Data Science Mini Project Report: Group T14

Anjani Upadhyay
Department of Engineering Mathematics
University Of Bristol
xx24941@bristol.ac.uk

Sahana Vishwanath Somani
Department of Engineering Mathematics
University Of Bristol
ag24971@bristol.ac.uk

Hangwei Pu
Department of Engineering Mathematics
University Of Bristol
ek24099@bristol.ac.uk

Shaoshuai Li
Department of Engineering Mathematics
University Of Bristol
bf24104@bristol.ac.uk

***Abstract*—Chemical analysis of potato plants is critical to optimise nutrient management and improve crop yield, but traditional methods are often resource intensive and invasive. This study explores the potential of ground-based spectral analysis as a non-destructive alternative for assessing nutrient concentrations in potato petioles and leaves, addressing the limitations of conventional tissue testing, which is labor intensive, costly and destructive. Despite advances in canopy-level reflectance for nitrogen estimation, there remains a gap in using leaf and petiole spectra to predict a broader range of nutrient contents in potatoes. This research aims to evaluate the accuracy of predicting nutrient concentrations using leaf spectra from fresh and dried potato leaves and to determine the efficacy of these spectra in estimating petiole nutrient levels. The leaves were equally divided into fresh and dried groups and spectral data was collected in the 400–2500 nm range. Stacking pipeline models were developed to estimate nutrient concentrations, incorporating various preprocessing techniques for feature standardisation, dimensionality reduction and handling missing values. Model performance was assessed using the Ratio of Prediction to Deviation (RPD), R-squared (R^2) and Root Mean Squared Error (RMSE) metrics.**

I. Introduction

Understanding the composition of soil nutrients is essential to monitor agricultural health and optimise crop yields. Traditional laboratory-based chemical analyses are accurate, but they can be time consuming, costly and often impractical for large-scale or real-time applications. With the increasing availability of high-resolution spectral data, reflectance based sensing has emerged as a promising alternative for estimating soil or plant nutrient concentrations. However, spectral data typically contain hundreds of variables and suffer from seasonal and environmental variability, posing challenges for robust and interpretable modeling.

In this project, our aim is to develop a generalisable and scalable pipeline for predicting five key nutrients: Nitrogen, Phosphorus, Potassium, Boron and Manganese based on reflectance spectra collected from dried and fresh plant samples across different seasons. Our approach evolved through extensive experimentation and critical evaluation of several methodologies. Early attempts at direct regression from the spectra yielded limited generalisability. We therefore explored a more sophisticated architecture that incorporates domain knowledge via correlation analysis and progressive learning over time.

Our final methodology consists of three main components. First, we use Partial Least Squares (PLS) regression to reduce the high-dimensional spectral input to informative latent features. Second, we identify the top two micronutrients most correlated with each target nutrient using Pearson correlation, and include their predicted values as mediators in the learning process. Finally, a two-tier regression model is built: base regressors estimate each target nutrient using both PLS components and predicted micronutrient values, while a meta-level Ridge multi output regressor helps us to stack these predictions.

To further tackle the issue of season-to-season variability, we implemented a progressive transfer learning strategy. Training was carried out in a sequence across all the seasons with a fraction of the previous data retained at each stage. A final refinement step consolidates all prior knowledge. This approach allowed the model to accumulate useful general patterns while gradually adapting to new season specific variations. The results demonstrate that our hybrid architecture, when combined with transfer learning, delivers strong predictive performance (especially for fresh mode) across most stages. In particular, we achieved high R^2 scores for Nitrogen and Potassium after 3 stages of transfer learning in the fresh mode (both exceeding 0.72), indicating excellent model reliability. The general framework exhibits robustness, interpretability and practical viability for real-world nutrient prediction from spectral data.

II. Literature Review

Vis–NIR spectroscopy has been widely applied for rapid plant nutrient analysis. Conventional tissue testing (e.g. petiole analysis in potatoes) is laborious, destructive and costly, whereas reflectance spectroscopy (400–2500 nm) offers a fast, non-destructive alternative [1]. Recent studies demonstrate the

feasibility of predicting multiple nutrients from leaf spectra. For example, foliar Vis–NIR reflectance can estimate potato petiole nutrient concentrations with strong predictive performance across most macro and micro nutrients (ratio of prediction deviation values rated "reasonable" to "excellent" for all elements except Na) [1]. Notably, dried leaf samples tended to yield higher correlation with laboratory measurements than fresh leaves [1]. These results underscore the potential of Vis–NIR spectroscopy, especially in prepared (dried) samples, to rapidly assess nutritional status of the plant in a multi element context.

