Multiple nutrient sensing for smart agriculture

John White
School of Engineering Mathematics and Technology
University of Bristol
Bristol, England
aa22951@bristol.ac.uk

Thomas Wilson
School of Engineering Mathematics and Technology
University of Bristol
Bristol, England
fj23813@bristol.ac.uk

Raj Dhakan
School of Engineering Mathematics and Technology
University of Bristol
Bristol, England
xn23972@bristol.ac.uk

Ross Duncan
School of Engineering Mathematics and Technology
University of Bristol
Bristol, England
lw23535@bristol.ac.uk

***Abstract*—NIR Spectroscopy (NIRS) is an established method used for predicting properties of plant matter. In potato agriculture, spectroscopy provides the potential to predict nutrient levels in plants efficiently, leading to quicker and more accurate treatment responses. Supervised multivariate regression models can be used in this setting to predict multiple nutrients simultaneously based on samples with NIRS reflectance and chemical measurement data. Such datasets normally have high dimensionality and small sample size. This study finds that preprocessing NIRS readings through baseline correction and extracting peaks for features can improve the performance of Multi Linear Regression (MLR) and Random Forest models (and depending on whether the samples are dried or fresh), while having a nominal to detrimental affect on PLS Regression. Multi Task Lasso and Elastic Net were also compared along with these to a MLR baseline, but optimised with minimal regularisation penalty using the preprocessed data. Overall, Random Forest performed the best based on R^2 and RMSE metrics for both dried (R^2=0.423, RMSE=0.0836) and fresh (R^2=0.025, RMSE=0.1027) samples.**

I. Introduction

Near-Infrared spectroscopy (NIRS) is a technology used in 'precision agriculture' [1], classified as a remote sensing technique [2], which can produce data to help maintain optimal nutrient levels in potato plants [3]. Using the reflectivity readings of leaf samples produced by NIRS (as features) together with measurements of nutrients (as targets), supervised machine learning (ML) techniques can be used to predict plant nutrient levels faster, more efficiently and cost-effectively when compared to traditional laboratory tests. NIRS produces high-dimensional reflectivity readings characterised by $p \gg n$ (with p features and n samples), and there are multiple nutrients of interest in the health/yield of potato plants. The purpose of this study is to extend the research of Abukmeil et al. [3], which focused on the effectiveness of using floral spectroscopy to predict plant nutrient concentrations using a Lasso multi-linear regression (MLR) approach, to:

- Analyse the effect of data preprocessing methods on spectral data and on predictive performance.
- Establish the impact of the elected data processing strategy and its suitability against multivariate models commonly used in similar settings.

Initially, the data is parsed through a preprocessing pipeline featuring baseline correction of a dataset and extracting the features of interest, which is compiled over multiple seasons to maximise generalisation. A univariate MLR baseline is used to assess the performance of basic data reduction techniques such as binning and Principle Component Analysis (PCA), along with more tailored techniques of baseline correction coupled with peak feature extraction. The resulting pipeline was compared to the performance of a multivariate application of a Partial Least Squares Regression (PLS or PLSR) model, Multi Task Lasso (MTL), Multi Task Elastic Net (MTEN) and Random Forest (RF) models. The effectiveness of the pipeline for fresh and dried sample modes is examined separately, as they present different reflective signals.

II. Literature Review

A. Problem Domain

Precision agriculture is an evolution in agricultural practices that uses technology to improve productivity, efficiency [4], and sustainability [5]. The growth in diverse and accessible sensing technologies provides the opportunity to use data to help predict conditions and respond to them with more automated and precise feedback into the agricultural process [6]. Among these sensing technologies is the measurement of plant sample reflectivity over different light wavelengths for various applications, including measuring contaminants, assessing plant quality and nutrient levels [7]. Micro and macro nutrient balance in potato plants directly and indirectly impacts tuber quality and yield [8]. Naumann et al. [9] address the impacts of macronutrients and the complex nature of the relationships between them on quality and yield, and Koch et al. [10] illustrate that the importance of balance, the 'law of the minimum' and the nature of a saturation curve, where no further benefit to plant development can be

realised, or impairment to quality can occur. Such evidence stresses the importance of accuracy when assessing nutrient levels in plants and any supplementary response. The use of spectroscopic reflectivity readings as independent variables for the concentrations of nutrients in plants is not new. The first studies of NIRS on plant tissue began in the mid-1970s [11] and have since shown promise under the later umbrella of precision agriculture as a means of maintaining optimal nutrient balance.

B. Review of Methods

Different ML models have been experimented with for their efficacy in predicting nutrient concentrations in soil and plant material. Waegeman et al. [12] propose the family of multivariate regression techniques for settings where multiple dependent variables are sought. PLSR has become a popular method in chemometrics [13], [14] and is related to methods like Principal Component Regression (PCR) and Ridge Regression, unified by an approach termed Continuum Regression [15]. PLS-2 regresses multiple response variables and finds a set of latent variables that explain the covariance between independent and dependent variables simultaneously [16], [17]. Such methods minimise the "curse of dimensionality" using inherent dimensionality reduction (DR) or regularisation processes. While PLSR does not use explicit regularisation (more implicit as part of its DR), Lasso Regression is a formal regularisation algorithm that offers multivariate [18], or further, multi-task methods [19]. An extension that combines both Lasso and Ridge Regression is ElasticNet that has been found to provide better interpretability over PLSR [20], and can overcome limitations of some regression models that are less successful in settings with high multicollinearity and propensity to overfitting [21]. Regression methods that can find non-linear relationships and are evidenced as giving positive results in similar settings include Gaussian Process Regression [22], [23], Support Vector Regression [24], [25] and Deep Learning [26], [27]. Ensemble methods such as Random Forest [25], [28] and Gradient Boosting [28], [29] have also been the subject of recent research with NIRS prediction settings. Research suggests the use of stacked regression (capable of combining the strengths of different models) shows encouraging results over single-method approaches [30].

