Furthering Deep Learning in Near-Infrared Spectroscopy for Fruit Quality Assessment

A thesis

submitted in fulfilment

of the requirements for the Degree

of

Doctor of Philosophy in Computer Science

at

The University of Waikato

by

Mark Wohlers

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2026

Abstract

Near-infrared (NIR) spectroscopy is widely used to assess fruit quality in the horticulture industry. It enables non-destructive estimation of key fruit quality measures from spectra, including dry matter content (associated with taste) and soluble solids content (associated with ripeness). Traditionally, partial least squares regression (PLSR) has been the dominant modelling method. However, more recently, deep learning (DL) has shown promise due to its ability to learn features automatically and model non-linear patterns. However, there are several challenges DL faces when applied to NIR. Labelled datasets are complicated, expensive, and time-consuming to obtain at the size required to fit these models. Deciding on the appropriate architecture and hyperparameters can also be challenging when validation data is sparse. Additionally, a problem of great practical importance in NIR spectroscopy is the difficulty of generalising across different devices of the same model or under different conditions, such as temperature.

This thesis addresses these challenges through three complementary methods. The first uses a data augmentation technique that samples from a multivariate normal distribution with a covariance matrix designed to simulate spectral differences observed across devices. The experiments investigate whether the augmentation improves generalisability and training with small sample sizes.

The second method is a metric based on model stability to diffeomorphic transformations relative to uncorrelated perturbations of similar magnitude. The experiment evaluates the appropriateness of this method for model selection tasks and compares its performance with standard validation methods.

The third method adapts the Barlow Twins contrastive learning method to enable semi-supervised learning in the NIR setting. The Barlow Twins loss function allows unlabelled data to compensate when labelled data is scarce. This method also improves generalisability by encouraging multiple measurements on the same fruit to be similar in the encoded latent space.

Evaluation of these methods is conducted on two datasets: a new dataset containing 5418 kiwifruit sampled across five devices and three seasons, and a previously published dataset of 4675 mangoes measured across four seasons.

The results show that the methods improve predictive performance, especially for small labelled datasets and calibration transfer problems. This allows for the easier application of deep learning to NIR spectroscopy by reducing the requirements for labelled data, improving model generalisability across devices, and enabling model selection under data constraints.

Acknowledgements

I would like to thank Dr Geoffrey Holmes, Dr Eibe Frank, and Dr Andrew McGlone for their constant guidance, support (and patience). I very much enjoyed discussing the work, getting their laser-sharp insights and advice, and benefiting from their wide range of knowledge to keep me on course. I couldn't have asked for better supervisors. Dr McGlone helped secure the funding and had faith in my ability to pursue a PhD. Without him, none of this would have been possible.

It has been a long journey to this point, and I have benefited from much encouragement along the way from my employer the Institute of Bioengineering Sciences (previously Plant & Food Research). My previous line managers, Peter Alspach, Linley Jesson, and the late Nihal De Silva encouraged me to work towards this. Dr Harpreet Kaur and the wider Bioengineering group helped me understand the practical aspects of the field, including taking measurements myself. I am also grateful to Dr Paul Johnstone for his understanding and support, and for offering sage advice.

Special thanks to all those involved in collecting the data, sharing code, and making the mango dataset available for reuse.

I would also like to thank my former school teachers, Anne Scott, Lyndon Coppin, and Alan Jewell, for making maths and physics exciting subjects to study. Not an easy task at a small rural school. My parents, David Wohlers and Margaret Ereckson, for their encouragement and support along the way.

Finally, I want to thank my wife, Evelyn Pino, and my children, Maite, Santiago, and Isidora. Your constant support and understanding have made this possible.

List of Abbreviations

AE	Autoencoder
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DL	Deep Learning
DMC	Dry Matter Content (percentage of fruit mass excluding water)
ELU	Exponential Linear Unit (activation function)
EMSC	Extended Multiplicative Scatter Correction
GAN	Generative Adversarial Network
GBM	Gradient Boosted Machine
GP	Gaussian Process
GPU	Graphics Processing Unit
LS-SVM	Least Squares Support Vector Machine
LSTM	Long Short-Term Memory
LV	Latent Variable
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multiple Linear Regression
MSC	Multiplicative Scatter Correction
MSE	Mean Square Error
MTS	Minimum Taste Standard
MVN	Multivariate Normal (distribution)
NIR	Near-Infrared (780–2500 nm range)
NIRS	Near-Infrared Spectroscopy
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLSR	Partial Least Squares Regression
R_f	Relative stability metric (ratio of diffeomorphic to uncorrelated transformation variability)
RBF	Radial Basis Function
ReLU	Rectified Linear Unit (activation function)
RF	Random Forest

RMSE	Root Mean Square Error
SELU	Scaled Exponential Linear Unit (activation function)
SG	Savitzky-Golay (filter)
SNV	Standard Normal Variate (preprocessing method)
SSC	Soluble Solids Content (measured in °Brix)
SVM	Support Vector Machine
VICReg	Variance-Invariance-Covariance Regularization
VIS	Visible Spectrum (approximately 380–780 nm)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Context

Consumers are increasingly demanding high-quality fruit at competitive prices. This is particularly challenging for export nations like New Zealand, where the distance to its major apple and kiwifruit markets of Europe, Asia and North America (United Fresh New Zealand, 2024) means significant shipping costs that hinder the ability to compete on price alone. The export crop value of kiwifruit and apples was $3.6 billion to the New Zealand economy in 2024 (United Fresh New Zealand, 2024). Protecting and growing this market can be achieved by consistently providing high quality and nutritious fruit. At the same time, reducing costs, including by ensuring that packed fruit meets industry standards, and is not wasted.

Current practice in New Zealand requires kiwifruit growers to ensure the profile of their crop meets specific standards prior to harvesting. Traditionally, this is done through destructive testing of at least 200 fruits (Zespri Group Limited, 2024). This can be time-consuming even for these small samples and is subject to sampling variability, or potentially biased estimates if the sampling is not conducted randomly. If a batch of harvested fruit is later destructively sampled and found not to meet the given standard, then it would be useful to recover the acceptable individual fruit within that batch. Al-

ternatively, this could also be done prior to testing in the packhouse to give the batch a high chance of passing quality control. These decisions need to be based on non-destructive predictions of the quality measures of interest; otherwise, the fruit we are attempting to recover is destroyed in the process.

1.2 Fruit Quality Assessment

1.2.1 Quality Metrics: SSC and DMC

Fruits are made up of various components. For most fruit, including kiwifruit and mangoes, water is by far the largest. Dry matter content (DMC) refers to the percentage of the fruit that is not water, the majority of which is carbohydrates, including fibre, starch, and sugars. Other major components include minerals, as measured by ash, protein, and lipids (Table 1.1).

As the majority of the DMC is made up of carbohydrates, and sugars in particular, this serves as a proxy for taste with numerous studies finding a relationship between DMC and consumer responses (Harker et al., 2003; Jaeger et al., 2011; Palmer et al., 2010; Serra et al., 2019). Often, it is measured through the time-consuming method of weighing a sample, such as a slice from the fruit, then drying and weighing again.

Another component, Degrees Brix, or soluble solids content (SSC), is the amount of soluble solids measured in extracted juice by a refractometer (Scalisi and O'Connell, 2021). In fruit, the majority of the soluble solids are made up of sugars and is used as a proxy for fruit ripeness (Schotsmans et al., 2007). In kiwifruit, SSC will increase over time as starch is converted into SSC (McGlone and Kawano, 1998). Because of this, DMC at harvest is correlated with ripe SSC so can give an indication of how sweet the fruit will be when ripe (Woodward, 2007) and has been shown to predict consumer preference in kiwifruit (Harker et al., 2009). Due to this relationship, SSC and DMC are used as indices to confirm harvest timing. The standards vary by fruit type, cultivar and even size (Kinhal, 2024). For example, the European Union

Table 1.1: Nutritional composition comparison of ripe kiwifruit varieties and mango (per 100g fresh weight). Note: Organic acids (primarily citric acid) are included in the total carbohydrate value by difference. These are approximately 1-2 g/100g for kiwifruit (Marsh et al., 2009) and 0.13-0.71 g/100g FW in mangoes (Maldonado-Celis et al., 2019)

Component (g)	**Kiwifruit Hayward**	**Kiwifruit SunGold**	**Mango (various)**
Water	83.1	82.4	78.9–82.8
Protein	1.14	1.02	0.36–0.40
Total lipid (fat)	0.52	0.28	0.30–0.53
Ash	0.61	0.47	0.34–0.52
Carbohydrate, by difference	14.7	15.8	16.2–17.18
Fiber, total dietary	3.0	1.4	0.85–1.06
Sugars, total	9.0	12.3	≈15.0

Sources: Kiwifruit data from Richardson et al. (2018); Mango data from Maldonado-Celis et al. (2019)

require imported kiwifruits to have a minimum SSC of 6.2° Brix and DMC of 15% at packing shortly after harvest. Sungold and Ruby Red cultivars require SSC of 8° and 9° Brix, and DMC of 16.1% and 17.2% respectively, with early harvest Sungold allowing for a lower SSC of 5° Brix (Kinhal, 2024). Similarly, 'Rocha' pears grown in Portugal must meet a number of harvest requirements, including SSC (Martins et al., 2023).

1.2.2 NIR Spectroscopy for Non-Destructive Estimates

Near-infrared (NIR) spectroscopy has been successfully used to estimate quality measures for various fruit types non-destructively, including DMC and SSC. This has become an increasingly important tool in the horticulture sector, where consumers demand high-quality produce.

Improving fruit quality predictions based on NIR spectroscopy could allow for more complex outcomes, such as estimating fruit storage potential. This, in turn, would allow for poor storing fruit to be taken to market while still at an acceptable standard, and so reduce fruit wastage and improve quality of fruit reaching the consumers. Using models that are more robust to these effects than those currently used and that can incorporate information from supplementary data could improve such predictions.

NIR refers to the 780 to 2500 nm range of the electromagnetic spectrum. Over this range, there are regions that display absorbances "related to overtones and combinations of -CH, -NH, -OH (and -SH) functional groups" (Reich, 2005). Its use in providing a non-destructive analysis technique in the agricultural setting was first popularised in the 1960s by Karl Norris (Reich, 2005).

Since then it has matured as a field in terms of technologies and data-processing techniques.

The equipment, called a spectrophotometer, uses a monochromator to select and output the NIR bands of a light source such as a halogen bulb. The emitted NIR radiation is then reflected by or transmitted through the fruit and measured (See Figure 1.1). The radiation is partially altered by the fruit's composition, and the resulting spectral signature is used to predict this composition using chemometric methods.

Currently, high-throughput fruit grading methods, including those based on NIR, help to segregate the kiwifruit such that only those deemed to be of export quality are sent to foreign markets (K. B. Walsh et al., 2020).

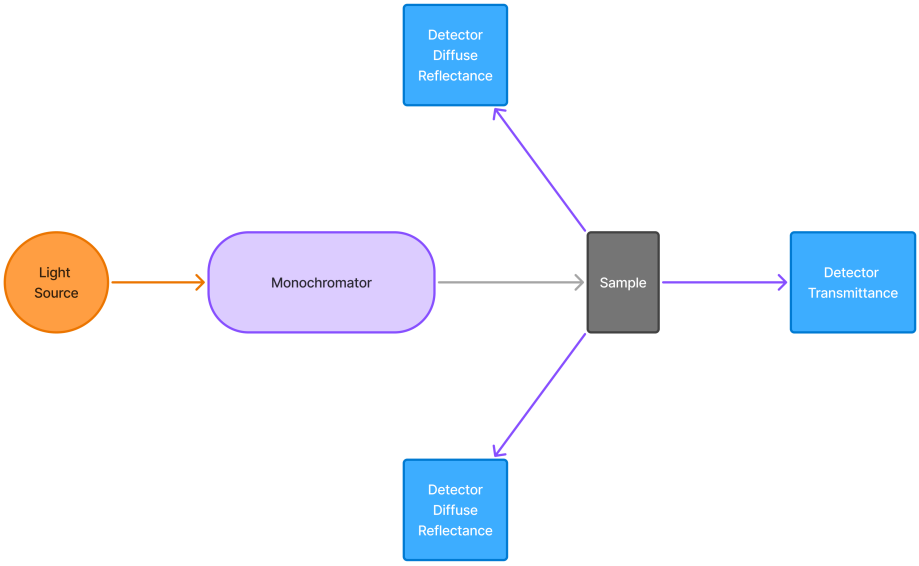


Figure 1.1: A simple spectrophotometer setup. Light from the source passes through the monochromator to select specific NIR wavelengths. This light then interacts with the sample, and detectors measure the diffuse reflected light or the transmitted light that passes through. Note that in certain configurations the monochromator may be placed between the sample and detector instead.

1.3 Research problem

1.3.1 Challenges in NIR Spectroscopy

NIR Spectroscopy has proved to be a useful tool in non-destructive fruit evaluation. However, several difficulties associated with NIR spectra can lead to poor results.

In their review of the use of spectroscopy for estimating fruit quality Wang et al. (2015) state that the NIR spectrum "has a low signal-to-noise ratio and high overlap of combination bands and overtones". This is further complicated by the influence of light scatter effects, such as the light level received by the sensor changing (diffuse reflected) or in the case of shiny surfaces, specular reflectance (Dixit et al., 2017). To overcome this a number of pre-processing

and modelling techniques have been developed over the years. However, the most appropriate method for a given fruit type or attribute is not always clear and often needs to be investigated (Wang et al., 2015).

The quality of these predictions using standard modelling techniques can be influenced by temperature (Diaz-Olivares et al., 2024), device calibration (Bouveresse & Massart, 1996), and batch effects (Alagappan et al., 2023). This is also an issue when attempting to extend NIR based models to predict fruit storage potential based on at-harvest measurements. Batch effects such as different geographic locations, growers, and harvest times have been shown to influence storage potential in kiwifruit, even in fruit with otherwise similar characteristics (Burdon et al., 2014). If the predictive models are trained on datasets containing only a small subset of such batch effects, then overfitting can lead to poor generalisation on unseen batches. Wang et al. (2015) conclude that a large database of training data is needed to achieve models that generalise well across these batch effects.

Additionally, while a fruit is measured during high-throughput grading, multiple NIR point estimates are taken, some of which may be of poor quality as scattering effects are more likely in these less controlled environments than in lab settings. Other spectra may vary due to the heterogeneity of the fruit composition. For example, soluble solids are non-uniformly distributed throughout the fruit (Martinsen and Schaare, 1998; Peiris et al., 1999). Spectra measured at different points along this distribution could be influenced, with the signal not completely matching the attribute of interest, which is measured at the whole-fruit level. This problem is often addressed by either averaging the spectra (Nicolaï et al., 2007) for each fruit or assigning the same overall fruit label to the individual spectra.

Another source of variation is the diversity of the devices used to obtain measurements, which inhibits the ability to use a predictive model trained on one device on another. So-called calibration transfer problems are a current

area of research with a number of proposed methods available (Folch-Fortuny et al., 2017; Workman, 2018).

As more advanced machine learning models become available, there is an opportunity to evaluate their appropriateness for modelling NIR spectra and to improve upon the current solutions to the problems listed above.

1.4 The Case for Deep Learning

Methods for translating NIR spectra into predictions of DMC and SSC have varied, with Partial Least Squares Regression (PLSR) perhaps the most established. More recently, Deep Learning (DL) has shown potential benefits over existing methods by being less reliant on preprocessing techniques and achieving continued performance gains as the size of the training data increases (Mishra et al., 2022). Anderson et al. (2021) found that while PLSR performed well at predicting DMC by fitting individual models to each mango cultivar in the study, performance degraded when fitting a global model (root mean square error (RMSE) of 0.86% and 1.01%, respectively). However, deep learning was able to learn a global model (RMSE 0.89%). Throughout this thesis, RMSE is used as a measure of predictive performance. Lower RMSE indicates predictions are closer to the observed values (see Section 3.3 for more details).

Despite Deep Learning offering automatic feature engineering and the ability to learn complex non-linear relationships (Mishra & Passos, 2021a), PLSR remains more popular. Several reasons explain this. PLSR generally works well even with small training datasets, which is common with handheld devices. Even high-throughput NIR units that generate a lot of NIR data often lack the matching destructive DMC and SSC measurements. While PLSR does require more feature engineering, such as the type of preprocessing, than DL, it is a well studied technique with a small set of standard methods that can be assessed for applicability. Comparing these methods on a given problem is easier because PLSR is computationally inexpensive compared to many other

ML methods, such as DL. It has fewer hyperparameters to tune, with only the number of latent variables to be specified. It is a widely used technique and so is implemented in a number of packages, making it more accessible than the more software environment dependent DL frameworks. The results are essentially a multiple linear regression equation, making it very simple to interpret and productionise.

Deep Learning, on the other hand, shows much promise in NIR spectroscopy. It generally outperforms PLSR when the appropriate architecture is defined and properly trained, particularly when large amounts of labelled data are available (Mishra et al., 2022). However, selecting this architecture and tuning sensitive hyperparameters is challenging. For example, even small changes in the learning rate parameter can lead to very different performance. Optimising these parameters is a much more involved task and can be time-consuming, particularly for large datasets. Still, as the field of research continues to mature, many of these challenges will become less daunting. The choice of architecture is one area that has developed a number of suitable methods for NIR spectroscopy. This is discussed in more detail in Chapter 2.

> **Thesis statement:** Self-supervised learning and appropriate data augmentation can significantly improve the performance and generalisation of deep learning for near-infrared spectroscopy.

1.5 Research Questions and Objectives

Three main research questions will be investigated in this thesis:

RQ1: Does data augmentation that includes the spectral correlation structure improve model robustness and enable calibration transfer between different measurement devices?

- What types of data augmentation techniques are suitable for NIR spectra?

- Can augmentation assist in calibration transfer and generalise to unseen devices?

RQ2: Are model selections based on relative stability to diffeomorphic transformations a viable option to replace or supplement traditional validation approaches when labelled validation data is sparse?

- Does relative stability to diffeomorphisms correlate with model test set performance?

RQ3: Is it possible to use self-supervised learning techniques to use unlabelled spectral data to improve deep learning performance in NIR spectroscopy when there is limited labelled data?

- Can the Barlow Twins contrastive learning framework, developed for computer vision, be used for NIR spectroscopy by treating multiple measurements of the same sample using different devices as multiple "views"?
- How much performance improvement can be expected by combining the Barlow Twins with semi-supervised learning, and under what conditions are the greatest performance gains observed?

1.6 Contributions

This thesis makes the following original contributions to the fields of chemometrics, deep learning, and fruit quality assessment:

- **Data augmentation method** based on sampling from multivariate normal distributions with empirically estimated covariance matrices from multiple measurement data.
- **Novel model selection criterion** (R_f) based on relative stability to diffeomorphic transformations versus uncorrelated perturbations.

- **Comparable performance** between R_f and cross-validation for model selection and early stopping tasks.
- **Novel adaptation of Barlow Twins** from the computer vision field to NIR regression by treating multiple measurements of the same fruit as different "views".
- **Semi-supervised framework** combining Barlow loss and MSE regression loss, enabling learning from both labelled and unlabelled data.
- **Theoretical connections** established between Barlow Twins objectives and PLS regression latent variables.

1.7 Thesis Structure and Publication Status

This thesis follows the PhD with Publication format. It includes two peer-reviewed research papers and one currently under review that address different aspects of improving NIR spectroscopy predictions using deep learning, along with chapters that provide context, background, synthesis, and conclusions.

1.7.1 Structure

The thesis is organised as follows:

Chapter 1: Introduction (this chapter) establishes the motivation for this research, describes the problem context, articulates the research questions and objectives, and summarises the contributions.

Chapter 2: Background and Literature Review gives an overview of related work in traditional chemometrics, machine learning for spectroscopy, semi-supervised learning, data augmentation, and model selection.

Chapter 3: Methodology gives an overview of the datasets and computational resources used across the three studies.

Chapter 4: Augmenting NIR Spectra in Deep Regression to Improve Calibration presents a data augmentation approach based on sampling from a

multivariate normal distribution to simulate variation amongst devices of the same model. This chapter addresses Research Question 1 and shows how augmentation enables robust cross-device generalisation. This paper has been published (Wohlers et al., 2023), "Augmenting NIR Spectra in Deep Regression to Improve Calibration," *Chemometrics and Intelligent Laboratory Systems 240 (2023) 104924.*

Chapter 5: Assessing Machine Learning Models for Near-Infrared Regression by Measuring Stability Towards Diffeomorphisms introduces a model selection criterion based on stability to diffeomorphic transformations. This chapter addresses Research Question 2 and provides a practical tool for architecture selection when validation data is limited. This chapter has been published (Wohlers et al., 2025), "Assessing Machine Learning Models for Near-Infrared Regression by Measuring Stability Towards Diffeomorphisms," *Chemometrics and Intelligent Laboratory Systems 264 (2025) 105449.*

Chapter 6: Barlow Twins for Semi-Supervised Learning in NIR Spectroscopy applies contrastive learning to the use of unlabelled data and multi device measurements. This chapter answers Research Question 3, and provides evidence that semi-supervised learning can significantly improve NIR modelling when the labelled training dataset is small. This chapter is currently under review at *Chemometrics and Intelligent Laboratory Systems* (submitted October 2025).

Chapter 7: Synthesis and Conclusions integrates the findings from all three studies, discusses practical implications, acknowledges limitations, and identifies future research possibilities.

Chapter 2

Background and Literature Review

This chapter gives an overview of the current machine learning methods used for NIR spectra based regression problems. While many of the methods can also be used for classification, the current focus is on predicting continuous fruit quality attributes. Because the spectral data are highly correlated across wavelengths, the models need to account for this. Wang et al. (2015)'s review of evaluating various types of fruit using NIR spectroscopy listed PLS as the most common calibration technique employed. This was followed by Least Squares Support Vector Machines (LS-SVM), Principal Components Regression (PCR), and Multiple Linear Regression (MLR). In some instances, other popular machine learning methods have been applied to NIR data after the dimension of the output has been reduced, for example Y. Liu et al. (2010) fit an Artificial Neural Network to the loading of a Principal Component Analysis (PCA) to predict SSC in navel oranges.

The choice of models used seems to have been very dependent on the domain, for example, while ANNs have been applied to NIR in various fields. Dixit et al. (2017) found no evidence that they had been applied to NIR spectroscopy of meat products at that time. This could be due to model suitability or, in part, to researchers' familiarity with certain techniques. Therefore, it is

also important to investigate models that have been successful in NIR spectroscopy studies in general, rather than limiting the scope to those involving fruit.

2.1 Traditional Methods in NIR Spectroscopy

2.1.1 Preprocessing

NIR spectroscopy employs several popular preprocessing techniques to remove scattering and noise prior to model fitting. Of these, the most popular are Multiplicative Scatter Correction, Standard Normal Variate, and derivation techniques including Savitzky-Golay Filters (Rinnan et al., 2009). The following will focus on these, but they are not the only methods used. For a more complete summary of the various methods, see Wang et al. (2015), Rinnan et al. (2009), and Nicolaï et al. (2007).

2.1.1.1 Standard Normal Variate

In the Standard Normal Variate (SNV) technique, the NIR spectra for a given sample have their respective mean subtracted and then divided by its standard deviation. That is, each spectrum is normalised to have a mean of zero and unit variance (Rinnan et al., 2009). For spectrum j of sample i the corrected spectrum x^*_{ij} is:

$$x^*_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_i}$$

2.1.1.2 Multiplicative Scatter Correction

Multiplicative Scatter Correction (MSC) aims to remove multiplicative scatter effects by using a reference spectrum. If no suitable reference is available, then the mean spectrum can be used. The uncorrected spectrum is then regressed on this reference using a linear regression model to obtain the intercept and slope coefficients. The corrected spectra are then simply the original spectra minus the intercept, and then divided by the slope.

$$x_{ij}^{*} = \frac{x_{ij} - a_i}{b_i}$$

where a_i and b_i are the least-squares solution to:

$$x_i = a_i + b_i \bar{x_j} + e$$

with x_i the spectrum vector for sample i, $\bar{x_j}$ as the reference spectrum, in this case the spectrum mean vector, and error term e. This method has been further refined into Extended Multiple Scatter Correction (EMSC), which uses polynomial coefficients in the regression model and also adjusts the scatter correction to account for known spectral regions where absorption may be large (Martens & Stark, 1991; Martens et al., 2003).

2.1.2 Savitzky-Golay Filters

The Savitzky-Golay (SG) Filter is a convolution method that smooths the data and can produce smooth derivatives. It was first proposed in 1964, although the original paper contained errors in the convolution coefficient tables, which were later corrected Madden (1978). The smoothing aims to reduce noise and is especially important if derivatives are taken as otherwise they may amplify the noise (Wang et al., 2015). Using the 1st derivative of the spectra rather than the original smoothed spectra removes any constant baseline shift. Taking the 2nd derivative also removes a linear multiplicative effect. As SG uses a polynomial to fit over a symmetric window of the data around a given point, the highest order derivative possible is therefore equal to the order of this polynomial (Rinnan et al., 2009). An additional consequence of using SG is that the output sequence will be shorter than the input spectra by the size of the smoothing window minus 1, with the start and end spectra trimmed equally. While the first and second derivatives are the most popular choices, there does not seem to be agreement over which is more appropriate (Wang et al., 2015), so it may be situation dependent. Similarly, as Nicolaï et al. (2007) notes, while SG and MSC often share the goal of removing additive

and multiplicative noise, the choice of which to use seems rather arbitrary.

2.1.3 Outlier Detection

Outliers in the spectral data can affect model performance. As with all model development, data cleaning is a required step. In the setting considered in this thesis, this involved removing outliers from both the NIR spectra and the measurement of interest (e.g., DMC or SSC) prior to analysis. The latter is often more straightforward as the data is univariate, where unrealistic labels can be easily identified by examining the tails of the distribution. Detecting outliers for multivariate data is more challenging. For example, an observation may appear normal when examining each variable individually, but the pattern of the variables together suggests it is very different from the other samples. Thus, multivariate outlier detection requires more complex techniques.

2.1.3.1 Mahalanobis Distance

Outliers can be detected by looking at how far they are from the centre of the distribution. The Mahalanobis distance accounts for the correlation among variables, unlike the Euclidean distance. When applied to detecting NIR outliers, the Mahalanobis distance is calculated based on a robust estimate of the covariance matrix (e.g., Leys et al. (2018)). This is done by taking random subsets of the data, often 75% of the total number of samples, and selecting the subset that gives the covariance matrix with the lowest determinant. This aims to remove the outliers that are to be detected from influencing the covariance matrix estimation by selecting the most central subsample.

This method can also be calculated on PLS or PCA scores rather than the observed spectra.

2.1.3.2 Hotelling T^2 and Q Residuals

Another common technique outlined in Pelliccia (2018) is based on two metrics, Hotelling's T^2 and Q Residuals. First, a PLS regression, or PCA, is fit and

scores calculated. The Hotelling T^2 measures how far the PLS (or PCA) scores for a given sample are from the centre of the model (Eigenvector Research, 2025).

The Q residual is the squared difference between the spectrum and the reconstructed spectrum from the PLS (or PCA) scores and loadings. The Q residuals and Hotelling T^2 are complementary metrics that measure how well the model describes the sample (and the variability of the samples' projected scores (Eigenvector Research, 2025)).

2.1.4 Traditional Regression Methods

2.1.4.1 Multiple Linear Regression

MLR is a regression technique where the attribute of interest is predicted by multiplying each wavelength response by a given weight, also called a coefficient and summing along with an offset term. More formally if x_{ik} is the spectral response relating to the i_{th} sample and k_{th} wavelength then the prediction of the attribute y_i is

$$\hat{y_i} = \beta_0 + \sum_{k=1}^{n} \beta_k x_{ik}$$

The coefficients β are often estimated by least squares or maximum likelihood methods and minimise the mean square error loss function.

One drawback of MLR is that the coefficients can be sensitive to changes in the training set when there is correlation among the features Franke (2010). This problem of multicollinearity is present in the NIR setting, where wavelengths are highly correlated, and indeed Wang et al. (2015) found that the method appeared unstable. Peiris et al. (1998) and Jaiswal et al. (2012) used MLR with reasonable predictive performance on the calibration dataset but found this degraded when applied to the validation dataset, implying overfitting.

2.1.4.2 Principal Component Regression

PCR addresses multicollinearity by fitting an MLR to a set of orthogonal latent variables derived from the spectral data. These latent variables are constructed using Principal Component Analysis (PCA).

PCA itself could arguably be called the most popular multivariate technique across a multitude of disciplines (Abdi & Williams, 2010). The first latent variable is obtained by projecting the wavelengths onto a single dimension in the direction of maximum variance.

The second principal component is constructed in the same way, with the restriction that it must be orthogonal to the first component. This process is repeated up to a predefined number of latent variables, which cannot exceed the number of wavelengths.

The PCR then performs an MLR using a subset of the PCA latent variables. In general, this might be the first n loadings, but there are times when this might not be appropriate. While the first dimension explains more variation in the dataset than any other single dimension, it may not be associated with the attribute of interest.

2.1.4.3 Partial Least Squares

Partial Least Squares (PLS) has been used extensively in the prediction of internal fruit quality measures. It was invented by the econometrician Herman Wold in the 1960s (Vrasti et al., 1998). Svelte Wold, Herman's son and Harald Martens later modified the technique for the chemometric setting (Wold et al., 2001a). It is similar to PCR, but instead of the dimensions being constructed to maximise the variance explained, it maximises the covariance with the attributes of interest. An important parameter to tune is the number of components, n, to include in the final model. This is often achieved through cross-validation, where n gives the lowest validation MSE (Wold et al., 2001a). It has been successfully used to predict attributes in fruit, including kiwifruit (Feng et al., 2011; McGlone et al., 2002, 2007), apples (Bureau et al., 2012),

and oranges (Y. Liu et al., 2010). Due to its popularity and proven performance, PLS is often used as a benchmark for comparing potential alternative modelling techniques. One limitation of PLSR is that, being a linear model, it can have difficulty modelling non-linear systems, although there do exist non-linear PLS methods (Rosipal & Trejo, 2000).

2.2 Advanced Machine Learning Methods

2.2.1 Artificial Neural Networks

In certain circumstances, Artificial Neural Networks (ANNs) have advantages over PLSR due to their non-linearity (Bampi et al., 2013). Wang et al. (2015) reported that the classical (shallow) neural network, being a network with only one hidden layer, was effective in a number of studies related to NIR fruit measurements. In this situation, the number of neurons in the hidden layer is one parameter that needs to be tuned to give reasonable performance. In the simplest type of ANN, all neurons from one layer are fully connected to each neuron in the next layer. The first layer, or input layer, is the input data, such as spectra. Each neuron in the hidden layer receives a weighted sum of the inputs plus a bias, where the weights differ across the hidden layer, and outputs a signal via an activation function such as the Rectified Linear Unit (ReLU) (see Figure 2.1).

These signals are then summed again with a different set of weights to a single output neuron with another activation function, usually a linear activation for regression. Popular activation functions are given in Figure 2.2. Bampi et al., 2013 found that ANNs provided better performance in predicting droplet size in biodiesel emulsions compared to PLSR but this trend was reversed when predicting water content. ANNs are prone to overfitting and generally need more intensive training than PLS.

