Group T12 Summative Assessment Report

Chieh-Yin Lee
School of Engineering Mathematics and Technology
University of Bristol
Bristol, United Kingdom
ha24847@bristol.ac.uk

Yexi Li
School of Engineering Mathematics and Technology
University of Bristol
Bristol, United Kingdom
mq24228@bristol.ac.uk

Shu-Shan Chiang
School of Engineering Mathematics and Technology
University of Bristol
Bristol, United Kingdom
pp24621@bristol.ac.uk

Yichi Liu
School of Engineering Mathematics and Technology
University of Bristol
Bristol, United Kingdom
rq24150@bristol.ac.uk

Abstract - Near-infrared spectroscopy (NIRS) offers a fast and non-destructive way to assess crop nutrients, providing an alternative to traditional methods. This project used spectral reflectance data from potato plants to build machine learning models predicting multiple nutrient concentrations. After hybrid missing value handling, spectral resampling, and PLSR dimensionality reduction, a stacking model combining PLSR and Ridge regression was developed. The model showed good performance for key nutrients like nitrogen, phosphorus, and potassium. However, cross-season predictions were still limited by environmental variability and concept drift. Future work will explore cross-season adaptation and more physically meaningful data augmentation to improve generalisation.

I. Introduction

Optimising nutrient levels in crops is crucial for improving agricultural productivity and promoting sustainability. Traditionally, nutrient testing has mainly relied on soil and plant tissue analyses. Although tissue analysis offers higher accuracy, the processes of sample collection and analysis are time-consuming and labour-intensive, limiting their application in large-scale and frequent monitoring. Therefore, developing fast, efficient, and cost-effective testing techniques has become a pressing need.

In recent years, non-destructive techniques such as near-infrared spectroscopy (NIRS) have emerged as promising alternatives. However, the high dimensionality of spectral data, the interactions and co-dependencies among different nutrients, and the variations between fresh and dried samples pose significant challenges to building accurate predictive models.

In this project, we explore spectral reflectance data collected from potato plants to develop predictive machine learning models capable of estimating multiple nutrient concentrations. Our goal is to build a robust and efficient nutrient prediction system, contributing to smarter and more sustainable agricultural practices.

II. Literature Review

Nowadays, hyperspectral and near-infrared (NIR) spectroscopy have been adopted in nutrition study of agricultural industries as a non-destructive way of detecting and predicting the level of nutrients [1]. Recent studies have further explored application in nutrient assessment in potato plants. Abukmeil et al.[2] developed a novel approach to estimate macro- and micronutrients in potato plants by analysing foliar spectral reflectance. By applying Lasso regression, the model effectively identified relevant wavelengths for different nutrition and their correlations. The research demonstrated notable predictive accuracy (by the indication of RPD >2) for most nutrition. Furthermore, this study highlighted significant wavebands in the visible and near-infrared ranges (400–1100 nm) for macronutrients, and across the full range (400–2500 nm) for micronutrients, supporting the feasibility of using spectral data to monitor nutrient levels in potato cultivation.

While many spectral studies focus on predicting individual traits or nutrients, some advanced studies explore the need to model multiple correlated targets simultaneously, particularly when working with high-dimensional data [3]. In this context, Rauschenberger and Glaab introduced multivariate regression models in high-dimensional data. Their work highlights the advantages of applying stacked generalisation to multivariate Lasso and Ridge regression, enabling predictions of multiple correlated targets. The studies demonstrated that this method better enhance predictive performance and interpretability in hyperspectral settings.

Multivariate regression model sometimes can include too many features, which could lead to poor efficiency and overfitting. To tackle this problem, we need to perform dimensionality reduction and feature selection. Based on the study of Jafarbiglu, 2022 [1], projection-based methods, such as Principal Component Analysis (PCA), are often applied to reduce data dimensionality by transforming features into a smaller set of latent components. Also, the study of Burnett, 2022 suggests that partial least square (PLS) is widely used alternative in plant phenotyping studies involving high-dimensional spectral dataset [4]. The model enables the dimensionality of plant dataset to be reduced and enables the trait to be estimated from hyperspectral optical reflectance data. This technique is of interest and importance in a wide range of contexts including crop breeding and ecosystem monitoring. However, although the transformed features in the reduced dimensional space maintain the connection to the original features, much of the original feature interpretation and their relationships are often lost, as noted in [5].