Dimensionality reduction and ensemble learning further enhance such spectral models. Raw Vis–NIR spectra consist of hundreds of highly collinear wavelength features, so chemometric techniques such as partial least squares (PLS) regression are commonly employed to distil spectra into a few informative latent variables [8]. PLS can relate the full spectral matrix to nutrient concentrations while handling noise and multicollinearity, making it a standard tool for extracting relevant variance in high-dimensional spectral data [8]. Advanced regression strategies further leverage correlations between multiple nutrient outputs. Multivariate (multi-target) regression can outperform single-target models by exploiting inter-nutrient relationships [6]. A stacked multivariate regression approach (using stacked generalisation with lasso/ridge base learners) has achieved superior accuracy in predicting correlated outcome variables [6]. The learning of the group through stacking has also shown benefits in agricultural applications: By integrating hyperspectral indices of multiple stages of wheat growth into a stacking group, researchers obtained more stable and accurate yield predictions than from any model of any individual stage [4]. Building on these insights, a stacked regression framework that combines intermediate predictions of correlated micro nutrients with reflectance features can further improve macro nutrient estimation.

III. Data Description and Preparation

A. Data Collection

For our data science mini-project conducted as part of the curriculum at the University of Bristol, we worked in a group of four to analyse datasets provided by Dalhousie University, Canada. The data sets comprised seven distinct collections, categorised into two modes: dried and fresh. These data sets included measurements of nutrient concentrations and spectral reflectance values from leaf and petiole samples of potato crops, specifically designed to explore relationships between chemical composition and spectral properties.

The dried data sets consisted of four seasonal collections, in which the leaves and petioles were subjected to a drying process at 55-60 ° C for 16-24 hours to reduce the water content and achieve a constant weight. In contrast, the fresh data sets included three seasonal collections, with samples analysed without drying. Across all datasets, the number of data points varied, ranging from a minimum of 41 to a maximum of 145. Nutrient concentrations measured included Nitrogen (N), Phosphorus (P), Potassium (K), Calcium (Ca), Magnesium (Mg), Sulfur (S), Manganese (Mn), Zinc (Zn), Iron (Fe), Sodium (Na), Boron (B), Copper (Cu), Aluminum (Al) and Chloride (Cl), although not all nutrients were present in every dataset. These chemical results were paired with spectral reflectance values of the leaves, measured using a Near-Infrared Spectroscopy (NIRS) Analyser (DS2500, Metrohm USA Inc.) across wavelengths from 400 nm to 2500 nm at intervals of 0.5 nm, yielding a total of 4200 readings per sample.

B. Feature Standardisation

In the dataset, significant variation is observed in magnitudes, units and ranges, with some nutrients measured in percentage concentrations (%) and others in parts per million (ppm). Without proper handling, these algorithms prioritise the magnitude of features, while their units are ignored, resulting in skewed outcomes. For example, features with higher magnitudes, such as those in percentage units, would be disproportionately weighted in distance computations compared to features with lower magnitudes, such as those in ppm.

To address this issue, a two-step process was employed to ensure uniformity and comparability across all features. First, all features were converted into a single unit by addressing the difference between the ppm and percentage measurements. Since 1% equals $10,000$ ppm, ppm values (Manganese, Boron, Zinc, Iron, Copper and Aluminum) were divided by $10,000$ to be expressed as percentages, aligning them with the other features (Nitrogen, Phosphorus, Potassium, Calcium, Magnesium, Sulphur, Sodium and Chloride).

Subsequently, standardisation was applied to redistribute the feature values. Each value was transformed into its corresponding z score, calculated as

$$z = \frac{x - \mu}{\sigma}$$

where μ represents the mean and σ denotes the standard deviation of the feature.

Through this process, the features were rescaled to have a mean of 0 and a standard deviation of 1, effectively normalising their distributions and ensuring equal contribution of each feature to the distance-based computations.

C. Missing Value Imputation Strategies

In the given datasets, missing values were addressed using specific strategies tailored to each season. For FreshSeason01 and DriedSeason01, missing values were observed in the Boron, Chloride and Aluminum columns. Boron and Aluminum each had three missing values, while Chloride had six missing values. An XGBoost regressor was employed to impute these missing values due to its ability to factor in missing data and optimise imputation values based on training loss reduction. While reasonable performance scores were

achieved for Boron and Chloride, Aluminum yielded a poor R^2 Score of -8.95, prompting the use of the column mean for imputation instead.

For FreshSeason02 and DriedSeason02, no missing values were present, eliminating the need for imputation. In FreshSeason03, one missing value was detected in the Nitrogen column and was imputed using XGBoost. Additionally, rows containing entirely missing data were removed. Similarly, in DriedSeason04, one Nitrogen value was imputed with XGBoost and one row lacking wavelength reflectance values (while containing concentration values) was excluded from analysis.

