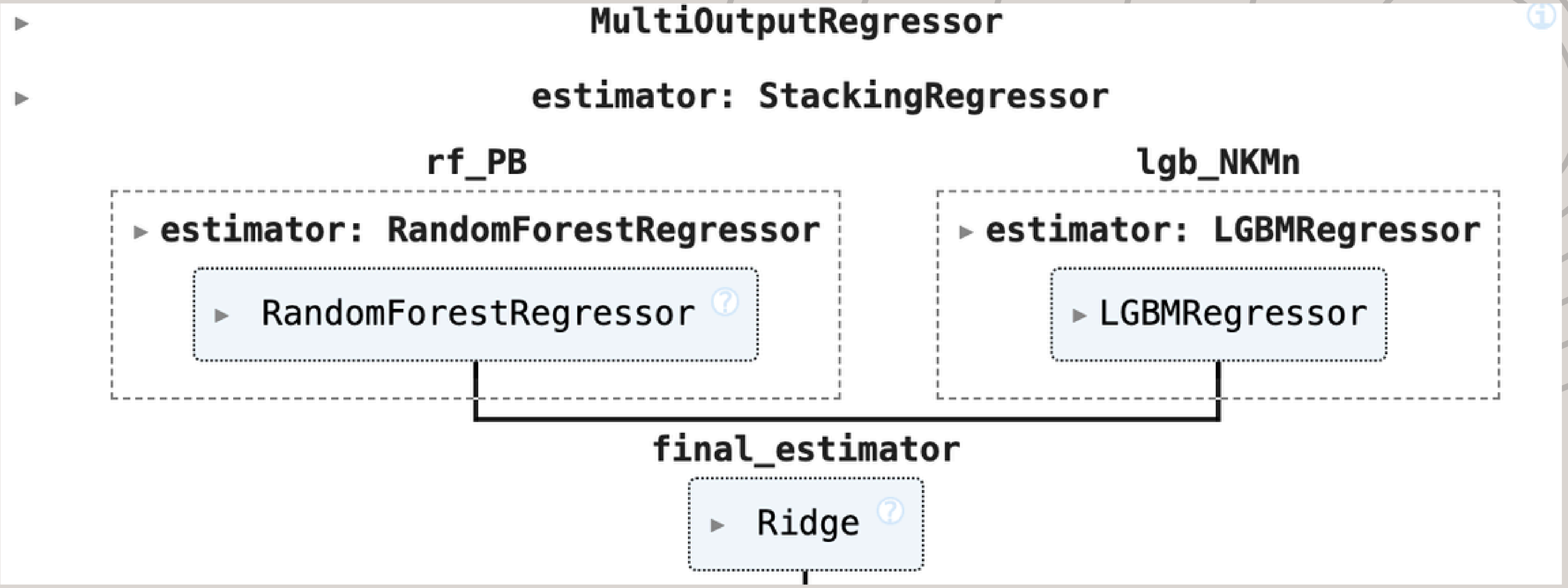
Dimensionality reduction: aggregated adjacent spectral bands into 8 nm average bins, smoothing noise while preserving key signal