III. Dataset

The data in this study consist of NIRS reflectance readings (features) and 14 (macro and micro) nutrient concentrations (targets), where concentration is either measured in parts per million (PPM) or percentage (PCT) as described by the matrix in Table I. Reflectance is measured over a spectrum of wavelengths between 400nm-2500nm at either 0.5nm or 1nm intervals. The sample size $n = 674$ (excluding null records) can have features $p = 4200$ (at the higher resolution of 0.5nm), defining the dataset as high-dimensionality low-sampling size (HDLSS). The data is further categorised by its season (of which there are 4) and the two sampling modes (fresh and dried). Fig. 1 shows the distribution of all targets (`MinMax` scaled) in season 1, split into the two sampling modes.

TABLE I: Description of nutrients.

	Macro	**Micro**
PPM	-	Al, B, Cu, Fe, Mn, Zn
PCT	Ca, Mg, N, P, K, S	Cl, Na

The graph in Fig. 1 shows that the distribution is not normal but is generally highly (positively) skewed, with some exceptions showing bimodality. In addition, when the same nutrient is compared to itself across different seasons, the distribution is dissimilar. Moreover, the distribution is identical between the two sample modes, which can be explained through the principle of mass conservation, whereby the concentration of nutrients is preserved after the petiole sample is dried. Another key finding is that the targets are correlated. Although some pairwise combinations of targets are consistently highly correlated regardless of the season, some are inconsistent. Both of these observations are evidenced by a small subset of pairwise nutrient combinations in Fig. 2. Finally, Fig. 3a shows the variations in the reflectance values between samples across the spectrum used for ML.

IV. Methodology

A. Data Partitioning

Given the data is HDLSS, a stratified test-train split (of 20:80) was implemented to maximise the training data, where stratification occurs on the season for each sampling mode, resulting in separate train-test datasets and ML models for each. This allows for a proportional representation of each season to account for the fact that targets in each season have a dissimilar distribution. By preserving the same ratios in the test set, it is ensured that no one season is over-represented due to an unbalanced number of samples from each, which may form biased ML models.

B. Multivariate Regression

Given the targets are not only related to the features but also intercorrelated, this setting demands the prediction of multiple targets simultaneously (multi-target prediction - MTP) over a continuous feature set, characterised as Multivariate Regression (MVR). As formally described by Waegman et al. [12], where instances of $x \in \mathcal{X}$ and targets $t \in \mathcal{T}$, the training set $\mathcal{D}$ is a set of $(\mathbf{x}_i, \mathbf{t}_j, y_{ij})$ where $y_{ij} \in \mathcal{Y}$ is a score that represents the relationship between $\mathbf{x}_i$ and $\mathbf{t}_j$. $\mathcal{Y}$ is a vector of real-number scores, which can be arranged in an $n \times m$ matrix ($\mathbf{Y}$), where n are the instances and m are targets. The aim of MTP and thus its subclass MVR is to predict the scores ($\mathcal{Y} = \mathbb{R}$) for permutations $(\mathbf{x}, \mathbf{t}) \in \mathcal{X} \times \mathcal{T}$. To conduct MVR, the below models were implemented in the study using Python's Scikit-Learn Library:

1) **MLR** is adopted as a baseline for univariate analysis. The implementation used in the study fits a separate regression model for each target whilst making use of all features. Each model learns a distinct set of coefficients tailored to its specific target. As this approach treats targets independently, it does not capture interdependencies

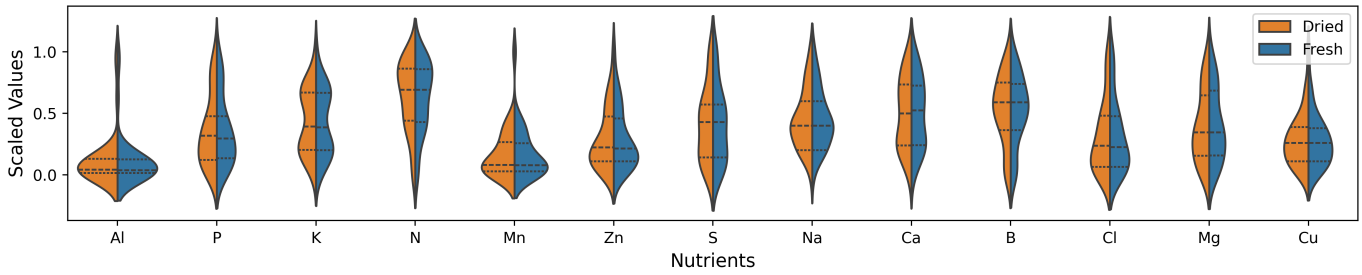

Fig. 1: Distribution of nutrient concentrations in season 1.

between them. This makes it a useful baseline for quantifying the added value of more complex multivariate models considered in this study.

2) **PLSR** is a cross-decomposition algorithm that extracts components that best explain the shared variance between data sets **X** and **Y**. PLSR derives orthogonal components from X that capture the covariance between X, Y, converting the correlated variables into a set of uncorrelated components. This underlines its strength in handling high multicollinearity. Although related, it differs from PCA in that instead of maximising variance in X, it maximises covariance in X, Y. By eliminating multicollinearity, it prevents overfitting by reducing the bias of highly correlated variables, while finding linear combinations of the X variables that have high correlation with Y.
3) **RF** is an ensemble machine learning method, combining predictions from multiple individual decision trees (hence 'forest'). Multiple subsets of data are selected using the bootstrapping method, with which decision trees are trained on. Regularisation can be set by determining the maximum depth of trees, the minimum number of samples required to split a node, and the minimum required in a leaf node.
4) **MTL** uses L_1 regularisation, which reduces dimensions by forcing some coefficients to zero [31], allowing for the selection of the same subset of features across all tasks. The strength of this regularisation is dictated by the λ term - the larger the term, the greater the regularisation [3].
5) **MTEN** incorporates both L_1 and L_2 regularisation, which is effective in the $p \gg n$ case where features may be strongly correlated [31]. L_2 (Ridge) regularisation uses a penalty term for the loss function, which may reduce coefficients minimising the impact of multicollinearity where it may exist. Both reduce over-fitting by removing noise and multicollinearity by balancing L_1 and L_2 regularisation on regression coefficients. MTEN and MTL are designed with the intention of multi-task learning but can be applied to multi-target prediction, by utilising a joint regularisation term to encourage feature selection relevant to multiple outputs.