III. Methodology

A. Data Cleaning and Preprocessing

To ensure data quality and comparability across datasets, a series of preprocessing steps were implemented prior to model development. The datasets, collected over multiple growing seasons from both dried and fresh leaf samples of potato plants, were handled separately due to the known differences in their spectral characteristics.

Firstly, variable naming inconsistencies across files were resolved by standardising the nutrient concentration column names. Datasets were partitioned by sampling mode, with four seasons of data available for fresh samples and three for dried samples. Each season was modelled independently to mitigate the impact of concept drift across growing periods.

In this project, missing values were addressed through a hybrid imputation strategy. Nutrients with less than 10% missing data were imputed using the median of the respective variable. For variables with a higher proportion of missing data, as well as all spectral reflectance values, K-Nearest Neighbours (KNN) imputation was employed using 5 neighbours [6], [7]. Prior to KNN imputation, all numerical features were standardised using z-score normalisation to ensure balanced distance calculations.

Outlier detection was conducted using the interquartile range (IQR) method, applied uniformly across all nutrient variables and spectral features [8].

In terms of spectral resolution, the original hyperspectral data—spanning from 400nm to 2500nm at a granularity of 0.5nm— was resampled to 1nm intervals by averaging adjacent points. This adjustment was made primarily to harmonise the data across all seasons, as one of the datasets was originally collected at a 1nm resolution. As a secondary benefit, this process also contributed to a moderate reduction in feature dimensionality while preserving the major signal trends [9].

Lastly, to avoid data leakage during model evaluation, all preprocessing steps, including imputation, transformation, and binning, were fitted exclusively on the training subset before being applied to validation or test data.

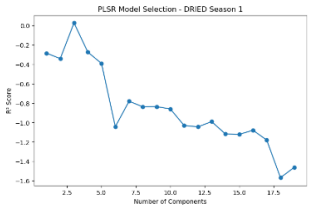
B. *Feature Aggregation and Dimensionality Reduction*

Due to the high-dimensional nature of spectral data, dimensionality reduction is crucial to reduce the risk of overfitting and improve the generalization ability of the model. The method we chose in this experiment is partial least squares regression (PLSR).

Compared with unsupervised methods, PLSR, as a supervised dimensionality reduction technique, combines the information of the predictor variables and the response matrix, so it is more effective when facing multi-target regression problems with highly correlated outputs. After discussion, it is currently found to be very effective in modeling nutrient concentrations using spectral data, among which there are many applications in leaf reflectance analysis in agricultural applications [2], [10].

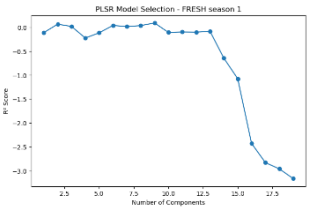
At the same time, we performed 10-fold cross-validation on the training data to optimize the optimal number of potential components in the PLSR model. For each candidate number of components, we trained and evaluated the PLSR model by the average R^2 score of the cross-validation. Considering the problem of overfitting, we finally selected the number of components that gave the highest average R^2 score and avoided selecting components that exceeded the actual number of available features.

PLSR was evaluated to assess its effectiveness in representing the original data with reduced complexity. Importantly, all transformations were fitted solely on the training data and then applied to the corresponding validation and test sets to avoid any risk of information leakage.



Best n_component = 1

Figure 1: PLSR Model Selection – FRESH Season 1



Best n_component = 2

Figure 2: PLSR Model Selection – DRIED Season 1

C. *Model Selection and Training*

To evaluate the performance of different regression models in multi-target nutrient prediction, we explored several approaches, including traditional linear models (e.g., Ridge, Lasso), kernel-based models (e.g., SVR), and ensemble techniques such as Random Forest and Gradient Boosting Regressor (GBR) [11]. To promote generalisability, all models were trained on the full set of nutrient targets simultaneously using multi-output regression frameworks (e.g., MultiOutputRegressor, MultiTaskLassoCV), rather than building separate models for each nutrient.

To further compare model performance, we evaluated both dried and fresh samples based on the average RPD values for key nutrients (N, P, K), as shown in Figure 3. The hyperparameters for each model were set based on empirical rules and are summarised in Table 1.