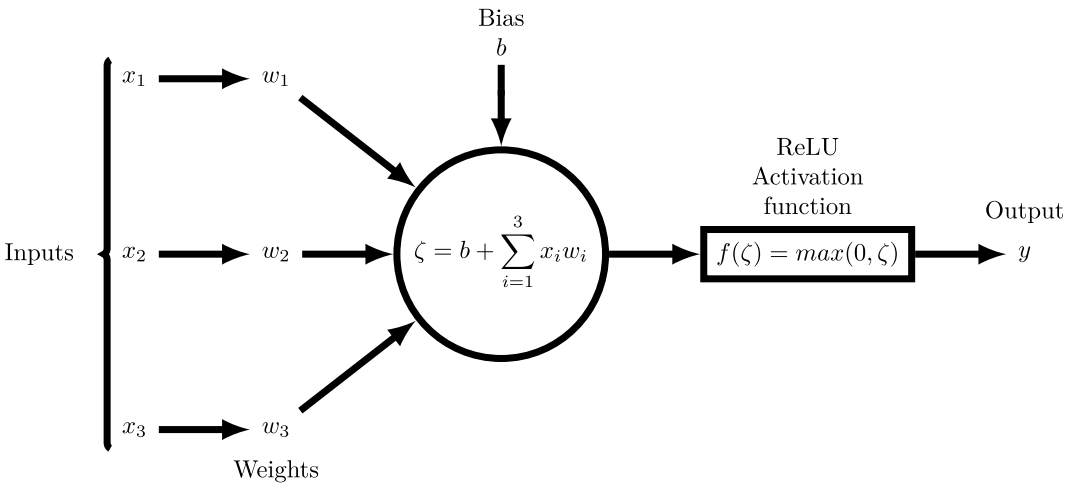


Figure 2.1: An example of a simple neuron with three inputs and a ReLU activation function

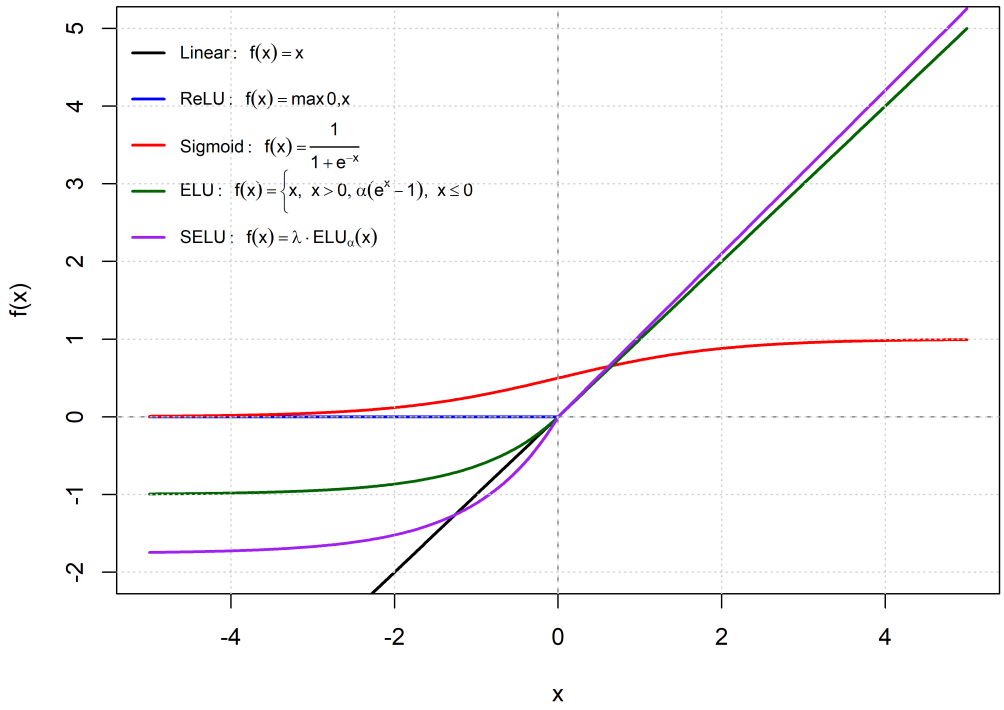


Figure 2.2: Comparison of Deep Learning Activation Functions

2.2.2 Gaussian Processes

Gaussian processes (GP), also known as kriging in the geostatistics field, where this approach first came to prominence, have evolved to become a popular

machine learning technique (Snelson, 2007). This approach is formulated in a Bayesian setting where, instead of having priors over parameters, there is a prior over functions. The variable of interest, y is assumed to be related to inputs x by an underlying function $f(x)$ plus Gaussian noise σ_n^2. That is

$$y = f(x) + \mathcal{N}(0,\, \sigma_n^2)$$

$f(x)$ itself is thought of as a point from an infinite-dimensional Gaussian distribution of which the observed values are a finite subset of dimensions. This GP has a set mean, often zero (Ebden, 2015) and given covariance K. The following definitions closely follow those presented in Ebden (2015). K is derived by the covariance function $k(x, x')$, which calculates the covariance between two attributes based on the respective two inputs. They give an example of Radial Basis Function (RBF) kernel, also known as the Squared Exponential or Gaussian kernel, with the Gaussian noise folded in by using the Kronecker delta function $\delta(x, x')$:

$$k(x, x') = \sigma_f^2 exp\left[\frac{-(x - x')^2}{2\ell^2}\right] + \sigma_n^2 \delta(x, x')$$

where ℓ is the length scale parameter.

Given a kernel function $k(x, x')$, we now define the covariance matrix as:

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{bmatrix}$$

For new inputs x_*, the method assumes that the corresponding unobserved attributes y_* are part of the same single point as the observed attributes y, but in other dimensions. That is:

$$\begin{bmatrix} y \\ y_* \end{bmatrix} = \mathcal{N}\left(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right)$$

where $K_* = k(x, x_*)$, and $K_{**} = k(x_*, x_*)$. Given this, the means of the unobserved attributes are estimated as:

$$\bar{y_*} = K_* K^{-1} y$$

with variance

$$var(y_*) = K_{**} - K_* K^{-1} K_*^T$$

As is seen above, the prediction involves inverting the covariance matrix of the training set K. When the training set becomes large, the GP is known to scale poorly (McIntire et al., 2016) and due to its $\mathcal{O}(N^3)$ training time it can become unusable for large datasets (Snelson & Ghahramani, 2005). There have been a number of proposals to address this problem and provide scalable GPs (H. Liu et al., 2020).

Applying GP's to an NIR dataset to predict nitrogen in 1240 seeds, Sow et al. (2022) found that GP outperformed PLSR in terms of root mean-squared error on the validation set (RMSEV), except when they limited the size of training data to 70 observations or less.

2.2.3 Least Squares Support Vector Machine

Support Vector Machines (SVM) are a popular machine learning technique for classification and regression problems (Wang & Hu, 2005). The basic idea is to map the inputs to a higher-dimensional space and then perform linear regression. To simplify the problem, a kernel function is specified, and the so-called "kernel trick" reduces the complexity of the model fitting. This is possible as the kernel function allows for calculating the inner product of points in the higher-dimensional space without having to calculate the actual position of points in that higher dimension.

Least Squares Support Vector Machines (LS-SVM) are a slight modification of the standard SVM. They were first proposed by Huang et al. (2014) to improve efficiency when applied to large data problems by reducing the problem to solving linear equations as opposed to quadratic programming in the standard SVM. Using the notation from Wang and Hu (2005) Given a dataset consisting of inputs X such as spectra, $X \in R^n$ and an attribute of interest y,

$y \in R$, the aim is to predict y using the function

$$f(x) = \langle \omega, \varphi(x) \rangle + b$$

Where b is a bias coefficient, $\langle \cdot, \cdot \rangle$ is the dot product; $\omega \in R^{n_h}$ are the regression coefficients in the higher, n_h dimensional space; $\varphi(\cdot)$ is a non-linear function that maps $R^n \rightarrow R^{n_h}$. In LS-SVM the optimisation problem is now:

$$\min_{\omega,b,e} J(\omega, e) = \frac{1}{2}\|\omega\|^2 + \frac{1}{2}\gamma \sum_{i=1}^{N} e_i^2$$

with the restriction that:

$$y_i = \langle \omega, \varphi(x) \rangle + b + e_i$$

where $\gamma \geq 0$ is a regularization constant. More details are given in Wang and Hu (2005) specifying the conditions for the optimal solution, but after defining the kernel function K, for example the Radial Basis Function (RBF):

$$K(x, x') = exp\left(\frac{-\|x - x'\|^2}{2\sigma^2}\right)$$

the LS-SVM for predicting the attribute of interest based on a new input spectra x^* is

$$f(x^*) = \sum_{i=1}^{N} \alpha_i K(x^*, x_i) + b$$

where $\alpha_i = \gamma e_i$

As noted in Wang et al. (2015), LS-SVM was found to be the best performing calibration model in a number of studies. More recently, Sow et al. (2022) compared LS-SVM with PLSR and GP in predicting nitrogen in seeds. LS-SVM always provided superior performance when compared to PLSR when varying the size of the training set. Performance compared to the GP was superior at low training sizes but similar when 100 or more observations were used.

2.2.4 Tree-based Ensembles

Tree-based ensembles are popular for machine learning from tabular data because of their ability to model non-linear problems, but they do not appear to

have been widely used for modelling fruit attributes based on NIR. However, two of these methods, namely Gradient Boosted Machines (GBM) and Random Forests (RF) have been successful in NIR spectroscopy relating to soil properties (L. Liu et al., 2017; Nawar & Mouazen, 2017).

2.2.4.1 Gradient Boosted Machine

GBM sequentially fits regression trees, or other weak learners, to the dataset, or a random subset of it, with each iteration, giving more weight to the residuals. The final estimate is a weighted average of the predictions. For a more complete summary of the method see Rogozhnikov (2016). Recently L. Liu et al. (2017) used PLSR as a dimension reduction technique by training a GBM on the PLS loadings to improve performance in predicting various soil properties. There is variation in the algorithms used to implement these model, with the tree boosting system XGBoost (T. Chen & Guestrin, 2016) being one of the most popular.

2.2.4.2 Random Forest

First proposed by Leo Breiman (Breiman, 2001), Random Forest (RF) has become a popular machine learning technique. Here, multiple datasets are generated from the original by bootstrapping the samples and taking random sub-samples. Separate regression trees are fit to each of these new datasets, with the final prediction being the average prediction of all trees. For a more complete summary of the method, see Gorman (n.d.). While comparing RF to ANN and GBM Nawar and Mouazen (2017) found that RF provided the best performance in predicting nitrogen and carbon in soil.

2.3 Deep Learning

Deep learning (DL) has become increasingly popular in part due to its state of the art performance in areas such as image classification (Krizhevsky et al.,

2017), natural language processing (Collobert & Weston, 2008), and speech recognition (Graves et al., 2013). The most basic description of DL is that it is an ANN with multiple hidden layers. However, this is somewhat simplistic, as the diverse areas of research in which DL has been applied have led to varied architectures that can be used for each hidden layer. For example, Convolutional layers for images (Lawrence et al., 1997; LeCun & Bengio, 1995), and long-short term memory in sequential datasets (Hochreiter & Schmidhuber, 1997), while other layers, such as dropout, which randomly set a proportion of outputs from neurons in a layer to zero, are commonly used to reduce overfitting.

With the increase in popularity, the field has become far more accessible with frameworks such as TensorFlow (Abadi et al., 2019) and PyTorch (Paszke et al., 2017) being freely available. In addition, API's such as Keras (Chollet et al., 2015) simplify the model building and training process. Improvements have also been made in the optimisation algorithms used (Qi et al., 2017). This, together with improved computing resources through cloud computing and GPUs, has greatly sped up the model fitting stage. As it was previously prohibitive in terms of resources and time for most researchers, it could explain why DL is now becoming an increasingly popular method to analyse NIR data. One reason that has inhibited its more widespread use is that it requires large amounts of data, although this has been somewhat improved through data augmentation methods. Another is that results can vary depending on the architecture and hyperparameters used, making model fitting a more involved process when compared to standard methods such as PLSR. At any rate, there has been an explosion in research in the area.

2.3.1 Deep Learning Architectures for NIR Spectroscopy

2.3.1.1 One-dimensional Convolutional Neural Networks (1D-CNNs)

In one of the first applications of deep learning to NIR data, Bjerrum et al. (2017a) utilised a Convolutional Neural Network (CNN) on an NIR pill dataset,

employing data augmentation to increase the training dataset size. A CNN contains at least one convolutional layer, in this case, two were used, and is applied to spatially or temporally ordered inputs. In terms of architecture, the convolutional layer is not fully connected, and instead, only a subset of consecutive inputs are fed to each neuron in the next layer. This "filter" slides across the data with the same weights used each time (see Figure 2.3). For

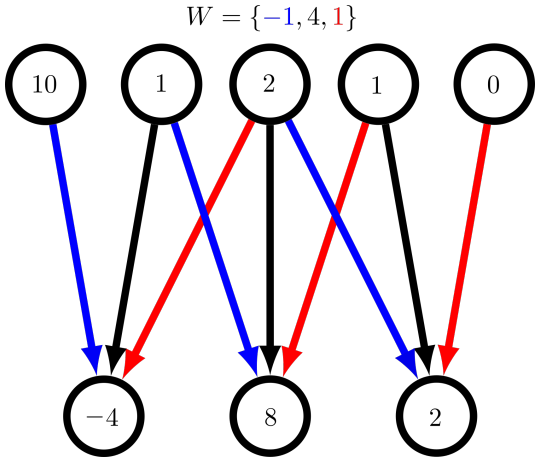


Figure 2.3: An example of a simple convolutional layer. Note how the weights summarised in vector W, are repeated across the network as indicated by their colour.

1-dimensional inputs such as NIR spectra, this can be thought of as a weighted moving average across the sequence, which is then transformed via the activation function. There can be a number of such filters in a given convolutional layer, the number of which must be decided. The architecture used requires a number of hyperparameters to be tuned (see Table 2.1) which was done using Bayesian optimisation techniques to minimise the Huber loss function rather than the more common MSE. Their results showed the CNN outperformed PLSR across datasets using various preprocessing methods. Interestingly, they also provided some evidence that the CNN learned some preprocessing techniques itself as did Cui and Fearn (2018). Acquarelli et al. (2017) also used a similar model for classification applied to a number of food, beverage, and tablet datasets. They found that the CNN was less reliant on preprocess-

Hyperparamter Search Space			
Layer	Parameter	Search Space	Type
Convolutional 1	Number of Kernels	2-40	Integer
Convolutional 1	Filter Size	5-150	Integer
Convolutional 2	Number of Kernels	2-140	Integer
Convolutional 2	Filter Size	5-150	Integer
Dropout	Proportion Dropout	0-0.5	Float
Dense	Number of Neurons	4-1000	Integer

Table 2.1: Hyperparameter search space for the CNN as presented in Bjerrum et al. (2017a)

ing than PLSR. Cui and Fearn (2018) used CNNs with a single convolutional layer with one filter, followed by three fully connected layers on three different datasets. They also found that the CNN achieved superior performance and demonstrated robustness for the two larger datasets. Performance on the third dataset, with only 415 training samples, was comparable to that of the PLSR model. Interestingly, the output from the convolutional layer was similar to the SG pre-processing. Since the SG pre-processing method itself is a convolution, it would be possible to set the weights of a convolutional filter in the CNN to achieve this.

X. Zhang et al. (2019) extended the convolutional network to include inception modules in their DeepSpectra model. This change is based on computer vision work presented in Szegedy et al. (2015), which utilised convolutional filters of varying sizes within the same layer. Specifying the convolution filters in parallel rather than in series allows the layer to extract features at different scales. DeepSpectra included three convolutional layers, with the second and third including inception modules. This was compared to three CNNs, one of which included a single inception layer and was found to give a lower MSE

across the four datasets tested.

An alternative method was introduced in Puneet Mishra and Passos (2021). Instead of using an inception module on the same input, they partitioned the raw spectra into two blocks of 450nm to 697nm, and 700nm to 1030nm, respectively. These blocks were then connected separately to convolutional layers, then concatenated before the fully connected layers. They found that this reduced the RMSE of the Dry Matter Content (DMC) prediction for Mango fruit from 0.855 to 0.818, compared with a single-block CNN on the same data.

More recently L. Zhang et al. (2025) used an attention-based network with the raw spectra partitioned into two blocks and found that it improved DMC prediction in mangoes compared to single block architectures.

2.3.1.2 Residual Networks

Residual networks have become increasingly popular. First described in He et al. (2016), they overcome a key challenge in training deep networks. These networks utilise skip connections, Figure 2.4, to aid in training deep networks which typically encounter problems with exploding gradients as the optimisation routine backpropagates through the network.

Martins et al., 2022 applied residual networks to maintain good performance at very low training set sizes, including as low as 125, with 1.3 million parameters when predicting SSC on an Orange dataset. Later Martins et al. (2023) used a similar but smaller residual network to predict SSC and temperature of the fruit at sampling time of Rocha pears.

2.3.1.3 Transformer Networks

Transformer networks were originally developed for natural language processing (NLP) Vaswani et al. (2017) but have recently been applied to NIR spectroscopy. In their recent review of NIR spectroscopy for estimating mango quality Chaudhary et al. (2025) found that while the applications of deep

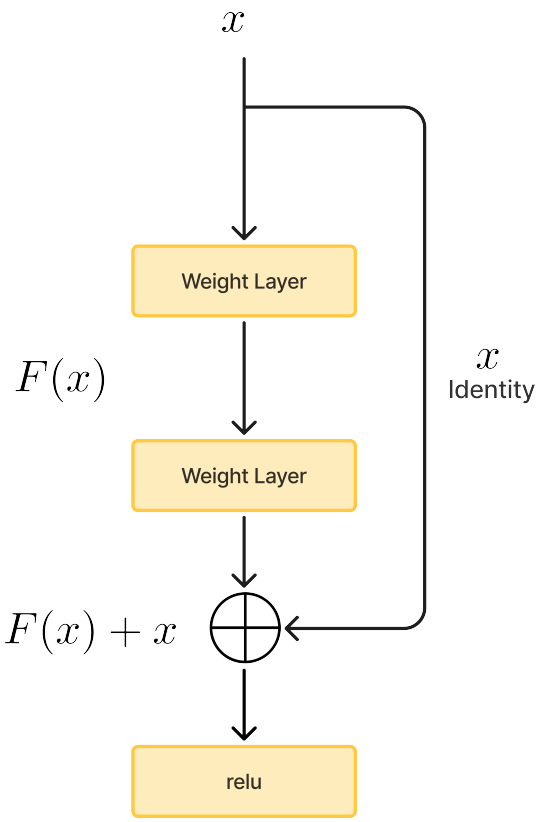


Figure 2.4: The building block of a residual network as presented in He et al. (2016). The $\oplus$ operator denotes elementwise addition.

learning attention-based models were still in the exploratory phase they were poised to gain traction. The models in theory have strengths in being able to focus on the most relevant features across the NIR spectrum, where CNNs focus on local features. Outside of fruit quality prediction, there have been a number of recent applications of transformer models. Singh et al. (2024) used a transformer model to predict protein content in lablab beans. L. Zhang et al. (2025) used an attention head along with residual network and convolutional layers, while Z. Chen et al. (2024) used a transformer network for grain classification.

2.3.1.4 Multimodal models

Another recent trend is to use multimodal models to improve performance by combining NIR with other spectra as input features. Peng et al. (2025) used a residual network to combine NIR with Raman spectroscopy to predict

volatile organic compounds in water. The use of multimodal spectral fusion aims to address the limitations of each. NIR can have overlapping absorption bands and water interference, while Raman spectroscopy has weak signal at low concentrations Peng et al. (2025). Gutiérrez et al. (2023) applied a multi-sensor model to spectra measured from two sensors of 570 - 1000nm and 1100-2100nm, to estimate grape composition, including SSC. Expanding this idea, there is potential to use multimodal models to combine NIR data with other measures, not limited to spectral data. For example, SSC changes over time as starch is converted to sugars so including time since harvest with the NIR spectra in a multimodel network could improve predictions on SSC.

2.4 Data Augmentation

Data augmentation techniques generate additional training data to improve machine learning models' performance on small, poorly representative datasets (Mumuni & Mumuni, 2022).

Blazhko et al. (2021) found that there have been very few studies that have employed data augmentation techniques with deep learning in the NIR spectroscopy field. Bjerrum et al. (2017a) used data augmentation through random variations in the offset, slope and multiplication of spectra from a pharmaceutical pill dataset. They found that CNN models trained on the augmented data outperformed PLS models and that the convolutional filters resembled standard preprocessing techniques. Blazhko et al. (2021) later extended this augmentation method and found that it improved classification accuracy for deep learning models applied to four datasets.

2.4.1 Contrastive Learning

Contrastive learning is a promising self-supervised learning approach that aims to learn features that are close amongst similar unlabelled samples while far apart for dissimilar instances Hu et al. (2024). There are a number of con-

trastive methods, including the Barlow Twins (Zbontar et al., 2021) that are discussed in Chapter 6.

2.5 Calibration Transfer

Calibration transfer is a set of methods used to transfer NIR models between different devices (Mishra et al., 2021) or under different conditions. These techniques are varied but can be categorised into five core approaches (Ramadan et al., 2025). These approaches are summarised in Figure 6.3 along with their respective subgroups. This thesis uses the robust modelling approach, in particular global modelling, for calibration transfer. However, it is not the only suitable deep learning technique. Model adjustment via transfer learning is a common method in deep learning. Mishra and Passos (2021b) had success with calibration transfer by training a deep learning model on one device, then freezing weights for some of the network's layers before fine-tuning the remaining layers on a second instrument's data. This included transferring between two handheld devices for an olive dataset and between two benchtop units.

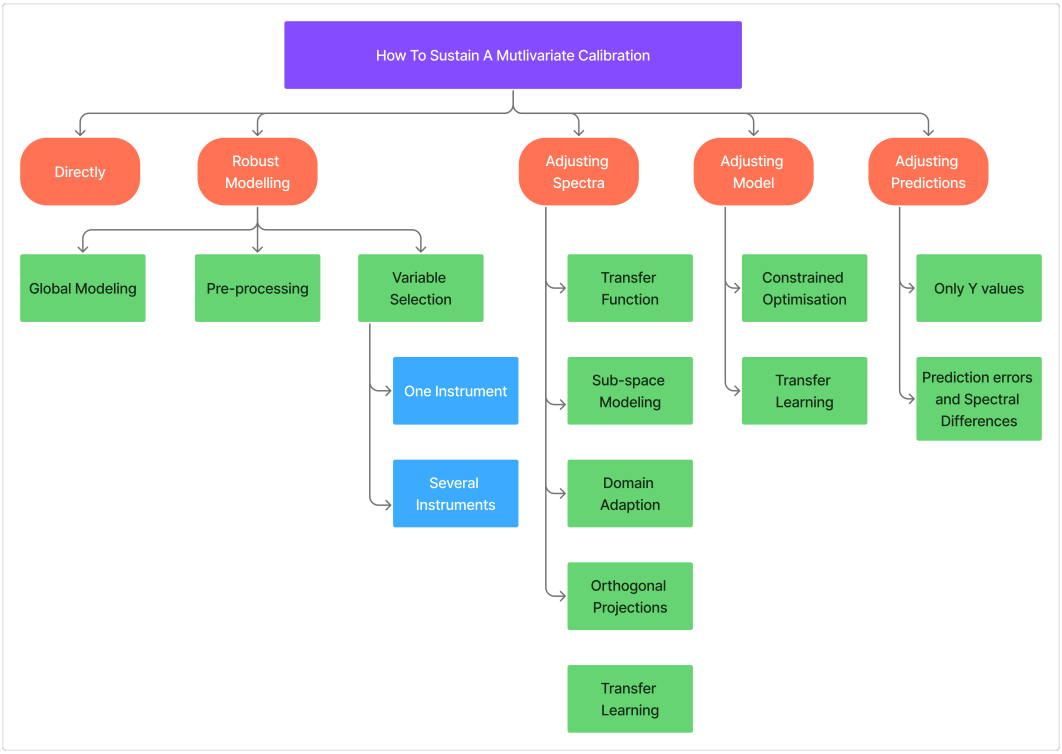


Figure 2.5: Hierarchal categorisation of calibration transfer approaches as presented in Ramadan et al. (2025)

2.6 Cross-Validation Techniques

Validation techniques are used to assess how well the model will perform on data other than the data used for training. This is done by dividing the available data into separate partitions. In general, this involves one partition for training and another for testing. However, this can also include a validation partition for tuning hyperparameters during training. There are several ways these partitions can be defined (see Allgaier and Pryss (2024) for a visual guide to these CV methods). The data used in this thesis involves two datasets, each partitioned into train, validation, and test sets. Further, these datasets are split by time, ensuring the model is validated on data collected at different times and providing an indication of how robust the model is to temporal effects such as device drift.

2.7 Summary of Research Gaps

This chapter identifies a number of key challenges to the wider adoption of deep learning models for NIRS that this thesis directly addresses. The first is that deep learning models generally need large labelled datasets for training. This is particularly time-consuming and expensive in the horticultural setting, where destructive measures, e.g., DMC and SSC, must be collected. Training deep learning models on small datasets is underdeveloped in NIRS applications. Secondly, there has been little work done to leverage the easier to collect unlabelled data or take advantage of the structure of the data collection, for example, multiple devices measuring the same fruit. Finally, there have been some applications of data augmentation in deep learning models for NIR data, but they have focused on individual spectral variations rather than between-device effects. Augmentation based on this may improve the fit of global models. This thesis aims to address these research gaps in Chapters 4, 5, and 6.

Chapter 3

Methodology

The following describes the methodology that is common across the studies outlined in Chapters 4 to 6.

3.1 Datasets

3.1.1 Kiwifruit dataset

This thesis uses a dataset comprising of 5418 kiwifruit measured by five devices across two sites and three seasons from 2017 to 2019. In total, there are 11,982 NIR scans with some fruit being measured by multiple devices. These total numbers differ slightly from what was used in the published papers included in this thesis, as a small number of sampling time points were later added to the dataset. Specific outlier detection and removal for the different experiments are discussed in the respective chapters. The data is divided into training, validation and test sets (see Table 3.1) by sampling date for use in the experiments. Chapter 6 combines all of the 2019 data for testing. All data were measured using the F-750 Produce Quality Meter (Felix Instruments, 2019) (see Figure 3.1), over the 402 to 1137nm range. This was reduced to the 459-1059nm range in the experiments due to missing values and variability. Each fruit had respective soluble solids (SSC) and dry matter content (DMC) destructively measured, although exact methodological details remain

confidential for commercial reasons.

Figure 3.1: An example of a F-750 Produce Quality Meter

Data collection often involved NIR measurements from multiple devices on the same fruit. For the training set data, the particular devices measuring together were variable. A detailed summary of fruit numbers measured by combinations of more than one device is presented in Table 3.2. Note that the later dates, including the validation (yellow) and test (blue), settled into the pattern of TP1, TP2, and TP3, all measuring the same fruit, as do KK1 and KK2.

Date	**KK1**	**KK2**	**TP1**	**TP2**	**TP3**	**Total Scans**	**Fruit**
2017-03-28	90	90	90	90	90	450	90
2017-04-05	100	0	99	100	100	399	255
2017-04-11	100	0	100	0	0	200	100
2017-04-19	100	0	100	100	100	400	200
2017-04-26	100	100	100	100	100	500	200

Continued on next page

Table 3.1 – Continued from previous page

Date	KK1	KK2	TP1	TP2	TP3	Total Scans	Fruit
2017-05-03	100	100	100	100	100	500	200
2017-05-09	100	0	100	0	0	200	100
2017-05-10	0	100	0	100	100	300	100
2017-05-17	100	100	100	100	100	500	200
2017-05-23	100	0	100	0	0	200	100
2017-05-30	100	0	100	0	0	200	100
2017-05-31	0	85	0	85	85	255	85
2017-06-07	100	0	100	0	0	200	100
2018-03-07	0	0	100	100	100	300	100
2018-03-08	100	0	0	0	0	100	100
2018-03-13	100	0	0	0	0	100	100
2018-03-20	100	0	0	0	0	100	100
2018-03-21	0	0	100	100	100	300	100
2018-03-27	100	0	0	0	0	100	100
2018-04-04	100	0	100	100	100	400	200
2018-04-10	100	0	0	0	0	100	100
2018-04-11	0	0	100	100	100	300	100
2018-04-17	100	0	0	0	0	100	100
2018-04-18	0	0	100	100	100	300	100
2018-04-24	100	100	0	0	0	200	192
2018-05-01	0	100	0	0	0	100	100
2018-05-02	0	0	99	99	99	297	99
2018-05-08	100	100	0	0	0	200	100
2018-05-09	0	0	100	100	100	300	100
2018-05-15	100	100	0	0	0	200	100
2018-05-16	0	0	100	100	100	300	100

Continued on next page

Table 3.1 – Continued from previous page

Date	KK1	KK2	TP1	TP2	TP3	Total Scans	Fruit
2018-05-22	100	100	0	0	0	200	100
2018-05-23	0	0	100	100	100	300	100
2019-03-06	100	100	0	0	0	200	100
2019-03-13	100	100	0	0	0	200	100
2019-03-19	100	100	0	0	0	200	100
2019-03-20	0	0	100	100	100	300	102
2019-03-26	100	100	0	0	0	200	100
2019-03-27	0	0	94	94	94	282	94
2019-04-02	100	100	0	0	0	200	100
2019-04-03	0	0	101	100	100	301	101
2019-04-10	100	98	100	100	100	498	200
2019-04-16	100	100	0	0	0	200	100
2019-04-17	0	0	100	100	100	300	100
2019-04-24	100	100	100	100	100	500	200

Table 3.1: Number of fruit measured by each device for training (white rows), validation (yellow rows), and test (blue rows) datasets.

Date	KK1 KK2	TP3, KK1	TP1 TP2	TP1&2, KK2	TP1&2 &TP3	KK1&2, TP1&2&3
2017-03-28	0	0	0	0	0	90
2017-04-05	0	99	45	0	0	0
2017-04-11	0	100	0	0	0	0
2017-04-19	0	100	100	0	0	0
2017-04-26	0	100	0	100	0	0
2017-05-03	0	100	0	100	0	0

Continued on next page

Table 3.2 – Continued from previous page

Date	KK1 KK2	TP3, KK1	TP1 TP2	TP1&2, KK2	TP1&2 &TP3	KK1&2, TP1&2&3
2017-05-09	0	100	0	0	0	0
2017-05-10	0	0	0	100	0	0
2017-05-17	0	100	0	100	0	0
2017-05-23	0	100	0	0	0	0
2017-05-30	0	100	0	0	0	0
2017-05-31	0	0	0	85	0	0
2017-06-07	0	100	0	0	0	0
2018-03-07	0	0	0	0	100	0
2018-03-08	0	0	0	0	0	0
2018-03-13	0	0	0	0	0	0
2018-03-20	0	0	0	0	0	0
2018-03-21	0	0	0	0	100	0
2018-03-27	0	0	0	0	0	0
2018-04-04	0	0	0	0	100	0
2018-04-10	0	0	0	0	0	0
2018-04-11	0	0	0	0	100	0
2018-04-17	0	0	0	0	0	0
2018-04-18	0	0	0	0	100	0
2018-04-24	8	0	0	0	0	0
2018-05-01	0	0	0	0	0	0
2018-05-02	0	0	0	0	99	0
2018-05-08	100	0	0	0	0	0
2018-05-09	0	0	0	0	100	0
2018-05-15	100	0	0	0	0	0
2018-05-16	0	0	0	0	100	0

Continued on next page

Table 3.2 – Continued from previous page

Date	KK1 KK2	TP3, KK1	TP1 TP2	TP1&2, KK2	TP1&2 &TP3	KK1&2, TP1&2&3
2018-05-22	100	0	0	0	0	0
2018-05-23	0	0	0	0	100	0
2019-03-06	100	0	0	0	0	0
2019-03-13	100	0	0	0	0	0
2019-03-19	100	0	0	0	0	0
2019-03-20	0	0	2	0	98	0
2019-03-26	100	0	0	0	0	0
2019-03-27	0	0	0	0	94	0
2019-04-02	100	0	0	0	0	0
2019-04-03	0	0	0	0	100	0
2019-04-10	98	0	0	0	100	0
2019-04-16	100	0	0	0	0	0
2019-04-17	0	0	0	0	100	0
2019-04-24	100	0	0	0	100	0

Table 3.2: Number of fruit measured by device combinations for training (white rows), validation (yellow rows), and test (blue rows) datasets.