Methodology: Overall Workflow

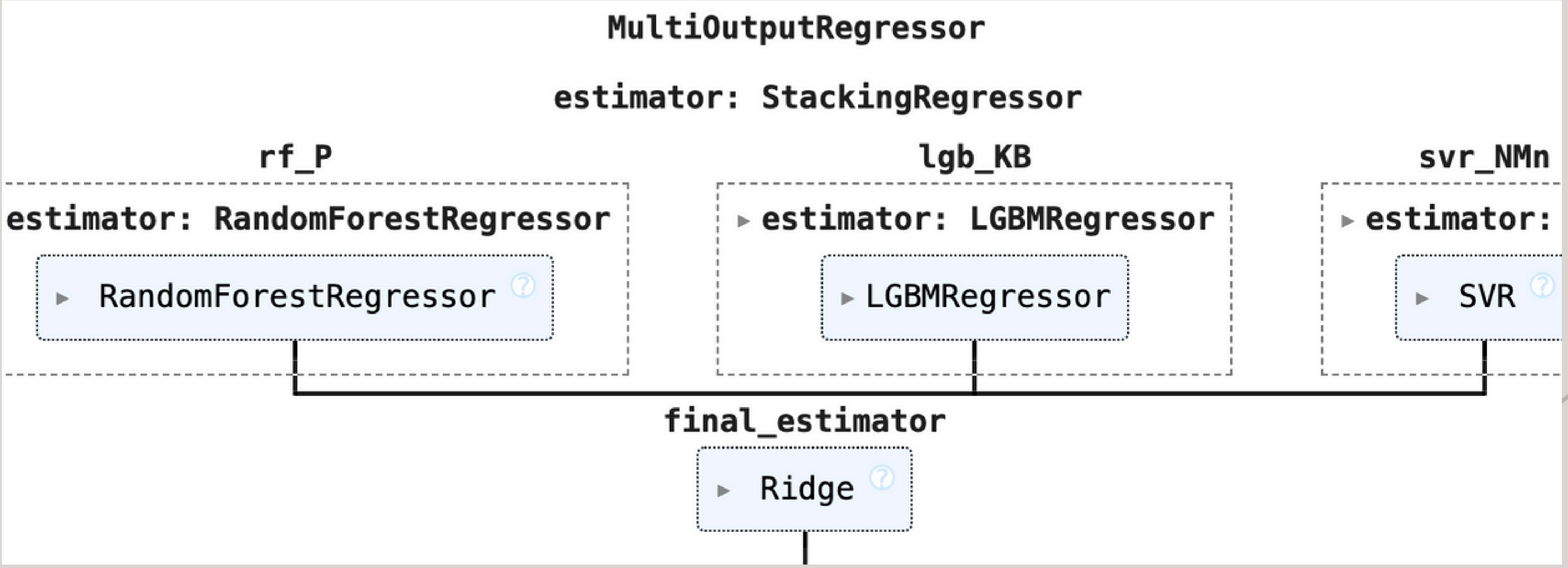
- Use Partial Least Squares (PLS) to project high-dimensional spectra into a small set of latent features.
- Identify the two most correlated micronutrients (via Pearson correlation) for each target nutrient; use their predicted values as additional inputs.
- Build a two-tier regression: base models predict each nutrient (using PLS features + micronutrient mediators), then a Ridge-meta model stacks these predictions

Methodology:
Stacking Pipeline & Transfer Learning

- Base models (e.g. Random Forest, Gradient Boosting, etc.) make nutrient predictions; these are combined by scikit-learn’s StackingRegressor with a Ridge meta-regressor.
- The entire stacking pipeline is wrapped in a MultiOutputRegressor, so all five nutrients are predicted simultaneously under a shared cross-validation scheme.