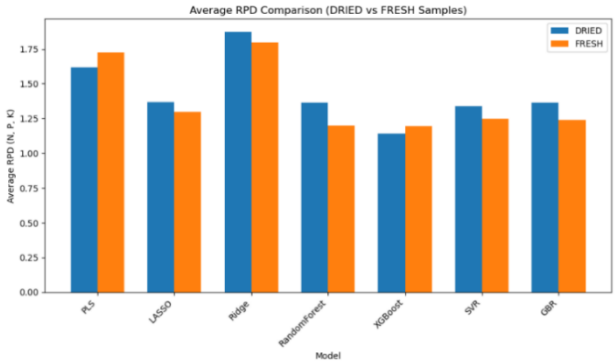


Figure 3: Average RPD Comparison

Table 1: Model Hyperparameter Settings

Model	Core Hyperparameter Settings
PLS	n_components=min(10, features)
MTLasso	alpha=0.1, max_iter=10000
Ridge	alpha = 1.0
RF	n_estimators = 100
XGB	n_estimators = 100
SVR	Default parameters
GB	n_estimators = 100

Among all tested models, PLS and Ridge regression showed higher average RPD scores in both dried and fresh datasets, indicating better predictive capability compared to other models. Therefore, in the final model, we applied a stacking ensemble strategy, selecting PLS and Ridge as base learners in order to combine their complementary strengths and further improve the overall predictive performance.

The results show that among all the tested models, Ridge regression and PLS regression consistently achieved higher

and more stable average RPD values on both dried and fresh samples. Overall, the Ridge model demonstrated the best performance, followed by PLS regression. These findings suggest that, compared to other methods such as Random Forest, XGBoost, SVR, and GBR, Ridge regression and PLS regression are more effective at capturing the complex relationships between spectral features and nutrient concentrations. Based on this preliminary screening, Ridge regression and PLS regression were selected as the primary candidate models for further in-depth study and optimization.

In addition to base learners, a stacking ensemble strategy was implemented to leverage the strengths of different models. Weighting adjustments were explored to improve the performance of stacked predictions [12], [13].

Feature sets used in training included both the original spectra and PLS-reduced spectral data, as described in Section 3.2. Comparisons were made between models trained on each of these feature spaces, as well as on the untransformed spectra, to assess the impact of dimensionality reduction techniques on overall performance.

To better reflect the importance and variability of different nutrients, we adopted a weighted loss strategy during model training. We assigned higher weights (3.0) to nitrogen (N), phosphorus (P), and potassium (K), as their concentrations are more critical for plant development. Boron (B) and manganese (Mn) were given moderate weights (2.0), while the remaining nutrients were treated equally with a baseline weight of 1.0. This weighting approach allowed us to focus model accuracy on the most important elements without sacrificing overall balance.

The data were split into training and test sets in an 80:20 ratio, using a fixed random seed to ensure reproducibility. All model selection and hyperparameter tuning steps were confined to the training set using 5-fold cross-validation.

We conducted hyperparameter tuning by combining grid search and manual adjustment. For the Ridge regression model, the regularisation strength (alpha) was optimised. For the PLS-Ridge stacking model, three sets of hyperparameters were tuned: the number of PLS components (n_components), the regularisation strength of the first-level Ridge model (alpha), and the regularisation strength of the final estimator (alpha). We used a custom weighted RMSE metric as the scoring function, giving higher importance to key elements such as Nitrogen (N), Phosphorus (P), and Potassium (K) [15].

D. *Evaluation Strategy*

As mentioned before, to ensure the consistency of the results, our data division uses 80% training set and 20% test set. Since all the previous steps are only performed on the training set and 5-fold cross-validation is used, in the evaluation stage, we can directly use the previously reserved test set for the final performance evaluation.

Model selection was guided primarily by the Ratio of Performance to Deviation (RPD), a robust metric for regression quality in spectroscopy tasks. R^2 and RMSE were also reported to provide complementary insights into prediction accuracy and error magnitude [14].

As we all know, RPD is the most commonly used and recognized standard in spectral modeling. When its value is higher than 2.0, it can be considered that the model has good predictive performance and high accuracy [2].

RPD is authoritative, but R^2 and RMSE are equally important. We can use these two indicators to supplement the model accuracy and display the residual distribution. We also conducted an overall evaluation of the average values of each nutrient and a single nutrient. Through data comparison, we can have a more comprehensive understanding of the advantages and disadvantages of the model.

All evaluations were carried out using a single random seed to ensure reproducibility.

E. *Implementation Details*

All data preprocessing, modelling, and evaluation tasks were conducted using Python. The primary libraries employed in this study included NumPy, Pandas, Scikit-learn, Matplotlib, and Seaborn.