3.1.2 Mango Dataset

The experiments also used a dataset comprising 4,675 mangoes and 11,691 scans, as made available by Anderson et al. (2020). This dataset complemented the kiwifruit dataset, as the measurements were recorded with the same device model (F-750) and included DMC, allowing easier comparison across datasets. It has also been widely applied to various modelling techniques, including deep learning, providing useful baseline results for comparison with the methods presented in this thesis. The same training, validation, and test sets specified

in the dataset are used in this thesis. The original dataset comprised a single device measuring fruit from ten cultivars, across two growing regions and four seasons and is what is used here. More recent versions of the dataset (updated May 2024) have greatly increased the number of samples.

3.2 Software and Hardware

Initially, experiments were conducted on a local Windows 10 machine (see Chapter 4 for more details). However, as more recent versions of TensorFlow have stopped Windows support, later work was conducted using Google Colab (Google Research, 2024) instead. This enabled a preconfigured environment with TensorFlow, without the complex setup. A subscription was required to reliably access GPU resources, which greatly sped up the experiments. However, these were limited by monthly credit caps and had to be managed by using cheaper CPU backends to refine experimental code before running the final experiment. One unexpected difficulty encountered was that Google Colab automatically updated packages such as TensorFlow from time to time, requiring code to be rewritten to account for the changes. As such, TensorFlow varied through versions 2.40, 2.13.0 and 2.18.0 across the experiments. Details of other parameters, including learning rates, activation functions, and optimisers, are included in the methods sections of the three papers.

3.3 Evaluation Metrics

Deep learning models in this thesis are trained using the mean squared error (MSE), a common loss function for regression tasks.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.1}$$

where y_i is the observed value (DMC or SSC) for the i^{th} observation, $\hat{y}_i$ is the respective model prediction, and n is the number of samples being pre-

dicted. Lower MSE values indicate better predictive performance.

Results are reported as the root mean squared error (RMSE), which is simply the square root of the MSE.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.2}$$

RMSE is easier to interpret as it is on the original scale (percentage for DMC, °Brix for SSC). If the residuals (prediction errors) are unbiased and approximately normally distributed, then it is expected that around 68% of predictions would be within ±1 RMSE.

3.4 Baseline Model Performance

The mango dataset has been widely analysed and often uses PLSR as a baseline for comparison with deep learning. The previous Chapter outlined several other possible models, which are compared here. The experiments were conducted in R 4.5.1 using the caret, kernlab, pls, proscpectr, randomforest, and xgboost packages. Each model had its hyperparameters optimised via grid search based on performance on the validation (tuning) dataset. Figure 3.2 summarises the results of the optimised model using no, MSC, or Savitzky-Golay (SG) 2nd derivative preprocessing. PLSR performs well across different preprocessing techniques, confirming that it is a reasonable baseline. Interestingly, while SVM performed poorly without preprocessing, it had the lowest overall RMSE of 0.866 with SG (sigma = 0.001 and C = 10).

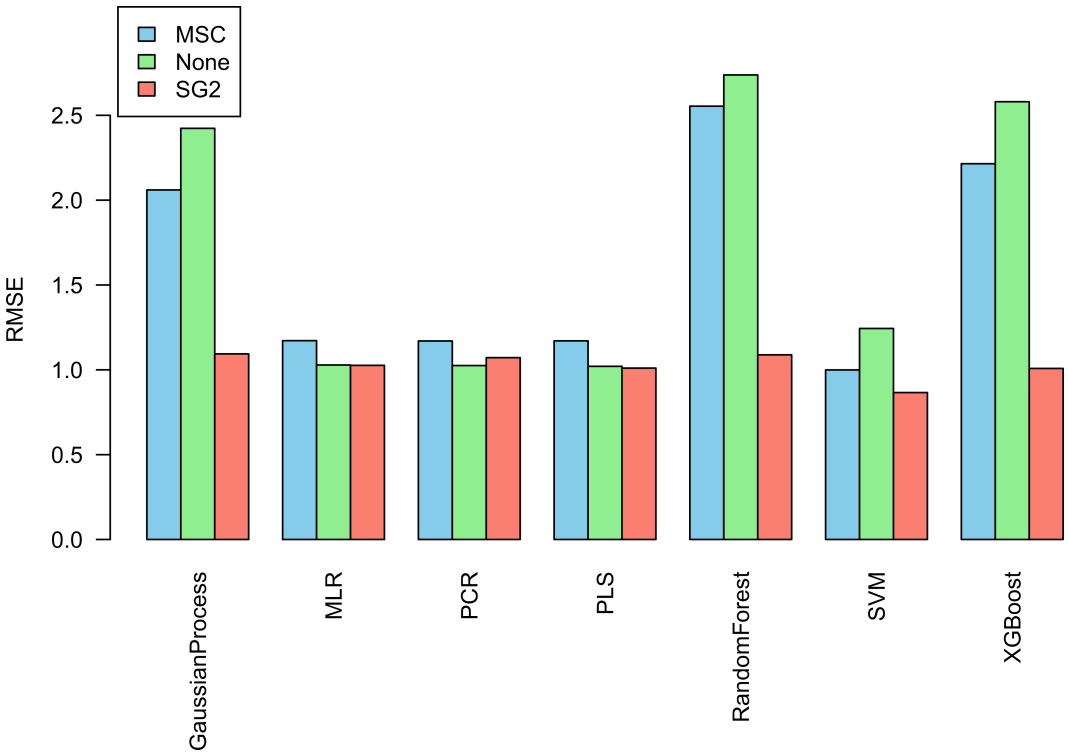


Figure 3.2: Baseline performance on the mango DMC dataset comparing various models in combination with different preprocessing techniques. Each model had its hyperparameters optimised by gridsearch using the validation dataset.

Similar results are observed for the kiwfruit dataset for both DMC (Figure 3.3) and SSC (Figure 3.4). PLSR gives consistent performance across the different preprocessing methods, while SVM gives the best performance when tuned. Using PLSR as a baseline for this research is reasonable due to its popularity and consistent performance demonstrated here. The Savitzky-Golay SVM performance implies that there are performance gains to be made and that properly optimised deep learning models should provide better performance than PLSR on these datasets.

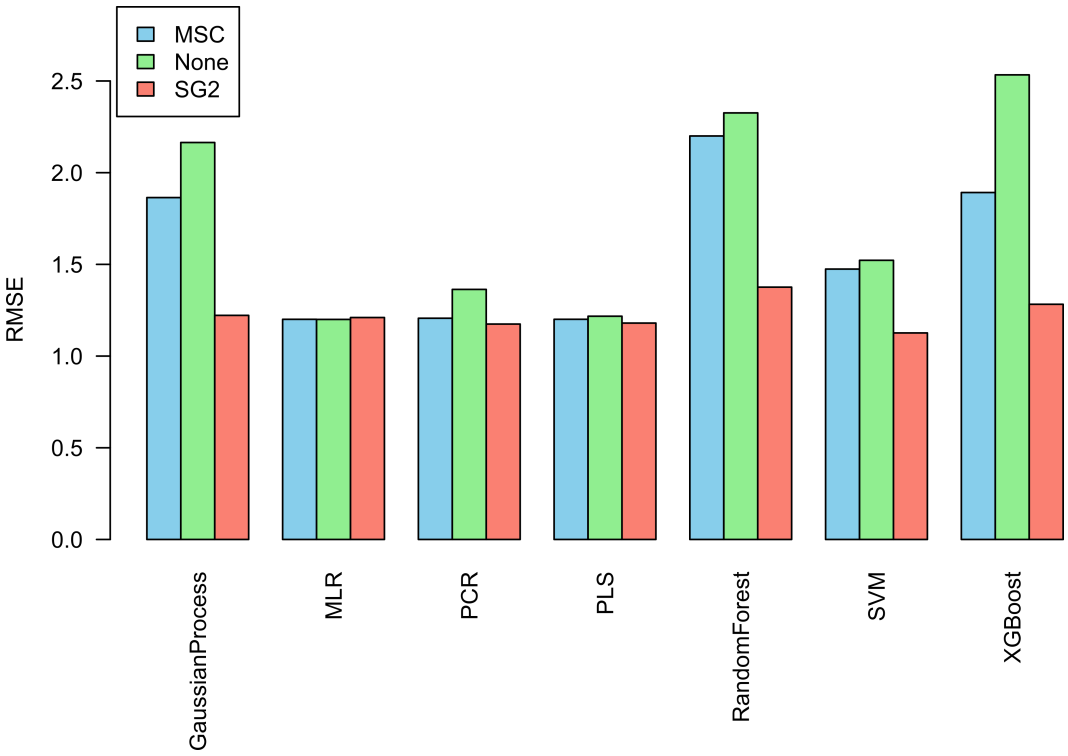


Figure 3.3: Baseline performance on the kiwifruit DMC dataset comparing various models in combination with different preprocessing techniques. Each model had its hyperparameters optimised by gridsearch using the validation dataset.

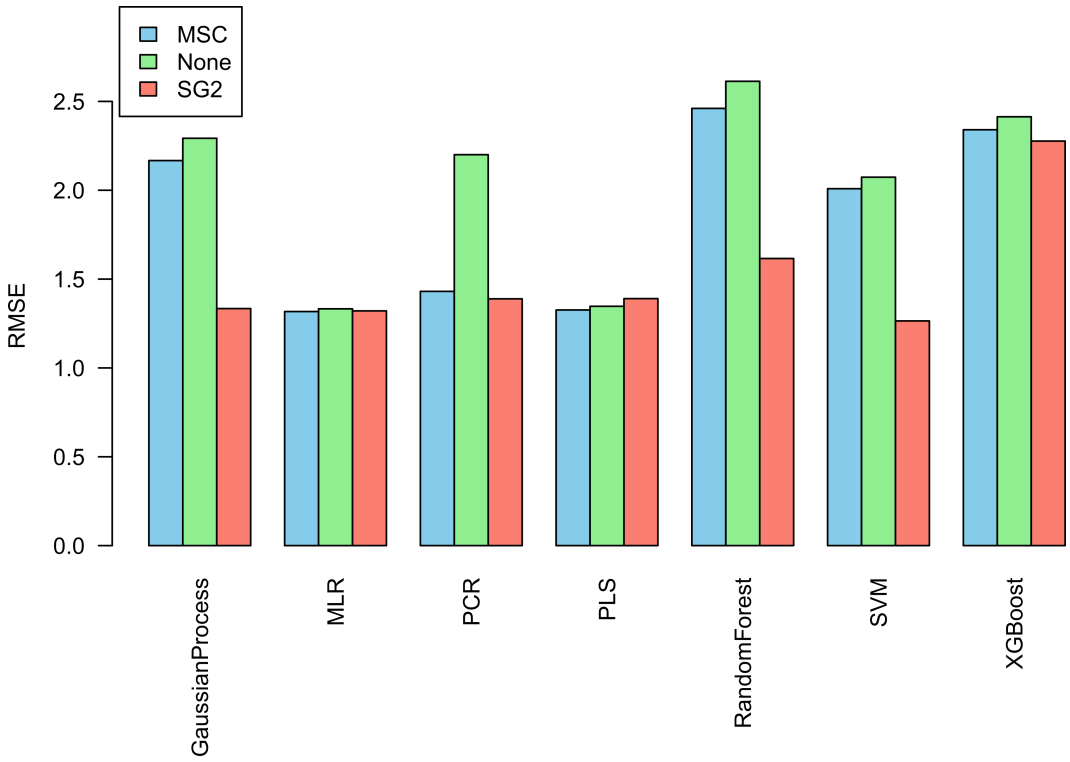


Figure 3.4: Baseline performance on the kiwifruit SSC dataset comparing various models in combination with different preprocessing techniques. Each model had its hyperparameters optimised by gridsearch using the validation dataset.

3.5 Wavelength Range Selection

The wavelength ranges used across the three studies (459-1062 nm in Chapters 4 and 5, 402-1002 nm in Chapter 6.1) were wider than typically recommended in the literature when using PLSR. Anderson et al. (2020) recommended a range of 684-990 nm for the mango dataset, while McGlone et al. (2002) recommended a range of 800-1100 nm when predicting kiwifruit SSC. Similarly, Goke et al. (2018) predicted SSC and DMC for d'Anjou and Bartlett Pear using a narrow range because wavelengths below 700 nm are "noisy and uninformative, likely because of the absorbance by chlorophyll and other pigmentation in the visible range (400–700 nm)". However, this thesis tested whether DL could benefit from using information contained in a wider wave-

length range, such as maturity related colour, through automatic feature selection and noise removal. Baseline experiments were conducted to compare the performance of various ML and preprocessing techniques for the wider and narrower wavelength ranges. Table 3.3 presents the results for PLSR and SVM with Savitzky-Golay second-derivative preprocessing. SVM showed improved performance across the wider range for both the mango and kiwifruit measures, as did PLSR for the mango DMC and kiwifruit SSC, but not for kiwifruit DMC. Recently, J. Walsh (2024) revisited the application of DL to the mango dataset and found benefits to using a wider range, as did L. Zhang et al. (2025). These baseline results supported the decision to use a wider range of wavelengths than is standard for PLSR throughout this thesis.

Table 3.3: Baseline model performance comparing narrow and wide spectral ranges with Savitzky–Golay second derivative preprocessing. Bold values indicate best performance for each task.

Dataset	**Attribute**	**Range**	**Type**	**PLS RMSE**	**SVM RMSE**
Mango	DMC	684-990 nm	Narrow	1.06	0.95
Mango	DMC	465-1065 nm	Wide	**1.01**	**0.87**
Kiwifruit	DMC	702-1002 nm	Narrow	**1.16**	1.15
Kiwifruit	DMC	465-1065 nm	Wide	1.17	**1.12**
Kiwifruit	SSC	702-1002 nm	Narrow	1.43	1.37
Kiwifruit	SSC	465-1065 nm	Wide	**1.39**	**1.26**

Chapter 4

Augmenting NIR Spectra in deep regression to improve calibration

This chapter contains the published paper titled "Augmenting NIR Spectra in deep regression to improve calibration" which appeared in Chemometrics and Intelligent Laboratory Systems 240 (2023) 104924.

Chemometrics and Intelligent Laboratory Systems 240 (2023) 104924

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Augmenting NIR Spectra in deep regression to improve calibration

Mark Wohlers [a,b,*], Andrew McGlone [a], Eibe Frank [b], Geoffrey Holmes [b]

[a] The New Zealand Institute for Plant and Food Research Limited, Auckland, New Zealand
[b] Department of Computer Science, University of Waikato, Hamilton, New Zealand

ARTICLE INFO

Dataset link: https://doi.org/10.6084/m9.figshare.21747983, https://github.com/mwohlers/NIR_augmentation

Keywords:
Near-infrared spectroscopy
Convolutional neural networks
Partial least squares regression
Data augmentation

ABSTRACT

Deep learning, particularly with convolutional neural networks, shows promise in modelling near-infrared spectroscopy (NIRS), but the lack of robust generalisation across instruments often affects performance in practice. Here, we investigate a method to increase the robustness of this approach. The proposed method involves using a simple data augmentation technique during the training process. The performance of convolutional neural network regression is compared to partial least squares regression (PLSR) using kiwifruit data collected from multiple handheld devices over three seasons and mango data collected from a single device over four seasons. The results suggest that data augmentation for NIR spectra can prevent overfitting. In particular, augmenting the training data to mimic spectra collected over multiple devices results in a neural network model with improved performance over PLSR.

1. Introduction

Near-infrared spectroscopy (NIRS) has been a useful tool in non-destructive measurements of fruit quality [1]. The levels at which the NIR wavelengths are absorbed or scattered vary based on the sample's chemical makeup and structure. Using machine-learning modelling methods, recorded spectra can be used to make predictions of various fruit quality attributes such as dry matter content (DMC) or soluble solids content (SSC). The kiwifruit industry uses a minimum DMC as minimum taste standard (MTS) with NIR sorting being successfully employed to recover high DMC fruit from populations failing this standard [2]. Other fruits have similar attribute specifications, for example, it is recommended that mango have an at-harvest DMC of at least 14 [2]. The use of deep learning models to make predictions based on NIR spectra has become increasingly popular [3]. However, fitting the models can require extensive optimisation to arrive at a final architecture and tuning parameters [4], which risks overfitting and poor generalisation, especially if the data set is small.

Most traditional techniques, such as partial least squares regression (PLSR) [5], rely on pre-processing methods [6] to remove nuisance effects such as the confounding influence of light scattering due to variation in internal tissue structure. However, there are other effects such as temperature [7], operator, and random noise that can also influence the absorbances critical to generating accurate models. The devices themselves can be variable, between devices of the same model, or even across time for the same device [8]. The exact sampling position of fruit measurements can greatly affect the spectra as the internal composition and structure are much less uniform than what is typically observed in industrial NIRS analysis, such as that of liquid or powder samples. All of these effects can influence the quality of model predictions and should be taken into account.

There is evidence that deep learning models learn the appropriate pre-processing techniques automatically [3], but they may require larger data sets for training. In practice, access to such data may be limited due to situations where data collection is time-consuming, for example, when measuring fruit with what are typically slow handheld NIR devices [9]. One possible solution that is commonly employed in other domains, such as image classification, is to augment the observed data using synthetic generation to produce a larger training set [10], where training images are routinely altered, for example, by transformations such as rotation, clipping, and zooming, among others. While data augmentation has become increasingly popular and mature in other deep learning applications, there are fewer examples of applying it in the area of NIRS.

Convolutional neural networks can be naturally applied to NIRS data by treating this data as a 1-dimensional signal [11], preserving the wavelength structure in the data. The idea of augmenting NIR spectra is not new and has previously been used with PLSR by adding Gaussian noise [12]. More recently, Bjerrum et al. [13] used convolutional neural networks to classify pharmaceutical samples based on NIR spectra. They found improved performance by augmenting the data using random offset, multiplicative, and slope effects. This technique was later expanded into a more general form [14] and found to perform

* Corresponding author at: The New Zealand Institute for Plant and Food Research Limited, Auckland, New Zealand.
E-mail address: mark.wohlers@plantandfood.co.nz (M. Wohlers).

https://doi.org/10.1016/j.chemolab.2023.104924
Received 18 December 2022; Received in revised form 17 July 2023; Accepted 25 July 2023
Available online 28 July 2023
0169-7439/

well on a number of classification data sets. Mishra et al. [15] used an alternative method whereby raw spectra were altered by various permutations of pre-processing techniques. However, this can perhaps be viewed more accurately as a form of pre-processing rather than data augmentation, as the process does not increase the size of the training set.

Not all applications of augmentation to NIR data provide a positive result: Acquarelli et al. [16] reported that data augmentation did not improve performance. The authors speculate that this may be due to the difficulty of modelling noise as indicated by the range of pre-processing techniques needed to model the various data sets in their experiments. Details of the augmentation used were limited to adding perturbed samples.

An advantage of being able to realistically augment spectra is that it can produce robust models. Currently, variations amongst devices can lead to PLSR and other models being trained for each device individually. A robust model that generalises across devices would allow for new devices to be put to use earlier, reducing the need to collect a large amount of data to train an individual model specifically for each device.

This paper presents an alternative augmentation method for model training that improves generalisation to other devices. This is achieved by simulating changes in absorbances observed when measuring fruit from multiple devices. Augmented data are generated from a multivariate normal distribution (MVN) and incorporated into a data input pipeline API for easy implementation while training deep learning models. The impact on training stability and the performance of trained models is assessed, including a comparison to PLSR.

2. Data augmentation using correlated Gaussian noise

According to Blazhko [14], data augmentation "should strive to produce observations as close as possible to what could be obtained in reality". Gaussian noise has been used to augment NIR spectra in prior work [12]. However, the generation of the noise in that method is equivalent to sampling from a multivariate normal distribution with a diagonal covariance matrix. While the amount of noise added to each wavelength is based on the respective observed standard deviation, the noise between wavelengths is considered statistically independent. This is clearly not ideal when the aim is to produce realistic data: when repeated measurements are taken from the same fruit, it can be seen that spectral deviations from the mean are not independent. This is demonstrated in Fig. 1 where spectra from multiple devices measured on the same fruit are roughly parallel over bands of wavelengths.

For example, spectra with higher than average absorbance at 550 nm will also likely have higher than average absorbance at 650 nm. The left panel of Fig. 2 presents spectra from 205 kiwifruits with the fruits' average spectra subtracted. For the augmentation to generate spectra close to what "could be obtained in reality", a non-diagonal covariance matrix should be used, yielding samples such as the ones shown in the right panel of Fig. 2.

The specification of an appropriate covariance structure for the investigation presented in this paper is based on observed spectra from a previously collected multi-instrument data set. A sample of 205 kiwifruits were each measured once by ten separate devices. The point of measurement across devices was taken from the same side at the equator of the fruit but may vary slightly in exact position. The spectra for each fruit were mean-centred (as in Fig. 2: left panel). The Gaussian noise added for data augmentation was then generated by sampling from an MVN with mean 0 and covariance structure Σ (as in Fig. 2 right panel), where Σ is estimated as the sample covariance matrix of this centred data set. More formally, $\boldsymbol{\Sigma} = [\sigma_{jk}]$ with $\sigma_{jk} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)$ where x_{ij} is the j_{th} wavelength of the i_{th} sample, and $\bar{x}_j$ is the sample mean of wavelength j.

Note that this means the augmentation is based on within-device measurement error as well as spectral variation between devices.

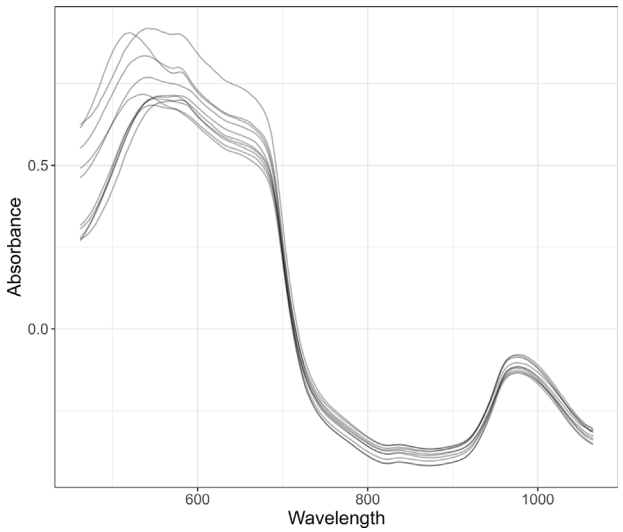


Fig. 1. Example of NIR measurements from multiple devices on the same fruit.

3. Materials and methods

3.1. Datasets

Two data sets are used for the empirical results presented in this paper. Both contain NIR measurements that were recorded using handheld F-750 produce quality meters by Felix instruments. The recorded spectra contained wavelengths between 402 nm and 1137 nm. This was further trimmed to 202 wavelengths from 459 nm to 1059 nm in steps of 3 nm. Outliers were removed from the training set based on Hotelling T2 scores from a PLSR with 20 latent variables and confirmed via visual inspection.

3.1.1. Kiwifruit dataset

The first data set contains spectra from 4956 kiwifruits collected over three seasons from two sites approximately 500 km apart. Depending on the site, each fruit was measured once by two or three devices. After the non-destructive NIRS was performed, two destructive industry-standard fruit quality measures of DMC and SSC for each fruit were recorded. These measurements have been shown to be related to consumer liking responses [17]. The DMC was recorded as a percentage of the fresh weight of an equatorial slice (3 mm thick) taken from the fruit and measured prior to drying in a convection oven at 60 °C for 24 h. The SSC was recorded as the average °Brix, measured by refractometry, of the juice squeezed separately from the stem and stylar ends of the fruit. The data are generally made up of groups of five fruit from the same vine at each time point. A separate analysis indicated that there is a high amount of variation amongst vines, which is possibly due to phenotypic differences such as tricome (hair) density on the skin, shape, core size, etc. The large geographical distance between the two data collection sites can also produce variation in DMC and SSC, among other measures. In total, there were five handheld devices used to measure the fruit across the two sites: two at a Kerikeri site and three at a Te Puke site. Kerikeri is located in the Bay of Islands region of New Zealand, and Te Puke is located in the Bay of Plenty region. The training set was taken as observations collected between 21st March 2017 and 6th March 2019, the validation set was the data collected immediately after the training set until 27th March 2019, and the test set was collected after the 3rd of April 2019. These sets comprise of 3762, 594, and 600 fruit with 8392, 1382, and 1500 total spectra, respectively.

The distributions of DMC measured at each site were slightly different, and as the devices were nested within site, there may be some

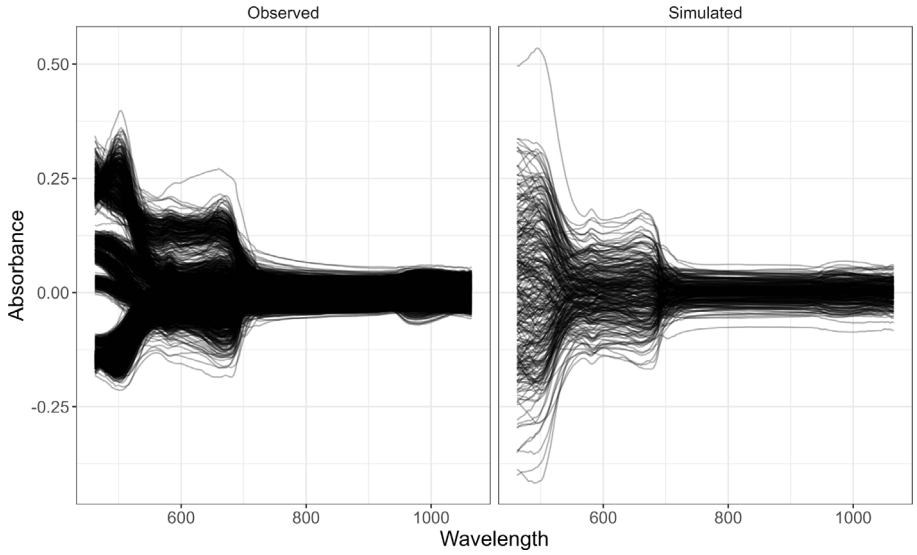


Fig. 2. Example of observed (left) and generated noise from an MVN distribution with the respective empirical covariance matrix (right).

confounding of device and site effects. This could result in a model erroneously using information about a site-specific device to predict SSC and/or DMC and may negatively affect the cross-site predictive performance. This was investigated by training models based on the training data above from each of the two sites and testing on the respective other site.

The estimation of the covariance structure used for augmentation of the training data, discussed in the previous section, is based on a historical data set collected prior to 21st March 2017. More specifically, we had access to a data set where all five devices measured 205 fruit in 2017. These data were not included in the training, validation, or testing sets to retain the independence of the covariance matrix.

3.1.2. Mango dataset

The second data set contains DMC measurements from 4675 mangoes with multiple scans per fruit, resulting in 11,691 NIRS spectra, taken from supplementary material in [18,19]. The data were collected from ten cultivars over four seasons from two Australian growing regions using one F-750 device. The data set was analysed using deep learning in [15], which provides a useful baseline to compare our augmentation method to. However, Mishra and Passos [15] used additional outlier removal that we did not replicate here, so our results using their methods, which we re-implemented, likely differ from the source text. The dataset used in [15] is freely available from the author's github site.[1] In our investigation, the training, tuning, and validation sets were used as specified in the source data set [18].

The mango dataset contained spectra from a single device so the covariance could not be estimated (as was done for the kiwifruit dataset). With the assumption that the variation amongst devices would be similar for different fruit relative to the absorbance levels, we used a scaled version of the data set 1 augmentation method for the mango data set. More specifically, we scaled the covariance matrix used in the augmentation of the mango data: each wavelength was scaled by its respective standard deviation in the kiwifruit training set. The mango data set was then normalised, as further outlined below, so that augmentation was done on the same relative scale as for the kiwifruit data set.

[1] https://github.com/dario-passos/DeepLearning_for_VIS-NIR_Spectra/tree/master/notebooks/Tutorial_on_DL_optimization/datasets.

3.2. Methods

Deep learning models were fit using TensorFlow 2.4.0 on a desktop computer containing an Intel Xeon E3-1270 CPU, 64 GB of main memory, and an NVIDIA Quadro M4000 GPU with 8 GB of device memory. Partial least squares regression (PLSR) was conducted using the Scikit-learn package [20]. The Hyperopt [21] package performed the hyperparameter optimisation over 100 iterations for each outcome measure. Each deep learning model was run for 100 epochs using the Adam [22] optimiser. During hyperparameter tuning, the He Normal method was used to initialise model weights. However, this was changed to He Uniform for the final models as it was later found to give more consistent convergence. All analyses were conducted in Python 3.6.

3.2.1. Pre-processing

Prior work presented in [3] suggests that the convolutional layers applied to NIR data learn the appropriate pre-processing operations and, as such, convolutional neural networks do not require the methods usually employed with such data. However, we did apply column-wise normalisation to both data sets, such that every feature had zero mean and a standard deviation of one. This normalisation was found to improve model convergence. When using the method of Bjerrum et al. spectra were insteadaltered by subracting the grand mean and dividing by twice the overall standard deviation as described in [13]. For the PLSR models, a Savitzky–Golay second-derivative filter, itself a convolutional filter, with a window size of 17, and polynomial order 2, was used for the kiwifruit data set. Other methods tried were varying window size, taking the first derivative, and using unaltered spectra. These methods were not found useful.

3.2.2. Data augmentation

The data augmentation was implemented using the TensorFlow Dataset API. This created a pipeline where the training data are read from disk, shuffled, and divided into random (mini-)batches to fit on the GPU for training. A random sample from the specified MVN is added by applying the TensorFlow probability library together with the tf.data.Dataset data pipeline. This allows the random addition to be repeated a number of times for each instance in the batch, with only augmented instances being used in model training. For hyperparameter tuning, 50 augmented spectra were generated per observed instance due to time constraints of model fitting. This was increased to 100 for

Table 1
Hyperparameter search space.

	Parameter	Search space
1	Convolutional layers	{1, 2}
2	Dense layers	{0, 1, 2}
3	Batch size	{100, 500, 1000}
4	Learning rate	0.0001–0.1
5	Generated data size	{1, 10, 30, 50}
6	Conv. layers: no. kernels	2–40
7	Conv. layers: filter size	5–150
8	Dense layers: no. neurons	4–1000

final model training with optimised parameters. For comparison, the method described in [13], where spectra are altered through random offset, slope, and multiplication transformation, was also used.

3.2.3. Deep learning

The architecture used in the deep-learning models consisted of a combination of convolutional layers followed by dense layers, with a final dense layer with a single neuron exhibiting a linear activation function, as is typical when neural networks are applied to regression problems. All other layers used a rectified linear unit (ReLU) activation function.

The models used the approach presented by [13], where the number of layers was chosen through optimisation. This is discussed in more detail in the next subsection. Kiwifruit final models were trained for 1000 epochs at a set learning rate. Performance on the validation set was inspected to ensure the appropriateness of this training regime. For the mango data set, a learning rate scheduler was used to be consistent with the analysis used in [15]. This meant the learning rate was halved if there was no improvement in the validation loss after 25 epochs, and training halted if no improvement was found after 50.

3.2.4. Hyperparameter tuning

The performance of convolutional networks can be influenced by the particular architecture used and other hyperparameters such as the learning rate, batch size, and the amount of augmentation performed. In particular, finding an appropriate network architecture involves tuning a number of hyperparameters, which can be difficult and time-consuming. Bayesian optimisation was used for this task as it is more time-efficient when model training is slow compared to methods such as grid or random searches [21]. A full list of the search space for the hyperparameters is given in Table 1. It is similar to the one presented in Bjerrum et al. [13] with the exception that it also includes the number of convolutional and dense layers to be used in the model as parameters to be tuned. The minimum validation set mean square error (MSE) over the full number of epochs was used as the criterion to be optimised. To reduce the variability of this measure, a moving average smoother with window size 10 was applied to the MSE prior to calculating the minimum. The stability of the optimisation process was further improved by running each hyperparameter configuration three times and taking the mean performance. The Bayesian optimisation was conducted using the Hyperopt package [21] in Python. Only the training and validation data were used during hyperparameter optimisation, not the test set, to avoid optimistic performance estimates on the test set. Note that due to the large training time, a relatively small number of hyperparameter iterations is performed considering the size of the search space, which may result in non-optimal solutions presented here. The optimisation time was highly dependent on the architecture being evaluated. It took approximately a week to evaluate the 100 iterations, with the bulk of the time spent on those models with a large number of convolutional filters across multiple layers.