The dataset was initially split into features and the target column, with missing value columns designated as targets. Data was divided into training and test sets to support effective XGBoost model training, evaluated using metrics like Mean Squared Error (MSE) and R^2 Score. Once trained, the model predicted missing values for the target column based on associated features. These imputed values were integrated back into the original datasets, and updated CSV files were created and exported for each dataset.

D. Dimensionality Reduction Techniques

Given that the wavelength reflectance data is between 400 to 2500 nm at 0.5 nm intervals, the resultant feature space quickly becomes very large. To manage this complexity, an aggregation step was introduced in which multiple adjacent wavelengths were combined into single features. The key was to choose the bin size carefully, as too coarse an aggregation risks losing valuable information, while too fine an approach may retain excess noise and redundancy. During the first half of our project, we experimented and discovered that a bin size of 8 nm usually produced more stable models that were both computationally efficient and could still capture the critical wavelength dependent information needed for accurate nutrient concentration predictions. This approach effectively reduced the total number of predictors while smoothing out high-frequency noise in the original data. By grouping reflectance values in bins of around 16 consecutive points (given the increments of 0.5 nm), the dimensionality was cut substantially and broad spectral patterns more relevant to nutrient predictions were preserved.

Another technique called Principal Component Analysis (PCA) was initially considered for further reducing the dimensionality because it identifies and projects the data in the directions of maximal variance in the feature space alone. However, PCA does not take into account any relationship between the predictors (spectral data) and the target variables (nutrient concentrations). Consequently, while PCA can be effective in compressing large feature spaces into a smaller number of uncorrelated components, it can overlook the spectral features most relevant to predicting specific nutrient levels. This led to suboptimal performance when dealing with complex, high-dimensional datasets characteristic of near-infrared spectra.

On the contrary, Partial Least Squares (PLS) aligns the dimensionality reduction process more closely with the predictive goal. PLS extracts latent components that maximise the covariance between the predictors and the response variables. As a result, PLS naturally focuses on those spectral features that are most influential in explaining nutrient variability, significantly increasing prediction accuracy. In practice, the ability of PLS to balance feature reduction and target correlation makes it more suitable than PCA for these datasets. Consequently, while both PCA and PLS can shrink the dimensionality of large spectral datasets, PLS was chosen over PCA for nutrient concentration modelling in this context.

E. Distribution Plots for Nutrients

While visualising the nutrient distributions for one of our datasets (FreshSeason02), several patterns stood out that are broadly representative across other datasets as well. Macro nutrients such as Nitrogen (N), Phosphorus (P) and Potassium (K) usually exhibit relatively wide value ranges. In contrast, micro nutrients such as Boron (B) and Manganese (Mn) often appear with narrower distributions but can still show right skewed tendencies, where the majority of samples cluster around moderate values and a small subset of samples contain disproportionately high readings. Recognising this skewness was essential because it was required to perform statistical transformations to normalise the data before feeding it into prediction models.

Beyond sheer distributional shape, a closer look at the correlation patterns among these nutrients reveals meaningful relationships. From our correlation matrix, it is evident that certain pairings are closely coupled (for example, N and P often show a strong positive correlation, while K and Ca exhibit a high negative relationship). For micro nutrients specifically, the correlation results guide us in filtering which ones to include in our predictive modelling. Only those with moderate to strong links to the target macro nutrients (N, P and K) or to the primary micro nutrients of interest (B and Mn) are carried forward into model building. This not only highlights the biological relevance of selected micro nutrients but also helps keep the overall feature set manageable, reducing the likelihood of overfitting. By combining insights from both the distribution plots and correlation analyses, we tailored our data preprocessing and model architectures to capture the most salient, nutrient specific signals in each season's dataset.

IV. Methodology

A. Micro-Nutrients' Concentration Prediction

We begun by identifying the two most strongly correlated micronutrients for each target nutrient via Pearson's correlation across all seasons. Next, we aggregate the original 400–2500 nm reflectance into 8 nm bands, standardise each band and apply Partial Least Squares (PLS) regression to predict each selected micronutrient from the spectra. PLS projects the high-dimensional spectra into a low-dimensional latent space that maximises covariance with the micro nutrient measurements.