Development and experimentation were carried out within Jupyter Notebook environments, accessed both through standalone Jupyter interfaces and through integrated support within Visual Studio Code (VSCode). A consistent random seed, which is set to 42 was applied across data splitting, model training, and evaluation to ensure reproducibility of results.

IV. Data Description & Preparation

A. *Dataset Overview and Exploratory Analysis*

During the preliminary analysis of the dried samples, the distribution of sample quantities across different seasons was summarised, as shown in the figure. Overall, there were considerable differences in the number of samples between seasons. Season 2 had the largest number of samples, approximately 145, which was significantly higher than the other seasons. The sample counts for Season 3 and Season 4 were relatively similar, with around 98 and 105 samples respectively. The first season has the least number of samples, only about 40. This sample imbalance may affect subsequent modeling. More samples can help the model better learn stable and generalizable features, while seasons with fewer samples are prone to overfitting or unstable model performance. Therefore, when building a model, the difference in sample size between seasons needs to be paid special attention, and the validation strategy should be paid special attention to minimize the impact of sample imbalance.

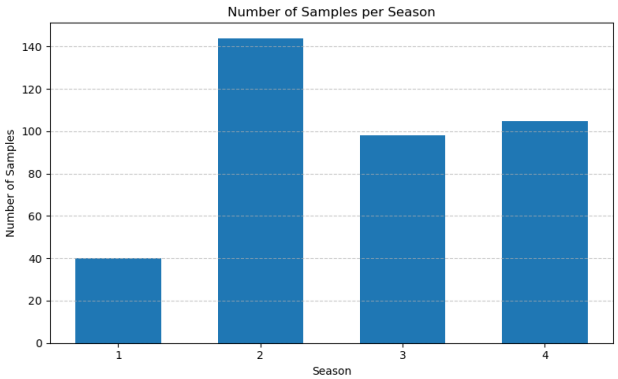


Figure 4: Sample Distribution Across Seasons

From the figure, we can see the distribution of missing rates of elements in dry and fresh samples in different seasons. In general, the missing rates of most elements are low and the

data are relatively complete. However, in dry samples, the missing rates of elements such as chloride, boron and aluminum are more significant in the first season, with the highest being close to 15%. Fresh samples also show a similar trend. The missing rates of a few elements are slightly higher in the first and fourth seasons, which may be related to seasonal changes or different sample processing methods. Given that the missing rates of most elements are less than 10%, we used a combination of median interpolation and K nearest neighbor (KNN) interpolation in data processing to minimize the impact of missing data on model performance.

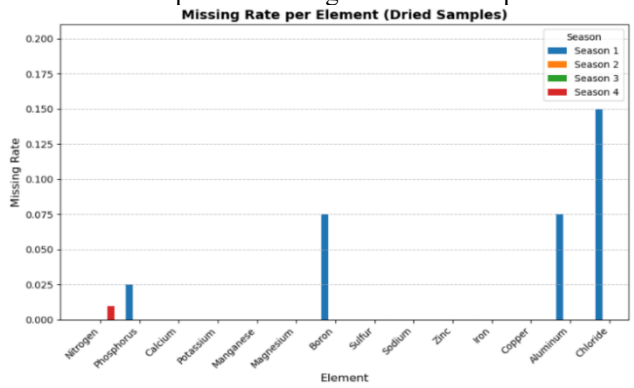


Figure 5: Element-wise Missing Data Rates by Dried Season

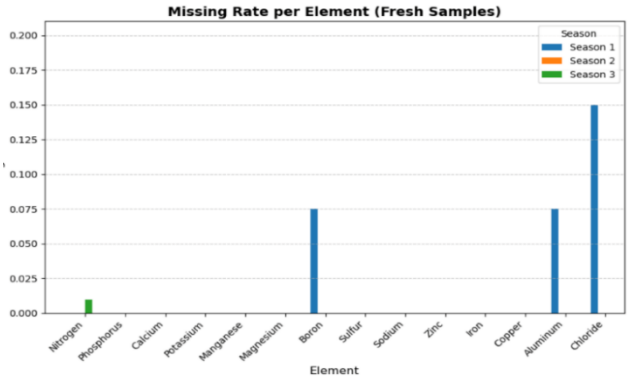


Figure 6: Element-wise Missing Data Rates by Fresh Season

To investigate the relationships between elements across different seasons, a heatmap of element concentration correlations was generated. The results indicate that although the overall pattern of correlations remained relatively consistent across seasons, with major elements such as nitrogen, phosphorus, and potassium generally showing strong positive correlations, certain seasonal differences were observed. In particular, in Season 4, the correlations between elements weakened significantly, and correlation coefficients tended to approach zero. This suggests a notable reduction in the synergistic behavior among elements. Such changes reflect the influence of environmental conditions on nutrient dynamics within plants and present challenges for subsequent modeling efforts, leading to a decline in predictive performance during Season 4.