Fresh mode

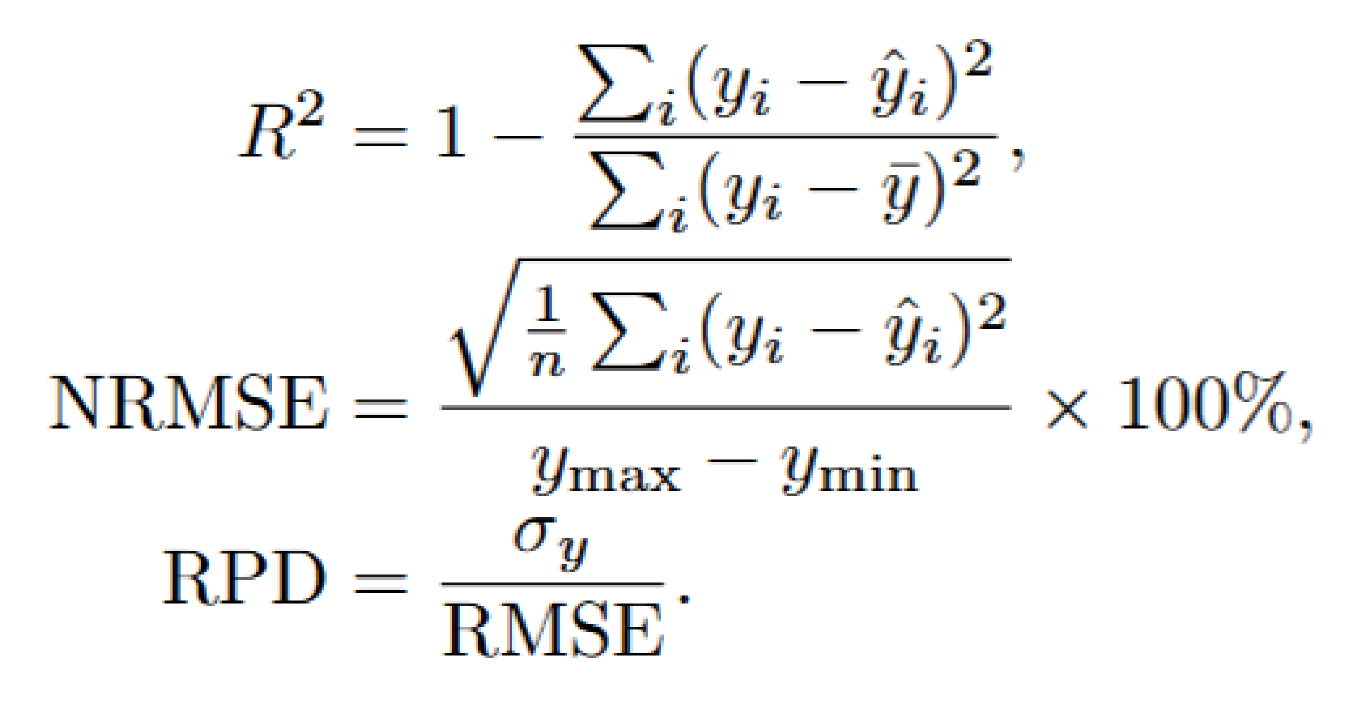


Dried mode

Methodology: Transfer Learning & Introduction for metrics

- Transfer learning: train on Season 2 first; then sequentially add Season 3, Season 1, and Season 4 data (retraining pipeline each time) to adapt to multi-season variability

- R² measures the proportion of data variance explained by a model; NRMSE is the RMSE divided by a reference value (e.g. range or mean) for scale-free error comparison; RPD is the standard deviation of the reference data over RMSE, with higher values indicating stronger predictive performance.