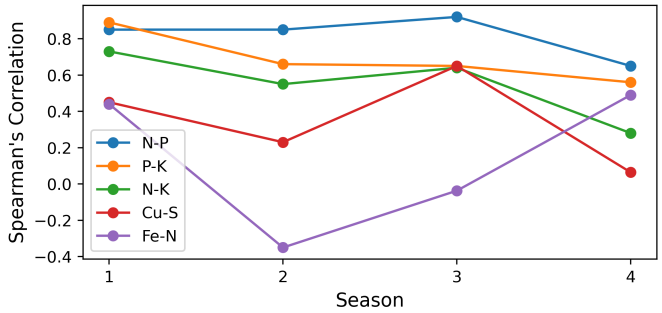

Fig. 2: Spearman's correlation across seasons.

C. Model Evaluation

To evaluate and compare model performance, the characteristics of the targets must be considered first. Since the targets are measured in different units and also have different ranges within the same units, the values are scaled. Section IV-D5 outlines how target scaling avoids the need for normalised error metrics at the point of evaluation.

The two metrics used for evaluating the models are:

1) Root Mean Square Error (RMSE) is the square root of the Mean Squared Error (MSE), and provides an error measure in the same units as the variable being predicted [32]. The benefits for MVR are that it gives better interpretability across all targets, it's a more intuitive scale and focuses on magnitude. Given that it is derived from the MSE, it is sensitive to outliers [32], though less so, given that large values will carry less weight in measurement. It is defined as [32]:

$$RMSE = \sqrt{\frac{\sum(y - \hat{y})^2}{n}} \tag{1}$$

2) Coefficient of Determination (R^2) is the proportion of variation in the target that is predicted from the input and is a measure of how well observed outcomes are predicted by a model [32]. R^2 provides values in a consistent range [0,1], with larger values implying a more successful model, thus more intuitive for comparison. It is defined as [32]:

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \overline{y})^2} \tag{2}$$

To evaluate the effectiveness of each preprocessing step and also for tuning hyperparameters of models, the above metrics

were used along with the cross validation method. Particularly for model tuning, SciKit Learn's `GridSearchCV` was adopted, using a consistent 5-fold approach over the training set, taking the average performance of prediction scores against each fold, for each permutation of hyperparameters. The most performant permutation of hyperparameters was selected based on the minimal loss over those tested. RMSE (of scaled targets) was the metric used as the loss measure.

D. Preprocessing

1) Imputation: Of the 14 nutrients, Cl was omitted as more than 90% of the samples, throughout the dataset, were null. Regarding the feature set, there are also missing values due to the discrepancies in the resolution (0.5nm and 1nm) between seasons. Therefore, imputation of missing values was crucial to maximise the number of usable samples.

The data was imputed separately on training and test data to avoid data leakage. Missing nutrients were imputed using K-Nearest Neighbour (KNN) model as it preserves the multivariate structure, but also because it is non-parametric. To impute the missing 0.5nm values for seasons with a lower resolution of 1nm, linear interpolation was implemented because of its simplicity in approximation for a small wavelength range.

2) Outlier Removal: Outliers can be naturally occurring extreme values or anomalies resulting from human/machine error that distort the underlying data patterns. Removing only true anomalies is important, as preserving the naturally extreme values allows the ML model to link feature ranges to extreme target values. Four possibilities were evaluated using cross-validation scores:

- **No outlier removal**.
- Outlier removal in **targets only**: this assumes human error when calculating the concentration of petiole samples in the laboratory. Here, the Isolation Forest algorithm was used, given that it is non-parametric, can handle multi-dimensional data with no preprocessing and is specifically designed for outlier detection [33].
- Outlier removal in **features only**: this assumes machine error when collecting reflectance data. The feature set is first reduced to two principal components for each sample using PCA. Based on the centres of each season cluster, any data points residing past the 95th percentile distance from the centre are classed as an outlier.
- Outlier removal using **both targets and features**: this assumes that true anomalies are the non-overlapping outlier sample records from targets only and features only analysis. Any sample records that match across both sets are assumed to be natural extreme values.

The outlier removal implementation was done only on the training set, as it was hypothesised that it would make the learning data cleaner. Cross-validation results showed that not removing any outliers produced better results across all ML models tested in the study.

3) Baseline Correction: Baseline correction is a preprocessing step in spectroscopy to remove unwanted background variations that can obscure important spectral features. Prior studies have demonstrated that effective baseline correction significantly improves the interpretability and predictive performance of models built on spectral data [34]–[36]. Unwanted variations in the baseline may arise from several sources. Instrumental drift causes gradual changes in the spectrometer's response over time. Scattering effects, especially common in NIR and Raman spectroscopy, occur when the sample's morphology influences the reflectance. Another contributor is fluorescence, where broad overlapping emission signals can obscure genuine absorption features. Lastly, environmental factors such as changes in temperature, humidity, and other measurement conditions during acquisition can also lead to baseline fluctuations.

These effects introduce smooth, large-scale trends in the spectra, which can dominate and obscure the fine, sharp absorption features linked to material composition.

Fig. 3a shows some of the sample spectra and it is observed that the reflectance spectra for all samples follow a similar general shape dominated by these smooth baseline variations. This similarity between samples is unhelpful for ML models, which rely on capturing differences between samples to make accurate predictions. The following assumptions have influenced the chosen baseline correction method. First, smooth variations across wavelengths are primarily interpreted as background noise, as previously described. Second, sharp localized troughs in reflectance (corresponding to increases in absorbance) are taken to indicate the presence of target elements. Finally, differences in the amplitudes of reflectance in these regions are considered meaningful for modelling element abundances.