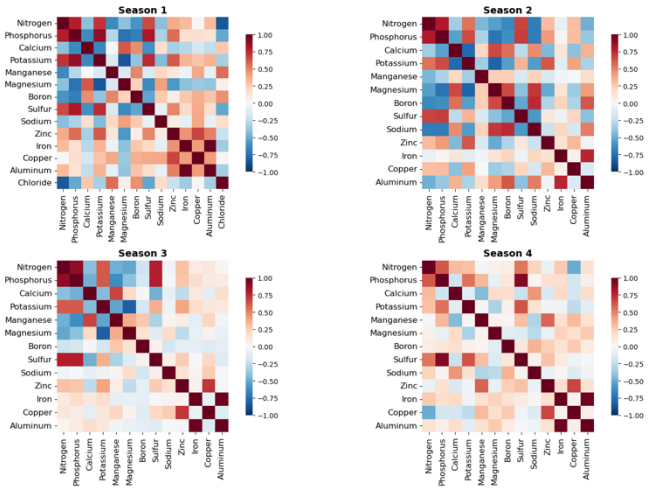


Figure 7: Element Correlation Heatmaps Across Seasons

The violin plots illustrate the distribution of element concentrations in dried and fresh samples. Overall, the distribution trends of most elements were largely consistent between the two sample types. However, elements such as Iron, Manganese, Boron, and Aluminum exhibited a noticeable positive skew, with a few extremely high values observed. In contrast, elements such as Nitrogen, Potassium, and Sulfur showed more concentrated distributions with smaller variation ranges. Given the high similarity in distribution patterns between fresh and dried samples, Winsorization was applied to elements with severe skewness during the data preprocessing stage in order to enhance model robustness and improve predictive stability.

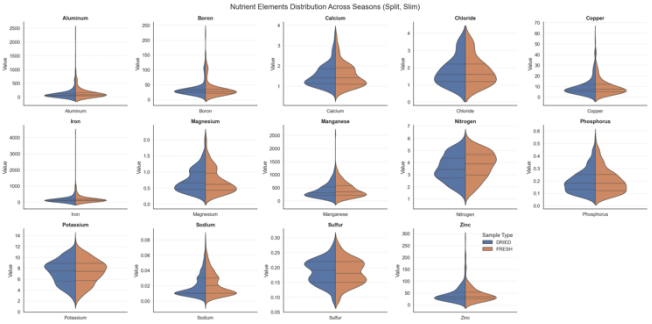


Figure 8: Element Concentration Distributions in Dried and Fresh Samples

B. Data Preparation

The nutrient concentration column names in the NIRS data from multi-season potato plants were standardised, and rows with all missing nutrient values were removed. Missing values were handled based on their proportion: nutrients with less than 10% missing data were imputed using the median, while those with a higher missing rate were filled using KNN imputation. Outliers were identified using the IQR method and adjusted with log transformation to minimise their impact on the model.

C. Methods

Spectral data underwent z-score normalisation to ensure consistency across wavelengths. A PLS-LASSO model was trained using 5-fold cross-validation and GridSearchCV to optimise hyperparameters, improving prediction accuracy and enhancing seasonal generalisation.

V. Results and Discussion

A. Exploratory Data Analysis

First, our analysis reveals that nutrient correlations vary across seasons. Strong correlations, particularly among Nitrogen, Phosphorus, and Potassium, are observed in Seasons 1 and 2 but weaken in Seasons 3 and 4. This trend may be influenced by environmental factors, though further research is needed to confirm the underlying causes.
Second, our study using PLS regression reveals that leaf reflectance significantly impacts nutrient concentrations, particularly within the 400-800 nm range in Seasons 1 and 2. However, in Season 3, the key wavelengths shift to different ranges. These findings suggest that specific wavelengths may serve as important features for nutrient identification.