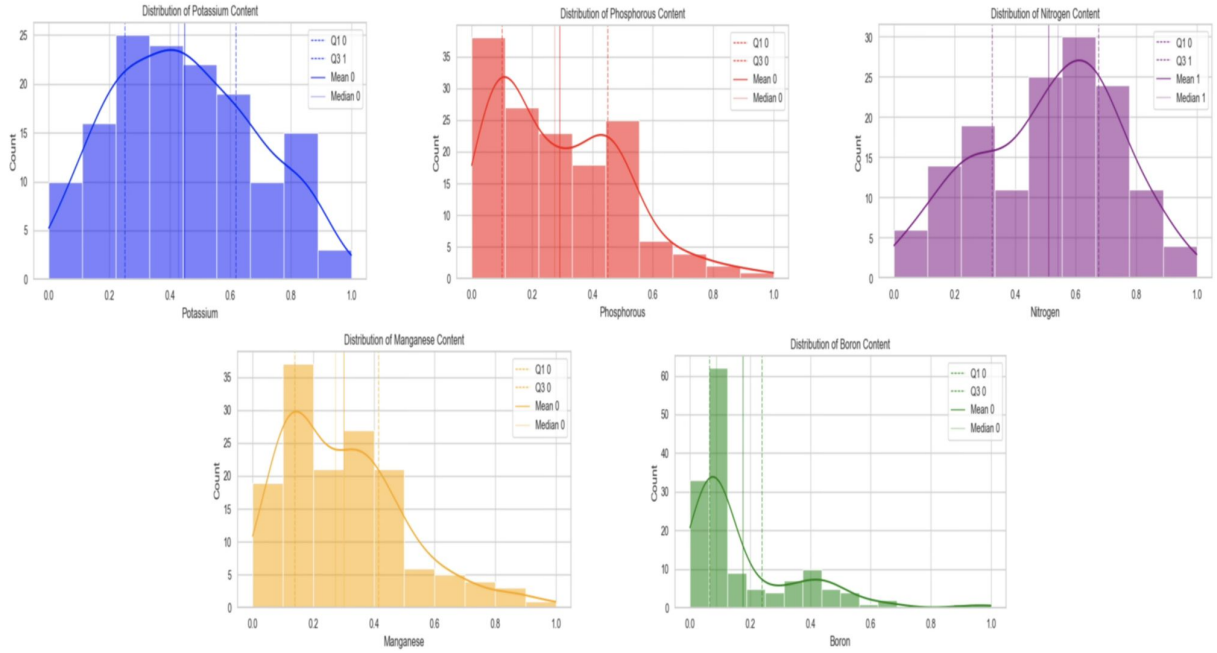

Fig. 1. Distribution plots for all the target nutrients

The adaptive selection of latent components ensures the capture of the most predictive spectral variance for each nutrient while preventing overfitting. This approach significantly compresses the data, reducing 263 spectral features to approximately 5–15 latent features. This compression enhances model training speed and mitigates multicollinearity issues among spectral bands.

$$T = XW, \quad \hat{y}^{(m)} = T\,c,$$

where X is the standardized reflectance matrix, W the PLS weights, T the latent scores and c the regression coefficients.

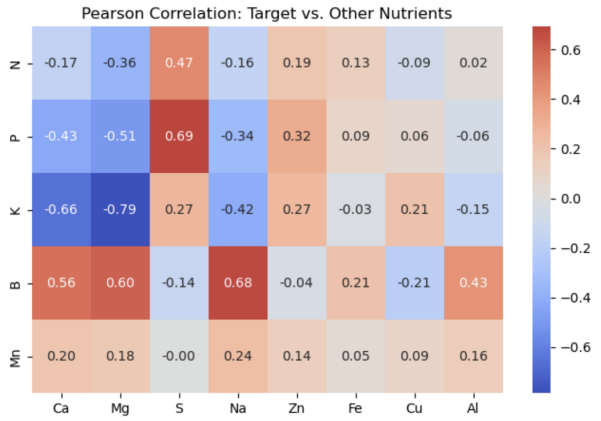

Fig. 2. Correlation Heatmap used to select top 2 micro nutrients for each target based on absolute Pearson Correlation

B. Target Nutrients' Concentration Prediction

A regression model is trained to predict the value of each chosen micro nutrient feature based solely on the reflectance spectra. In particular, the aggregated reflectance is used to fit a PLS regression model that predicts the micronutrient. To prevent information about the target nutrient from leaking, each micronutrient model is trained exclusively on the training-set reflectance and the associated lab values. These models are used to predict micronutrient values for any sample after they have been trained. The micronutrient predictions can essentially be stacked on top of the spectral features when these predictions are utilised as features in the final target nutrient models. Every micronutrient model is trained solely on training data and they serve as extra sensors based on the leaf spectra during prediction.