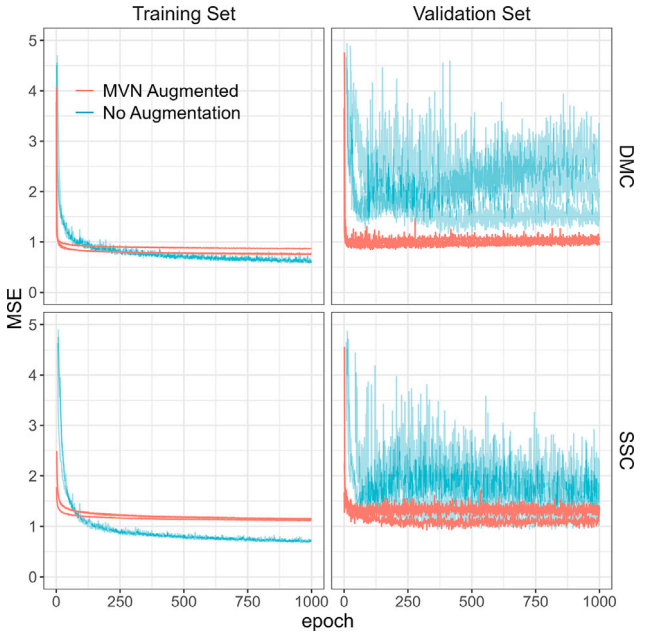


Fig. 3. Comparison of results with and without data augmentation. The left column shows MSE on the training data, the right column MSE on the validation data.

3.2.5. Partial least squares regression

PLSR with second derivative Savitzky–Golay pre-processing was used as a baseline to compare against deep learning. The number of latent variables used in the PLSR models was selected based on the minimum MSE of the respective validation dataset. We also considered unaltered spectra with no pre-processing or augmentation applied. Unless specified otherwise, PLSR results include second derivative Savitzky–Golay preprocessing.

4. Results and discussion

4.1. Kiwifruit results

First, we briefly discuss the outcome of hyperparameter tuning. Then, we study the effect of data augmentation on training the neural networks and PLSR, respectively, before comparing them. Cross-validation of the kiwifruit data by site is also considered.

4.1.1. Selection of network architectures

For dry matter, the optimal architecture for the convolutional neural network found during the optimisation required only a single convolutional layer of 122 filters with a kernel size of 13, followed by a single linear output neuron. Interestingly, after training this architecture for 200 epochs, only 10 of the convolutional filters ever fired over all training, validation, and test data sets. Based on this result, a reduced model was trained with only ten filters and a lower learning rate (0.001). It provided similar performance but was more robust to overfitting.

The hyperparameter tuning for predicting SSC gave a solution with two convolutional layers with 113 filters of size 32 and 93 filters of size 28, respectively. Again, no dense layer was needed. Investigating the output of the filters, all in the first convolutional layer provided a non-zero output for at least one observation of the training set. However, the second convolutional layer consisted of only two filters that gave non-zero output and so the model could be pruned. However, it was decided to leave this architecture unchanged.

As discussed above, hyperparameter optimisation was based on the mean of three runs of a given configuration. During this process, it was noted that there were occasions where two of the runs had a

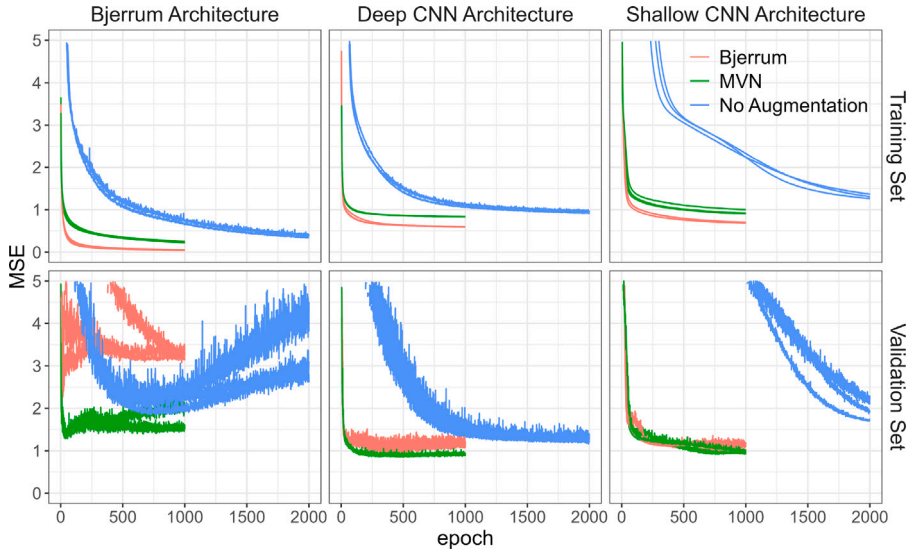


Fig. 4. Comparison of the effect of different data augmentation methods on model fitting to predict DMC using different CNN architectures: deep, shallow, and that outlined in Bjerrum et al.. The top row shows MSE on the training data for three runs of each data augmentation method; the bottom row shows MSE on the validation data.

Table 2
Results of PLSR trained on non-augmented data and PLSR trained on MVN-augmented data, from either Te Puke (TP1, TP2, TP3 training sets) or Kerikeri (KK1, KK2 training sets), based on both, the covariance excluding all devices from the test site (excl test) and including all devices. The number of latent variables (LVs) for each of the three models is presented separated by "/". RMSE presented in brackets is from the respective PLSR model with 30 LVs. All PLSR models used Savitzky–Golay 2nd derivative pre-processing.

Dataset			PLSR RMSE			
Training	Test	Measure	No-Aug	MVN Aug (excl test)	MVN Aug	LVs
Kerikeri	TP1	DMC	1.26(1.29)	1.49(1.36)	1.13(1.14)	29/84/89
Kerikeri	TP2	DMC	1.15(1.15)	1.72(1.06)	1.14(1.19)	29/84/89
Kerikeri	TP3	DMC	1.48(1.48)	2.48(1.75)	2.00(2.34)	29/84/89
Te Puke	KK1	DMC	0.90(0.87)	0.96(0.99)	0.96(0.95)	18/16/16
Te Puke	KK2	DMC	1.06(1.04)	1.36(1.76)	1.05(1.04)	18/16/16
Kerikeri	TP1	SSC	4.15(4.23)	3.22(1.26)	1.19(1.25)	38/86/81
Kerikeri	TP2	SSC	6.31(5.00)	3.29(1.28)	1.2(1.29)	38/86/81
Kerikeri	TP3	SSC	5.98(4.49)	3.8(2.71)	1.46(1.49)	38/86/81
Te Puke	KK1	SSC	1.27(3.37)	2.33(1.45)	1.18(1.22)	12/43/48
Te Puke	KK2	SSC	2.42(4.99)	2.28(2.10)	1.48(1.58)	12/43/48

Table 3
Results of Deep CNN trained on non-augmented data, data augmented using the method of Bjerrum, and MVN-augmented training data, from either Te Puke (TP1, TP2, TP3 training sets) or Kerikeri (KK1, KK2 training sets), based on both, the covariance excluding all devices from the test site (excl test) and including all devices.

Dataset			Deep CNN RMSE			
Training	Test	Measure	No-Aug	Bjerrum Aug	MVN Aug (excl test)	MVN Aug
Kerikeri	TP1	DMC	2.63	1.14	1.43	**1.09**
Kerikeri	TP2	DMC	2.15	2.7	**1.08**	1.17
Kerikeri	TP3	DMC	2.75	2.85	**1.93**	2.88
Te Puke	KK1	DMC	1.43	1.33	**0.92**	0.98
Te Puke	KK2	DMC	1.46	1.78	1.8	**1.07**
Kerikeri	TP1	SSC	1.6	3.81	2.17	**1.2**
Kerikeri	TP2	SSC	1.71	5.65	1.48	**1.29**
Kerikeri	TP3	SSC	1.64	6.41	2.12	**1.35**
Te Puke	KK1	SSC	2.35	1.84	1.73	**1.13**
Te Puke	KK2	SSC	2.51	1.66	2.14	**1.44**

low MSE, but the third did not converge with both high training and validation errors resulting in a poor overall loss. On these occasions, it was obvious that the model optimisation process was not progressing as the training error was stuck at a value far greater than if all predictions were set to the grand mean. This does cast some doubt on the optimised parameters as perhaps a better configuration that was missed could be superior if different initialisation or optimisation routines were used. Because the hyperparameter space did not permit very deep architectures, a deeper model was also trained to evaluate the suitability of the MVN augmentation for different architectures. This consisted of five convolutional layers using exponential linear unit (ELU) activation functions with 4, 4, 8, 16, and 24 filters of sizes 9, 6, 7, 5, and 3 respectively. Each convolutional layer was followed by batch normalisation. This was connected to a single dense layer with a single neuron and linear activation function.

4.1.2. Augmentation for deep learning

Training the shallow and deep convolutional models was greatly improved using the MVN augmented data for both DMC and SSC. Fig. 3 shows the training history for a deep architecture predicting DMC and SSC on the training and validation sets with three runs for non-augmented and MVN augmented data. Training with observed data alone quickly leads to overfitting and poor performance on the validation data.

Additionally, the data were augmented using the method described in Bjerrum [13] for comparison, see Fig. 4. Here we compare three architectures: the one used in Bjerrum et al. [13], our (optimised) shallow CNN architecture, and a deep architecture. For both the optimised and reduced architectures, the data set augmented with our method (MVN) consistently reached a lower validation MSE and reached it in fewer epochs. This was not true of the training set, where the Bjerrum augmentation gave the lowest MSE, implying that MVN augmentation provided some protection against overfitting. Similar observations can be made for the Bjerrum architecture.

4.1.3. Kiwifruit cross-validation by site

Fitting PLSR models to the MVN augmented data set was not consistently better than fitting on the Savitzky–Golay second derivative pre-processed data, see Table 2, in contrast to the consistently positive effect of augmentation in the case of CNNs (Table 3). Note that the results are based on PLSR models trained on devices from the training site only and tested on all data from the excluded site. The MVN augmented data considered thus far uses a covariance matrix for all devices, including those from the test site. Here, augmented results based on a covariance matrix excluding the test set devices are also presented. The results are inconsistent for DMC prediction, as some

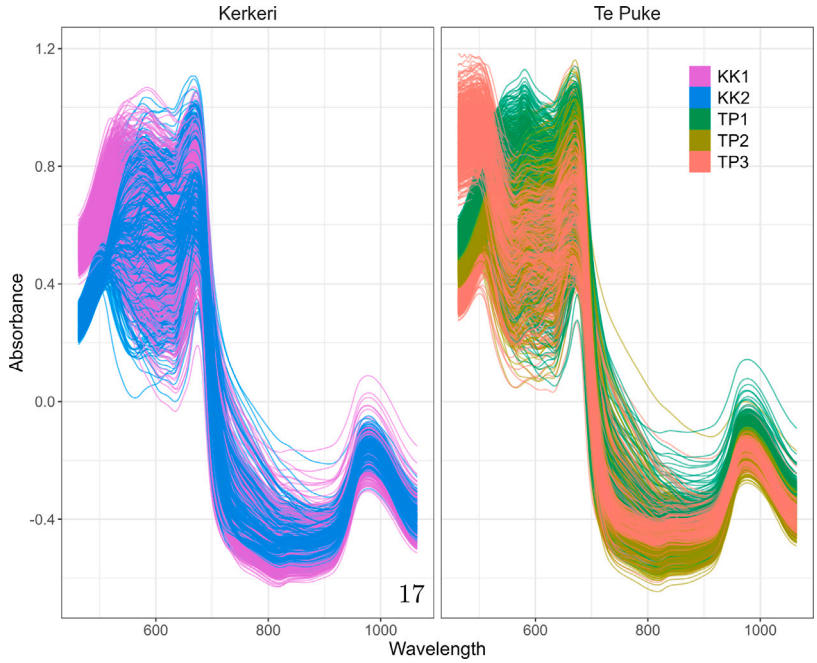


Fig. 5. NIR spectra by device. Scans within site are on the same fruit.

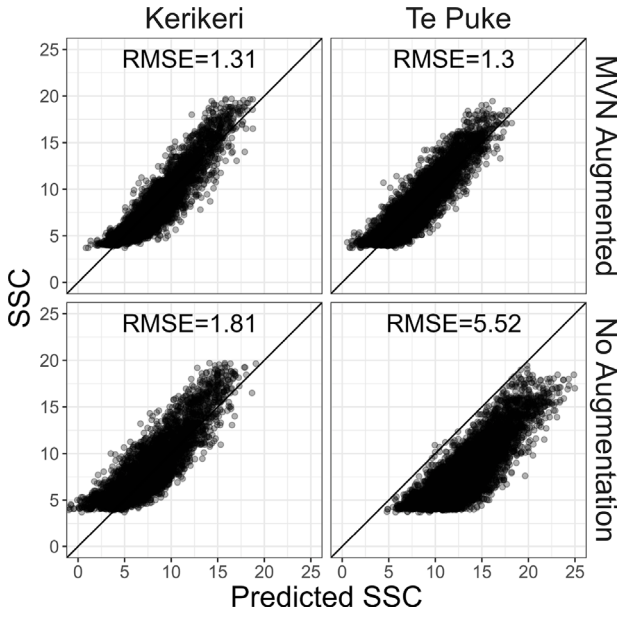


Fig. 6. PLSR test set SSC predictions based on non-augmented (bottom) or augmented (top) data calibration. In the left column are models trained on Te Puke data and tested on Kerikeri data. In the right column are models trained on Kerikeri data and tested on Te Puke data.

devices have lower RMSE for models trained on the MVN augmented data, while others are higher when compared to models trained on the non-augmented data. TP3 in particular had poorer DMC predictions when MVN augmentation was applied. In general, the MVN augmentation performs superior predicting SSC. In almost all tests, the full covariance augmentation performed better than the covariance with the test devices excluded. This was consistent with the results from the deep CNN models shown in Table 3. It indicates that the augmentation method is sensitive to the covariance matrix, and it is beneficial to include the devices of interest in its estimation and/or estimate on a wide range of similar devices.

Table 4
Kiwifruit DMC and SSC test set results on 600 individual fruit (300 per site). PLSR includes Savitzky–Golay 2nd derivative pre-processing. CNN's both use MVN augmentation.

Device	N	DMC RMSE			SSC RMSE		
		Shallow CNN	Deep CNN	PLSR	Shallow CNN	Deep CNN	PLSR
TP1	300	**1.23**	1.32	1.29	1.47	1.22	**1.21**
TP2	300	**1.28**	1.36	1.38	1.38	**1.33**	1.54
TP3	300	1.29	1.35	**1.00**	1.60	**1.28**	1.29
KK1	300	**1.00**	1.07	1.28	1.31	**1.30**	1.33
KK2	298	0.91	0.96	**0.82**	1.35	**1.31**	1.40

Interestingly, when applying PLSR to augmented data, the optimum number of latent variables (LV) found to minimise the RMSE of the non-augmented validation data set was often large (Table 2). Including excessive LVs in the model has the risk of overfitting; this appears true when augmenting with the covariance matrix excluding the test site devices. The test RMSE improved when using fewer LVs, such as the optimal number found with the non-augmented data. For comparison, results are also presented in brackets using 30 LVs. This number is based on observed performance in previous analyses with the same devices. Using the full covariance matrix in the augmentation seemed more robust and did not see a dramatic improvement when reducing the number of LVs. A possible explanation for the overfitting is that the validation data includes data from the same devices as the rest of the training data while the test set contains data from a different device. Nevertheless, caution is recommended when selecting the number of latent variables while applying PLSR to augmented data.

Large differences occurred between sites and devices in terms of the spectra produced. In particular, the Te Puke site had one device, TP3, that had substantially higher absorbance levels than the other devices in the lower wavelengths (Fig. 5). This is reflected in the performance estimate obtained when performing cross-validation on a per-site basis. When TP3 was not included in the training set, the subsequent validation predictions on that device were far poorer than others. When it was included in the training, it did not have a detrimental effect on predicting the Te Puke sites. However, using the MVN augmentation method, PLSR improved significantly, see Fig. 6, by reducing the offset in the predictions.

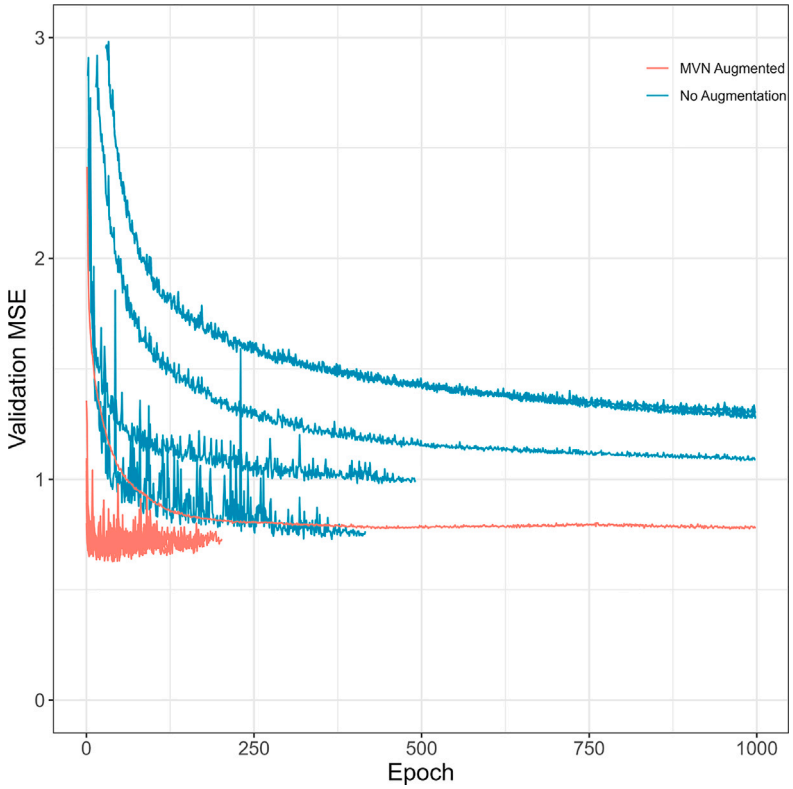


Fig. 7. Comparison of validation MSE for shallow model trained on non-augmented and MVN augmented mango data. Training includes early stopping in some cases. The figure includes the histories of ten runs for each model trained on the observed and augmented data. Note that some lines overlap.

Table 5
DMC and SSC RMSE of the test set for PLSR, shallow CNN models, and deep CNN models, trained on data that was not augmented or pre-processed, Savitzky–Golay 2nd derivative data, data augmented using Bjerrum et al.'s method, or data augmented using MVN.

Model	Measure	No Aug	Savitzky–Golay 2nd	Bjerrum Aug	MVN Aug
PLSR	DMC	1.20	1.18	1.22	**1.16**
Shallow CNN	DMC	19.38	18.61	1.45	1.20
Deep CNN	DMC	1.28	1.47	1.23	1.22
PLSR	SSC	1.33	1.36	1.34	1.37
Shallow CNN	SSC	1.77	2.25	1.71	1.40
Deep CNN	SSC	1.72	1.44	1.78	**1.29**

4.1.4. Comparison of CNN and PLSR

The results of the PLSR and deep learning models, when trained on the entire training set of all devices and tested on the respective test data, applying augmentation based on the covariance matrix established using all devices, are summarised in Table 4. In general, the deep convolutional neural network resulted in lower RMSE across the SSC test set, and to a lesser extent, DMC test set, compared to the PLSR model.

Results of models with and without augmentation and pre-processing trained on the full training set, with data from both sites included, are summarised in Table 5. Both the deep models as well as the shallow models found using hyperparameter optimisation benefit from MVN augmentation. To a lesser extent the augmentation method of Bjerrum et al. improved performance in the DMC predictions for the CNN models. PLSR however showed little change in performance across the different pre-processing techniques. This is somewhat surprising as Savitzky–Golay pre-processing proved effective when separately fitting individual PLSR models to each device (data not shown). We observed that CNNs in general worked better than PLSR which is consistent with other studies [3,13,23].

4.2. Mango results

Similar to the kiwifruit results, training on the augmented data led to faster convergence on the Mango data. The validation training history of the shallow architecture, inherited from the kiwifruit experiments for DMC, is shown in Fig. 7. The difference in the length of the lines is due to the learning rate scheduler, which will reduce the learning rate and later terminate training if no further improvement in the validation MSE is found. The augmented data improves the training of the network. This improvement was also observed while training the deep network albeit with a higher RMSE than the shallow network. This is consistent with the kiwifruit data, where a simpler model was sufficient in predicting DMC. The simpler architecture of a single convolutional layer performed well compared to the PLSR models. Of the PLSR models, the 2nd derivative Savitzky–Golay method produced slightly better overall results and is reported here.

Fig. 8 shows the validation and test performance of repeated training of shallow and deep networks with and without augmentation. The PLSR baseline RMSEs are also included for reference. It can be seen that the deep-learning models without data augmentation are variable in terms of prediction RMSE. The shallow deep-learning model with only one convolutional layer had consistently better predictions when augmentation was used. However, the more complex deep model took far longer to train, providing worse results, if only slightly, than the PLSR model. A comparison of the test set predictions for the PLSR and shallow network is shown in Fig. 9, with the shallow CNN model providing a better fit when comparing RMSE.

5. Conclusion

We have presented a data augmentation method based on sampling spectra from a multivariate normal distribution with empirically estimated covariance matrices. The primary benefit was observed when training convolutional neural networks: faster convergence and better

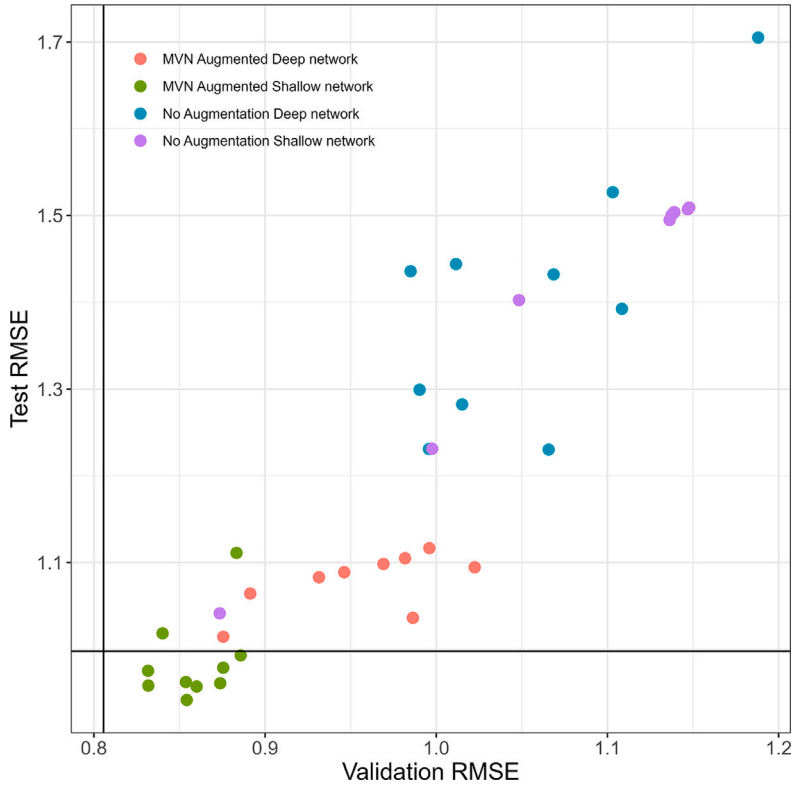


Fig. 8. Validation vs. test RMSE, lines represent PLSR RMSE. Each point represents a randomly initialised model for the given architecture, with and without augmentation. The horizontal and vertical black lines indicate the PLSR test and validation RMSEs respectively.

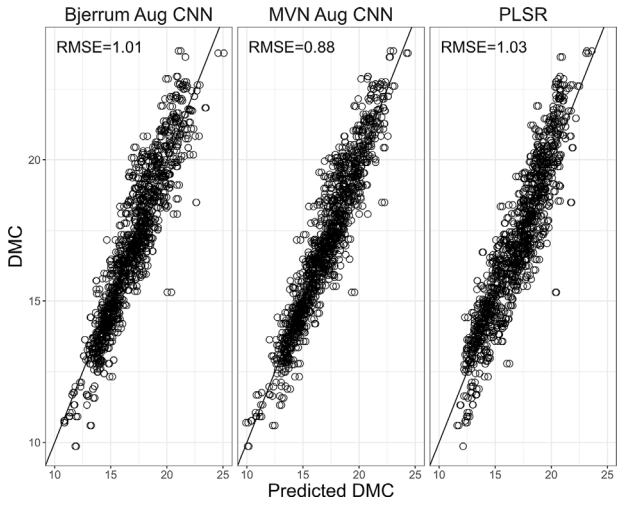


Fig. 9. Comparison of model fits on the Mango test set for a shallow CNN trained on Bjerrum augmented data (left), MVN augmented data (middle) and PLSR with Savitzky–Golay second derivative pre-processing (right).

validation performance were observed. The PLSR analysis also benefited from the same augmentation when the covariance matrix included information relating to between-device information. This proved useful when predicting on devices not included in the training set, albeit with most of the performance increase observed when the test devices were included in the covariance matrix estimation.With a reasonable estimate of the covariance matrix, there is potential to apply the method to other devices of the same type, irrespective of the measurement unit. This potential was demonstrated by successfully applying the technique to the independent Mango dataset.

The augmentation was particularly useful in training convolutional neural networks, both shallow and deeper architectures which is consistent with the results found by [13].

The inclusion of other sources of variation in the covariance matrix should be investigated. As the devices are nested within the site, the effects of operators and sites, and the cultivars measured there, are confounded with device. Therefore, while the augmentation simulates variation due to devices, it would be more robust to include other sources. This would require a new set of repeated measurements on fruit under various scenarios.

The suggested augmentation technique may help improve the quality of non-destructive measures of fruit quality, which will aid recovery of fruit that would otherwise be deemed to not meet an industry export standard.

CRediT authorship contribution statement

Mark Wohlers: Conceptualization, Formal analysis, Methodology, Software, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Andrew McGlone:** Conceptualization, Resources, Writing – review & editing. **Eibe Frank:** Conceptualization, Writing – review & editing, Supervision. **Geoffrey Holmes:** Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used for the kiwifruit experiments can be found at https://doi.org/10.6084/m9.figshare.21747983 with code available at https://github.com/mwohlers/NIR_augmentation.

Acknowledgements

We thank Harpreet Kaur (The New Zealand Institute for Plant and Food Research Limited) and Liz Popowski at the kiwifruit breeding centre (KBC) for their help collating the kiwifruit measurements and overseeing the collection of such a rich data set. We also thank the KBC staff of Susan Murphy at Te Puke, Ann Krebs, Gustavo Hernandez-Grijota, Lisa Anderson at Kerikeri for collecting the NIR spectra over a number of seasons, as well as the KBC fast lab staff who collected the SCC and DMC data. The research received financial support from the New Zealand Ministry of Business Innovation & Employment (MBIE) as part of the Endeavour funded project *Perfecting storage life prediction for delivery of high quality fruit*.

References

[1] H. Wang, J. Peng, C. Xie, Y. Bao, Y. He, Fruit quality evaluation using spectroscopy technology: A review, Sensors (Basel Switzerland) 15 (5) (2015) 11889, http://dx.doi.org/10.3390/S150511889, URL /pmc/articles/PMC4481958/ /pmc/articles/PMC4481958/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4481958/.

[2] K.B. Walsh, V.A. McGlone, D.H. Han, The uses of near infra-red spectroscopy in postharvest decision support: A review, Postharvest Biol. Technol. 163 (2020) 111139, http://dx.doi.org/10.1016/J.POSTHARVBIO.2020.111139.

[3] C. Cui, T. Fearn, Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration, Chemometr. Intell. Lab. Syst. 182 (2018) 9–20, http://dx.doi.org/10.1016/j.chemolab.2018.07.008.

[4] X. Zhang, J. Yang, T. Lin, Y. Ying, Food and agro-product quality evaluation based on spectroscopy and deep learning: A review, Trends Food Sci. Technol. 112 (2021) 431–441, http://dx.doi.org/10.1016/J.TIFS.2021.04.008.

[5] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (C) (1986) 1–17, http://dx.doi.org/10.1016/0003-2670(86)80028-9.

[6] A. Rinnan, F.v.d. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TRAC Trends Anal. Chem. 28 (10) (2009) 1201–1222, http://dx.doi.org/10.1016/J.TRAC.2009.07.007.

[7] K.B. Walsh, J. Blasco, M. Zude-Sasse, X. Sun, Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use, Postharvest Biol. Technol. 168 (2020) 111246, http://dx.doi.org/10.1016/J.POSTHARVBIO.2020.111246.

[8] E. Bouveresse, D.L. Massart, Standardisation of near-infrared spectrometric instruments: A review, Vib. Spectrosc. 11 (1) (1996) 3–15, http://dx.doi.org/10.1016/0924-2031(95)00055-0.

[9] H. Kaur, R. Künnemeyer, A. McGlone, Comparison of hand-held near infrared spectrophotometers for fruit dry matter assessment:, 25 (4) (2017) 267–277. http://dx.doi.org/10.1177/0967033517725530. URL https://journals.sagepub.com/doi/10.1177/0967033517725530.

[10] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 1–48, http://dx.doi.org/10.1186/S40537-019-0197-0/FIGURES/33, URL https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0.

[11] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, D.J. Inman, 1D convolutional neural networks and applications: A survey, Mech. Syst. Signal Process. 151 (2021) 107398, http://dx.doi.org/10.1016/J.YMSSP.2020.107398.

[12] A.K. Conlin, E.B. Martin, A.J. Morris, Data augmentation: an alternative approach to the analysis of spectroscopic data, Chemometr. Intell. Lab. Syst. 44 (1–2) (1998) 161–173, http://dx.doi.org/10.1016/S0169-7439(98)00071-9.

[13] E.J. Bjerrum, M. Glahder, T. Skov, Data augmentation of spectral data for convolutional neural network (CNN) based deep chemometrics, 2017, URL http://arxiv.org/abs/1710.01927.

[14] U. Blazhko, V. Shapaval, V. Kovalev, A. Kohler, Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra, Chemometr. Intell. Lab. Syst. 215 (2021) 104367, http://dx.doi.org/10.1016/J.CHEMOLAB.2021.104367.

[15] P. Mishra, D. Passos, A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit, Chemometr. Intell. Lab. Syst. 212 (February) (2021) http://dx.doi.org/10.1016/j.chemolab.2021.104287.

[16] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T.N. Tran, L.M. Buydens, E. Marchiori, Convolutional neural networks for vibrational spectroscopic data analysis, Anal. Chim. Acta 954 (2017) 22–31, http://dx.doi.org/10.1016/J.ACA.2016.12.010.

[17] F.R. Harker, B.T. Carr, M. Lenjo, E.A. MacRae, W.V. Wismer, K.B. Marsh, M. Williams, A. White, C.M. Lund, S.B. Walker, et al., Consumer liking for kiwifruit flavour: A meta-analysis of five studies on fruit quality, Food Qual. Pref. 20 (1) (2009) 30–41.

[18] N.T. Anderson, K.B. Walsh, P.P. Subedi, C.H. Hayes, Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content, Postharvest Biol. Technol. 168 (2020) 111202, http://dx.doi.org/10.1016/J.POSTHARVBIO.2020.111202.

[19] N.T. Anderson, K.B. Walsh, J.R. Flynn, J.P. Walsh, Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content. II. Local PLS and nonlinear models, Postharvest Biol. Technol. 171 (2021) 111358, http://dx.doi.org/10.1016/J.POSTHARVBIO.2020.111358.

[20] F. Pedregosa, R. Weiss, M. Brucher, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830, URL http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html%5Cnhttp://arxiv.org/abs/1201.0490.