B. Model Training and Validation Results

Based on the experiments we conducted in the previous stage, we selected the PLS-Ridge stacking pipeline as our final predictive model. The choice closely matched the characteristics of the provided data and the task of this project. The dimension of the feature is extremely high while the sample size is relatively small. Additionally, strong multicollinearity existed among adjacent wavelengths, and the output is expected to involve multiple nutrient concentrations to be predicted simultaneously.

Given these conditions, Partial Least Square (PLS) Regression was employed to perform dimensionality reduction and extract the most informative latent components from the features in original data. When predicting nutrient concentrations, it was equally important to ensure model stability and avoid overfitting, especially given the limited sample sizes. Ridge Regression played a crucial role by applying L2 regularisation, thus stabilising the model estimates and enhancing generalisation. Therefore, combining PLS and Ridge through a stacking approach allowed us to balance dimensionality reduction and regularisation, both essential for building robust predictive models under the large features, small samples condition.

Across different seasons, the overall performance of the PLS-Ridge stacking model demonstrated some variability. In general, the DRIED_season2 dataset showed the strongest results, achieving a decent RMSE of 27.40 and excellent predictive ability for key nutrients, as reflected by RPD values for Nitrogen (4.76), Phosphorus (2.53), and Potassium (2.63). Conversely, performance was noticeably weaker in DRIED_season4 and FRESH_season3, where weighted RMSE remained relatively high and the RPD values for Nitrogen, Phosphorus, Potassium fell below 1.5, indicating limited predicting capability. These results show that certain seasons posed more challenges, due to different data collection methods, higher sample variability or the phenomenon of concept drifting across seasons.

Notably, the RPD of calcium in the DRIED_season1 dataset is exceptionally high, reaching 7.94 — significantly higher than that of any other element in our experiments. We assume that this may be influenced by the limited amount of data (N=40) and the train-test split ratio, which resulted in only 8 samples in the testing set. Additionally, the distribution of calcium concentrations in DRIED_season1 is relatively even, and the variance is moderate, which may have reduced the difficulty of training. Therefore, despite the encouraging results, we should remain cautious when interpreting the generalizability of this model.

In terms of element-specific performance, the model generally performed best for predicting macronutrients such as Nitrogen, Phosphorus, Potassium. These nutrients achieved relatively high RPD values (greater than 2) across different seasons and in both DRIED and FRESH data. In contrast, the prediction of certain micronutrients, especially for Boron, Zinc and Iron, still remained challenging, with RPD often lower than 1.5 across different datasets, reflecting the limitation of this combined model.

The model outputs demonstrated strong predictive value for macronutrient levels, which is a good indication for the growth of potatoes. According to Davenport et al. (2019) , the availability and balance of macronutrients such as Nitrogen, Phosphorus, and Potassium are critical determinants of potato yield and tuber quality. Accurate estimation of these nutrient levels enables efficient and precise nutrient management practices, which are essential for optimising plant growth, maximising crop productivity, and mitigating environmental destruction due to over-fertilisation. Therefore, the model's strong predictive performance for these macronutrients has significant advantages for sustainable potato production.

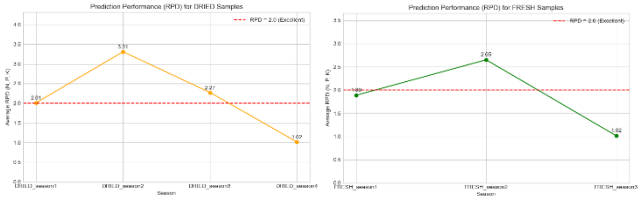


Figure 9: RPD by Season
(Left: Dried Samples; Right: Fresh Samples)

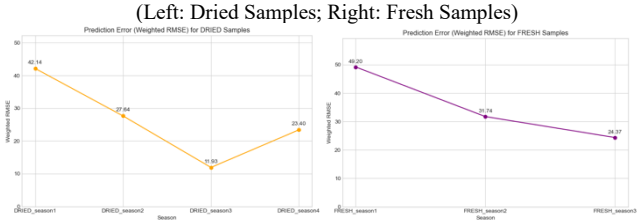


Figure 10: Weighted RMSE by Season
(Left: Dried Samples; Right: Fresh Samples)

C. Overall Model Performance Comparison

To further evaluate the effectiveness of different regression strategies, the overall predictive performance of the four modeling approaches explored in this study was compared: PLS regression, Ridge regression, a simple average of PLS and Ridge (PLS with Ridge Avg), and a stacked ensemble of PLS and Ridge (PLS with Ridge Stack). The average performance metric (mean RPD value) across all elements and all seasonal datasets were calculated for each model, with the results summarised in Table 2.