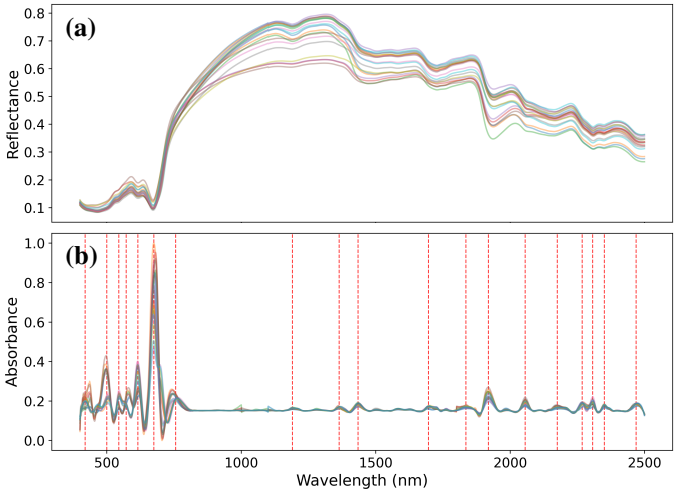

Fig. 3: Application of baseline correction: (a) raw reflectance spectra, (b) absorbance spectra after $\log(1/R)$ transformation followed by ALS using parameters $\lambda = 10^5$, $p = 0.01$, with key wavelengths highlighted.

In Fig. 3a, a transformation of $A = log(1/R)$ is applied, commonly this is found to improve results in similar studies as it is thought to convert the non-linear reflectance measure-

ments into linear absorbance measurements [37].

Given these assumptions, Asymmetric Least Squares (ALS) baseline correction is the chosen method and a common choice for chemometric analysis [35]. ALS is particularly well-suited for the requirements because it penalises deviations above the baseline more heavily than below and preserves sharp absorption features in the spectra while smoothing broader baseline trends.

ALS solves the following optimisation problem:

$$\min_{z} \left(\sum_{i=1}^{n} w_i (y_i - z_i)^2 + \lambda \sum_{i=2}^{n-1} (z_{i-1} - 2z_i + z_{i+1})^2 \right) \quad (3)$$

where:

- y_i is the observed absorbance at wavelength index i,
- z_i is the estimated baseline,
- w_i is a weight controlling the influence of each point,
- λ controls the smoothness of the baseline.

The weights w_i are updated iteratively as:

$$w_i = \begin{cases} p, & \text{if } y_i > z_i \quad \text{(peak region)} \\ 1 - p, & \text{if } y_i \leq z_i \quad \text{(background region)} \end{cases} \quad (4)$$

where $p \in (0, 0.5)$ controls the asymmetry between peaks and background points; the smaller the value the more asymmetric the baseline.

The resulting spectra after applying ALS in Fig. 3b has been flattened to an identical level, at an absorbance level of approximately 0.16 across all samples. The rest reveals clear sharp peaks which occur for the same wavelength across samples but with different amplitudes.

4) Dimensionality Reduction (DR): Due to the low sample count combined with high dimensionality and multicollinearity, any model would have a high risk of introducing noise and over-fitting. It would also significantly increase model complexity and slow down training. DR methods solve these issues and, depending on the model used, also help with interpretability. Various DR methods were applied to features when tuning the data pre-processing pipeline:

- **Binning** reduces the number of features by taking the mean of the reflectance values over the specified wavelength range.
- **PCA** handles the multicollinearity between features by reducing them to few uncorrelated principle components.
- **Peak Feature Extraction** leverages the peaks identified in Fig. 3b following baseline correction, for each sufficiently prominent peak, the minimum amplitude, maximum amplitude, and area under the curve were extracted. These extracted values serve as the features used as inputs for subsequent machine learning models.

Analysing binning shows a pattern of accelerated degradation in performance past an averaging over 4nm. Both sample modes show peaks of poor performance at dried (bin size of 8) and fresh (bin size of 11), after which the validation performance starts to reduce asymptotically to a lower RMSE score, a concerning phenomenon as it is clear at some point important information may be lost. This implies the need for a less crude method to reduce noise in the data.

During cross-validation, PCA provided no easily discernible optimal number of components. Given the small sample size, the maximum number of components for dried and fresh were 77 and 58, respectively. Scree plots determined that the first two components explain the majority of variance for both dried and fresh samples. Likely losing finer details or subtle patterns in the data, PCA was omitted as a data-reduction method in the absence of more in-depth investigation. Although PCA produces features that maximise variance, it does not necessarily preserve meaningful features as it is an unsupervised model and does not know about the targets.

Compared to more general approaches, such as binning and PCA, peak feature extraction preserves the physical interpretability of the features along with the chosen method going forward. From the samples, 19 peaks were identified and from those 57 features were extracted. Therefore, the final DR method chosen was peak feature extraction.

5) Target Scaling: Analysis in Section III highlights the disparity in scales of nutrient concentrations, even within each unit type. Scaling of the targets is necessary for stability and convergence (of gradient-based algorithms such as SVR), to avoid issues with regularisation being dominated by large values, and particularly to help with meaningful performance evaluation across different units and scales. To preserve relationships between targets, it is important to apply the same scaling method to all targets. This is difficult as each have different distributions; skewed, normal and multimodal. Methods such as MinMax, Standardisation and Log transforms were considered, as these handle normal and skewed distributions well, but settled on Quantile scaling as it works for all of these distributions with the added benefit of converting them to a normal distribution. This gives the flexibility for predictive models, as some require normal distributions. It is recognised that information is lost with this technique as targets are scaled into quantile bins. As such, the inverse transform may not produce precise results, so it should be considered as part of any application of the predictions, rather than how the results are compared in the study.