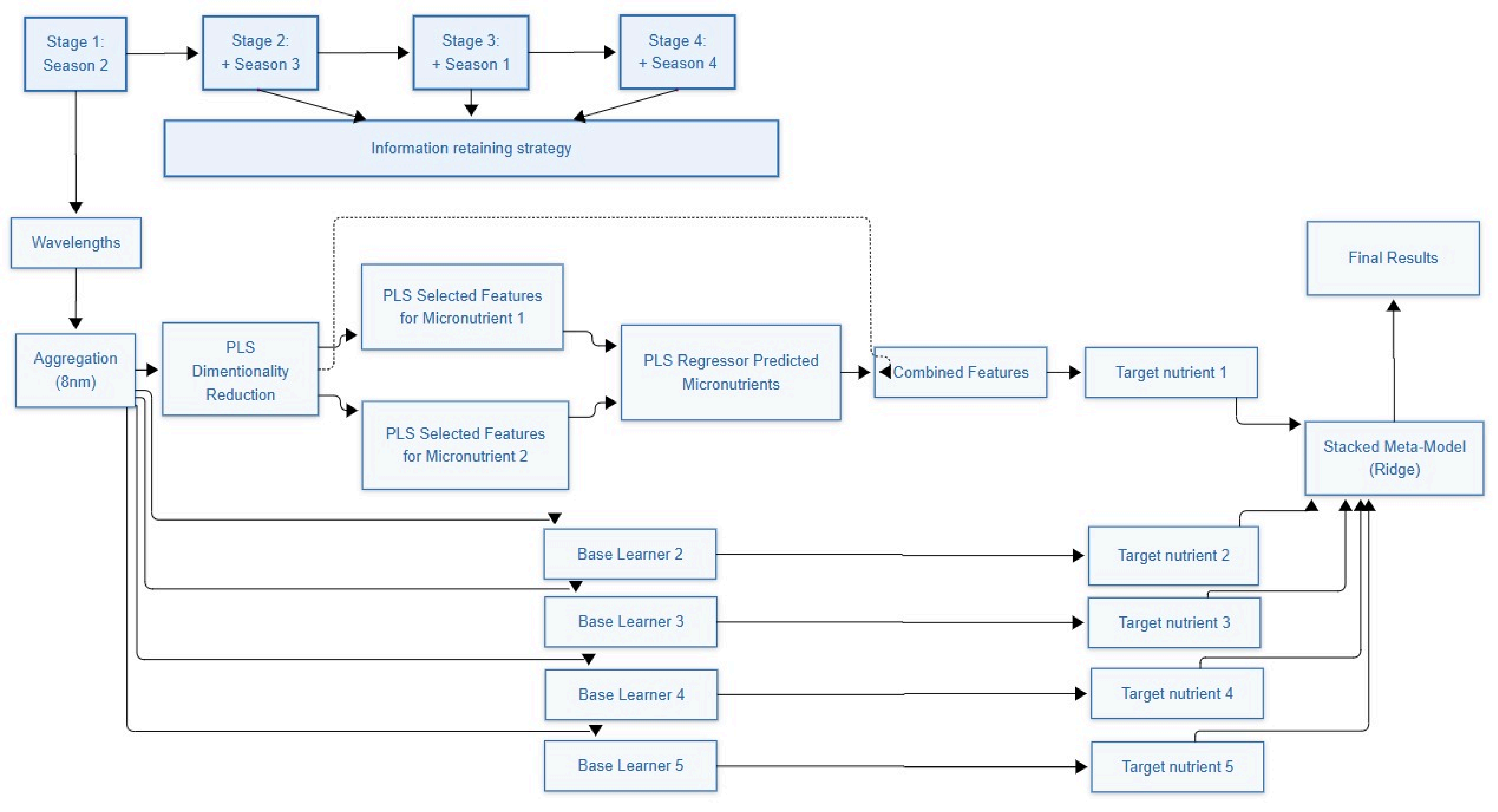


$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}}{y_{\max} - y_{\min}} \times 100\%,$$

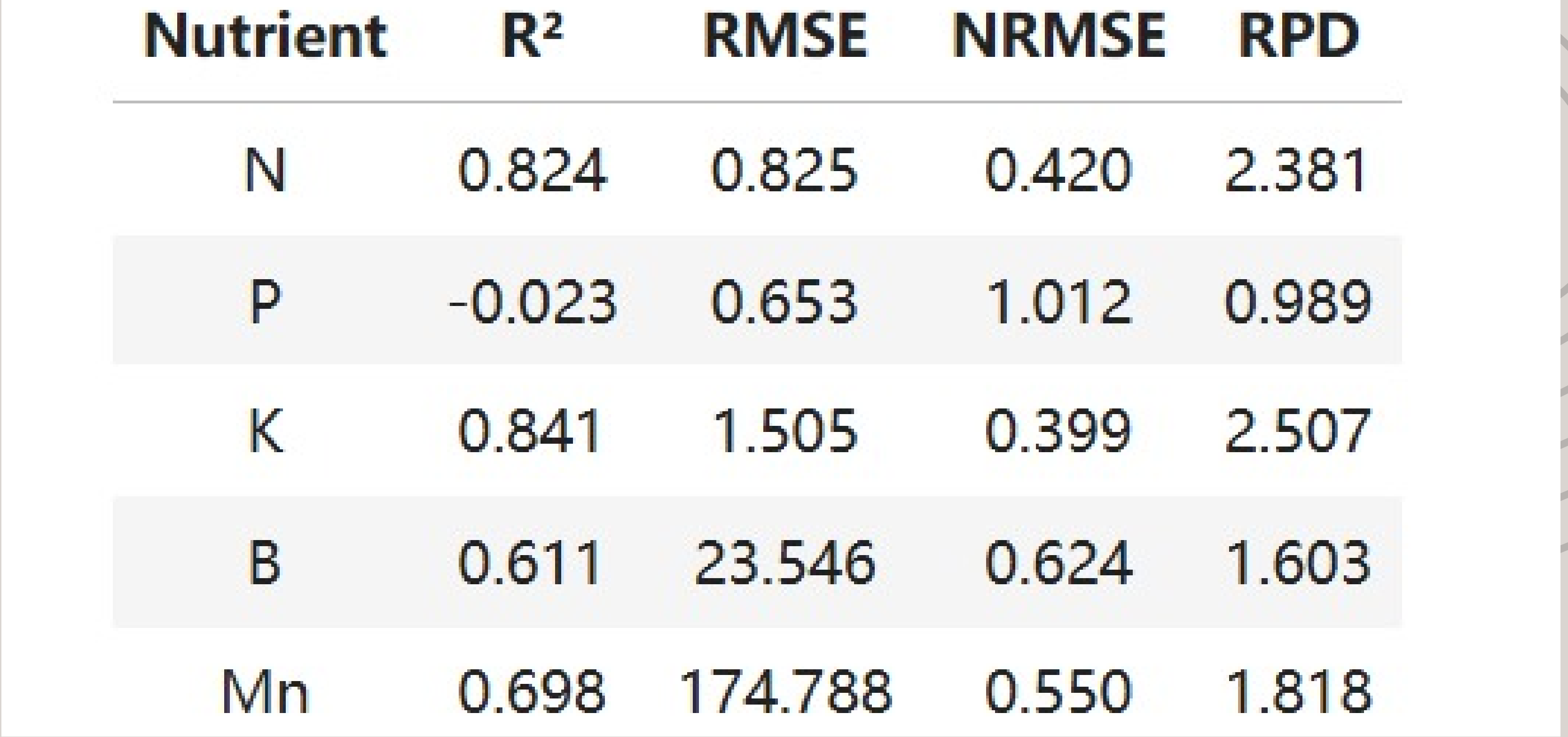
$$\mathrm{RPD} = \frac{\sigma_y}{\mathrm{RMSE}}.$$