[21] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, D.D. Cox, Hyperopt: A Python library for model selection and hyperparameter optimization, Comput. Sci. Discov. 8 (1) (2015) 014008, http://dx.doi.org/10.1088/1749-4699/8/1/014008, URL https://iopscience.iop.org/article/10.1088/1749-4699/8/1/014008 https://iopscience.iop.org/article/10.1088/1749-4699/8/1/014008/meta.

[22] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, 2015, URL https://arxiv.org/abs/1412.6980v9.

[23] P. Mishra, D. Passos, F. Marini, J. Xu, J.M. Amigo, A.A. Gowen, J.J. Jansen, A. Biancolillo, J.M. Roger, D.N. Rutledge, A. Nordon, Deep learning for near-infrared spectral data modelling: Hypes and benefits, TRAC Trends Anal. Chem. 157 (2022) 116804, http://dx.doi.org/10.1016/J.TRAC.2022.116804.

Chapter 5

Assessing machine learning models for near-infrared regression by measuring stability towards diffeomorphisms

This chapter contains the published paper titled "Assessing machine learning models for near-infrared regression by measuring stability towards diffeomorphisms" which appeared in Chemometrics and Intelligent Laboratory Systems 264 (2025) 105449.

Chemometrics and Intelligent Laboratory Systems 264 (2025) 105449

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Assessing machine learning models for near-infrared regression by measuring stability towards diffeomorphisms

Mark Wohlers [a,b,*], V.A. McGlone [a], Eibe Frank [b], Geoffrey Holmes [b]

[a] The New Zealand Institute for Plant and Food Research Limited, Auckland, New Zealand
[b] Department of Computer Science, University of Waikato, Hamilton, New Zealand

ARTICLE INFO

Keywords:
Near-infrared spectroscopy
NIR
Convolutional neural networks
Partial least squares regression
Data augmentation

ABSTRACT

Near infrared (NIR) spectroscopy is widely used as a tool for non-destructive assessment of fruit quality by applying measured spectra to predict quality parameters such as dry matter and soluble solids content using a suitable regression method. With continued advancements in deep learning, there is potential for improved predictive performance when neural network models are applied instead of partial least-squares regression, but choosing a model remains challenging as performance is sensitive to the model's architecture. Taking inspiration from work done in image classification, we propose model selection by assessing relative stability to diffeomorphic transformations, providing a complementary approach to standard validation methods. This is particularly useful when labelled validation data is limited. Our empirical results on several NIR regression problems indicate that the proposed approach is comparable to the use of independent validation sets. In addition to the choice of deep learning architecture, we also consider the selection of the number of components in partial least-squares regression to demonstrate the method's generality.

1. Introduction

Near infrared (NIR) spectroscopy is an important tool used in fruit grading to meet consumer demand for high-quality foods. This technology allows for a non-destructive and rapid prediction of the characteristics of individual fruit at scale. In particular, dry matter content (DMC) and soluble solids content (SSC), which are associated with consumer responses [1,2], have been successfully predicted using NIR spectroscopy [3,4].

The recent emergence of deep learning models has shown promise in improving the quality of these predictions through automatic feature extraction and robustness to biological, environmental and instrumental effects [5]. However, results can vary depending on the architecture used and can require extensive tuning based on validation data [6]. This validation is commonly based on evaluating the model on an independently sampled validation set. This has the advantage of being similar to how the model will be used to predict future observations and may give a more realistic measure if there is variation due to sampling, e.g., seasonal effects, model drift, etc.

Alternatively, validation can be performed through k-fold cross-validation, whereby the training data is randomly divided into k subsets. Then, k models are fit, each one using all training data excluding the respective kth subset, which is used to evaluate the kth model. These results are then aggregated and used to determine the final model to be used in the analysis.

These validation methods do have some drawbacks. The process of k-fold cross-validation, while common for other models, is rarely used for deep learning as it requires multiple models to be trained and evaluated, which can be time-consuming if training is computationally expensive. Alternatively, external validation data can be resource-intensive to collect and sometimes impractical. Recently, [7] provided an "a priori model performance measure", the Integral Error Correlation Index (IECI). The IECI metric is based on the covariance of spectral deviations in the training data due to effects such as the device used. In [7], standard pre-processing techniques for the training set of cannabinoid content measurements were selected to minimise the IECI prior to model fitting. This method was found to provide superior results compared to those in the literature.

While the approach we outline here is not strictly a priori, as the model is trained first, it does provide a measure of model stability to transformations that does not require a separate validation set. Our approach assesses the robustness of a given architecture to data with a diffeomorphic transformation applied. It has previously been theorised that deep neural networks learn to correctly classify image data by becoming stable to diffeomorphisms [8,9]. We investigate a model

* Corresponding author at: The New Zealand Institute for Plant and Food Research Limited, Auckland, New Zealand.
E-mail address: mark.wohlers@plantandfood.co.nz (M. Wohlers).

https://doi.org/10.1016/j.chemolab.2025.105449
Received 6 March 2025; Received in revised form 27 May 2025; Accepted 30 May 2025
Available online 21 June 2025

selection approach for regression of NIR data based on this assumption and evaluate it for PLS regression as well as deep neural networks.

It is important to note that the aim of this study is not to establish whether neural networks outperform PLS regression; rather, we investigate whether the proposed model selection approach is a useful performance indicator that applies across the two techniques. The data used in this study have previously been successfully modelled using PLS regression, with some improvement gains when applying deep learning models [6,10]. This indicates that reasonable model performance should exist for the neural network approach even if the learning problems may not take full advantage of the deep neural network's ability to model non-linear relationships.

Our approach is based on work considering the *relative* stability of deep networks for image classification to diffeomorphic transformations, comparing stability observed when exposing data to diffeomorphism to stability observed when applying uncorrelated transformations. This work has shown that this relative stability correlates with the performance of a neural network image classifier on test data [11]. The motivation for the work presented here is that we conjectured that this relationship should also apply to NIR spectroscopy. Both domains contain strong relationships between features. There are spatial relationships between pixels in images and between wavelengths for NIR spectra. The diffeomorphic transformations keep the relationships "as connected sets remain connected, disjoint sets remain disjoint, smoothness of anatomical features such as curves and surfaces is preserved, and coordinates are transformed consistently" [12]. These properties are similar to distortions seen in images due to the parallax effect or differences in view angle, and in NIR spectroscopy with temperature shifts, instrument differences, and light scattering effects. For example, [13] found that temperature variations in milk samples resulted in wavelength shifts in the spectral data. Consequently, we investigate whether relative stability to diffeomorphisms can be used as a model selection criterion for regression of near-infrared spectra. The performance of a range of architectures is tested on spectra altered by applying a diffeomorphism and compared to the results of an uncorrelated transformation of the same magnitude. We would like to emphasise that this method is not intended to replace the standard method of validating models on independent data, but rather to complement it in assessing model quality. Like all model selection techniques, the proposed method has limitations, but it also provides useful information on model stability. The proposed approach could be of particular value in situations where labelled validation data is scarce.

2. Theoretical framework

2.1. Relative stability to diffeomorphisms

The core idea is to quantify the effect of a diffeomorphism transformation on model predictions relative to a similar uncorrelated transformation. This is measured by taking the difference in predictions between transformed and untransformed data and calculating the ratio of summary statistics, such as MSE, for the two transformations. We use the same definition of relative stability to diffeomorphisms, R_f, as used in [11]. That is

$$R_f = \frac{\langle \|f(\tau x) - f(x)\|^2 \rangle_{x,\tau}}{\langle \|f(x+\eta) - f(x)\|^2 \rangle_{x,\eta}}, \tag{1}$$

where $f(x)$ is the model prediction for features x, τx is the observation transformed through a diffeomorphism, and η is uncorrelated Gaussian noise of the same magnitude, i.e., $diag(Cov(\eta)) = diag(Cov(\tau x - x))$. The angular brackets $\langle\ \rangle$ indicate the mean or median of the distribution. This investigation only used the mean. Fig. 1 shows an example spectra transformed using our diffeomorphisms (left) and uncorrelated Gaussian noise (right).

2.2. Diffeomorphisms

Diffeomorphisms are smooth, differentiable transformations. More formally, diffeomorphisms are bijections that also have a smooth inverse. In a real-world setting, NIR spectra can vary due to effects of temperature, variation amongst devices, etc. However, in these situations the resulting spectra still retain relationships amongst wavelengths. The diffeomorphic transformation also retains core relationships amongst wavelengths, so a model robust to these changes may indicate robustness to the effects found in practice. Wavelength shift, as observed in spectral data from changes in temperature [13], could be considered a diffeomorphism. It is bijective, provided the shift is not so extreme that spectral features are lost, e.g. peaks are shifted outside the measured range. This shift is smooth, and invertible by shifting back by the same magnitude, and its inverse is also smooth. Previous work on image classification [11] used a maximum entropy model of diffeomorphisms suited to two-dimensional images. In the setting considered here, we propose a different transformation based on sampling the covariance matrix.

The NIR spectra can be considered a sample of points in a high-dimensional space with a given mean vector μ_0 and covariance matrix Σ_0. If Σ_0 is positive definite, then the data can be transformed to have any other positive definite covariance Σ_1. In practice, Σ_0 may be singular due to highly correlated wavelengths, and especially if the sample size is small relative to the number of wavelengths. To overcome this, a small constant may be added to the diagonal of Σ_0 to ensure it is positive definite. More specifically, there exist lower triangular, non-singular matrices C_0 and C_1 satisfying $C_0 C_0^T = \Sigma_0$ and $C_1 C_1^T = \Sigma_1$ such that if

$$x \sim \mathcal{N}(\mu_0, \Sigma_0)$$

then

$$C_1 C_0^{-1} x = x' \sim \mathcal{N}(C_1 C_0^{-1} \mu_0, \Sigma_1)$$

The diffeomorphism we propose is a simple linear map of the form

$$f : X \to X'$$

$$x' = Ax,$$

where $A = C_1 C_0^{-1}$ is a square, lower-triangular matrix of dimension equal to the number of wavelengths in the spectra, and x is the original spectra. For the transformation to be a diffeomorphism, A must be non-singular. To ensure this, a "scatter matrix" S is sampled from a Wishart distribution with d degrees of freedom and covariance matrix equal to the training set covariance,

$$S \sim \mathcal{W}_p(d, \Sigma_0),$$

and we set $\Sigma_1 = S/d$. Then, C_0 and C_1 can be obtained by the Cholesky decomposition of Σ_0 and Σ_1, respectively.

We can avoid the Cholesky decomposition of Σ_1 by using Bartlett's decomposition [14] to generate samples from the Wishart distribution. More specifically, we sample a lower triangular matrix B given by

$$B = \frac{1}{\sqrt{d}} \begin{pmatrix} b_1 & 0 & 0 & \cdots & 0 \\ n_{21} & b_2 & 0 & \cdots & 0 \\ n_{31} & n_{32} & b_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{p1} & n_{p2} & n_{p3} & \cdots & b_p \end{pmatrix},$$

where $b_i^2 \sim \chi^2_{d-i+1}$ and $n_{ij} \sim \mathcal{N}(0,1)$ independently.

A is then calculated as $A = C_0 B C_0^{-1}$, and this transformation is used in Eq. (1), that is $\tau x = C_0 B C_0^{-1} x$

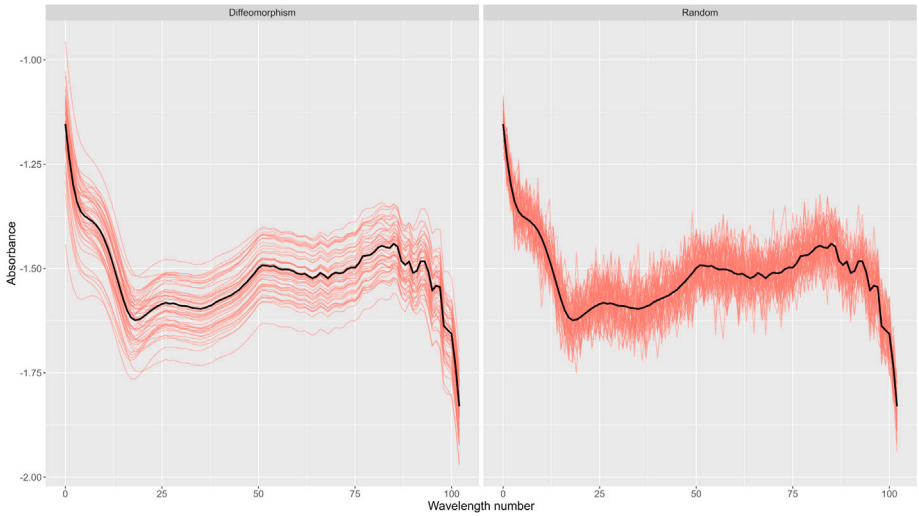


Fig. 1. Example comparison of diffeomorphisms (left) and uncorrelated transformations (right) of a single scan. The black lines represent the original spectrum.

2.3. Uncorrelated transformations

The method described in this paper requires a uncorrelated transformation of a magnitude similar to the diffeomorphism. For a given C_0 and sample x, we can calculate the variance of x' across sampled diffeomorphisms:

$$Var(x') = C_0^2\, Var(B)\, (C_0^{-1} x^T)^2$$

where

$$Var(B) = \begin{pmatrix} Var(b_1) & 0 & 0 & \cdots & 0 \\ 1/d & Var(b_2) & 0 & \cdots & 0 \\ 1/d & 1/d & Var(b_3) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1/d & 1/d & 1/d & \cdots & Var(b_p) \end{pmatrix}$$

and the variance of b_i is calculated from the variance of the χ distribution:

$$Var(b_i) = \frac{(d - i + 1) - \mu_i^2}{d}$$

with

$$\mu_i = \sqrt{\frac{2}{d}}\, \Gamma\left(\frac{d - i + 2}{2}\right) / \Gamma\left(\frac{d - i + 1}{2}\right).$$

If d is 150 or greater, the following approximation is used [15]:

$$\frac{\Gamma(y+1)}{\Gamma\left(y + \frac{1}{2}\right)} \approx \left(y^2 + \frac{y}{2} + \frac{1}{8}\right)^{1/4} \quad \text{, where } y = \frac{d - i}{2}.$$

The mean vector of transformed x' for a given x can be calculated as

$$\mu_{x'} = C_0^2\, \mu_B\, C_0^{-1} x^T,$$

where μ_B is diagonal with elements $\mu_1, \mu_2, \ldots, \mu_n$ along the main diagonal. Note that if $\mu_B = 0$, then $\mu_{x'}$ is also 0.

For a given x and C_0, we can now generate randomly transformed samples x' using a normal approximation with the above mean and variance:

$$x' \sim \mathcal{N}(\mu_{x'}, I(Var(x'))).$$

2.4. Effect of column order

Transforming the data through multiplication with A, a lower-triangular matrix, results in a transformed matrix where the first column is the original data rescaled by a constant, as illustrated in Fig. 2. This is undesirable if that column is highly correlated with the outcome measure. To reduce this effect, we shuffle the columns of the data before transformation and then re-sort to the original column order for model prediction. More specifically, a systematic shuffling of columns is conducted using a Williams design [16] using custom Python code translated from the crossdes R package [17]. We use the same design for the diffeomorphism and the uncorrelated transformation to enable a fair comparison. The Williams design is a Latin square balanced for position: each wavelength appears in each position with equal probability, and each wavelength column is immediately preceded and followed by every other wavelength column with equal probability. The design is an $m \times m$ row-column design if the number of wavelengths, m, is even, and 2 $m \times m$ if m is odd.

Table 1
Parameters and their sampling space for model generation.

Parameter	Parameter space
Preprocessing	None, Savitzky–Golay 0, 1st, or 2nd derivative with window size 13, 2nd order polynomial
Convolutional Layer	0 or 1 layer with 1, 10, or 50 filters of kernel size 13 or 27
Number of Dense Layers	1 to 4 with 256, 128, 64, 32, or 16 neurons in each layer (sorted descending)
Activation Function	ReLU or ELU
Learning Rate	0.005, 0.01, 0.05, or 0.1

3. Deep learning methodology

3.1. Model architectures

Experiments were conducted using Google Colab with TensorFlow 2.13.0 and TensorFlow probability 0.20.1. For each NIR dataset, 100 models with different architectures were trained. The full configuration space is given in Table 1. The total number of combinations in this space is prohibitive, so all experiments used a subset. The chosen 100 combinations that made up the subset were generated through a D-optimal design using the OptFederov function in the Algdesign R package [18]. For each model, performance was calculated at 100, 200, and 300 epochs.

A separate analysis recorded the metrics for a single model over each epoch to examine the feasibility of R_f providing a metric to stop training and prevent overfitting. In this case, the model architecture was fixed to have five convolutional layers, each followed by a max pooling layer (Fig. 3). The model was trained for 1500 epochs with the learning rate reducing by half during training if the validation MSE did not improve for 25 epochs. The training, validation, and test MSEs and R_f were compared across the training history.

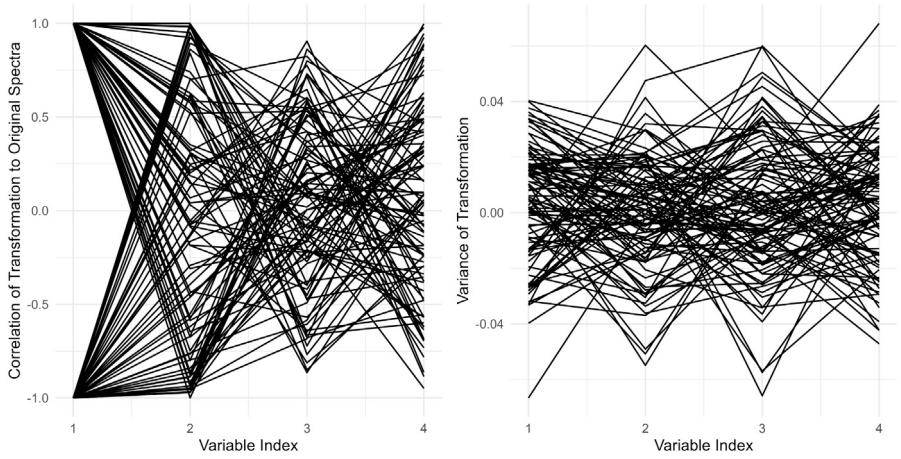


Fig. 2. Samples of a four-dimensional multivariate Gaussian with unit variances and all covariances equal to 0.3 transformed using a sampled diffeomorphism. The first dimension is always highly correlated (−1 or 1) with the original spectra. Note that the variance of the transformations (right) is constant and not affected by the order.

4. Data

The experiments were based on two datasets. The first dataset contained DMC for 4675 mangos, corresponding to 11,691 NIRS spectra from a single F-750 device, with multiple NIR scans per fruit. These data were previously made available as supplementary material in Anderson et al. 2020, 2021 [19,20]. The mangos include ten cultivars and are sampled across four Australian growing seasons. Training, validation, and test sets were previously specified in [19] and used here. The training and validation sets span the first three seasons, with the test set the final growing season.

The second dataset included 11,274 scans from 4956 kiwifruit sampled from two regions within New Zealand over three years and five devices. The training set comprises the first two years, while validation and test sets are the early and late part of the final season year, respectively. Further details are provided in [21].

To assist with training, all spectra were normalised wavelength-wise by subtracting the respective training set means and dividing by the standard deviation. For both datasets, the data was trimmed to the 459 to 1062 nm range for a total of 202 wavelengths.

5. Results

5.1. Choosing a model configuration

We first investigate how useful R_f is when choosing a model configuration for the mango data before considering the results for the kiwifruit data. Some of the model configurations did not converge, resulting in a high training MSE, frequently predicting a constant irrespective of the spectra. These divergent models were excluded from the analysis and would be easily picked up in a real-world training situation due to a large training loss. Prior to correlation analysis, MSE and R_f values were log-transformed. This transformation "commonly makes sense" for strictly positive variables, especially if they have large relative variation [22, p. 59, 65], and provides a more linear relationship among the metrics.

5.1.1. Mango

The R_f values for the different configurations, computing these values from spectra in the training set, decrease along with the test MSE, as shown on the left-hand side of Fig. 4. For reference, the right-hand side of the figure shows the correlation between MSE on the validation and MSE on the test data.

Alternatively, R_f values can be computed from spectra in the validation set or even the test set. Pearson's correlations between R_f values computed from these three subsets of data and MSE estimates obtained from the same three subsets are shown in Table 2. Interestingly, R_f based on the training spectra is highly correlated with R_f based on the validation or the test spectra, implying that using the training data alone to calculate R_f is sufficient. The correlation between the log validation MSE and test MSE (0.837) is higher than the correlation between log test MSE and R_f (0.704).

Table 2
Pearson correlation matrix for Mango models (log-transformed variables)

	Train log(MSE)	Val log(MSE)	Test log(MSE)	Train log(R_f)	Val log(R_f)	Test log(R_f)
Train log(MSE)	1.000	0.068	0.068	0.199	0.189	0.186
Val log(MSE)	0.068	1.000	0.837	0.614	0.616	0.630
Test log(MSE)	0.068	0.837	1.000	0.704	0.704	0.724
Train log(R_f)	0.199	0.614	0.704	1.000	0.997	0.994
Val log(R_f)	0.189	0.616	0.704	0.997	1.000	0.990
Test log(R_f)	0.186	0.630	0.724	0.994	0.990	1.000

Most importantly, using R_f as a criterion to select an appropriate model configuration yields results comparable to using the MSE on the validation dataset. The test MSE values of the configurations with the ten lowest R_f values, and the test MSE values of the configurations with the ten lowest MSE values on the validation data are presented in the right-most pair of distribution plots in Fig. 5. The distribution plots show comparable results with a slightly lower median for the R_f-based selection.

5.1.2. Kiwifruit

For the kiwifruit data, the R_f values for the different architectures, computed from the training spectra, positively correlate with the test MSE for both DMC and SSC, see Figs. 6 and 7. As the figures and tables show, the correlation is comparable to that observed between the validation MSE and the test MSE. There are 12 points, relating to four models at different epochs, in the bottom right of Fig. 6 that have relatively low test MSE given the high R_f. These models have some common features: a single convolutional filter, second derivative Savitzky–Golay preprocessing, and a large learning rate. However, there are also models with these same combinations that exhibit low test MSE and low R_f. Inspecting the numerator and denominator components of the R_f ratio, the numerator was low, indicating stability to diffeomorphisms, similar to other models with low test MSEs. However, the denominator, which measures the sensitivity to adding random noise, is unusually low, resulting in higher R_f than comparable models. One possible explanation is that the convolutional filter and Savitzky–Golay preprocessing have reduced the impact of the addition of unstructured noise. However, as previously mentioned this was not

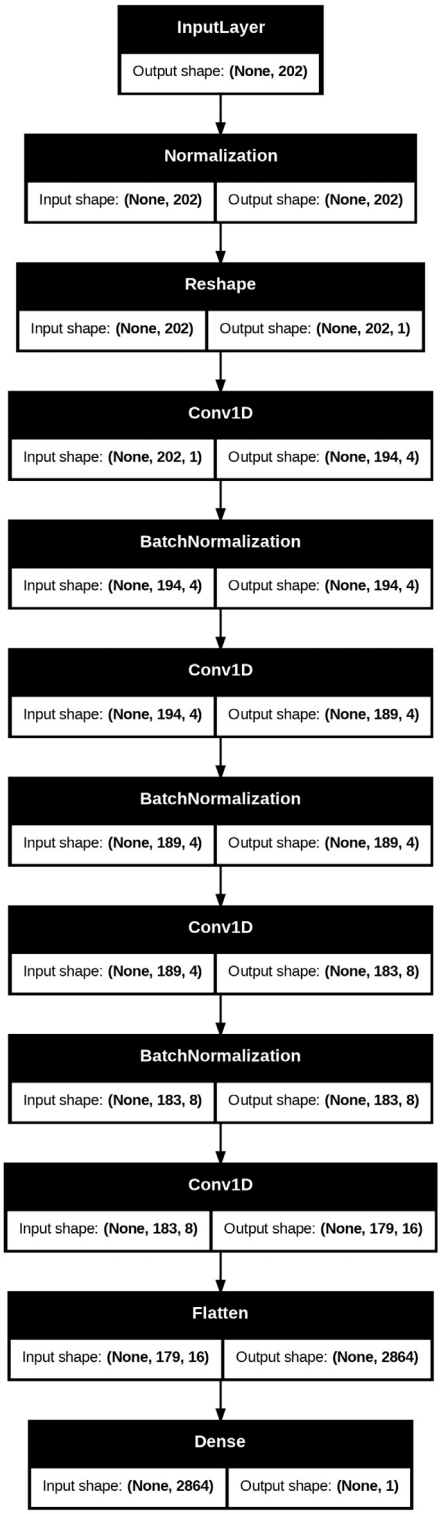


Fig. 3. Model architecture used to monitor R_f and losses throughout training.

always the case with similar architectures and was not observed in the other experiments for the SSC and mango datasets with the same architecture.

The validation MSE also has three similar outliers with low test MSE but comparably high validation MSE, although fewer and less pronounced.

Table 3
Pearson correlation matrix for DMC models (log-transformed variables).

	Train log(MSE)	Val log(MSE)	Test log(MSE)	Train log(R_f)	Val log(R_f)	Test log(R_f)
Train log(MSE)	1.000	0.032	0.064	0.206	0.215	0.228
Val log(MSE)	0.032	1.000	0.824	0.589	0.561	0.571
Test log(MSE)	0.064	0.824	1.000	0.763	0.729	0.745
Train log(R_f)	0.206	0.589	0.763	1.000	0.994	0.998
Val log(R_f)	0.215	0.561	0.729	0.994	1.000	0.996
Test log(R_f)	0.228	0.571	0.745	0.998	0.996	1.000

Table 4
Pearson correlation matrix for SSC models (log-transformed variables).

	Train log(MSE)	Val log(MSE)	Test log(MSE)	Train log(R_f)	Val log(R_f)	Test log(R_f)
Train log(MSE)	1.000	0.106	0.175	0.143	0.125	0.142
Val log(MSE)	0.106	1.000	0.736	0.700	0.732	0.693
Test log(MSE)	0.175	0.736	1.000	0.775	0.765	0.789
Train log(R_f)	0.143	0.700	0.775	1.000	0.994	0.999
Val log(R_f)	0.125	0.732	0.765	0.994	1.000	0.992

The relationship between Rf and test MSE is more variable for the SSC models than the DMC models, as seen in Figs. 6 and 7. This pattern is present in the comparable validation figure with the SSC models having a wider range of test MSEs at similar validation MSEs. There are two outliers with low Rf but large test MSE in the top left corner of Fig. 7. They belong to the same model architecture at different numbers of epochs, and investigating the residual plots showed that ten poor predictions significantly increased the test MSE. These predictions were outside the range of the training SSC, while the remaining test set predictions were in line with models of similar R_f. The validation MSE also included two poor predictions, which gives a more reasonable position in the right panel of Fig. 7. While R_f measures stability to input transformations, it does not necessarily detect if a model is giving unrealistic predictions.

The difference between SSC and DMC performance could also result from SSC measurements being less precise than those of DMC, as demonstrated by the poor test MSE.

For DM, Pearson correlations between the logarithms of R_f and the test MSE are slightly lower (0.763) but still comparable to the correlation between the validation set MSE and the test MSE (0.824) (Table 3). Conversely, the correlations between log train MSE for the SSC models (Table 4) where slightly higher for R_f (0.775) than validation MSE (0.736).

Again, R_f values based on the validation and test spectra provide minimal differences to R_f values calculated on the training set only.

As a criterion to select a model configuration, R_f is comparable to using the validation dataset for the DMC models (Fig. 5). For SSC, the selected ten configurations with the lowest R_f values exhibit somewhat higher test MSE than those selected with the ten lowest validation MSE values, and there are two models selected with very high test MSE. However, it should be noted that even the models selected using the validation MSE have a high test MSE.

5.2. Behaviour during training

We now consider the training history of the model in Fig. 3, trained as described above. We consider both the mango and the kiwifruit DMC data. We find that R_f provides a better prediction of the test MSE than the validation error. This can be seen in Figs. 8 and 9, where the test MSE values for the epochs with the 20 lowest validation MSE values, training MSE values, and R_f values, respectively, are shown. Clearly, the distribution for R_f is lower than for the validation MSE. Interestingly, on the mango data, selection based on the training MSE gives slightly lower test MSE.

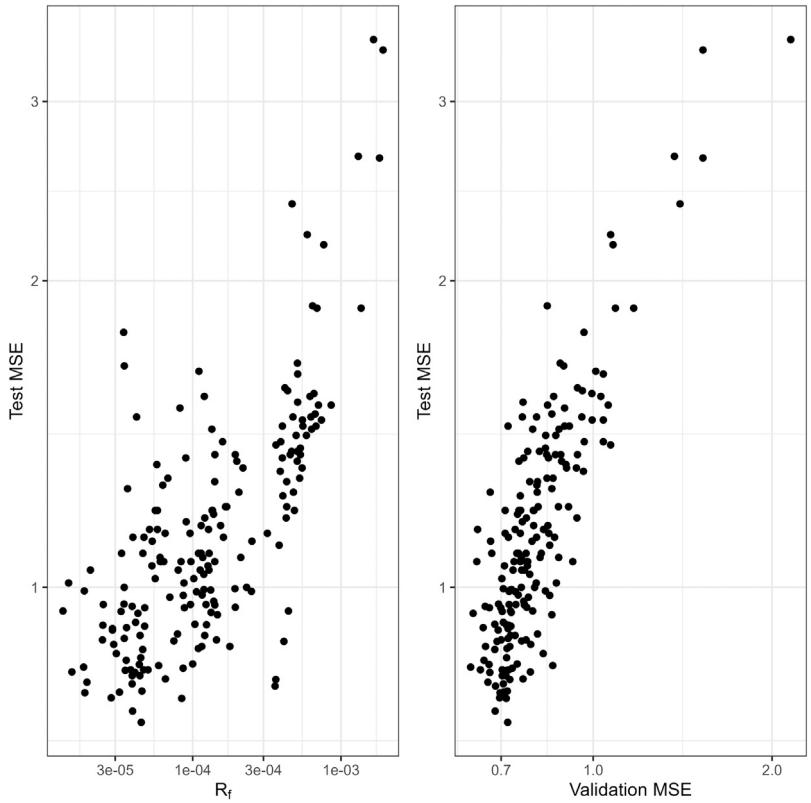


Fig. 4. Test MSE vs. training data R_f (left) and test MSE vs. validation MSE (right) for the Mango dataset. Models with large training error are excluded. Note that the axes use a logarithmic scale.

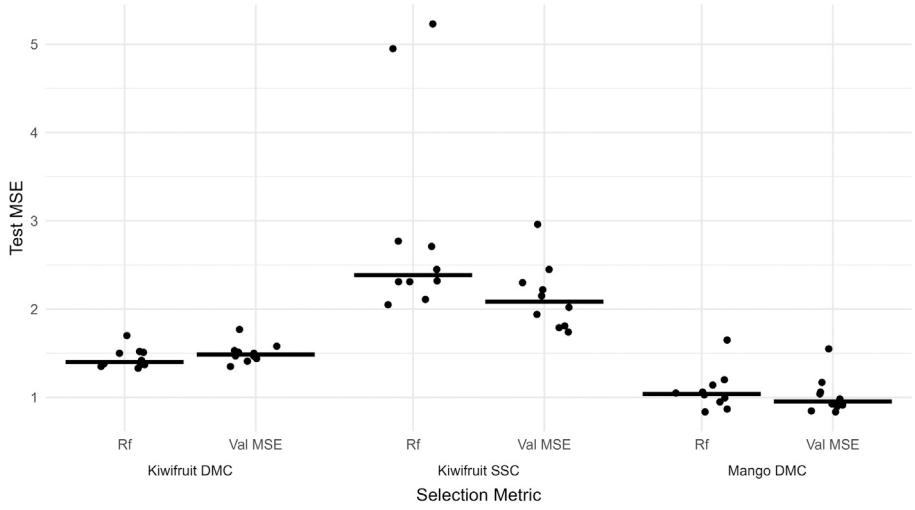


Fig. 5. Test MSE for the top ten models selected by validation MSE or training R_f.