V. Results and Discussion

On the raw spectral data, irrespective of the MVR algorithm explored, the test scores for both R^2 and RMSE were worse at predicting dry sample nutrients than fresh (for dried, R^2=-3.06, RMSE=0.215, for fresh, R^2=-1.65, RMSE=0.16). By applying baseline correction and peak feature extraction, the baseline models improved performance, though this was not consistent across the nutrients as shown in Fig. 4. The shape of improvement between RMSE and R^2 is synonymous.

Results from the MVR models showed that MTEN optimised at the same hyperparameters as MTL for fresh and dried models, thus presenting the same results (Table II). The $L1$ ratio of 1 meant that only $L1$ regularisation was adopted in MTEN, and optimised to a penalty term $\lambda = 0.0001$, as with

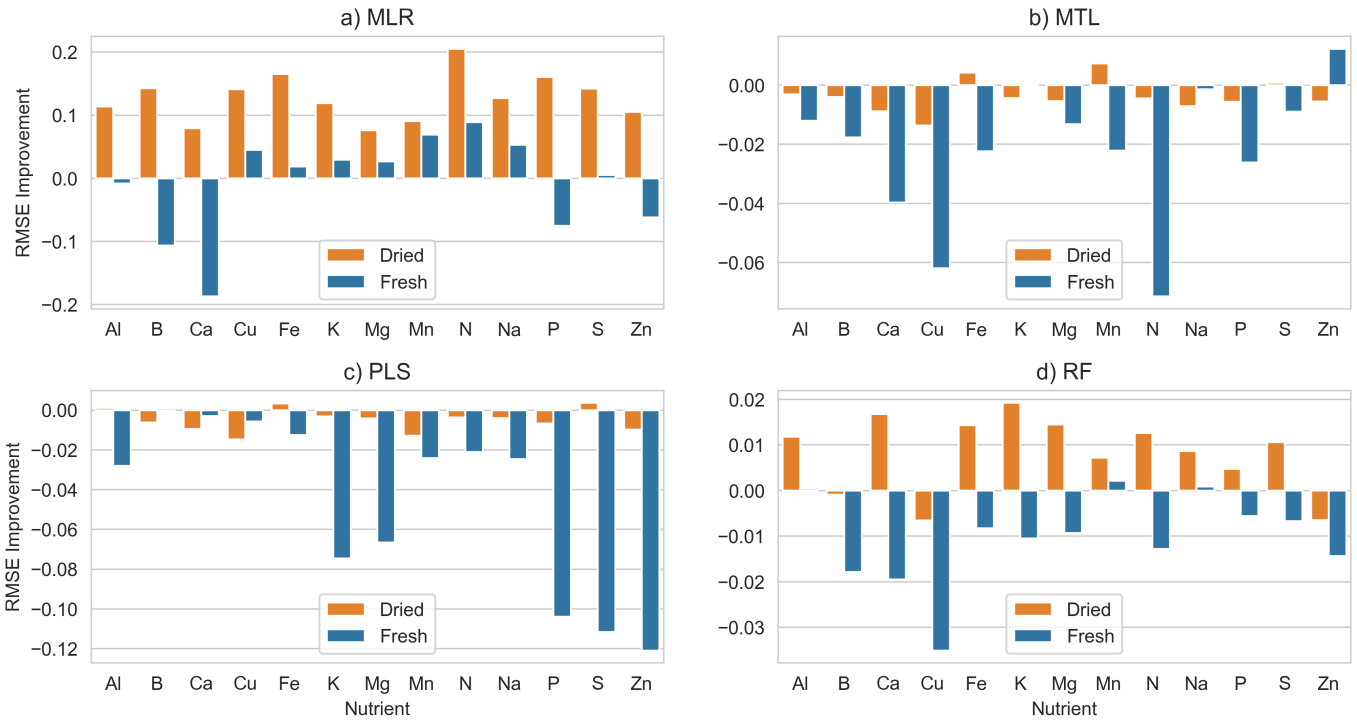

Fig. 4: RMSE impact from ALS baseline correction with peak feature extraction for MLR, MTL, PLSR and RF.

MTL. This penalty term means that there is minimal effect on the features. Though its uncertain why $L2$ regularisation was deferred and a minimal $L1$ penalty applied, it is expected that the applied feature selection meant the algorithm only found truly relevant features, which have a clear relationship with the target variables. Detailed results for MTEN are therefore omitted from Table III, which outlines all outcomes measured in R^2 and RMSE across the baseline model MLR against PLSR, MTL and RF on preprocessed data.

The PLSR models for fresh and dried optimised at 37 and 23 components respectively. RF optimised with 200 estimators and a maximum depth of 20 and 10 respectively.

There is a notable difference in performance between estimators for fresh and dried samples, across all models. This was anticipated based on the impact that moisture can have on infrared absorption at certain wavelengths [38], or baseline shifting due to factors such as refraction. No matter, the relationship between reflectance and moisture is complex and not necessarily linear [39], warranting further study in this context. The RF algorithm was the least impacted and the only model achieving a positive R^2 score on aggregate [Table II], with the least variance across nutrients.

The results of baseline correction and peak feature extraction have noticeably diminished the predictive effectiveness of an optimised PLSR algorithm for both sample modes, particularly the fresh data. Average RMSE for fresh predictions fell from 0.1 to 0.146, particularly attributable to Mg, P, K, S, Zn which increased in error by 101%, compared to only a 5% increase for dried samples. PLSR appears to be affected by losing the ability to exploit correlations between the variables in the same way, given that linear relationships (on which it relies) are largely removed or distorted by the feature selection. The RF method conversely does learn relationships between these features well and can exploit non-linear relationships between variables. Comparing trained RF models on pre/post baseline corrected data showed an improvement of 8.8% for dried samples (p value of 0.19), but an increased error of 11.1% on average for the fresh model (p value of 0.42). From a 95% confidence interval, this is a statistically significant result and shows a common outcome between these multi-variate models trained on pre-processed data, though at different magnitudes.