Methodology: Workflow Diagram



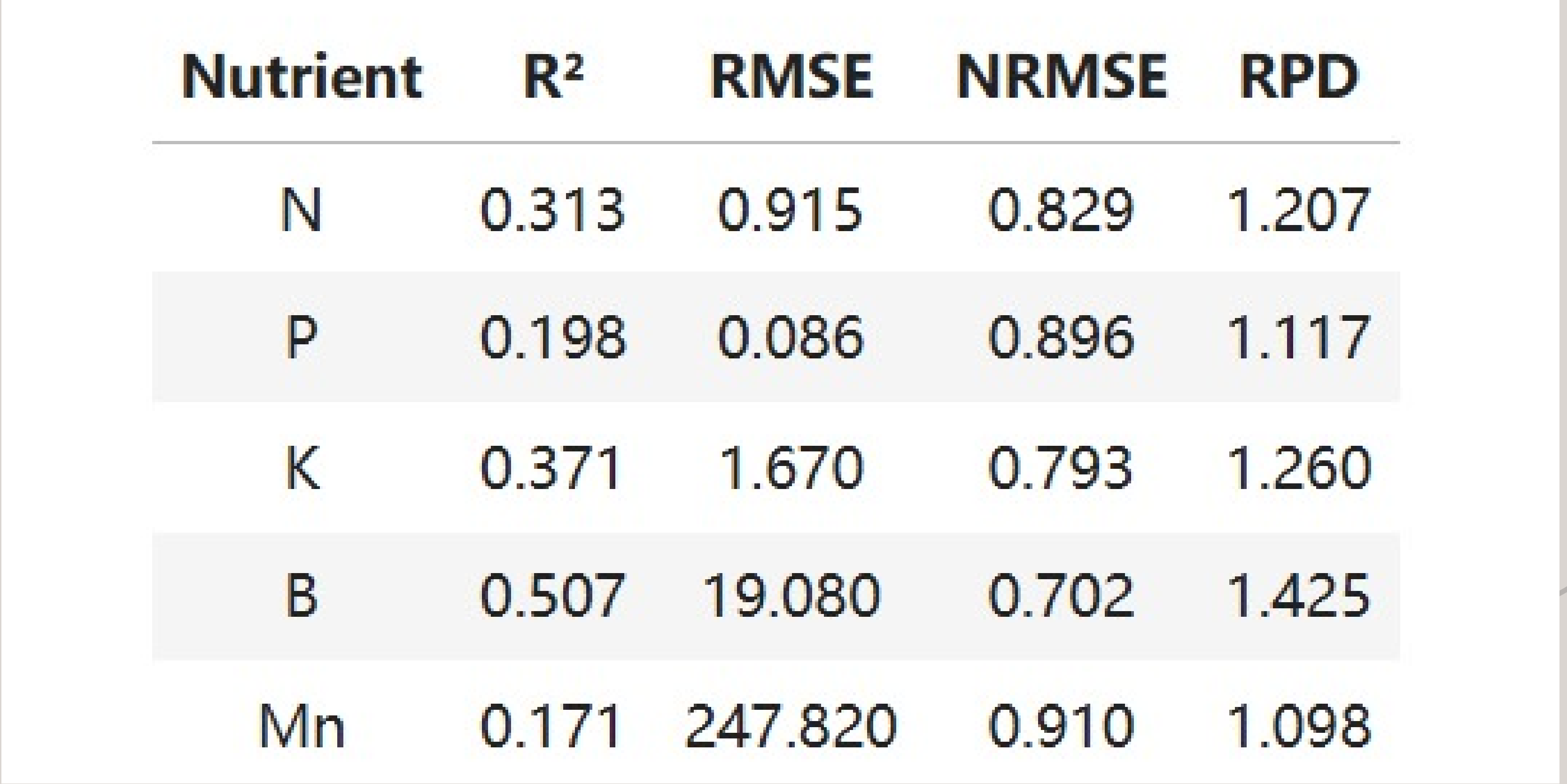
Results & Discussion (Stacking)