With the PLS scores and the two predicted micronutrient values in hand, we form a combined feature set to predict each target nutrient. We evaluate three regression algorithms (SVR, Random Forest, LightGBM) under FLAML's CASH framework, selecting both the best model and its hyperparameters by minimising cross-validation error.

We then build the combined feature matrix

$$Z = \left[T \;\middle|\; \hat{y}^{(m_1)} \;\middle|\; \hat{y}^{(m_2)}\right],$$

where $T \in \mathbb{R}^{n\times k}$ are the PLS latent components and $\hat{y}^{(m_j)} \in \mathbb{R}^n$ are the two predicted micronutrient values. The final target prediction is obtained by applying the optimised regression function f^*:

$$\hat{y}^{\text{target}} = f^*(Z),$$

which was selected and tuned via the FLAML CASH procedure to minimise cross-validation error on the training set.

C. Stacked Pipeline Model Building

To fully leverage both spectral signatures and inferred chemical signals, we construct a two-level stacking ensemble. In our stacking pipeline, we begin by creating specialised base learners each focused on a subset of the target nutrients. Rather than forcing a single model to predict all five nutrients at once, we wrap individual regressors with a custom class, 'MultiOutputSubsetRegressor', which neatly isolates each learner to its assigned targets. This wrapper inspects the target columns by name and ensures that during training and prediction the underlying estimator only sees the relevant portion of the nutrient matrix.

Once our base learners are defined, we assemble them into a 'StackingRegressor'. This meta-ensemble takes the predictions of each specialised model as new features and feeds them into a final estimator, which is a Ridge regressor. During fitting, the stacking regressor performs an internal cross-validation (we used five folds) so that each base learner generates out-of-fold predictions for the training set. These predictions form a "second tier" feature space from which the Ridge meta-learner learns the best linear combination of the base outputs. This layered approach allows each model to play to its strengths while the meta-learner resolves any residual errors by blending their forecasts.

To handle the fact that our final goal remains the simultaneous prediction of all five nutrients, we wrap the entire stacking assembly in a 'MultiOutputRegressor'. This scikit-learn wrapper simply applies the stacking pipeline independently to each target column but shares the same parameters and shared cross-validation scheme. The result is a single unified estimator object that accepts our original spectral features and returns a five-column matrix of predicted nutrient concentrations.

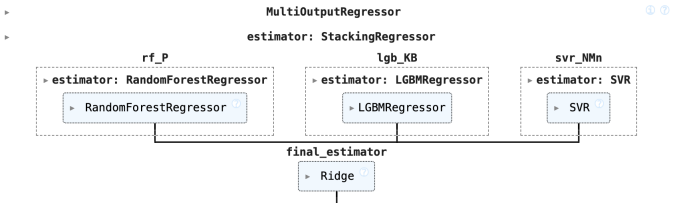

Fig. 3. Structure of the Multi Output Regressor for the dried mode

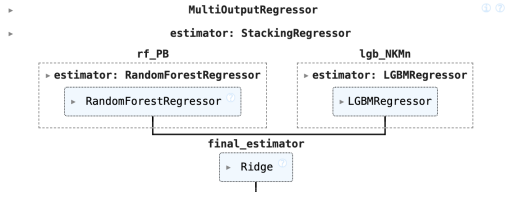

Fig. 4. Structure of the Multi Output Regressor for the fresh mode

By stratifying the learning, we gain several benefits. Each base model learns its "slice" of the problem without interference, producing more accurate sub-predictions than a single all-purpose regressor might. The meta-learner then integrates these specialised insights, smoothing out individual model biases and enhancing overall robustness. This design ensures the pipeline is transparent, modular and readily interpretable, while also leveraging the complementary strengths of random forests, gradient boosting and kernel-based methods in a coherent ensemble.

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

Fig. 5. Comparison of the performance metrics in dried mode

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

Fig. 6. Comparison of the performance metrics in fresh mode

D. Transfer Learning

To accommodate shifts in the relationship between spectral signatures and nutrient concentrations across different growing seasons, we employ a staged transfer-learning strategy. We begin by training the full stacking pipeline on Season 2 data alone. At each subsequent stage, we expand the training set to include one more season's samples, first Season 3, then Season 1 and finally Season 4, retraining the entire pipeline on the union of all data accumulated so far.

$$D_{\text{stage } t} \;=\; \bigcup_{s=2}^{\text{new}_t} D^{(s)}.$$

Here, $D_{\text{stage } t}$ is the combined training set formed by taking the union of all seasonal datasets from Season 2 up to the newly added season new_t.

$$(f^*, \theta^*) \;=\; \arg\min_{f,\theta} \sum_{(X,y)\in D_{\text{stage } t}} \mathcal{L}\big(f_\theta(Z), y\big).$$

At each stage, we jointly select the best model f^* and tune its hyperparameters θ^* by minimising the total loss over the aggregated training data.