Comparing results to that of Abukmeil et al. [3], they found that S had acceptable performance through Lasso MLR compared to other macronutrients as well as high r (Pearson's correlation) value. While this has not been a focus of this study, our R^2 score for S was the lowest for nutrients in the dried testing mode, and the second poorest in the fresh testing mode ranked over all models, therefore inconsistent with the previous study based on our pipeline. However, S did show a relatively low error (RMSE=0.0842) despite its poor R^2 score with RF, which further reinforces the idea that this model is leveraging non-linear relationships given that the R^2 score reflects the linearity of the model, shown in Fig. 5. With exception to S, the figure also highlights a difference in performance for macronutrients in the dried RF model. Their study also showed P and K having the highest r values in dried testing, and there is certain overlap with our results. K

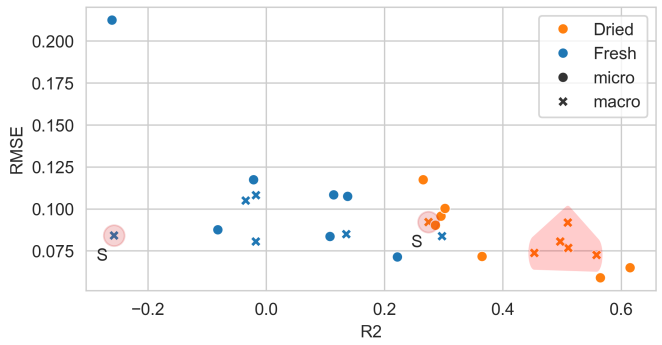

Fig. 5: R^2 and RMSE scores for RF, highlighting performance of S and clustering of macro nutrient scores for dried samples.

featured in the top 3 performing predictions for dried samples across all models, with a more accurate R^2 in the dried mode (0.57 with PLSR). With fresh, it had mediocre performance relative to other nutrients. P was the third poorest R^2 (-0.02, output from RF) in the fresh testing mode.

A noteworthy advantage PLSR has over RF is its efficiency in optimising its loss function over a high-dimensional dataset. PLSR can optimise over [2,45] components with 5-fold cross validation in 80 seconds, compared to training RF on a single set of hyperparameters (a maximum depth of 10, with 200 estimators) in 1hr 10min. This makes optimisation more impractical without prior DR for RF.

TABLE II: Aggregated metrics for each model showing optimal performance by RF, across dried and fresh sample modes.

	Dried		**Fresh**	
	R^2	RMSE	R^2	RMSE
MLR	0.37	0.0869	-2.23	0.168
MTL	0.374	0.0868	-0.394	0.1189
MTEN	0.374	0.0868	-0.394	0.1189
PLS	0.364	0.0876	-1.598	0.1457
RF	**0.423**	**0.0836**	**0.025**	**0.1027**

VI. Further Work and Improvement

Inevitably, the limited sample size of the dataset has its constraints on the ability of the trained model to generalise. Further work should be considered to extend the study with an increased sample size. This is especially applicable with fresh samples given that these were fewer and proved more difficult for the algorithms to find relationships.

Other more novel approaches, such as the use of stacked regression techniques (referenced in Section II-B) may help to exploit different types of linear and non-linear relationships together. As the initial study of collinearity shows, some of these relationships are obvious, yet others will be less so. Exploiting the strengths of algorithms with different approaches to modelling relationship seems obvious and appears to hold benefit in alternative spectroscopy settings [40], [41].

As part of the initial feasibility study, more models were assessed for their 'out-of-the-box' capability with respect to multivariate regression. Gaussian Process, Support Vector and Gradient Boosting regression model libraries assessed did not have native capability to exploit the relationships between target variables simultaneously and so were dropped. However, they all have potential and warrant more detailed assessment against the proposed pipeline given their different properties.

A final aspect that warrants further study is the ability of models to focus on particular nutrients of interest. One perspective considered as part of this study regards model training, where the importance of nutrients is weighted to focus the performance of a model biased to certain nutrients, be it through feature selection or other approach. Another perspective is the use of weighting in evaluation. Where models are assessed and compared based on an average across all nutrients, this would bias results based on chosen weights. Such approaches could help stakeholders focus treatment on nutrients they are particularly concerned with.

VII. Conclusion

The study reveals several important observations for the domain of nutrient prediction using NIRS in potato plant leaf samples. While preprocessing reduces the number of features in a HDLSS dataset and make prediction more effective appears an obvious choice, the results of binning and PCA provide no clear benefit and (in the latter case) can cause multivariate models to overfit.

A more advanced approach utilising baseline correction (via the ALS algorithm), which serves to remove noise in the dataset prior to training, and, when combined with peak feature extraction, has proven to reduce the dimensionality of the data, highlight desired signals, and improve results. However, this improvement was not consistent across dried and fresh samples: dried samples benefitted significantly, while fresh samples did not. This highlights a key challenge in the domain — different sample modes may require tailored preprocessing pipelines to achieve reliable predictive performance. In particular, the high water content in fresh samples causes broad absorption bands and large-scale shifts in reflectance, which obscure the sharp, localised spectral features our preprocessing approach relies on to exploit. As a result, key information is lost or distorted during transformation, limiting the models' ability to extract meaningful patterns.

The application of ALS baseline correction and peak feature extraction also had varied effects depending on the model. For models like MLR and RF, which operate on independent, well-separated features, the preprocessing steps improved accuracy by enhancing the signal-to-noise ratio and removing irrelevant spectral variation. In contrast, PLSR, which relies on capturing global linear relationships and covariation across the entire spectrum, suffered a performance drop — likely due to the removal of important latent structure during feature extraction. This suggests that while the feature selection methodology outlined in the study holds promise, its application must be carefully aligned with the modelling approach. Traditional MVR models may not be the optimal choice when key structural relationships are stripped during preprocessing.