- The stacked ensemble pipeline produced reliable, multi-nutrient predictions from the leaf spectra.
- Best performance was for Nitrogen and Potassium: test R² ≈ 0.8 and RPD ≈ 2.3–2.5 (approaching quantitative accuracy).
- Manganese and Boron had moderate accuracy (RPD ≈ 1.6–1.8), while Phosphorous was poorest (RPD ≈ 0.9-1.1), suggesting P may require alternate modelling



Nutrient	R²	RMSE	NRMSE	RPD
N	0.824	0.825	0.420	2.381
P	-0.023	0.653	1.012	0.989
K	0.841	1.505	0.399	2.507
B	0.611	23.546	0.624	1.603
Mn	0.698	174.788	0.550	1.818

Fresh mode

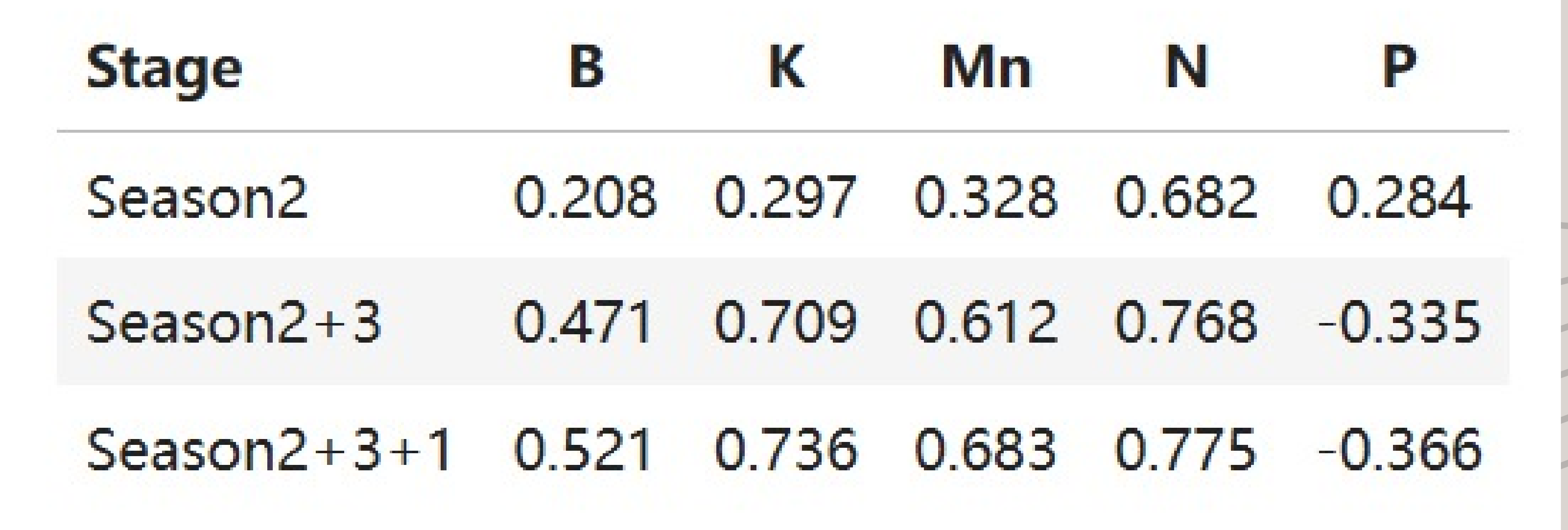


Nutrient	R²	RMSE	NRMSE	RPD
N	0.313	0.915	0.829	1.207
P	0.198	0.086	0.896	1.117
K	0.371	1.670	0.793	1.260
B	0.507	19.080	0.702	1.425
Mn	0.171	247.820	0.910	1.098