During each update, samples from the newly added season are given a slightly higher weight to promote rapid adaptation, while the model continues to optimise performance on earlier

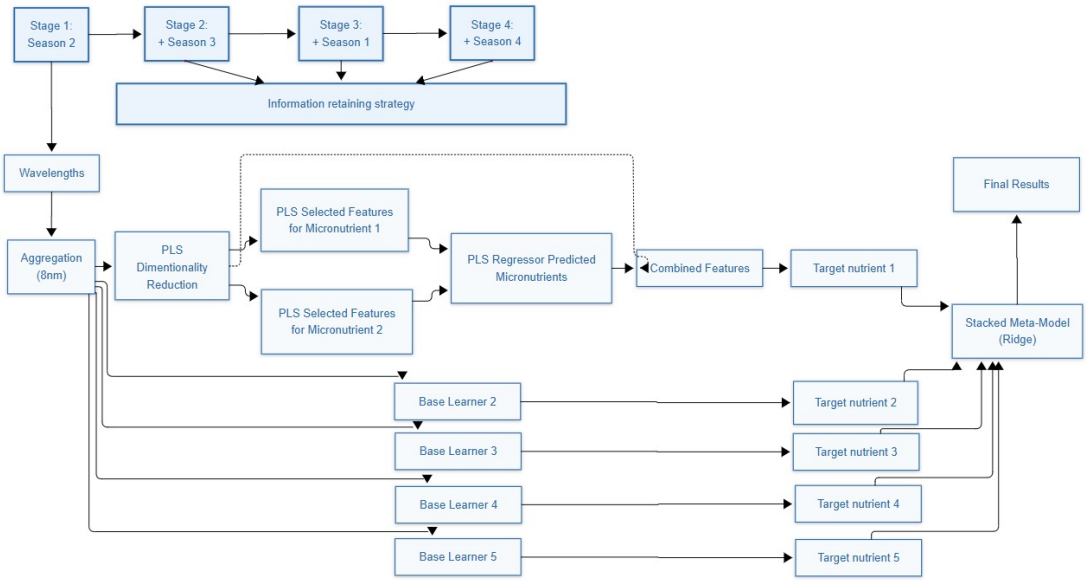

Fig. 7. Architecture of our stacking pipeline

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

Fig. 8. Evolution of test R^2 for the dried mode as Seasons 3, 1 and 4 are incrementally added to the training data.

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

Fig. 9. Evolution of test R^2 for the fresh mode as Seasons 3 and 1 are incrementally added to the training data.

seasons. This procedure produces a single unified model that gradually learns to generalise across all seasons. Figure 9 illustrates how the Nitrogen model's test R^2 on each season's test set evolves as Seasons 3 and 1 are added.

E. Model Evaluation

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

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

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

Final model performance is assessed on an independent test set for each nutrient using three complementary metrics. The coefficient of determination (R^2) measures the proportion of variance in the observed values that the model explains, with values closer to one indicating stronger predictive alignment. The normalised root-mean-square error (NRMSE) expresses the average prediction error as a percentage of the nutrient's observed range, allowing direct comparison of precision across nutrients with different scales. Finally, the ratio of performance to deviation (RPD) compares the natural variability of the data to the model's RMSE, indicating practical utility. RPD values above two suggest the model is suitable for quantitative prediction, while values between 1.4 and 2 are typically sufficient for screening purposes.

V. Results and Discussion

A. Performance Metrics Comparison

To identify the optimal predictive models for nutrient concentrations (N, P, K, B, Mn) across seasonal datasets and full-season spectral data, three performance metrics - R^2, NRMSE and RPD - were analysed from FLAML model selection (SVR, RF, LGBM), stacking and transfer learning. The goal was to select models with the highest R^2, lowest NRMSE and highest RPD for each nutrient, ensuring accurate and robust predictions.