In both datasets, inspection of the learning curves on the validation and test data revealed that training did not overfit significantly despite the training loss tending towards zero. Hence, to examine the behaviour of the selection criteria when training and test losses diverge through excessive training, an additional model was trained with a custom architecture, and Savitzky–Golay preprocessing was applied by fixing the weights of an initial convolutional layer appropriately. This meant the preprocessing was part of the model, making it simpler to evaluate by having the same features across the various configurations. The results in Fig. 10 show clear overfitting, with both the validation and test MSE exhibiting local minima around 200 and 250 epochs, respectively. The R_f also follows this trend with a minimum between the two, showing that an appropriate model would be selected using this criterion as well.

6. Extension to partial least squares

6.1. Methodology

To evaluate the method's applicability beyond deep learning models, we consider PLS regression next. When applying the proposed method to PLS, the denominator in Eq. (1) diverges, with random additive noise producing increasingly large variability in predictions as the number of latent variables increases compared to the diffeomorphism.

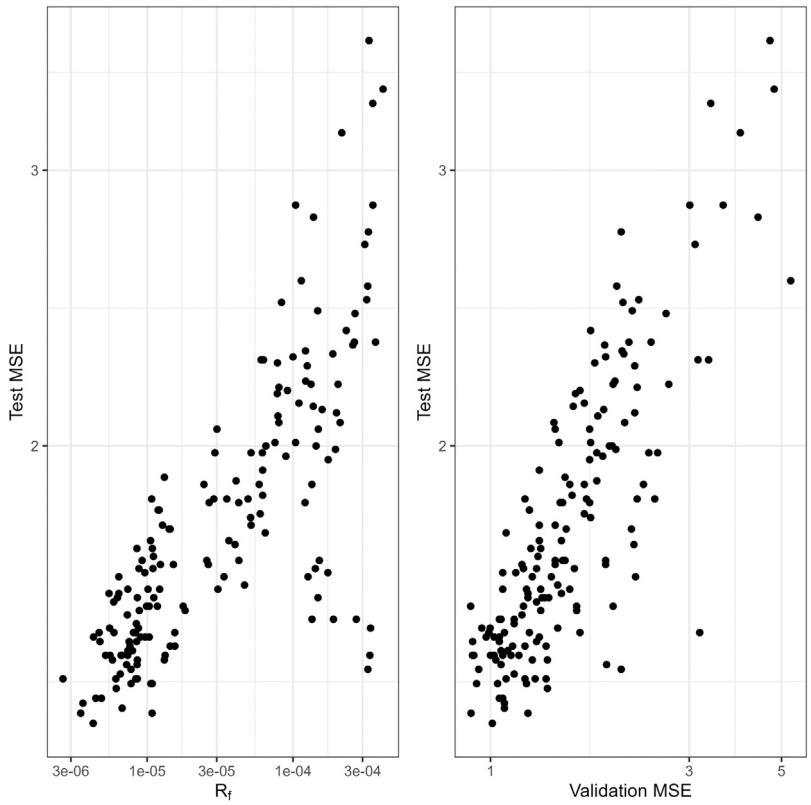


Fig. 6. Test MSE vs. training data R_f (left) and test MSE vs. validation MSE (right) for the Kiwifruit DMC dataset. Models with large training error are excluded. Note that the axes use a logarithmic scale.

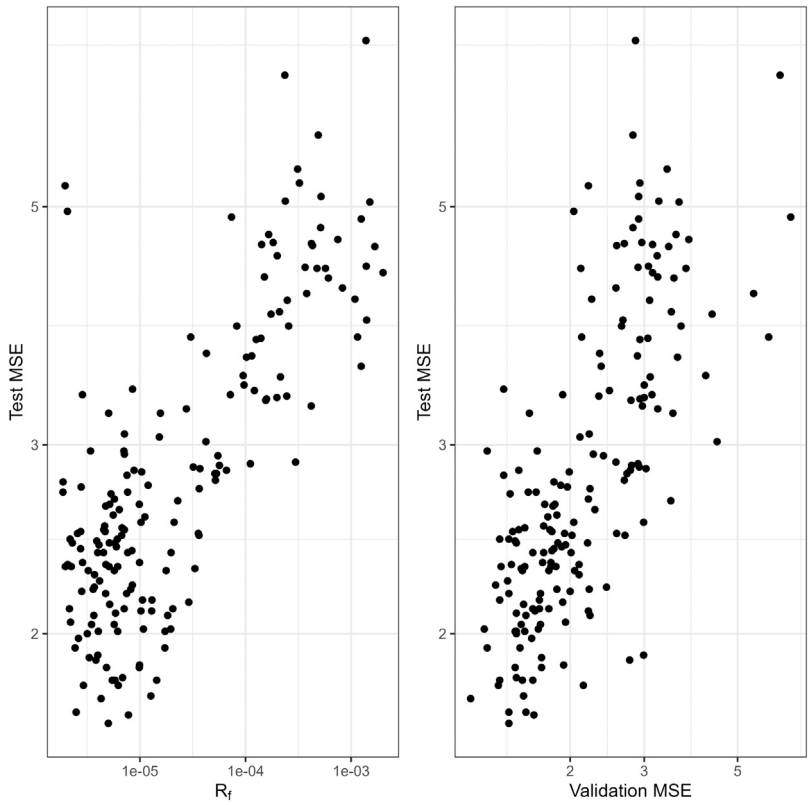


Fig. 7. Test MSE vs. training data R_f (left) and test MSE vs. validation MSE (right) for the Kiwifruit SSC dataset. Models with large training errors are excluded. Note that the axes use a logarithmic scale.

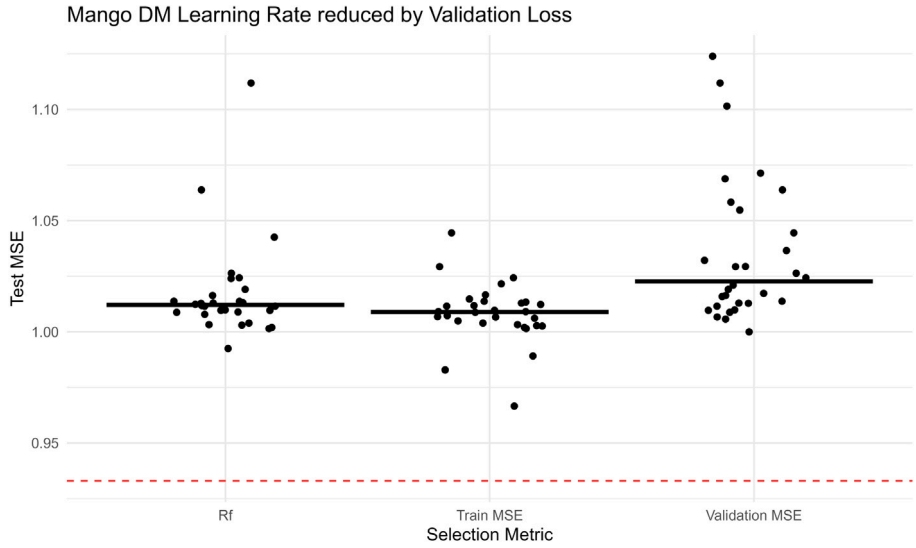


Fig. 8. Mango DMC Test MSE for training epochs with the 20 lowest training MSE, R_f, and validation MSE respectively.

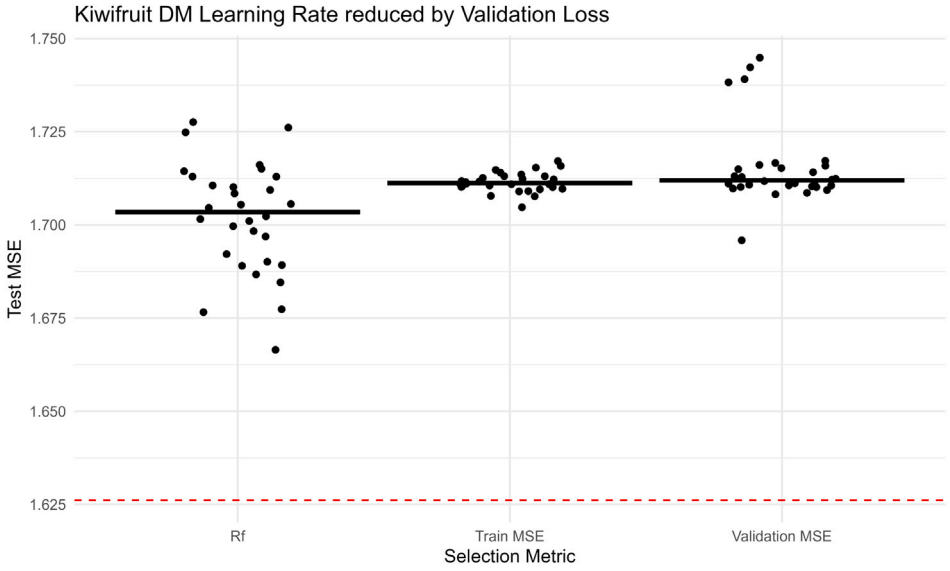


Fig. 9. Kiwifruit DMC Test MSE for training epochs with the 20 lowest training MSE, R_f, and validation MSE respectively.

The variance $f(x+\eta) - f(x)$ depends on the sum of the squares of the PLS regression coefficients. The sums of these squares increase or stay the same as the number of latent variables increases. The numerator, however, depends on the outer product of the coefficients and the covariance of $f(\tau x) - f(x)$. While this can increase and decrease as the number of latent variables increases, some of the contributions cancel, so changes occur at a lower rate. This is consistent with the feature of PLS that more components lead to a model more sensitive to noise [23]. When used as a model selection criterion, this means the criterion will always select the largest number of latent variables, which is clearly not a useful selection technique. Hence, the criterion was modified to employ a different noise model:

$$R_f^* = \frac{\langle \| f(\phi x) - f(x) \|^2 \rangle_{x,\phi}}{\langle \| f(\tau x) - f(x) \|^2 \rangle_{x,\tau}} \tag{2}$$

Here ϕ scales each wavelength by a respective constant that is equal in magnitude of the diffeomorphism. This has the effect of ϕx having the same correlation matrix as x and the same variance as τx. This change sets the original diffeomorphism as the denominator but keeps the central idea of the measure. A good model should perform better on transformations that preserve information than those that do not. While the denominator does retain much of the covariance of the original data, the ϕ transformation retains the correlation structure of the original data while altering the variance.

In our experiments with this criterion, the training sets were randomly divided into ten groups of roughly equal size. PLS models were then fit on increasingly larger training datasets ranging from 250 observations through to the full training dataset. As the kiwifruit data contained 5 unique devices, this sampling was stratified by device. For example, the subsample of 250 contained the first 50 observations from each device. As a baseline model selection criterion, random 10-fold cross-validation was used to give an estimated optimal number of dimensions. For consistency with [20] the optimum was taken as the maximum number of latent variables where each explain at least 1% of the training SSC or DMC variance. This mimics the situation where both unlabelled data (spectra in this case) and the labelled data are available for model training and selection. For each model, the optimal number of dimensions based on these statistics was compared using the RMSE of the test dataset.

All PLS analyses were conducted in R 4.31 using the pls package.

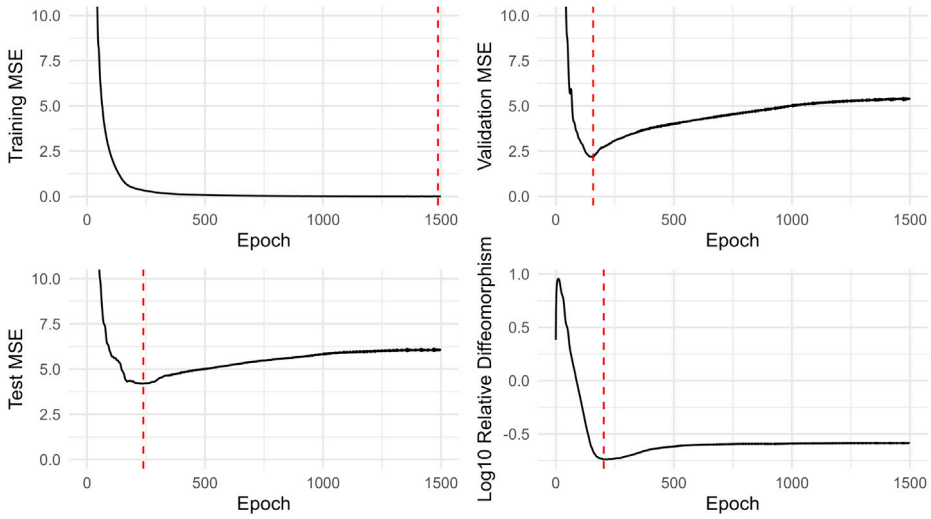


Fig. 10. Training, validation, and test MSE by training epoch compared to R_f for the mango DMC dataset. The red vertical line indicates the epoch corresponding to the minimum. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6.1.1. Mango

We first consider how the number of PLS components relating to the minimum R_f varies depending on the wavelength orders evaluated in the Williams design. This can be seen in the black boxplots in Fig. 11. Each boxplot corresponds to a model trained on a different number of spectra. It is clear that model selection is strongly dependent on the order of wavelengths used; hence, aggregation of these results is clearly necessary.

Calculating the overall R_f by combining the output from every permutation in the Williams design and using this for model selection (blue dot) closely matches the lowest test MSE model dimension (red dot), consistently across the models. Moreover, this method performs close to and, in some cases, slightly better than model selection by cross-validation and validation, as indicated by the blue and red crosses, respectively. The means and medians of the boxplots being close to the optimal test value indicate that the use of Williams' design to balance the wavelength orderings leads to a more stable result.

The number of latent variables selected by all methods in this study is considerably higher than the 7–10 LVs reported as optimal in previous work with the same mango data [19,20]. Two key differences explain this. Firstly, Anderson et al. [19] used Savitzky–Golay second derivative preprocessing, whereas these PLS experiments did not. Secondly, this study used a wider range of wavelengths, 459–1062 nm, compared to the narrower 684–990 nm.

Fig. 12 shows the cross-validation, validation and test MSE by the number of dimensions, alongside R_f for the mango dataset. As expected, the models based on smaller training sets had a more pronounced optimal number of latent variables with a clear minimum visible. Where as larger training sets flattened out as the number of components increased. The shape of the R_f curves resembled the cross-validation, validation, and test MSE trends. The exception is in the smaller training sets where the model becomes unstable at approximately 150 latent variables, resulting in large R_f.

6.1.2. Kiwifruit

The kiwifruit results were less consistent with the R_f criterion selecting lower than optimal dimensions for large training sets. This was also observed for the smallest training set of 250 observations, Fig. 13, in contrast to the cross-validation and validation set results (blue and red crosses, respectively), which were closer to the optimal dimension (red dot). Conversely, the R_f criterion selected appropriate components for the 500 and 1000 sizes but underestimated the components for the 750 observation set. Here, the R_f criterion selected a far lower number of dimensions, leading to a high test MSE in Fig. 14. The R_f value by dimension is plotted alongside the training, validation, and test MSEs in Fig. 14. Interestingly, in the 750 training set, the components resulting in the lowest test MSE correspond to local minima in R_f, with the global minimum R_f giving a much higher test MSE. It is possible that other information, such as training loss, could help inform which local minima may not be appropriate. In general, R_f had similar trends to the cross-validation and validation results, indicating that it captures model performance. However, there is no clear advantage over the more standard and simpler cross-validation method. Its usefulness for PLS may be limited to a supplementary method combined with the more established approaches.

7. Discussion and conclusion

Our experimental results show a relationship between R_f and the test MSE. This suggests that the relative stability to diffeomorphisms can be a useful indicator of performance in regression problems. However, performance varied across datasets and so may be influenced by the different spectral relationships present in the particular dataset used. R_f performed comparably to traditional model selection methods and stopping rules such as cross-validation, although this was more variable for the kiwifruit datasets. The main advantage of the new approach is that the method does not require separate validation datasets or the additional computational training effort required for cross-validation of the training data. However, while it shows promise as an evaluation method, it is more effective when used in conjunction with traditional validation approaches.

A modified version of the method was applied to PLS regression, which gave an indication of the optimal number of dimensions. While the optimal number of PLS latent variables to minimise the test MSE often coincided with a local minimum R_f, it was not always the global minimum R_f. This could possibly be overcome by using other metrics, such as training MSE when making decisions on the number of components.

There are some limitations to the proposed method. The sensitivity to the order of the wavelengths adds complexity. To compensate for this, balancing order for position and first-order carry-over through the use of a Williams design greatly increases computational overhead. For a dataset with n observations of m spectral wavelengths, the process requires generating mn transformed spectra, or $2mn$ if n is an odd number. For large datasets, this can be prohibitive.

Despite this issue, the proposed method provides a potential approach to model selection and deciding when to stop model training.

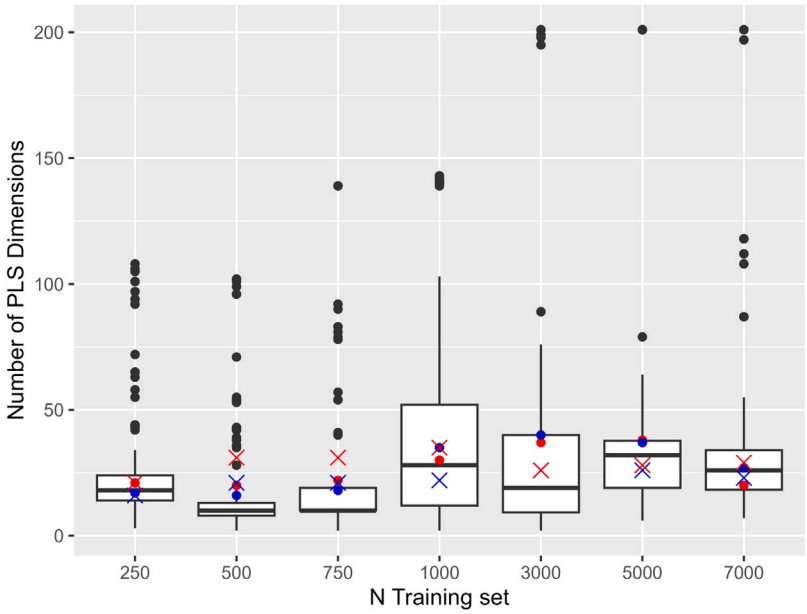


Fig. 11. Optimal dimension number selected to minimise R_f for each of the 202 Williams design permutations for the mango dataset. The blue dot indicates the optimal dimension selected after averaging across the 202 permutations. The red dot indicates the dimension with the lowest test error; red and blue crosses are the dimensions that explain more than 1% of the DMC training variance for validation and cross-validation MSE, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

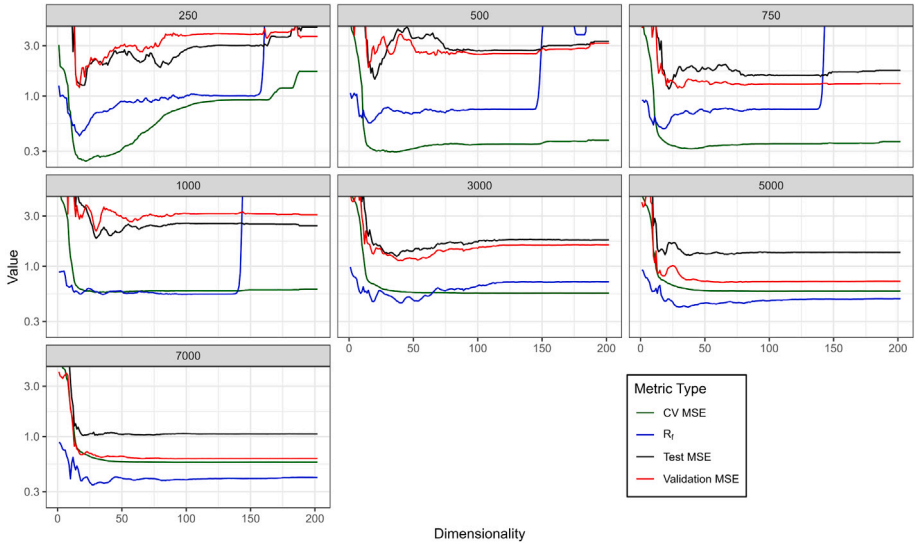


Fig. 12. Test MSE, validation MSE, cross-validation MSE, and Rf by the number of PLS components for the mango dataset across different training set sizes. A logarithmic scale is used to aid in visual comparison.

For additional robustness, it could be used in conjunction with other methods when appropriate, e.g., when a validation set is available.

The proposed method uses the R_f metric to measure how much a model's predictions vary from a diffeomorphism transformation relative to the prediction variation from uncorrelated transformations of a similar magnitude. R_f was found to correlate with test MSE, which offers a valuable measure to assist in model selection, particularly when applied to choose between deep learning models for NIR spectral data. While the proposed approach was not found to be superior compared to methods such as cross-validation for PLS regression, there are situations when it could provide a useful alternative, or be used in parallel to aid in training decisions or model selection.

Monitoring the R_f value while training was competitive with observation of the validation sets in our experiments. An increasing ratio indicated that the model's performance was degrading; however, there may be some lag before this becomes apparent.

Interestingly, the value of R_f did not change significantly when applied to the training, validation, or test datasets. This has the benefit of making the proposed approach applicable in situations when validation sets are unavailable. However, it is important to note that it does add a computational overhead: in our experiments, the data was permutated over 200 times. Nevertheless, it could be preferable in situations where model training, in the case of k-fold validation, or the collection of an independent validation dataset, is time- and resource-intensive.

One limitation of the method is that it requires the covariance matrix of the training set to be positive definite. In situations where the sample covariance is singular, a small constant can be added to the diagonal of the covariance matrix. The effect of this regularisation on

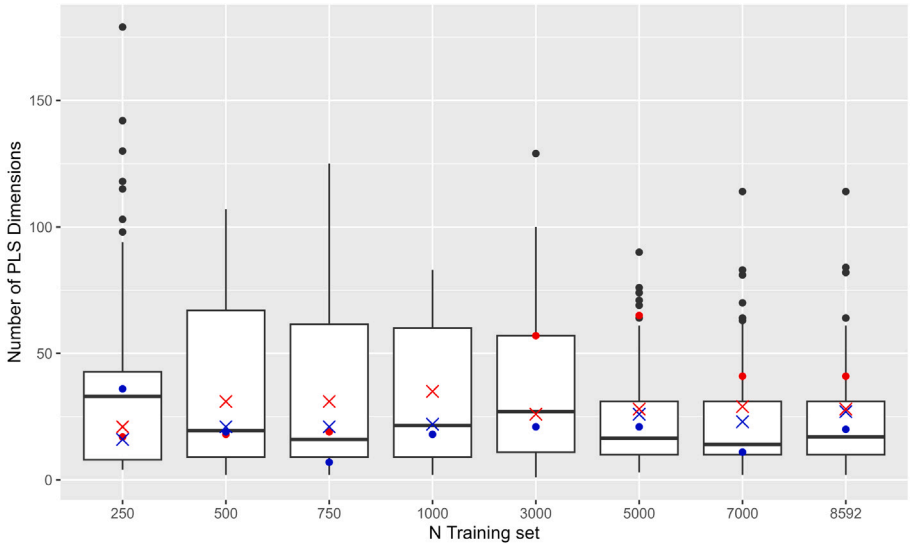


Fig. 13. Optimal dimension number selected to minimise R_f for each of the 202 William's design permutations for the kiwifruit DMC dataset. The blue dot indicates the optimal dimension selected after averaging the 202 permutations. The red dot indicates the dimension with the lowest test error, red and blue crosses are the dimensions that explain more than 1% of the DMC training variance for validation and cross-validation MSE, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

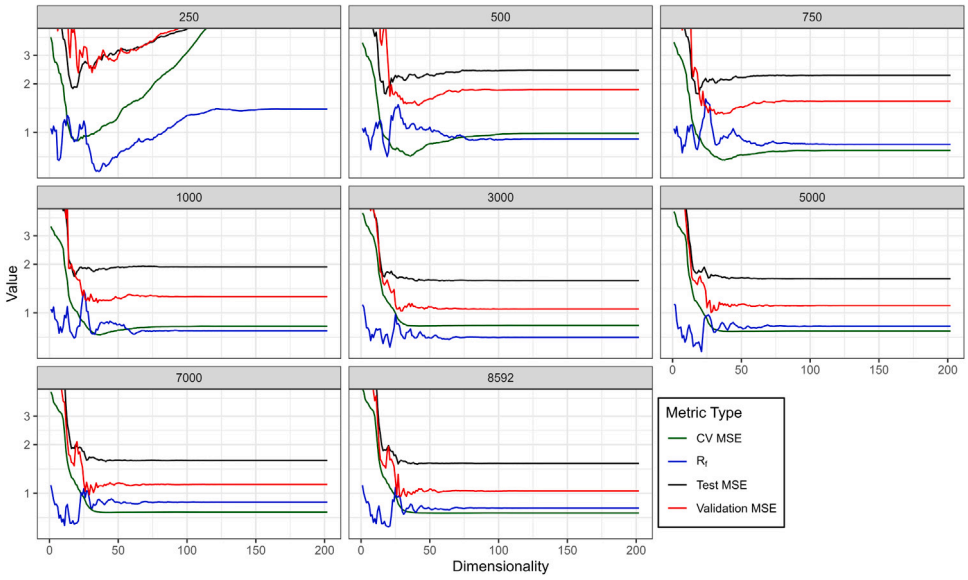


Fig. 14. Test MSE, validation MSE, cross-validation MSE, and Rf by the number of PLS components for the kiwifruit DMC dataset across different training set sizes. A logarithmic scale is used to aid in visual comparison.

R_f when the number of training samples is small should be investigated to assess our method's applicability in such situations.

The diffeomorphisms used in this investigation are not the only possible candidate transformations, and alternative methods of applying different diffeomorphisms and random noise may improve the current findings. In particular, PLS was very sensitive to the choice of uncorrelated transformation and required customisation. An alternative noise method was implemented where wavelengths were multiplied by a constant rather than being modified by additive noise. The suitability of alternative transformations should be explored more fully.

Data

All data used in this paper are available through Anderson et al. 2020 [19] and Wohlers et al. 2023 [21]. The code used in the analyses will be available upon acceptance.

CRediT authorship contribution statement

Mark Wohlers: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **V.A. McGlone:** Writing – review & editing, Resources, Conceptualization. **Eibe Frank:** Writing – review & editing, Supervision, Conceptualization. **Geoffrey Holmes:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research received financial support by the New Zealand Ministry of Business Innovation & Employment (MBIE) as part of the Endeavour funded project *Perfecting storage life prediction for delivery of high quality fruit*.

Data availability

Data is already available.

References

[1] S.R. Jaeger, R. Harker, C.M. Triggs, A. Gunson, R.L. Campbell, R. Jackman, C. Requejo-Jackman, Determining consumer purchase intentions: The importance of dry matter, size, and price of kiwifruit, J. Food Sci. 76 (3) (2011) S177–S184, http://dx.doi.org/10.1111/J.1750-3841.2011.02084.X, URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1750-3841.2011.02084.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1750-3841.2011.02084.x. https://ift.onlinelibrary.wiley.com/doi/10.1111/j.1750-3841.2011.02084.x.

[2] C.H. Crisosto, G.M. Crisosto, Relationship between ripe soluble solids concentration (RSSC) and consumer acceptance of high and low acid melting flesh peach and nectarine (Prunus persica (L.) Batsch) cultivars, Postharvest Biology Technol. 38 (3) (2005) 239–246, http://dx.doi.org/10.1016/J.POSTHARVBIO.2005.07.007.

[3] V.A. McGlone, R.B. Jordan, R. Seelye, P.J. Martinsen, Comparing density and NIR methods for measurement of Kiwifruit dry matter and soluble solids content, Postharvest Biology Technol. 26 (2) (2002) 191–198, http://dx.doi.org/10.1016/S0925-5214(02)00014-5.

[4] Y. Zhang, J.F. Nock, Y. Al Shoffe, C.B. Watkins, Non-destructive prediction of soluble solids and dry matter contents in eight apple cultivars using near-infrared spectroscopy, Postharvest Biology Technol. 151 (2019) 111–118, http://dx.doi.org/10.1016/J.POSTHARVBIO.2019.01.009.

[5] X. Zhang, J. Yang, Advanced chemometrics toward robust spectral analysis for fruit quality evaluation, Trends Food Sci. Technol. 150 (2024) 104612, http://dx.doi.org/10.1016/J.TIFS.2024.104612.

[6] D. Passos, P. Mishra, Deep Tutti Frutti: Exploring CNN architectures for dry matter prediction in fruit from multi-fruit near-infrared spectra, Chemometr. Intell. Lab. Syst. 243 (2023) 105023, http://dx.doi.org/10.1016/J.CHEMOLAB.2023.105023.

[7] J. Ezenarro, D. Schorn-García, M. Plans, O. Busto, R. Boqué, Quantification of spectral measurement errors to guide preprocessing method selection: A case study on cannabinoid prediction across multiple NIR instruments, Anal. Chim. Acta 1343 (2025) 343705, http://dx.doi.org/10.1016/J.ACA.2025.343705.

[8] S. Mallat, Understanding deep convolutional networks, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 374 (2016) 2065, http://dx.doi.org/10.1098/RSTA.2015.0203, https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0203.

[9] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1872–1886, http://dx.doi.org/10.1109/TPAMI.2012.230.

[10] P. Mishra, D. Passos, Synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit, Chemometr. Intell. Lab. Syst. 212 (2021).

[11] L. Petrini, A. Favero, M. Geiger, M. Wyart, Relative stability toward diffeomorphisms indicates performance in deep nets, 2021, http://arxiv.org/abs/2105.02468.

[12] M.F. Beg, M.I. Miller, A. Trouvé, L. Younes, Computing large deformation metric mappings via geodesic flows of diffeomorphisms, Int. J. Comput. Vis. 61 (2) (2005) 139–157, http://dx.doi.org/10.1023/B:VISI.0000043755.93987.AA/METRICS, https://link.springer.com/article/10.1023/B:VISI.0000043755.93987.aa.

[13] J.A. Diaz-Olivares, S. Grauwels, X. Fu, I. Adriaens, W. Saeys, R. Bendoula, J.M. Roger, B. Aernouts, Temperature correction of near-infrared spectra of raw milk, Chemometr. Intell. Lab. Syst. 255 (2024) 105251, http://dx.doi.org/10.1016/J.CHEMOLAB.2024.105251, https://www.sciencedirect.com/science/article/pii/S0169743924001916#fig1.

[14] M.S. Bartlett, XX—On the theory of statistical regression, Proc. R. Soc. Edinb. 53 (1934) 260–283, http://dx.doi.org/10.1017/S0370164600015637.

[15] J.S. Frame, An Approximation to the Quotient of Gamma Function, Am. Math. Mon. 56 (8) (1949) 529, http://dx.doi.org/10.2307/2305527.

[16] E.J. Williams, Experimental designs balanced for the estimation of residual effects of treatments, Aust. J. Chem. 2 (2) (1949) 149–168, http://dx.doi.org/10.1071/CH9490149.

[17] O. Sailer, Crossdes: A package for design and randomization in crossover studies, R News 5 (2) (2005) 24–27.

[18] B. Wheeler, M.J. Braun, Package 'AlgDesign', R Proj Stat. Comput 1 (2019) 1–25.

[19] N.T. Anderson, K.B. Walsh, P.P. Subedi, C.H. Hayes, Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content, Postharvest Biology Technol. 168 (2020) 111202, http://dx.doi.org/10.1016/J.POSTHARVBIO.2020.111202.