Dried mode

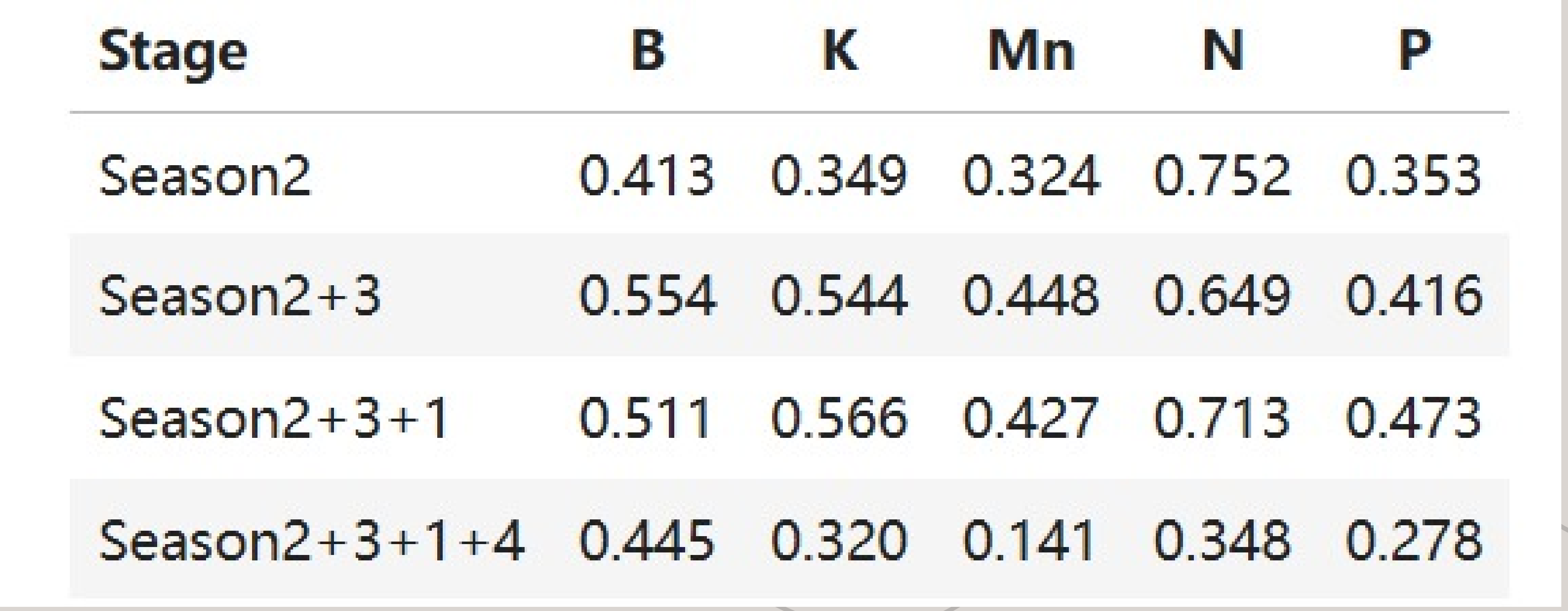
Results & Discussion (Transfer Learning)

- Progressive transfer learning (incremental multi-season training) further improved generalisation, especially for the fresh-leaf data.
- In the fresh mode after 3 transfer stages, Nitrogen and Potassium both achieved R² > 0.72, indicating excellent reliability
- The combined PLS-stacking model with transfer learning was robust and interpretable across seasons



Stage	B	K	Mn	N	P
Season2	0.208	0.297	0.328	0.682	0.284
Season2+3	0.471	0.709	0.612	0.768	-0.335
Season2+3+1	0.521	0.736	0.683	0.775	-0.366