Project code library: https://github.com/EMATM0050-2024/dsmp-2024-groupo2

REFERENCES

[1] A. M. Cavaco, A. B. Utkin, J. Marques da Silva, and R. Guerra, "Making sense of light: The use of optical spectroscopy techniques in plant sciences and agriculture," *Applied Sciences*, vol. 12, no. 3, 2022.

[2] H. Jafarbiglu and A. Pourreza, "A comprehensive review of remote sensing platforms, sensors, and applications in nut crops," *Computers and Electronics in Agriculture*, vol. 197, p. 106844, 2022.

[3] R. Abukmeil, A. A. Al-Mallahi, and F. Campelo, "New approach to estimate macro and micronutrients in potato plants based on foliar spectral reflectance," *Computers and Electronics in Agriculture*, vol. 198, pp. 1070–74, 2022.

[4] "Precision agriculture: Benefits and challenges for technology adoption and use," tech. rep., US Government Accountability Office, 1 2024.

[5] R. Bongiovanni and J. Lowenberg-Deboer, "Precision agriculture and sustainability," *Precision Agriculture*, vol. 5, no. 4, pp. 359–387, 2004.

[6] I. Herrmann and K. Berger, "Remote and proximal assessment of plant traits," *Remote Sensing*, vol. 13, no. 10, 2021.

[7] J. Workman, "A review of the latest spectroscopic research in agriculture analysis," *Spectroscopy Online*, 2024.

[8] J. Oivukkamäki, J. Atherton, S. Xu, A. Riikonen, C. Zhang, T. Hakala, E. Honkavaara, and A. Porcar-Castell, "Investigating foliar macro- and micronutrient variation with chlorophyll fluorescence and reflectance measurements at the leaf and canopy scales in potato," *Remote Sensing*, vol. 15, no. 10, 2023.

[9] M. Naumann, M. Koch, H. Thiel, A. Gransee, and E. Pawelzik, "The importance of nutrient management for potato production part ii: Plant nutrition and tuber quality," *Potato Research*, vol. 63, no. 1, pp. 121–137, 2020.

[10] M. Koch, M. Naumann, E. Pawelzik, A. Gransee, and H. Thiel, "The importance of nutrient management for potato production part i: Plant nutrition and yield," *Potato Research*, vol. 63, no. 1, pp. 97–119, 2020.

[11] J. A. Prananto, B. Minasny, and T. Weaver, "Chapter one - near infrared (nir) spectroscopy as a rapid and cost-effective method for nutrient analysis of plant leaf tissues," in *Advances in Agronomy* (D. L. Sparks, ed.), vol. 164 of *Advances in Agronomy*, pp. 1–49, Academic Press, 2020.

[12] W. Waegeman, K. Dembczynski, and E. Huellermeier, "Multi-target prediction: A unifying view on problems and methods," 2018.

[13] C. Liu, S. X. Yang, and L. Deng, "A comparative study for least angle regression on nir spectra analysis to determine internal qualities of navel oranges," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8497–8503, 2015.

[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," *Journal of Plant Nutrition*, vol. 35, no. 11, pp. 1725–1734, 2012.

[15] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection* (C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, eds.), (Berlin, Heidelberg), pp. 34–51, Springer Berlin Heidelberg, 2006.

[16] G. D. Garson, *Partial Least Squares: Regression and Structural Equation Models*. Statistical Associates Publishing, 2016.

[17] C. Liu, X. Zhang, T. T. Nguyen, J. Liu, T. Wu, E. Lee, and X. M. Tu, "Partial least squares regression and principal component analysis: similarity and differences between two popular variable reduction approaches," *General Psychiatry*, vol. 35, p. e100662, Jan. 2022.

[18] A. Rauschenberger and E. Glaab, "Predicting correlated outcomes from molecular data," *Bioinformatics*, vol. 37, pp. 3889–3895, 08 2021.

[19] G. Obozinski, B. Taskar, and M. Jordan, "Multi-task feature selection," tech. rep., UC Berkeley, 07 2006.

[20] G.-H. Fu, M.-J. Zong, F.-H. Wang, and L.-Z. Yi, "A comparison of sparse partial least squares and elastic net in wavelength selection on nir spectroscopy data," *International Journal of Analytical Chemistry*, vol. 2019, no. 1, p. 7314916, 2019.

[21] J. Liu, T. Geng, W. Jiang, S. Fan, J. Chen, C. Jia, and S. Ji, "A new application of elasticnet regression based near-infrared spectroscopy model: Prediction and analysis of 2,3,5,4-tetrahydroxy stilbene-2-o-beta-d-glucoside and moisture in polygonum multiflorum," *Microchemical Journal*, vol. 199, p. 110095, 2024.

[22] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[23] T. Chen, J. Morris, and E. Martin, "Gaussian process regression for multivariate spectroscopic calibration," *Chemometrics and Intelligent Laboratory Systems*, vol. 87, no. 1, pp. 59–71, 2007. Selected papers presented at the Conferentia Chemometrica 2005 Hajdúszoboszló, Hungary 28-31 August 2005.

[24] R. Tange, M. A. Rasmussen, E. Taira, and R. Bro, "Application of support vector regression for simultaneous modelling of near infrared spectra from multiple process steps," *Journal of Near Infrared Spectroscopy*, vol. 23, no. 2, pp. 75–84, 2015.

[25] C. J. Fearer, A. O. Conrad, R. E. Marra, C. Georskey, C. Villari, J. Slot, and P. Bonello, "A combined approach for early in-field detection of beech leaf disease using near-infrared spectroscopy and machine learning," *Frontiers in Forests and Global Change*, vol. 5, 2022.

[26] C. Cui and T. Fearn, "Modern practical convolutional neural networks for multivariate regression: Applications to nir calibration," *Chemometrics and Intelligent Laboratory Systems*, vol. 182, 07 2018.

[27] H.-H. Huang, J.-F. Luo, F. Gan, and P. K. Hopke, "Two revised deep neural networks and their applications in quantitative analysis based on near-infrared spectroscopy," *Applied Sciences*, vol. 13, no. 14, 2023.