Table 1: Model Comparison by Average RPD

Model	Average RPD (All Elements)
PLS	1.319
PLS with Ridge Avg	1.476
Ridge	1.518
PLS with Ridge Stack	1.610

The results in Table X clearly show that the PLS with Ridge stacked model achieved the highest overall mean RPD

(1.610), outperforming all other methods. Although Ridge regression alone already performed better than PLS regression, stacking the two models further enhanced generalisation ability.

This trend suggests that stacked ensembles can effectively leverage the complementary strengths of different base models, leading to more robust and accurate predictions across different seasons and sample conditions. As a result, the PLS with Ridge stacked model not only excelled in seasonal tests but also demonstrated superior overall performance across the full dataset.

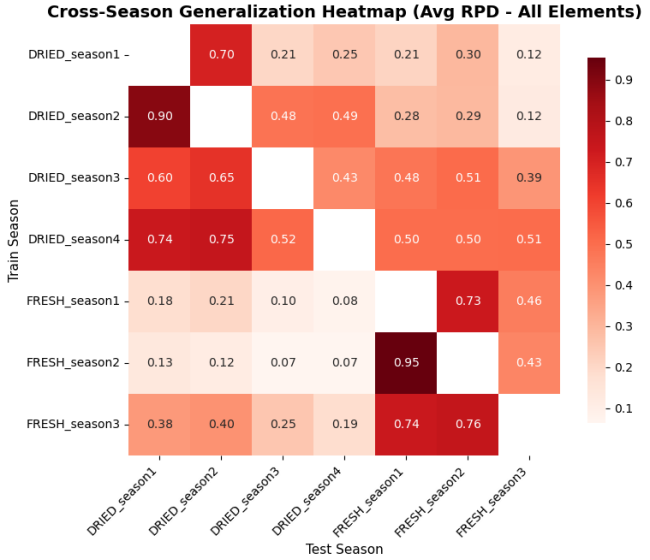


Figure 11 : Cross-Season Generalisation Heatmap

To assess the model's generalisation ability across different seasons, cross-seasonal testing was conducted. Specifically, models trained on one season's data were directly applied to the test sets of other seasons, and the average RPD across all elements was calculated. As shown in the heatmap, most cross-seasonal predictions yielded relatively low average RPD values, typically ranging between 0.2 and 0.8. Only a few cases involving the same sample type (e.g., DRIED_season2 to DRIED_season1) achieved RPD values close to 0.9. Overall, model performance dropped noticeably when transferred across seasons.

These results indicate that the current models have limited generalisation ability between seasons. In particular, a significant performance decline was observed when transferring between dried and fresh samples, further confirming the substantial differences in the relationship between spectral features and elemental concentrations across different sample types. Moreover, even within the same sample type, seasonal variations in environmental factors such as temperature, moisture, and light conditions may cause concept drift, altering the accumulation patterns of elements within the plants. This drift makes it difficult for models trained on one season's data to accurately adapt to another.

Several factors likely contribute to the observed limitations in cross-seasonal transfer performance. First, the relatively small sample size for each individual season restricts the models' ability to learn sufficiently generalised features. Second, changes in environmental conditions between seasons modify the underlying relationships among elements, further increasing modeling complexity. Therefore, based on the experimental results, achieving high prediction accuracy still requires training and optimising models separately for each season, rather than relying on a single model to generalise across all seasonal datasets.

VI. Further Work and Improvement

Future work will focus on strengthening model robustness and generalisation by integrating cross-seasonal joint training and transfer-learning strategies to counter the effects of seasonal drift. We will conduct systematic comparisons of various regression algorithms and assemble them into ensemble pipelines that capitalise on each method's strengths. Early experiments with data augmentation (e.g. Mixup) revealed that unrealistic spectral distortions can harm performance, so going forward we will adopt physics-informed techniques—such as adding controlled Gaussian noise—to more faithfully mimic real-world NIRS variability and bolster model resilience.

At the same time, the efficiency of Random Forest can be improved by hyperparameter tuning. Training time increases as the number of trees grows, and grid search further amplifies this by exploring all hyperparameter combinations. Our initial attempts were halted due to long processing times. While random search is quicker, it may miss the best hyperparameters. With more computing resources in the future, optimal results may be achievable.