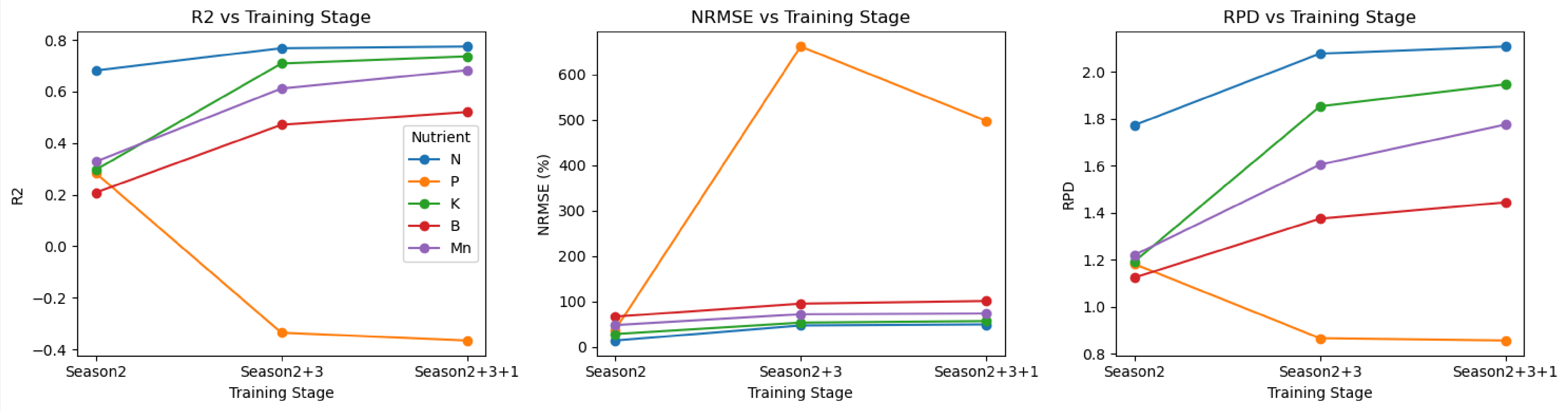
Fresh mode



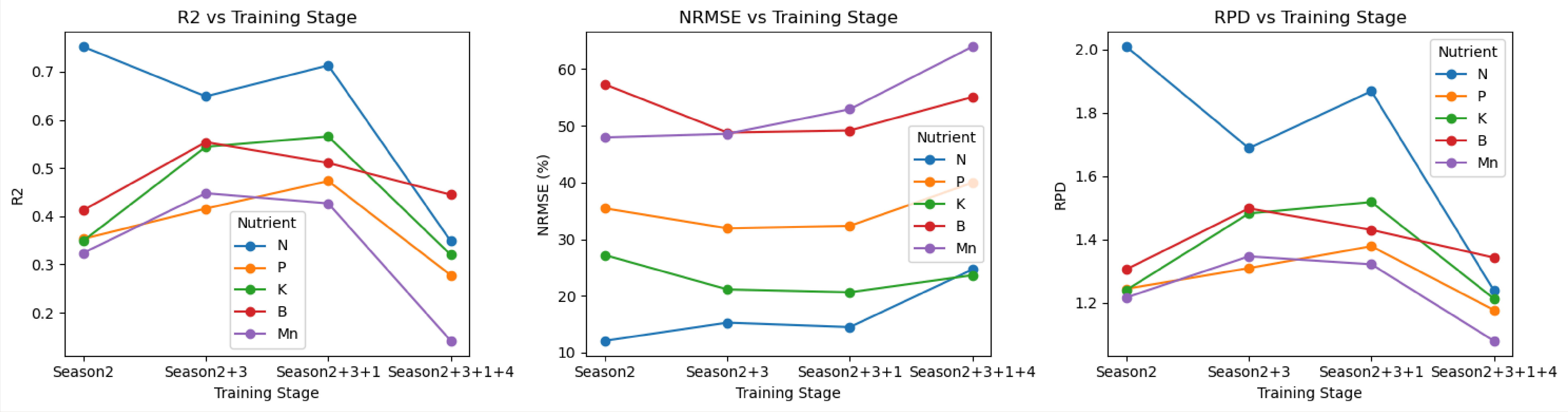
Stage	B	K	Mn	N	P
Season2	0.413	0.349	0.324	0.752	0.353
Season2+3	0.554	0.544	0.448	0.649	0.416
Season2+3+1	0.511	0.566	0.427	0.713	0.473
Season2+3+1+4	0.445	0.320	0.141	0.348	0.278

Dried mode

Results & Discussion (Transfer Learning)



Fresh mode



Dried mode

Future Work

1

Improve Phosphorous prediction: incorporate non-linear models or additional features since P RPD was low

2

Deploy as a continuously learning system: e.g. package inference in an API that can be updated with new data in the background.

3

Extend and validate the pipeline on new seasons/environments, leveraging progressive multi-season training to further boost robustness

Thank You