From Figure 11 below for dried mode, when training jointly using the data of Season2, Season3 and Season1, the R^2 of each nutrient reached the highest, and the NRMSE and RPD were optimal, indicating that the data of these three seasons were highly complementary. However, after simply incorporating the data of Season4, the model's degree of fit (R^2) for all nutrients dropped sharply, the error (NRMSE) rebounded and the predictive ability (RPD) declined. Particularly, the performance of Mn was the poorest, indicating that there was significant heterogeneity in the data distribution between

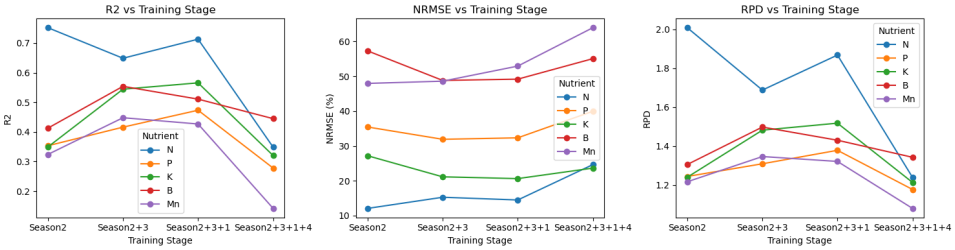

Fig. 10. Effect of transfer learning on dried mode

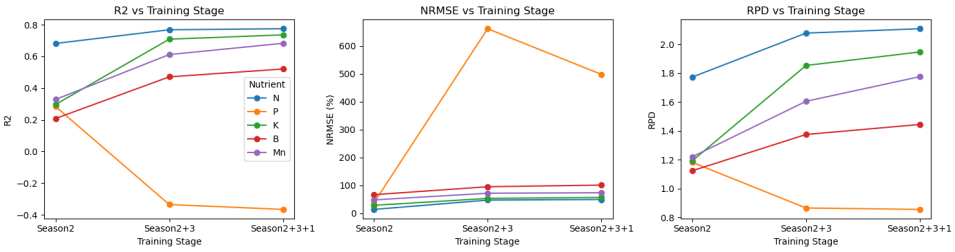

Fig. 11. Effect of transfer learning on fresh mode

Season4 and the previous seasons. Direct merging instead weakened the generalisation effect of the model.

From Figure 12 demonstrating the effect of transfer learning for fresh mode, with the change from using only Season2 to successively adding Season3 and then adding Season1, the R^2 of the other four nutrients except phosphorus (P) increased from approximately 0.3 to between 0.5 and 0.8. The RPD also increased from 1.1-1.2 to 1.4-2.1, indicating that transfer learning significantly enhanced the model's correlation and discriminative ability for nitrogen, potassium, boron and manganese. However, at the same time, NRMSE generally rose from tens of percent to over 50% and phosphorus not only had R^2 become negative, NRMSE soared to hundreds of percent, but RPD dropped to less than 0.9. This indicates that the effect of transfer training on phosphorus is extremely poor. Overall, although transfer learning can improve relative performance, it brings a significant increase in absolute error and targeted strategies are needed to improve it.

B. Interpretation of Key Patterns

The stacking ensemble approach, which integrates multiple base learners tailored to specific nutrients and employs a meta-learner to combine their predictions, demonstrates a robust ability to capture nutrient-specific patterns across a combined dataset. The performance metrics indicate that this method achieves varying levels of predictive accuracy across different nutrients, with some nutrients benefiting more from the ensemble's ability to leverage diverse model strengths. The stacking method's strength lies in its capacity to handle the heterogeneity of nutrient responses by assigning specialised models to subsets of targets, thereby optimising predictions for each nutrient. However, the results suggest that the approach may struggle with certain nutrients where the base learners' predictions are less complementary, leading to moderate performance improvements compared to individual models. This indicates that while stacking enhances generalisation across seasons, its effectiveness is contingent on the compatibility and diversity of the base learners.

In contrast, when we look at the transfer learning approach, which incrementally incorporates data from additional seasons to build a more comprehensive model, it reveals a dynamic pattern of performance evolution. The results show that predictive accuracy generally improves as more seasonal data is included, especially for the fresh mode, reflecting the model's ability to adapt to diverse conditions and capture broader spectral-nutrient relationships. However, for the dried mode, the performance tends to plateau or slightly decline in the final stage when all seasons are combined, suggesting potential overfitting or increased noise from integrating highly variable data. Transfer learning excels in scenarios where sequential data accumulation enhances model robustness, particularly for nutrients with consistent spectral signatures across seasons. The comparison highlights that transfer learning is more sensitive to data volume and diversity, offering advantages in adaptability but requiring careful management of data integration to avoid diminishing returns.

C. Technical Challenges

While evaluation aggregation methods, we learnt how the bin size significantly affects spectral model performance. Excessively large bins may lose critical spectral details leading to underfitting, while very small bins amplify noise causing overfitting. Systematic evaluation of bin sizes, using techniques like cross-validation, grid search or Bayesian optimisation, can achieve the optimal balance.

In seasonal transfer learning, balancing historical data retention with adaptation to new seasons is challenging. Overintegration of diverse seasonal datasets risks introducing negative transfer, reducing predictive accuracy. Adaptive weighting, hierarchical modeling or attention-based strategies are essential to selectively incorporate useful data while minimising adverse effects.