VII. Conclusions

The PLS–Ridge stacking ensemble demonstrated strong capability for the estimation of key macronutrients, achieving an overall mean RPD of 1.61 across multiple seasons and sample types. Macronutrient predictions (N, P, K) consistently exceeded the RPD>2 threshold. Seasonal drift and sample-type differences are limited cross-season generalisation and micronutrient accuracy. Addressing these challenges will require transfer-learning approaches, and physics-informed data-augmentation methods to reinforce model stability. With these enhancements, our NIRS-based workflow has clear potential to deliver rapid, cost-effective nutrient monitoring for precision agriculture.

References

[1] H. Jafarbiglu and A. Pourreza, 'A comprehensive review of remote sensing platforms, sensors, and applications in nut crops', Comput. Electron. Agric., vol. 197, p. 106844, Jun. 2022, doi: 10.1016/j.compag.2022.106844.

[2] R. Abukmeil, A. A. Al-Mallahi, and F. Campelo, 'New approach to estimate macro and micronutrients in potato plants based on foliar spectral reflectance', Comput. Electron. Agric., vol. 198, p. 107074, Jul. 2022, doi: 10.1016/j.compag.2022.107074.

[3] A. Rauschenberger and E. Glaab, 'Predicting correlated outcomes from molecular data', Bioinformatics, vol. 37, no. 21, pp. 3889–3895, Nov. 2021, doi: 10.1093/bioinformatics/btab576.

[4] A. C. Burnett et al., 'A best-practice guide to predicting plant traits from leaf-level hyperspectral data using partial least squares regression', J. Exp. Bot., vol. 72, no. 18, pp. 6175–6189, Sep. 2021, doi: 10.1093/jxb/erab295.

[5] A. Senawi, H.-L. Wei, and S. A. Billings, 'A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection

and ranking', Pattern Recognit., vol. 67, pp. 47–61, Jul. 2017, doi: 10.1016/j.patcog.2017.01.026.

[6] O. Troyanskaya et al., 'Missing value estimation methods for DNA microarrays', Bioinformatics, vol. 17, no. 6, pp. 520–525, Jun. 2001, doi: 10.1093/bioinformatics/17.6.520.

[7] F. Pedregosa et al., 'Scikit-learn: Machine Learning in Python', J Mach Learn Res, vol. 12, no. null, pp. 2825–2830, Nov. 2011.

[8] J. W. Tukey, Exploratory Data Analysis. Reading, MA: Addison-Wesley, 1977.

[9] Y.-P. Lin, W.-C. Lin, M.-Y. Li, Y.-Y. Chen, L.-C. Chiang, and Y.-C. Wang, 'Identification of spatial distributions and uncertainties of multiple heavy metal concentrations by using spatial conditioned Latin Hypercube sampling', Geoderma, vol. 230–231, pp. 9–21, Oct. 2014, doi: 10.1016/j.geoderma.2014.03.015.

[10] R. Abukmeil, A. A. Al-Mallahi, and F. Campelo, 'Multivariate stacked regression pipeline to estimate correlated macro and micronutrients in potato plants using visible and near-infrared reflectance spectra'.

[11] W. Waegeman, K. Dembczyński, and E. Hüllermeier, 'Multi-target prediction: a unifying view on problems and methods', Data Min. Knowl. Discov., vol. 33, no. 2, pp. 293–324, Mar. 2019, doi: 10.1007/s10618-018-0595-5.

[12] D. H. Wolpert, 'Stacked generalization', Neural Netw., vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

[13] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, 0 ed. Chapman and Hall/CRC, 2012. doi: 10.1201/b12207.

[14] H. Liao, J. Wu, W. Chen, W. Guo, and C. Shi, 'Rapid diagnosis of nutrient elements in fingered citron leaf using near infrared reflectance spectroscopy', J. Plant Nutr., vol. 35, no. 11, pp. 1725–1734, Aug. 2012, doi: 10.1080/01904167.2012.698352.

[15] J. Bergstra and Y. Bengio, 'Random search for hyper-parameter optimization', J Mach Learn Res, vol. 13, no. null, pp. 281–305, Feb. 2012.

APPENDIX

For the specific code part, please refer to our code repository: https://github.com/EMATM0050-2024/dsmp-2024-groupt12