[28] S. Nawar and A. M. Mouazen, "Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line vis-nir spectroscopy measurements of soil total nitrogen and total carbon," *Sensors*, vol. 17, no. 10, 2017.

[29] R. Zheng, Y. Jia, C. Ullagaddi, C. Allen, K. Rausch, V. Singh, J. C. Schnable, and M. Kamruzzaman, "Optimizing feature selection with gradient boosting machines in pls regression for predicting moisture and protein in multi-country corn kernels via nir spectroscopy," *Food Chemistry*, vol. 456, p. 140062, 2024.

[30] G. Dumancas and I. Adrianto, "A stacked regression ensemble approach for the quantitative determination of biomass feedstock compositions using near infrared spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 276, p. 121231, 2022.

[31] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, pp. 301–320, 03 2005.

[32] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," in *8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2022)*, Scipedia SL, 2022.

[33] F. T. Liu, K. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery From Data - TKDD*, vol. 6, pp. 1–39, 03 2012.

[34] X. Li, J. Liu, L. Xu, H. Xu, Y. Wang, and Y. Zhang, "Baseline correction for infrared spectra using relative absorbance-based independent component analysis," *Opt. Express*, vol. 32, pp. 47137–47153, Dec 2024.

[35] F. Zhang, X. Tang, A. Tong, B. Wang, J. Wang, Y. Lv, C. Tang, and J. Wang, "Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method," *Spectroscopy Letters*, vol. 53, no. 3, pp. 222–233, 2020.

[36] P. Eilers and H. Boelens, "Baseline correction with asymmetric least squares smoothing," *Unpubl. Manuscr*, 11 2005.

[37] M. Min and W. Lee, "Determination of significant wavelengths and prediction of nitrogen content for citrus," *Transactions of the ASAE*, vol. 48, no. 2, pp. 455–461, 2005.

[38] B. Chon, S. Xu, and Y. J. Lee, "Compensation of strong water absorption in infrared spectroscopy reveals the secondary structure of proteins in dilute solutions," *Analytical Chemistry*, vol. 93, no. 4, pp. 2215–2225, 2021. PMID: 33433190.

[39] L. Weidong, F. Baret, G. Xingfa, T. Qingxi, Z. Lanfen, and Z. Bing, "Relating soil surface moisture to reflectance," *Remote Sensing of Environment*, vol. 81, no. 2, pp. 238–246, 2002.

[40] E. Cobbinah, O. Generalao, S. K. Lageshetty, I. Adrianto, S. Singh, and G. G. Dumancas, "Using near-infrared spectroscopy and stacked regression for the simultaneous determination of fresh cattle and poultry manure chemical properties," *Chemosensors*, vol. 10, no. 10, 2022.

[41] M. I. Solihin, C. Yuan, W. Hong, L. P. Pui, C. K. Ang, W. Hossain, and A. Machmudah, "Spectroscopy data calibration using stacked ensemble machine learning," *IIUM Engineering Journal*, vol. 25, pp. 208–224, 01 2024.

APPENDIX

TABLE III: Table of prediction metrics of fresh and dried samples for the models studied, all trained on baseline-corrected data.

	Dried								**Fresh**							
	MLR		**MTL**		**PLSR**		**RF**		**MLR**		**MTL**		**PLSR**		**RF**	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Al	0.29	0.076	0.45	0.066	0.48	0.064	**0.56**	**0.059**	-1.38	0.137	-0.16	0.096	-0.62	0.113	**0.11**	**0.084**
B	0.58	0.068	0.59	0.067	0.59	0.067	**0.61**	**0.065**	-3.28	0.241	0.1	0.11	**0.36**	**0.093**	-0.02	0.117
Ca	0.33	0.09	0.41	0.084	0.4	0.085	**0.51**	**0.077**	-8.27	0.314	-0.41	0.123	**0.19**	**0.093**	-0.03	0.105
Cu	**0.34**	**0.111**	0.24	0.119	0.19	0.123	0.27	0.117	-0.09	0.197	-0.6	0.239	**0.07**	**0.182**	-0.26	0.213
Fe	0.15	0.105	0.26	0.098	0.22	0.101	**0.3**	**0.096**	-0.62	0.147	-0.45	0.139	-0.3	0.132	**0.14**	**0.108**
K	0.5	0.077	0.53	0.075	**0.57**	**0.072**	0.56	0.073	-0.58	0.115	**0.2**	**0.082**	-2.06	0.16	0.14	0.085
Mg	0.47	0.095	0.49	0.093	**0.51**	**0.092**	**0.51**	**0.092**	-0.2	0.11	**0.39**	**0.078**	-0.87	0.137	0.3	0.084
Mn	**0.5**	**0.085**	0.46	0.088	0.3	0.1	0.3	0.1	**0.32**	**0.095**	-0.14	0.123	-0.22	0.127	0.11	0.108
N	**0.46**	**0.073**	0.34	0.081	0.32	0.082	0.45	0.074	-0.45	0.096	-2.42	0.148	-0.53	0.099	**-0.02**	**0.081**
Na	0.28	0.076	0.24	0.079	0.27	0.077	**0.36**	**0.072**	-0.19	0.089	0.07	0.078	-0.69	0.105	**0.22**	**0.071**
P	0.42	0.086	0.4	0.088	0.4	0.088	**0.5**	**0.081**	-6.34	0.29	-0.81	0.144	-3.35	0.224	**-0.02**	**0.108**
S	0.23	0.095	0.22	0.096	**0.29**	**0.091**	0.27	0.092	-2.66	0.144	-0.98	0.106	-7.24	0.216	**-0.26**	**0.084**
Zn	0.25	0.092	0.23	0.094	0.19	0.096	**0.29**	**0.09**	-5.26	0.211	**0.09**	**0.08**	-5.54	0.215	-0.08	0.088