Also, proper weighting of predicted micronutrients in stacking models is essential to avoid overshadowing spectral data. Improper weights can weaken the predictive power of spectral features. Techniques like regularisation, adaptive weight assignment or meta-learners dynamically determine feature weights, while additional spectral indices or derivatives further enhance model accuracy.

VI. Further Work and Improvement

A. Broader Crop Application

Near-infrared spectral data predicts nutrient concentrations in potato leaves and can be applied to other crops like radish, onion or leafy greens. Despite differences in spectral signatures due to leaf morphology, pigment composition and structure, the physics of light absorption and reflectance remain consistent. Calibration experiments, such as wavelength adjustments and preprocessing, can address variations in surface texture and moisture content.

B. Application Optimisation

In environments with continuously updated crop data, system architecture must support rapid feedback. Decoupling the model's inference capability from its training component ensures seamless operation. Packaging inference as an API allows new data integration and model updates without disrupting predictions. This enables faster feedback cycles and minimises downtime by refreshing training parameters in the background while maintaining stable predictions.

VII. Conclusion

Our study's findings demonstrate how well a stacked ensemble pipeline predicts leaf nutrient concentrations from VIS/NIR reflectance data, producing reliable and accurate estimates for several nutrients at once. The pipeline greatly improved model performance by combining dimensionality reduction using 8nm spectral averaging and PLS with a two-step prediction method. Key micronutrients were first anticipated as intermediate features and the top two correlated nutrients were chosen by utilising nutrient correlations found in the heatmap (Fig. 2). In order to predict target nutrients, these predictions were combined with raw wavelength data and fed into a stacked ensemble. The final estimator in this ensemble was a ridge meta-learner, which effectively weighted contributions. The benefit of the transfer learning strategy was that a combined-training model was obtained for all seasons. By combining this architecture with progressive training over several seasons, generalisation and robustness to domain shifts were enhanced.

With R^2 values of 0.7 and RPD values of 1.7–2.0 on independent test data, the pipeline produced especially good results for nitrogen (N) and phosphorus (P), suggesting dependable screening capabilities and for N, approaching the quantitative prediction threshold (RPD $>$ 2, per calibrationmodel.com). Although less accurate, the predictions for potassium (K) and boron (B) still showed broad patterns (RPD 1.5–1.6), making them appropriate for preliminary screening. However, manganese (Mn) predictions performed worse (RPD 1.1), indicating that Mn might need nonlinear modelling, extra features or be unreliable from reflectance data alone in this dataset. The pipeline's practical utility is highlighted by the noteworthy fact that these results were obtained solely from spectral data, without the use of direct chemical measurements.

This method works well because it combines agronomic knowledge with statistical precision. While PLS addressed multicollinearity and dimensionality issues in hyperspectral data, feature selection based on nutrient correlations introduced domain knowledge in a data-driven way. Accuracy was increased by the stacked ensemble's combination of direct spectral features and inferred chemical features and robustness was guaranteed by multi-season training. The pipeline, which uses FLAML for model training and is implemented in Python in modular steps, is flexible and scalable for additional datasets. This process provides a repeatable framework for developing spectral-based agronomic applications by showcasing the effectiveness of model stacking and integrated prediction strategies for accurate, dependable nutrient screening.

GitHub Repository Link

You can access the GitHub repository at the following link: https://github.com/EMATM0050-2024/dsmp-2024-groupt14

References

[1] Abukmeil, R., Al-Mallahi, A. and Campelo, F., 2022. New approach to estimate macro and micronutrients in potato plants based on foliar spectral reflectance. *Computers and Electronics in Agriculture*, 198, 107074.

[2] https://arxiv.org/abs/2201.05340

[3] https://github.com/microsoft/FLAML

[4] Li, C., Wang, Y., Ma, C., Chen, W., Li, Y., Li, J., Ding, F. and Xiao, Z., 2021. Improvement of Wheat Grain Yield Prediction Model Performance Based on Stacking Technique. *Applied Sciences (Switzerland)*, 11(24), 12164.

[5] Prananto, J. A., Minasny, B. and Weaver, T., 2021. Rapid and cost-effective nutrient content analysis of cotton leaves using near-infrared spectroscopy (NIRS). *PeerJ*, 9, e11042.

[6] Rauschenberger, A. and Glaab, E., 2021. Predicting correlated outcomes from molecular data. *Bioinformatics*, 37(21), 3889–3895.

[7] https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/

[8] Wold, S., Sjöström, M. and Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), pp.109–130.