[20] N.T. Anderson, K.B. Walsh, J.R. Flynn, J.P. Walsh, Achieving robustness across season, location and cultivar for a nirs model for intact mango fruit dry matter content. ii. local pls and nonlinear models, Postharvest Biology and Technology 171 (2021) 111358, http://dx.doi.org/10.1016/J.POSTHARVBIO.2020.111358.

[21] M. Wohlers, A. McGlone, E. Frank, G. Holmes, Augmenting NIR Spectra in deep regression to improve calibration, Chemometr. Intell. Lab. Syst. 240 (2023) 104924, http://dx.doi.org/10.1016/J.CHEMOLAB.2023.104924.

[22] A. Gelman, J. Hill, Data Analysis using Regression and Multilevel/Hierarchical Models, Cambridge University Press, 2007.

[23] P. Geladi, E. Dåbakk, Computational methods and chemometrics in near infrared spectroscopy, Encycl. Spectrosc. Spectrom. (1999) 386–391, http://dx.doi.org/10.1016/B978-0-12-374413-5.00352-3.

Chapter 6

Barlow Twins for Semi-Supervised Learning in NIR Spectroscopy

Barlow Twins for Semi-Supervised Learning in NIR Spectroscopy

Mark Wohlers[a,b,*], Andrew McGlone[a], Eibe Frank[b], Geoffrey Holmes[b]

[a]*New Zealand Institute for Bioeconomy Science Limited, Auckland, New Zealand*

[b]*Department of Computer Science, University of Waikato, Hamilton, New Zealand*

[*]Corresponding author: `mark.wohlers@plantandfood.co.nz`

Status: Under review at *Chemometrics and Intelligent Laboratory Systems*

Submitted: October 2025 Accepted: February 2025

Abstract

Near-infrared (NIR) spectroscopy is a widely used technology in the horticulture industry for non-destructive fruit grading. Partial Least Squares (PLS) regression is the dominant method for producing fruit quality predictions from measured spectra. Alternative deep learning methods have shown promise, but often require large amounts of labelled data to train. This study proposes a semi-supervised method based on Barlow Twins to include unlabelled data in the training process. We adopt the Barlow Twins method by using repeated measurements on the same fruit from different devices as different "views" to encode into the same latent space and combine the encoder network with a regression head for prediction. Our approach demonstrates improved performance over PLS with up to 17% lower RMSE, especially when the labelled data is limited. The Barlow loss function also improves calibration transfer results.

6.1 Introduction

Partial Least Squares Regression (PLSR) is a popular and widely adopted method in NIR spectroscopy. While deep learning approaches, particularly convolutional neural networks (Cui & Fearn, 2018), (J. Walsh et al., 2023), (Bjerrum et al., 2017b), (Mishra et al., 2022) have shown promise, much of the research focuses on architectures and training methodologies for image-based data. In recent years, self-supervised learning methods (SSL) have enabled learning from unlabelled data, but these are underexplored for NIR data. Fortunately, standard deep learning frameworks such as PyTorch (Paszke et al., 2017) and TensorFlow (Abadi et al., 2019) allow for simple specification of network architectures suitable for NIR data and custom loss functions, including weighted combinations of losses, that are applied in SSL methods. This paper investigates the application of the Barlow Twins SSL approach to NIR data.

This approach applies a nonstandard loss function that consists of a weighted combination of terms based on the diagonal and off-diagonal components of the cross-correlation matrix between two embedded views of the same sample. By further combining the Barlow Twins loss with a regression mean square error (MSE) loss, we can train models that benefit from the information contained in both labelled data (spectra with reference values) and unlabelled data (spectra without reference measurements) within a semi-supervised framework.

6.2 Background

6.2.1 NIR Spectroscopy

Near-Infrared spectroscopy provides a nondestructive method to predict the chemical makeup of samples based on the absorbances of different wavelengths in the near-infrared range. It has been applied to many industries, including horticulture, for the prediction of fruit quality measures associated with consumer responses such as dry matter content (DMC) and soluble solids content (SSC).

6.2.2 Partial Least Squares

Partial least squares regression has been widely applied in the NIR spectroscopy space. Despite a large number of competing methods, including convolutional networks, PLS remains the dominant method for NIR regression and classification tasks (Makalesi et al., 2021). Due to its popularity and robustness, PLS is often used as a baseline method for comparisons of alternative modelling techniques.

Developed in 1975 by Herman Wold and later modified by Svante Wold and Harald Martens (Wold et al., 2001b), PLS regression addresses the problem of multicollinearity, which is present in NIR data, where responses at different wavelengths are highly correlated. It does this by finding linear combinations of the input variables (wavelengths) to form latent variables. These latent vari-

ables are fitted sequentially to maximise the covariance of the latent variable to the target variables (Y), subject to the constraint that the weight vectors have unit norm and successive latent variables are orthogonal. After each latent variable is determined, the X (wavelength) and Y (target) matrices are deflated before finding the next component.

6.2.3 Self-Supervised Learning

Self-supervised learning is a form of unsupervised learning that derives discriminatory features from unlabelled data (Gui et al., 2024). Examples include reconstructing images from mixed-up jigsaw pieces, which helps learn spatial features (Gui et al., 2024). There has been little work on self-supervised learning applied to NIR spectroscopy. L. Zhang et al. (2025) pretrained a dual-branch autoencoder on unlabelled VIS (450-780nm range) and NIR (783-1500nm) datasets of mango fruit. The trained encoder was then used to encode smaller labelled datasets for regression tasks. They found that their method achieved results within 99% of the best results using only 10% of the labelled data.

6.2.4 Barlow Twins

Barlow Twins (Zbontar et al., 2021) is a self-supervised learning method originally applied to image data. The method involves generating pairs of augmented views of a given image, which are then fed into an encoder. This encoder is trained so that the values of any given latent variable in the encoder's output are highly correlated between the two augmented pairs, while the values of different latent variables have low correlation. The loss function $\mathcal{L}_{BT}$ is calculated from the cross-correlation matrix of the two latent samples. To minimise the loss, the off-diagonal elements should tend to 0, and the diagonal elements to 1, approximating the identity matrix (Figure 6.1). The method has the advantage of providing latent variables that avoid redundancy and are invariant to the distortions used in the augmentation, while preventing

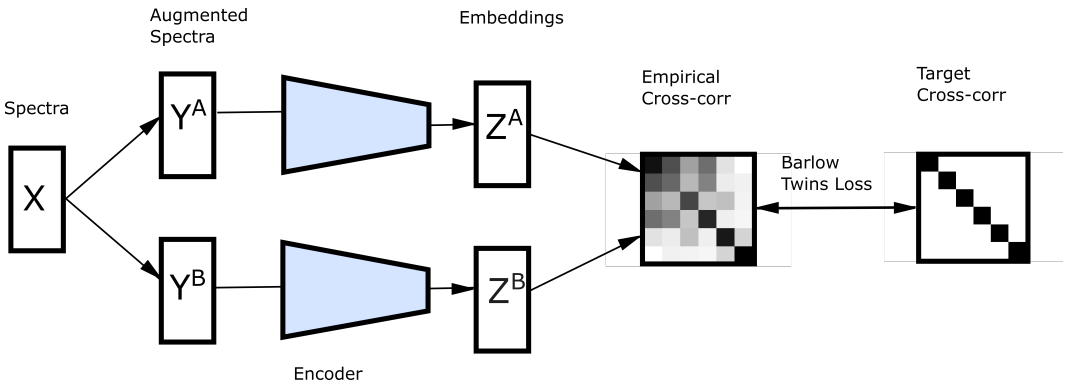


Figure 6.1: Barlow Twins loss measures the cross-correlation of embeddings from two related inputs and penalises for how different it is from the identity matrix

collaps to constant outputs.

The Barlow Twins loss function is defined as a weighted sum of the invariance and redundancy reduction terms. The λ parameter controls this weighting :

$$\mathcal{L}_{BT} \triangleq \underbrace{\sum_i (1 - C_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} C_{ij}^2}_{\substack{\text{redundancy} \\ \text{reduction term}}} \tag{6.1}$$

The cross-correlation matrix elements C_{ij} are calculated as:

$$C_{ij} \triangleq \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b (z_{b,i}^A)^2} \sqrt{\sum_b (z_{b,j}^B)^2}} \tag{6.2}$$

In the current study, we treat spectra from different devices measuring the same fruit as different views of the same object, similar to capturing the object with two different cameras.

6.2.5 Semi-Supervised learning

Semi-supervised learning is similar to self-supervised learning in that it derives information from unlabelled data (van Engelen & Hoos, 2020). The distinction is that it does this in combination with a (often smaller) labelled dataset. In a deep learning setting, unlabelled data could be used to train an encoder with a Barlow loss function, while a regression head is trained on the

labelled set with a mean squared error (MSE) loss function. Several recent publications have examined the potential of applying semi-supervised learning to NIR spectroscopy (Mishra & Woltering, 2023; Said et al., 2022). Said et al. (2022) looked at semi-supervised learning for predicting milk fat content. Their method used autoencoders (AE) trained on unlabelled data, while training a regression head connected to the encoder on a smaller subset of labelled data. The autoencoder and regression head were trained simultaneously, with the average of the autoencoder reconstruction MSE and the regression MSE as the loss function. They found improved results compared to the same deep learning regression model trained on the labelled data only, including a reduction in root mean square error (RMSE) from 0.287 to 0.221 with only 35% of the labelled and 65% unlabelled data. Mishra and Woltering (Mishra & Woltering, 2023) did not incorporate unlabelled data into their method and instead focused on robust regression by down-weighting outliers.

6.3 Proposed Framework

6.3.1 Barlow Twins for Spectral Data

We propose using the Barlow Twins loss to train an encoder model combined with a regression head that uses MSE loss. To adapt the Barlow method to NIR spectra, suitable augmentation methods are needed. Recently, Dhaini et al. (2024) examined contrastive learning for hyperspectral image classification using various augmentation techniques, including spectral shift, spectral flipping, and scattering using Hapke's model (Hapke, 1981). Spectra could also be augmented by adding random multivariate normal samples as in Wohlers et al. (2023). Alternatively, when samples are measured by multiple devices, pairs of measurements of the same sample by different devices can be used as different views.

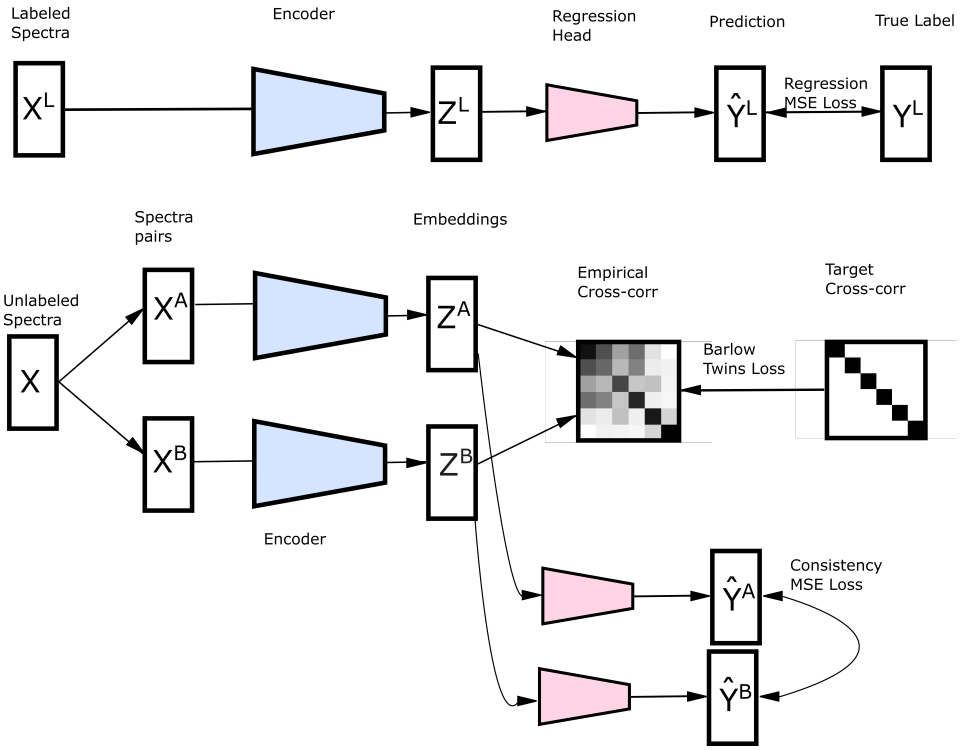


Figure 6.2: Barlow Twins model for Spectral data. X represents unlabelled pairs of spectra from devices A and B, XL spectra from a separate labelled dataset, with labels YL (dry matter content or soluble solids content). The Barlow loss reduces redundancy and enforces invariance, the regression loss ensures that the network can predict Y, and the consistency loss ensures that predictions from spectra measured on the same fruit are similar. The regression head (pink) benefits indirectly from unlabelled data through the shared encoder (blue), which learns robust device invariant spectral representations from all paired measurements (including unlabelled data via the Barlow loss) while being constrained to produce accurate predictions on the labelled subset (via the MSE loss). During training, all three losses contribute to updating the encoder weights, with the weighted combination determining each loss's relative influence.

6.3.2 Links to PLS

A PLS regression model can be approximated using this framework by employing an appropriately configured neural network. The encoder network should consist of a single linear dense layer with the number of neurons equal to the number of latent variables, and with a unit norm constraint on the weights. To train a PLS-type regression type model, no augmentation is used; that is, the pairs of spectra are identical. The Barlow loss is then based on the correlation matrix, rather than the cross-correlation matrix. The diagonal elements will always be one. Therefore, minimising the Barlow loss will require the off-diagonal elements to tend to zero, similar to orthogonal PLS latent variables. The other loss function, to be combined with the Barlow loss in a PLS-equivalent neural network approach, should maximise the covariance between the encoder outputs and the target variable. This approach is less straightforward and requires a custom loss function. For each latent variable, this is calculated as a partial covariance with the target variable, accounting for the previous latent variables.

6.4 Methods

6.4.1 Datasets

Our experiments employ a NIR dataset comprising kiwifruit across two sites in New Zealand (Te Puke and Kerikeri) and three seasons, with the last season being in 2019 season. Individual fruit were often measured by multiple devices. Only fruit with measurements from two or more devices were included in the study, resulting in a total of 4,316 fruit and 10,869 scans over the 402 to 1137nm range with a spectral resolution of 3nm. All devices were of the same model, the F-750 Produce Quality Meter produced by Felix Instruments (Felix Instruments, 2019). Reflectance values were measured with a xenon tungsten lamp. The measured range was trimmed to 222 wavelengths spanning 402-

1065nm and includes part of the visible (VIS, 402-780nm) as well as the near-infrared (NIR, 780-1065nm) regions. While wavelength selection methods are widely used in chemometrics to improve model performance (Yamashita et al., 2022), the wide range of wavelengths allows the proposed method to be assessed without specific feature engineering. Each fruit had its respective soluble solids content (SSC) and dry matter content (DMC) measured destructively. SSC was measured from a juice sample by a digital refractometer, expressed in °Brix. DMC represents the weight of a fruit tissue sample after drying as a percentage of its initial fresh weight. Summary statistics of these measures for the training and test sets are provided in table 6.1. The first two years were used for the training tasks, with the remaining year used as a test set. No validation data was used for hyperparameter tuning. Five NIR spectrometers were used to collect the data, TP1, TP2, TP3 for the Te Puke site, and KK1, KK2 for the Kerikeri site.

		DMC				SSC			
Set	N	Mean	SD	Min	Max	Mean	SD	Min	Max
Training	2924	16.48	2.14	9.01	23.6	8.37	3.19	3.70	19.4
Testing	1392	17.48	2.48	10.30	24.2	7.69	3.12	3.75	19.7

Table 6.1: Summary statistics for dry matter (DMC) and soluble solids content (SSC) in training and testing sets.

The training set, comprising of 2924 fruit (7495 total scans), had some overlaps among the various devices. Table 6.2 shows the number of fruit scanned for each device (diagonal counts) as well as pairs of scans from different devices measured on the same fruit (offdiagonal). There was some mixing of the devices in the training set, but not in the 2019 data, where devices were restricted to a certain site (Table 6.3).

Table 6.2: Number of fruit scanned per device in the training dataset. Diagonal elements indicate the total number of fruit measured by the respective device. Off-diagonal elements show the number of fruit measured by pairs of devices.

Device	**KK1**	**KK2**	**TP1**	**TP2**	**TP3**
KK1	**1397**	397	1089	90	90
KK2	397	**881**	90	574	573
TP1	1089	90	**1985**	986	986
TP2	90	574	986	**1616**	1615
TP3	90	573	986	1615	**1616**

Table 6.3: Number of fruit scanned per device in the 2019 test dataset. The diagonal elements indicate the total number of fruit measured by the respective device. Off-diagonal elements show the number of fruit measured by pairs of devices.

Device	**KK1**	**KK2**	**TP1**	**TP2**	**TP3**
KK1	**797**	797	0	0	0
KK2	797	**798**	0	0	0
TP1	0	0	**592**	592	591
TP2	0	0	592	**594**	593
TP3	0	0	591	593	**593**

6.4.2 Model Architecture

The model architecture used in these experiments was based on previous work on the same dataset. The encoder model consisted of an initial convolutional layer with bias set to zero and the weights set to the Savitzky-Golay second derivative, with a window size of 13 and second-order polynomial. This layer was frozen so that it was equivalent to applying the respective Savitzky-Golay preprocessing. Following this, a convolutional layer with 50 filters of size 13 and a linear activation function was used as in (Wohlers et al., 2025). The encoder was completed by flattening the convolutional layer and connecting a single linear dense layer of size 16. A regression head was then added to the model, which included a single linear neuron. The justification for the simple architecture was to enable a fair comparison between the Barlow method and PLSR, while avoiding confounding results with model complexity.

6.4.2.1 PLSR model

The PLSR regression was fit using scikit learn (Pedregosa et al., 2011) and used Savitzky-Golay second derivative, with a window size of 13, and a second-order polynomial. All models used 16 latent variables to provide a direct comparison with the neural network encoder's architecture. This choice was validated using 10-fold cross validation (CV) PLSR models across different training set sizes. For SSC prediction, 16 components minimised the CV RMSE for moderate training sizes (N=250 fruit, 648 scans; N=500 fruit, 1297 scans), with 17 optimal at N=1000 fruit (2122 scans). For DMC prediction, cross validation selected 20 and 21 latent variables at N=100 and N=250, respectively, but evaluating on the test set showed optimal numbers of latent variables to be 12 and 13, with 16 latent variables giving similar performance. The small training datasets (50 fruit, 150 scans) had a slightly lower optimal (13 latent variables) for the SSC and DMC prediction. The consistent use of 16 latent variables across experiments ensures that architectural changes to the models did not affect the Barlow Twins loss's performance.

6.4.3 Training Procedure

All analyses were conducted in Google Colab notebooks with a T4 GPU and high RAM backend. Software used included Python 3.11.13, Tensorflow 2.18.0, Tensorflow Probability 0.25.0, and scikit-learn 1.6.1. Models were trained for 1000 epochs using the Lamb optimiser (You et al., 2019) with a learning rate of 0.01.

A custom loss function was used for training with a weighted combination of 10 x Barlow loss + 0.5 consistency MSE loss + 0.5 prediction MSE loss. The λ parameter in the Barlow loss was set to 1/15 (the number of latent variables minus one), so that the invariance and redundancy terms in equation 6.1 are equally weighted. For n latent variables, the $n \times n$ cross-correlation matrix has n diagonal elements and $n(n-1)$ off-diagonal elements. For the experiments, the Barlow loss was computed using only the labelled subset of data (Barlow Labelled), or using all available paired device measurements (Barlow Full), regardless of whether the fruit had DMC or SSC labels.

6.4.4 Experiments

6.4.4.1 Semi-Supervised Learning

To investigate the effectiveness of semi-supervised learning we used different labelled training set sizes (50, 100, 250, 500, 1000, 2924 fruit) for separate models predicting DMC and SSC. These fruit were selected by taking the n most recent fruit and their respective scans in the training set. This approach was more realistic than sampling a random subset of scans from the full training set. The full 2924 fruit training set (7495 scans) was used as unlabelled data to train the model encoder using the Barlow and consistency loss functions. This semi-supervised approach (Barlow Full) was compared to a the same model with the Barlow loss calculated only on the subset of labelled fruit (Barlow Labelled, using the paired device measurements available for those fruit, rather than on the full unlabelled dataset. Both methods combined the Barlow loss

with the regression MSE loss, but differed in the size of the dataset used for computing the Barlow loss. Similarly, autoencoder models based on methods from (Said et al., 2022) were trained on the same datasets (AE Full and AE Labelled). Additionally, models trained with the MSE loss only and PLS regression with Savitzky-Golay preprocessing were included as baselines.

Six modeling approaches were compared based on their loss functions and use of unlabelled data:

- **Barlow Full:** Barlow Twins loss computed on full training set measurements (unlabelled) + MSE loss on labelled subset
- **Barlow Labelled:** Barlow Twins loss computed only on labelled paired measurements + MSE loss on labelled subset
- **AE Full:** Autoencoder reconstruction loss computed on full training set measurements (unlabelled) + MSE loss on labelled subset
- **AE Labelled:** Autoencoder loss computed only on labelled paired measurements + MSE loss on labelled subset
- **MSE:** Standard supervised learning with MSE loss on labelled subset only
- **PLSR:** 16-component PLS regression with Savitzky-Golay preprocessing

6.4.4.2 Calibration Transfer

Calibration transfer experiments involved excluding one device at a time from the training data. In each iteration, the excluded device served as the target device for testing, with the remaining four serving as source devices for training. The scans from the left-out device were completely excluded from calculating the Barlow loss to prevent data leakage.

Models were fit using different loss functions across increasing labelled training set sizes (N = 50, 100, 250, 500, 1000, 2924 fruit):

- **Barlow Full:** Barlow Twins loss computed on full training set measurements from source devices (unlabelled) + MSE loss on labelled subset
- **Barlow Labelled:** Barlow Twins loss computed only on labelled paired measurements + MSE loss on labelled subset
- **MSE:** Standard supervised learning with MSE loss on labelled subset only
- **PLSR:** 16-component PLS regression with Savitzky-Golay preprocessing

Model performance was assessed on the 2019 test data for the respective left out device and aggregated across all devices.

6.4.4.3 Augmentation

The original Barlow Twins paper used image augmentation to train the network. This is a viable alternative when collected data is only available for a single device. The augmentation method implemented here involves adding random samples from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the empirical covariance of differences among devices on kiwifruit measured prior to the current dataset on a wider range of devices. The dimension of $\mathbf{\Sigma}$ is equal to the number of wavelengths included in the model training. The augmented spectrum is then calculated as:

$$\mathbf{x}' = \mathbf{x} + \boldsymbol{\varepsilon}, \quad \text{where } \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{6.3}$$

See (Wohlers et al., 2023) for more details. If the Barlow loss is used for training on unlabelled scans from a single device, it is effectively only enforcing the latent variables to be orthogonal and has no invariance to device variation. For these experiments, the model was fitted on data from a single device, while augmenting the unlabelled spectra to train the encoder using the Barlow loss function and MSE loss to train the regression head on the labelled data. For comparison, equivalent models trained using only the labelled data were

also created. Model performance was assessed on the 2019 test data for the respective device and all other devices separately.

6.5 Results

6.5.1 Semi-supervised learning

Table 6.4 summarises the performance on the 2019 data of semi-supervised (SSL) and supervised learning (SL) approaches using different loss functions (Barlow, AE, MSE) compared to a PLSR baseline.

Generally, all models improved as the number of fruit included in the labelled training set increased. The main exception was for the autoencoder loss, which performed the worst until the full training set was used. However, at N = 2924 fruit, the SL and SSL methods are equivalent and there is considerable differences observed (RMSE 1.31 and 1.43 compared to 1.58 and 1.64 respectively), indicating it may be less stable than the Barlow loss.

Among the neural network approaches, the Barlow loss consistently achieved the best performance, especially with the smaller training sets. For large training sizes, there appeared to be no negative impact to using the Barlow loss, with the test RMSE being practically identical to that of training the same model using only the MSE loss. The SSL method with the Barlow loss reduced the RMSE at the smaller training sets for the DMC, with SSL and SL results converging as the sample size increases. This is expected as the data used to calculate the Barlow loss in the SL model is a subset of the full data used in the SSL.

These RMSE reductions were consistent across the range of observed DMC and SSC values.

Compared to the PLS baseline, the Barlow and MSE models outperformed PLSR for DMC. Interestingly, when training on 100 labelled fruit instead of 50, the RMSE increased for the SL neural networks, but not PLSR or the SSL neural networks. For SSC, PLSR had the lowest RMSE until the full training

set was used.

Table 6.4: Comparison of different loss functions tested on the 2019 data across different Training Set Sizes. N represents the number of labelled fruit samples in the training set. Full Data uses all training spectra (labelled and unlabelled) for calculating the Barlow and Autoencoder losses, while Labelled uses only the labelled spectra for these losses. All RMSE values are calculated on the same 2019 labelled test set.

Training	**Measure**	**2019 Test RMSE**					
		Full Data		**Labelled Data Only**			
N		**AE**	**Barlow**	**AE**	**Barlow**	**MSE**	**PLS**
50	DMC	4.23	**1.55**	3.45	1.69	1.60	1.87
100	DMC	3.19	**1.33**	3.72	2.17	2.45	1.54
250	DMC	2.18	**1.17**	2.05	1.21	1.27	1.27
500	DMC	2.40	1.20	2.14	1.24	**1.17**	1.38
1000	DMC	3.97	1.19	1.84	**1.15**	**1.15**	1.42
2924	DMC	1.31	1.17	1.58	1.14	**1.11**	1.42
50	SSC	3.19	2.07	2.77	1.97	4.61	**1.74**
100	SSC	2.91	2.26	2.28	**2.22**	2.91	2.25
250	SSC	2.29	1.98	2.34	1.92	2.03	**1.70**
500	SSC	2.73	1.96	2.32	1.84	2.22	**1.42**
1000	SSC	2.15	1.65	2.42	1.63	1.70	**1.39**
2924	SSC	1.43	1.49	1.64	**1.43**	1.46	1.62

6.5.2 Calibration transfer

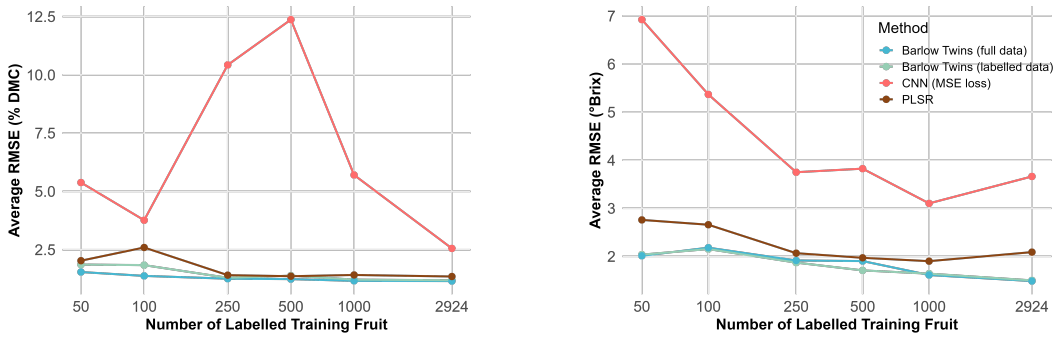
The results of leaving a single device out of the training set and testing on that respective device are summarised in Figures 6.3a and 6.3b. As expected, the RMSE generally decreased across the models as the number of fruit included in the training set increased.

For both DMC and SSC prediction the Barlow loss models consistently outperformed the PLSR baseline and prevented the extremely poor performance of the equivalent CNN with the MSE loss only. This CNN showed instability at the low to moderate training sizes with RMSE exceeding 12% for DMC at N=500 fruit and 7°Brix for SSC at N=50 fruit despite low training losses, indicating overfitting to source devices. In contrast, both Barlow models were stable across all labelled training sizes.

Including the full data for the Barlow Twins loss showed some improvement over using the labelled data only (Figure 6.3a) with lower RMSE for the smaller training set sizes (50 and 100 fruit). However, this was not observed for the SSC prediction (Figure 6.3b, where the two Barlow Twins models showed practically identical performance across all the training sizes. The Barlow Twins loss provides regularisation by constraining the structure on the encoder embeddings to have low redundancy in the latent variables and high device invariance, similar to PLSR latent variables. Neither of these methods had the extreme RMSE observed with the less constrained MSE training.

6.5.2.1 Augmentation

Results of training on augmented spectra from a single device and predicting on 2019 data from the same device are shown in Figures 6.4a and 6.4c. Results for predicting on the remaining four devices not used in training are shown in Figures 6.4b and 6.4d. Contrary to previous results, there appears to be little benefit in using the full unlabelled training set for a given device when minimising the Barlow loss over using the labelled data only. The RMSE values for the two Barlow Twins methods are consistently similar across the two measures and training sizes. The only notable exception is SSC at N=100 fruit when predicting on the same device, which appears to be driven by a single device at that training size(data not shown). The similarity in Barlow Twins results may be explained by the augmentation used, adding an independent random sample from a multivariate Gaussian distribution. Since this



(a) (b)

Figure 6.3: Calibration transfer results for DMC (a) and SSC (b) prediction using leave-one-device-out validation. Models were trained on four source devices and tested on the remaining target device using different loss functions and increasing numbers of labelled fruit in the training set. RMSE is averaged across all five target devices. Barlow Twins (full data) uses all training set data for the Barlow loss, while Barlow Twins (labelled data) uses only the labelled data.

is independent of the observed spectra being added to, there is less to gain in increasing the unlabelled data than if the augmentation depended on the observed spectra. More complex augmentation techniques that are related to the observed spectra may benefit from semi-supervised techniques.

The Barlow loss function provided, whether used with the labelled data only or the additional unlabelled data, saw lower RMSE than training with the MSE loss function at the smaller training sizes. This was more pronounced when testing on devices left out of training (Figures 6.4b and 6.4d).

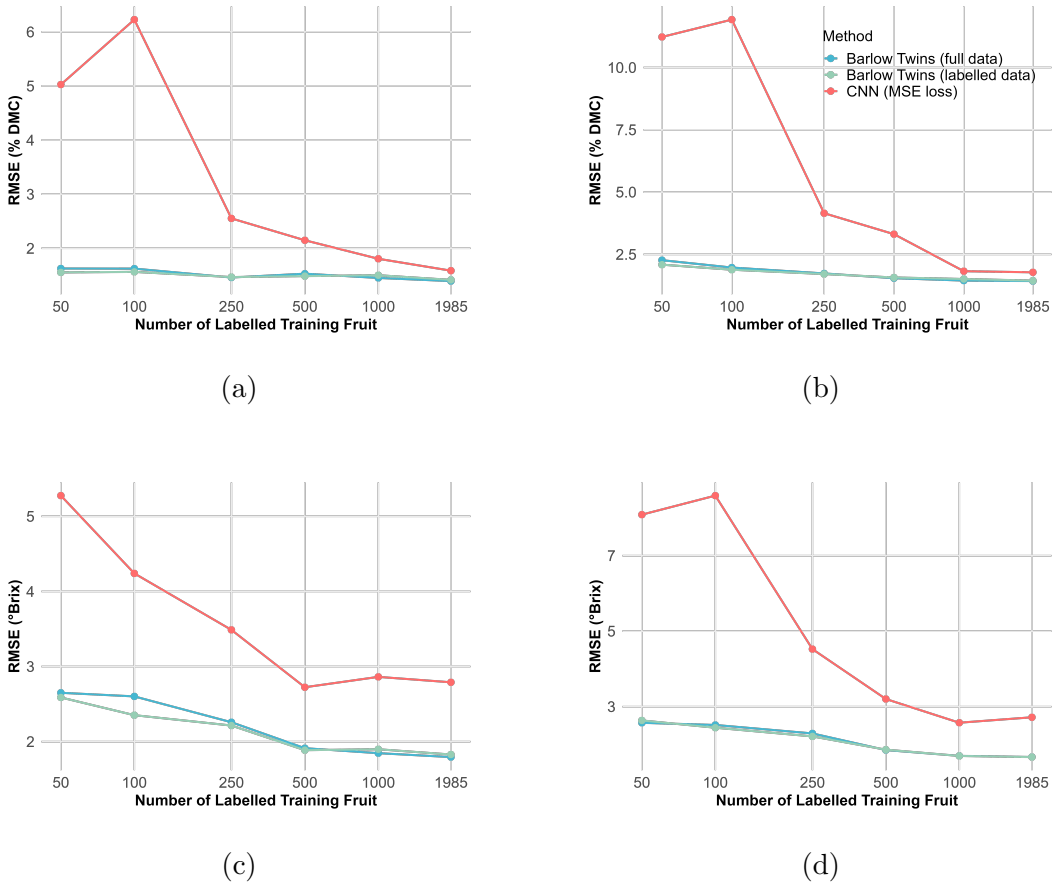


Figure 6.4: Models trained with different loss functions on augmented spectra from a single device. Models were evaluated on 2019 data for the same device (DMC in (a), SSC in (c)) and from the remaining four devices (DMC in (b), SSC in (d)). RMSE is averaged across the five training devices with varying numbers of labelled fruit. Barlow Twins (full data) uses all training set data for the Barlow loss, while Barlow Twins (labelled data) uses only the labelled data. The rightmost point (1985) represents the maximum available training data, which varied by device (881 to 1985 fruit).

6.6 Discussion

6.6.1 Advantages over traditional approaches

Barlow Twins improves upon PLSR by explicitly learning device-invariant representations. PLSR learns latent variables that maximise covariance with the target while enforcing orthogonality. However, it lacks a specific mechanism to ensure that these latent variables are invariant to device variability. The Barlow loss simultaneously reduces redundancy (like PLSR) while also enforcing that paired measurements of the same fruit produce similar embeddings, regularising the network to prevent overfitting to device specific features. This overfitting is observed in the calibration transfer results in Figure 6.3, where the CNN trained with the MSE loss overfits to the training devices, producing the highest RMSE for both SSC and DMC. In contrast, the models that included the Barlow loss were stable because of the invariance constraint, which penalises device-specific encodings. As the labelled data increase, PLSR's regularisation via dimensionality reduction and covariance maximisation becomes sufficient, and the benefits of semi-supervised learning diminish (Table 6.4).

6.6.2 Limitations and Challenges

The current research focused on the Barlow Twins method in a semi-supervised method, without optimising the architecture of the encoder or regression head. The results may vary for a more flexible architecture. Limited experiments with deeper encoders encountered training issues with exploding gradients, although this was overcome by using the SELU (Klambauer et al., 2017) activation function and LeCun norm initialisation for the encoder.

A potential issue is that fixing the number of latent variables to 16 may disadvantage PLSR models. We investigated this by repeating the experiments in Table 4 using 10-fold CV to select the optimal number of PLSR latent variables. The CV-optimised models showed performance similar to, or slightly worse than, that of the models with 16 components. Some improvement was

observed at the largest training size, where CV selected more latent variables, but this did not change the overall conclusions.

Another limitation is the augmentation used for NIR spectra; we employ the method from (Wohlers et al., 2023), but alternative augmentation methods could also be considered. Temperature variations, for example, have been shown to be associated with wavelength shift in raw milk (Diaz-Olivares et al., 2024). Augmenting the data by wavelength shifting for use with this model may improve model sensitivity to these effects.

DMC and SSC may vary throughout the fruit depending on the tissue location sampled. The NIR measurements were taken at consistent locations to minimise this variation. The success of the Barlow Twins method suggests that the paired measurements were based on sufficiently similar tissue.

While we investigated the Barlow Twins approach, other semi-supervised learning approaches could also perform well. The comparison to the autoencoder method used in (Said et al., 2022) was limited due to the lack of details of their model architecture. The autoencoder employed here was based on a separate paper looking at NIR spectra from the same brand of devices as used here (L. Zhang et al., 2025), but there may be alternative autoencoder architectures better suited to kiwifruit. Other self-supervised methods could also improve performance. For example, variance-invariance-covariance regularisation (VicReg) (Bardes et al., 2021) is a similar method to Barlow Twins. While preliminary experiments (not shown) with this did not improve over the Barlow Twins method, a more thorough investigation may yield improved results.

6.6.3 Practical Implications

Large quantities of unlabelled data may not always be available, but the method is reasonably straightforward to implement and showed benefits even with smaller datasets. The computational overhead is modest and mainly due to the increased number of calculations involved in comparing all pair-wise

device combinations of the same fruit for the Barlow Twins loss. For optimal results, unlabelled data should be collected in a systematic way to improve robustness to various effects. This could be multiple scans on the same fruit at different temperatures, multiple devices scanning the same fruit (as observed here), or multiple measurements at different locations on the fruit, as is seen in commercial fruit graders.

6.7 Conclusions

The results show the potential benefit of self-supervised learning for NIR spectroscopy, through adapting the Barlow Twins method for regression tasks. The greatest benefit over traditional PLS regression in terms of RMSE is consistently observed when the labelled data is limited. The robustness of the method to device variation could be particularly useful in practice. Labelling enough spectra to train a new device using standard supervised methods, such as PLSR, can be time-consuming and expensive. For example, each kiwifruit DMC measurement requires two equatorial slices to be dried for 24 hours at 65 ° C (McGlone & Kawano, 1998). Conversely, spectral measurements of the same kiwifruit are simple and quick, taking about 4-6 seconds per scan for the Felix f-750 handheld devices used to collect the data in this study (Felix Instruments, 2019).

Future work could explore alternative self-supervised learning methods, find optimal model architectures for this application, and appropriate augmentation methods for the model to learn invariance. Additionally, measuring the same fruit from the same device multiple times would allow for an investigation into whether methods such as the one presented in this paper are able to reduce measurement error.

6.8 Data

All data used in this paper are available through Anderson et al. 2020 (Anderson et al., 2020) and Wohlers et al. 2023 (Wohlers et al., 2023). Code is available at https://github.com/mwohlers/BarlowTwins_NIR.

6.9 Acknowledgements

The research received financial support by the New Zealand Ministry of Business Innovation & Employment (MBIE) as part of the Endeavour funded project *Perfecting storage life prediction for delivery of high quality fruit.*

Chapter 7

Synthesis and Conclusions

This thesis introduces three methods to assist with deep learning models trained on small datasets. The methods described here resulted in significant improvements when used compared to standard partial least squares regression (PLSR) models. The previous recommended approach to NIR spectral modelling with deep learning is shown in Figure 7.1. This is updated in Figure 7.2 to include the contributions presented here.

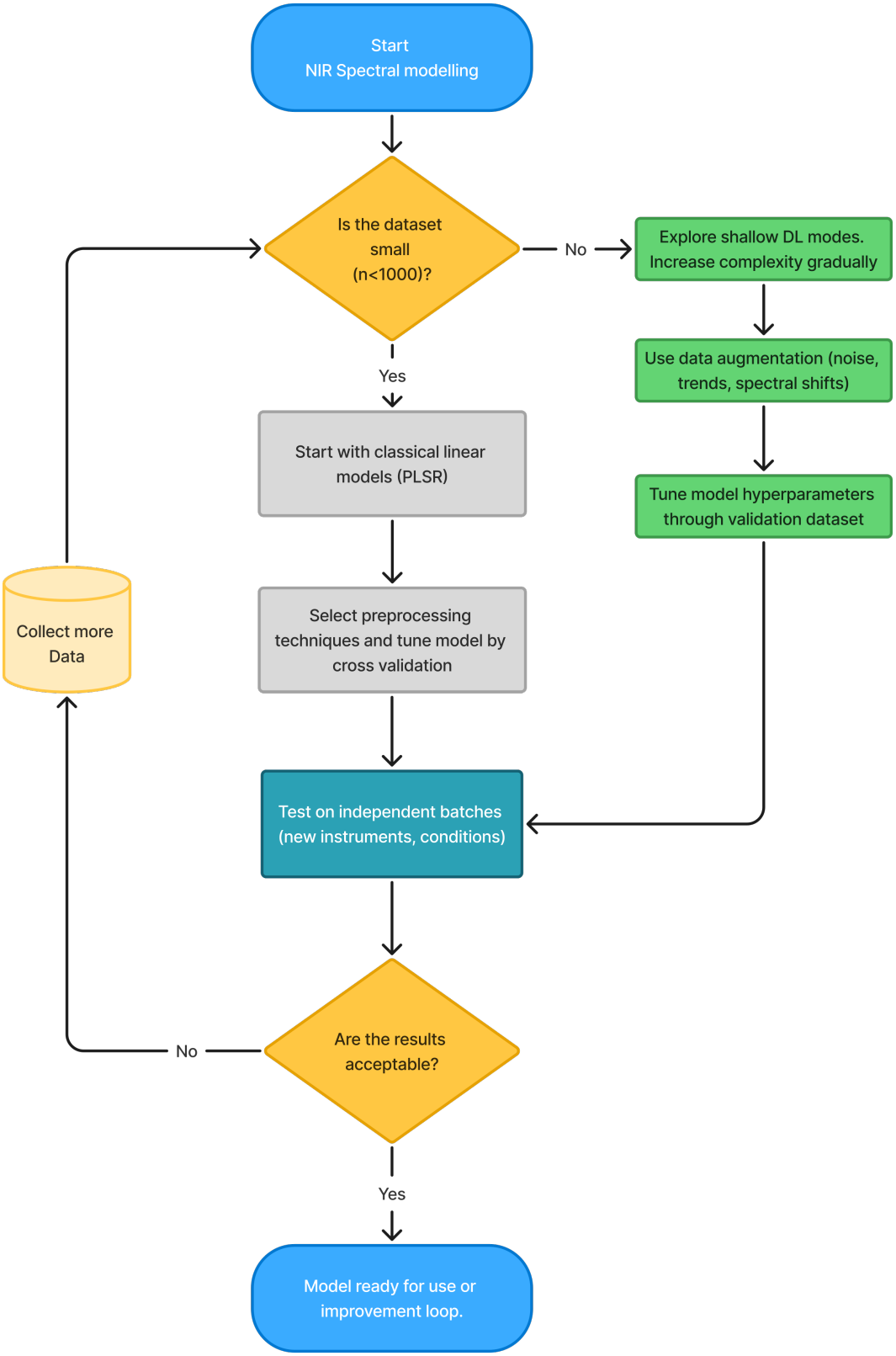


Figure 7.1: Recommended workflow for modelling NIR spectral data based on advice presented in Mishra et al. (2022)

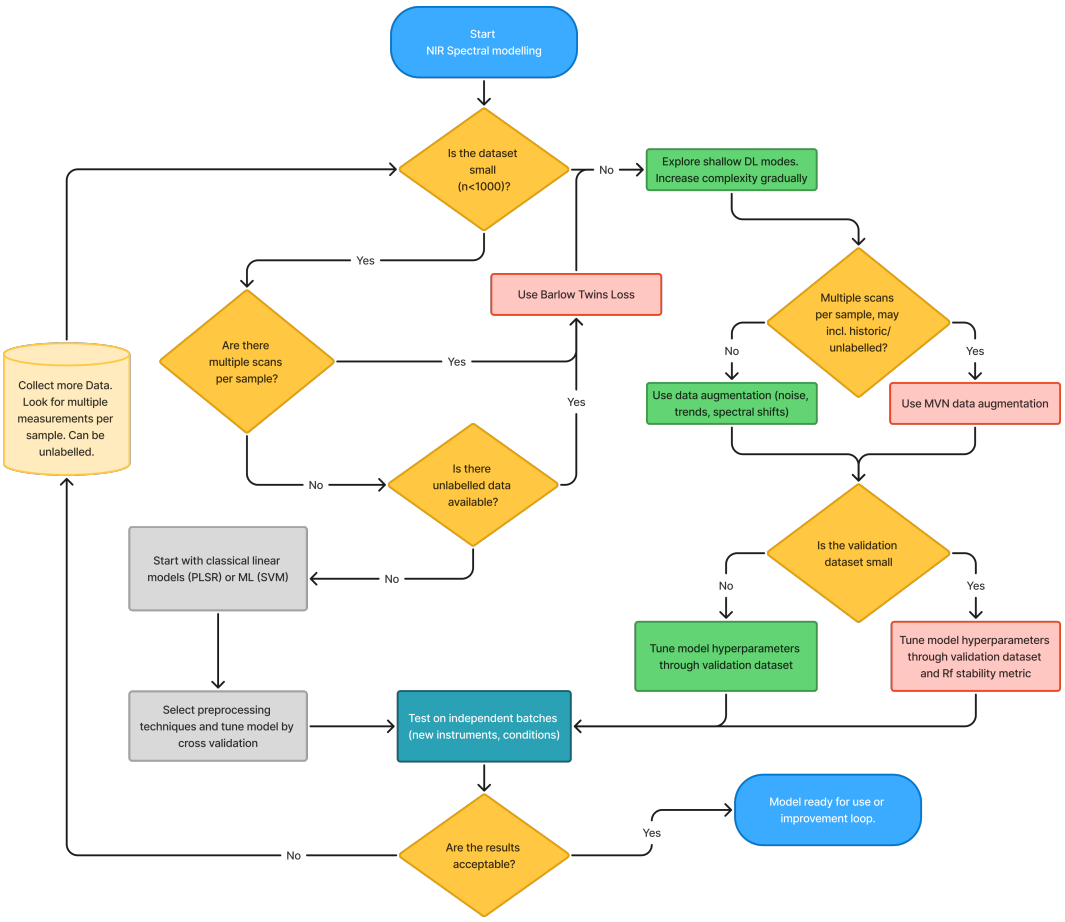


Figure 7.2: Recommended workflow for modelling NIR spectral data updated based on the research presented in this thesis. Direct contributions are highlighted in red.

7.1 Overview and Key Findings

This thesis demonstrates the potential of deep learning for near-infrared (NIR) spectroscopy to produce global models that generalise across multiple devices and also addresses the related challenges of training and validating these models when labelled data is limited. This was achieved through three methods: a semi-supervised learning approach using the Barlow Twins loss function, data augmentation that retains the spectral relationships, and a relative stability metric to assist in model selection. With these methods, the research demonstrated that deep learning improved predictions over the standard partial least squares regression (PLSR) on small datasets, which was consistent with previous findings on larger datasets.

The main finding is that deep learning can perform well for NIR spectroscopy when methods are used to address the data constraints common in the domain. Instead of requiring large, labelled datasets to train and tune deep learning, the approaches described in the thesis leverage the structure of NIR data. This includes using multiple measurements on the same sample to enable contrastive learning and using the correlation of these measurements to generate realistic augmented data for training. Additionally, the highly correlated nature of the NIR data was used to construct a metric measuring the stability of predictions when spectra were perturbed by a smooth transformation relative to a random one.

Chapter 4 examined whether augmentation that retained the relationships amongst spectra could improve robustness and calibration transfer. The results showed that preserving spectral correlation structure is important. Augmentation by sampling from multivariate normal distributions based on empirically estimated covariance matrices improved training stability and prevented overfitting for both shallow and deep convolutional neural network (CNN) architectures compared with augmentation based on the transformations outlined in Bjerrum et al. (2017a) or independent Gaussian noise. The MVN augmentation enabled successful calibration transfer, with models trained on devices

from one site, to predict dry matter content (DMC) and soluble solid content (SSC) on unseen devices, particularly when the covariance matrix was estimated from a wide set of devices, including the test devices. The approach successfully transferred from kiwifruit to a mango dataset measured with the same make of device, by scaling the covariance matrix. This suggests broader applicability than the kiwifruit dataset alone.

Chapter 5 explored whether the relative stability to diffeomorphisms could be used for model selection and stopping rules for training as had been observed in computer vision tasks. This involved defining an appropriate transformation based on Bartlett's decomposition of the Wishart distribution. The mean square error (MSE) between the predictions with and without this transformation was compared to the prediction variations when adding uncorrelated Gaussian noise of similar magnitude by taking the ratio of the respective MSEs (defined as R_f). R_f was found to correlate with the validation and test dataset MSEs across a wide range of deep learning model architectures, suggesting it could be used as a model selection metric. R_f results were comparable to those based on the validation dataset MSEs, with best model selections using both metrics in combination. Similarly, R_f followed a similar trace over epochs as the validation MSE and gave comparable performance as a stopping rule. Interestingly, it was also found that R_f based on the training dataset was almost identical to that of the validation and test datasets, suggesting that it could be useful when validation data is limited.

Chapter 6 investigated the application of Barlow Twins contrastive learning to NIR spectroscopy regression tasks. The main innovation was treating spectra from different devices measuring the same sample as paired "views", rather than augmenting the spectra to generate them. The Barlow loss was combined with a standard MSE regression loss for labelled data. This semi-supervised learning approach outperformed PLSR. The biggest improvements were observed when training on small (50-500 fruit) labelled datasets, with diminishing returns as the sample sizes increased. Another benefit of the Barlow

loss function is that it provides regularisation by rewarding orthogonality and reducing redundancy in the latent space. This appeared to help with calibration transfer tasks when applying models to new devices not included in the training.

The method also outperformed deep learning models trained without the Barlow loss (MSE loss only) and an autoencoder based semi-supervised approach. This showed that contrastive learning is a natural fit for NIR data when multiple measurements are taken on the same sample.

7.2 Limitations

7.2.1 Limitations of the Data Augmentation Approach

The augmentation used in the study was based on observed differences among devices measuring the same kiwifruit, with samples drawn from a multivariate normal distribution (MVN) added to the observed spectra. This method essentially assumes that the spectral variation due to the devices is independent of the sample being measured i.e., that it is additive. While the results were promising, this assumption seems unreasonable and given the limited scope of the datasets may not transfer to other devices, fruit, seasons etc. It would be useful to study alternative augmentation techniques and to devise a method that accounts for the sample being measured, e.g., multiplicative effects, as well as sources of other variation, such as temperature.

Another limitation of the study is that it relies on the quality of the covariance matrix estimate. As discussed in the Mahalanobis outlier detection method, the empirical covariance estimate can be sensitive to outliers. The study did not examine the effect of the covariance matrix's quality or whether more robust covariance estimation techniques could improve results. There was also no recommendation of the required sample size needed to estimate the matrix reliably. Similarly, the experiments used 50 to 100 augmented samples per observation, resulting in significant computational overhead. It would

be useful to show optimal augmentation levels by presenting results across different numbers of augmented samples.

7.2.2 Limitations of the R_f Stability Metric

The R_f metric correlated with test performance at a level similar to the validation MSE, but less so for the mango dataset. This could indicate that the usefulness of R_f as a model selection metric may depend on the dataset or attribute of interest. A wider range of datasets would need to be investigated to understand if this is true and, if so, what the possible causes of the variation are. Some models had low R_f but large test MSE, suggesting they were very sensitive to uncorrelated noise. This could be explored in future work to understand when it occurs and whether modifications to the method are needed to account for it. Another shortcoming is that the method does not detect if a model's predictions are unrealistic. This implies that while the method does provide useful information, it should be used alongside other information, such as training and validation MSE, to make model selection decisions.

This method also lacks a strong justification for the particular linear transformation diffeomorphism used. It is not the only diffeomorphism, and investigating other methods could provide better model selection. Similarly, while it is understandable that the proposed method does not work for PLSR, the alternative version designed for PLSR lacks a strong theoretical justification. More empirical results would be needed before recommending its use with PLSR models.

Another drawback, similar to augmentation, is the computational overhead. The diffeomorphism used a Williams design matrix, resulting in 202 times the original dataset size. This greatly increased training time from minutes to hours. There may be more efficient methods to balance the spectral order influence on the results.

7.2.3 Limitations Semi-Supervised Learning via Barlow Twins

The biggest improvement from the Barlow Twins application to NIR data was achieved through pairs of observed spectra from the same kiwifruit. This requirement for multiple measurements per sample could be seen as a limitation if collecting such data is not realistic. For example, only one device is available or perhaps it is too logistically difficult to measure the fruit at different temperatures. While augmentation is presented as an alternative in such cases, the data augmentation used is the same as that used above, with the same limitations. This includes the need for historical data from multiple devices to estimate the covariance matrix. An augmentation technique less dependent on multiple measurements would allow the Barlow loss to be used in these cases.

While comparisons with other semi-supervised methods, including autoencoder based models, were provided, further semi-supervised approaches could be investigated. There was a limited comparison with another contrastive learning approach, VICReg, but it was not included in the results because it was not competitive with the Barlow model. Further tuning of VICReg hyperparameters could improve performance and should be explored, as should other contrastive learning methods. Another difficulty was trying to replicate competing methods from the literature as they either lacked full details of the architecture or the description was not accurate and had to be slightly modified to accommodate the datasets used here.

PLSR was included for comparison and used the same number of latent variables (16) as the CNN encoder. These were not optimised, and it may be that the optimal number of dimensions for the Barlow Twins CNN is smaller or larger than that for PLSR. This would need to be explored, especially when using more complex CNN architectures, to achieve a more meaningful comparison.

7.2.4 Limitations Across All Studies

The current research was limited in scope, as it included only two similar datasets. While these are relevant to the area of application, it would be beneficial to validate the methods on a wider range of datasets. This includes a more varied set of devices rather than only the F-750 used here, and samples measured under a wider range of conditions such as temperature, locations, different fruits, e.g., apples and avocados. Including data from industrial fruit graders would also test whether the methods are suitable for production, as the data are high-throughput, with more scans per sample. It is also worth noting that the methods developed in this thesis should translate to classification problems, but this was not tested in the current settings.

Another limitation of the work is that the experiments mainly used simple architectures and did not assess the methods' suitability for more complex deep learning models, such as the increasingly popular transformer networks. While some hyper-parameter optimisation was used, it was not exhaustive. Different activation functions and weight initialisation methods were compared in a heuristic rather than a systemat ic way. More research here could lead to improved results.

All three studies utilised wider wavelength ranges than what is typically recommended in the literature for PLSR. Although baseline experiments supported this choice, this likely resulted in a larger number of PLSR latent variables than comparable analyses in the literature. Including information from the visible range could make the results less generalisable. For example, if colour correlates with maturity, this relationship may vary depending on the cultivar, grower region, or season.

Preprocessing approaches varied across the studies for PLSR, with Chapters 4 and 6.1 employing Savitzky-Golay second derivatives while Chapter 5 used minimal preprocessing to test the stability metric suitability for PLSR. This inconsistency may have affected the comparability of PLSR baselines across papers.

7.3 Future Work

The methods developed in this thesis show potential for furthering deep learning in NIR spectroscopy. Future research should build on these while addressing the limitations outlined in the previous section.

7.3.1 Enhanced Data Augmentation

More advanced methods of augmentation could be investigated, particularly to extend beyond the assumed additive device-to-device variation.

One possible solution would be to embed the spectra into a lower-dimensional space and apply the augmentation in the latent space. The augmented latent scores would then be reconstructed back to the original scale as the augmented spectra. The encoder and decoder would be non-linear and so as to remove the additive restriction. A natural fit for this problem would be to investigate variational autoencoders (VAE) for generating the data. Rather than encoding observations into a single point, VAEs embed observations into a distribution (usually Gaussian) in the latent space, which can then be sampled to generate the augmented spectra. During this thesis, VAEs were initially investigated for their suitability, but were found to be difficult to train.

Other priorities should be to refine the guidance on the current approach. This includes experimenting with different numbers of augmented samples across training sizes and model architectures. This would provide clearer guidance on the optimal number of generated samples, potentially reducing computational overhead by avoiding the generation of unnecessarily large samples.

Given the reliance on the covariance matrix in the augmentation method, the effects of outliers and the benefits of robust covariance estimation could be systematically studied. These refinements would result in clearer guidance on the use of the proposed MVN augmentation method.

7.3.2 Alternative Stability Metrics

The development of R_f focused on a single diffeomorphism, so future work would evaluate alternative smooth transformations. Including non-linear transformations that better reflect actual spectral changes, such as baseline shifts due to temperature, could potentially improve model selection performance. This could be combined with the data augmentation work as a potential transformation. Systematically comparing the different techniques and their use in combination with other metrics, such as validation loss, would give improved strategies for model selection. For example, investigate whether using R_f as a metric to reduce the learning rate and as an early stopping metric during training, and the validation loss for final model architecture selection. The current calculation of R_f involves the use of a Williams design to balance for the order effect of the Cholesky decomposition, and so alternative methods that do not use this could be more computationally efficient.

7.3.3 Semi-Supervised Learning Extensions

The Barlow Twins approach showed significant benefits when training on small labelled datasets. Evaluation of other contrastive methods would be an interesting avenue to explore, including a more thorough exploration of the VICReg with appropriate tuning. Additionally, a study on the effect of various batch sizes could also be undertaken. Generally, large batch sizes are used with these methods, but with the current NIRS setting being based on observed data, these may not be feasible. The trade-off of using small batch sizes when calculating the Barlow loss could be investigated.

Running additional experiments on data comprising multiple measurements at different temperatures would also assess the applicability of the method on problems outside device variation. Being able to combine environmental and systematic effects with the Barlow Twins loss rather than treating them separately would also be of interest.

It would be interesting to evaluate additional data augmentation techniques

with the Barlow Loss for situations were the data only includes a single device, and compare to the MVN results presented in this thesis.

7.3.4 Advanced Architectures and Hyperparameter Optimisation

Recent trends in NIR spectroscopy include more advanced architectures, including transformer networks, and multimodal networks. A systematic evaluation of these techniques to evaluate their appropriateness and potential performance gains using the techniques outlined in this thesis would be of interest. These results should also inform any potential modifications or advances needed for these thesis techniques. For example, the Barlow Twin method may need to be modified to adapt to inception or multimodal networks.

Similarly, a more systematic evaluation of hyperparameter tuning could be conducted. Given the complexity and number of parameters, Bayesian optimisation could be more widely implemented in future experiments, and guidelines for its use should be provided. The specification of networks often uses a standard initialisation method for the activation function, e.g., He normal. There were a number of occasions when model training failed due to exploding gradients. Alternative methods to reduce or avoid this could also be investigated. One option is to use the results from a simpler model to initialise the network. It is reasonably simple to encode a PLSR model into a standard CNN network for the training data. Data outside of this distribution will diverge from the PLSR depending on the activation functions. More complex networks would be more difficult to initialise this way, but may only need an approximation to improve training stability.

7.3.5 Datasets

A high priority is to evaluate the methods across a wider range of datasets, including different fruit varieties, products (e.g., grains), and spectrophotome-

ters (including industrial graders). This includes available open datasets used elsewhere in the literature for regression and extended to classification tasks, where possible, extending to data for multimodal models, such as Raman or MIR spectroscopy. Having more diverse datasets in terms of devices, seasons, and grower regions would also strengthen the method evaluations.

7.4 Concluding Remarks

Returning to the thesis statement, the work contained here demonstrates that self-supervised learning and appropriate data augmentation can significantly improve the performance and generalisation of deep learning for near-infrared spectroscopy. The thesis answers the three research questions through appropriate augmentation, stability-based model selection, and contrastive learning. In doing so, it provides methods for applying deep learning to NIR spectroscopy when labelled data is limited and cross-device generalisation is required.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2019). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, Version 1.15*. https://github.com/tensorflow/tensorflow/releases/tag/v1.15.0

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Acquarelli, J., van Laarhoven, T., Gerretzen, J., Tran, T. N., Buydens, L. M., & Marchiori, E. (2017). Convolutional neural networks for vibrational spectroscopic data analysis. *Analytica Chimica Acta*, *954*, 22–31. https://doi.org/10.1016/j.aca.2016.12.010

Alagappan, L., Chu, J. E., Chua, J. H., Ding, J. W., Xiao, R., Yu, Z., Pan, K., Elejalde, U., Lim, K. J., & Wong, L. (2023). Class-specific correction and classification of NIR spectra of edible oils. *Chemometrics and Intelligent Laboratory Systems*, *241*, 104977. https://doi.org/10.1016/J.CHEMOLAB.2023.104977

Allgaier, J., & Pryss, R. (2024). Cross-Validation Visualized: A Narrative Guide to Advanced Methods. *Machine Learning and Knowledge Extraction 2024, Vol. 6, Pages 1378-1388*, *6*(2), 1378–1388. https://doi.org/10.3390/MAKE6020065

Anderson, N. T., Walsh, K. B., Flynn, J. R., & Walsh, J. P. (2021). Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content. II. Local PLS and nonlinear models. *Postharvest Biology and Technology*, *171*, 111358. https://doi.org/10.1016/J.POSTHARVBIO.2020.111358

Anderson, N. T., Walsh, K. B., Subedi, P. P., & Hayes, C. H. (2020). Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content. *Postharvest Biology and Technology*, *168*, 111202. https://doi.org/10.1016/J.POSTHARVBIO.2020.111202

Bampi, M., Scheer, A. D. P., & De Castilhos, F. (2013). Application of near infrared spectroscopy to predict the average droplet size and water content in biodiesel emulsions. *Fuel*, *113*, 546–552. https://doi.org/10.1016/j.fuel.2013.05.092

Bardes, A., Ponce, J., & LeCun, Y. (2021). VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *ICLR 2022 - 10th International Conference on Learning Representations*. https://arxiv.org/pdf/2105.04906

Bjerrum, E. J., Glahder, M., & Skov, T. (2017a). Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics. http://arxiv.org/abs/1710.01927

Bjerrum, E. J., Glahder, M., & Skov, T. (2017b). Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics. http://arxiv.org/abs/1710.01927

Blazhko, U., Shapaval, V., Kovalev, V., & Kohler, A. (2021). Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, *215*, 104367. https://doi.org/10.1016/J.CHEMOLAB.2021.104367

Bouveresse, E., & Massart, D. L. (1996). Standardisation of near-infrared spectrometric instruments: A review. *Vibrational Spectroscopy*, *11*(1), 3–15. https://doi.org/10.1016/0924-2031(95)00055-0

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burdon, J., Wohlers, M., Pidakala, P., Laurie, T., Punter, M., & Billing, D. (2014). The potential for commonly measured at-harvest fruit characteristics to predict chilling susceptibility of 'Hort16A' kiwifruit. *Postharvest Biology and Technology*, *94*, 41–48. https://doi.org/10.1016/j.postharvbio.2014.03.005

Bureau, S., Ścibisz, I., Le Bourvellec, C., & Renard, C. M. (2012). Effect of sample preparation on the measurement of sugars, organic acids, and polyphenols in apple fruit by mid-infrared spectroscopy. *Journal of Agricultural and Food Chemistry*, *60*(14), 3551–3563. https://doi.org/10.1021/jf204785w

Chaudhary, R. K., Neupane, A., Wang, Z., & Walsh, K. (2025). Mango Quality Assessment Using Near-Infrared Spectroscopy and Hyperspectral Imaging: A Systematic Review. *Agronomy 2025, Vol. 15, Page 2271*, *15*(10), 2271. https://doi.org/10.3390/AGRONOMY15102271

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., Zhou, R., & Ren, P. (2024). Spectraformer: deep learning model for grain spectral qualitative analysis based on transformer structure. *RSC Advances*, *14*(12), 8053–8066. https://doi.org/10.1039/D3RA07708J

Chollet, F., et al. (2015). Keras. https://keras.io

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing, 160–167. https://doi.org/10.1145/1390156.1390177

Cui, C., & Fearn, T. (2018). Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. *Chemometrics and Intelligent Laboratory Systems*, *182*, 9–20. https://doi.org/10.1016/j.chemolab.2018.07.008

Dhaini, M., Berar, M., Honeine, P., & Van Exem, A. (2024). Contrastive Learning for Regression on Hyperspectral Data. https://arxiv.org/pdf/2403.17014v1

Diaz-Olivares, J. A., Grauwels, S., Fu, X., Adriaens, I., Saeys, W., Bendoula, R., Roger, J. M., & Aernouts, B. (2024). Temperature correction of near-infrared spectra of raw milk. *Chemometrics and Intelligent Laboratory Systems*, *255*, 105251. https://doi.org/10.1016/J.CHEMOLAB.2024.105251

Dixit, Y., Casado-Gavalda, M. P., Cama-Moncunill, R., Cama-Moncunill, X., Markiewicz-Keszycka, M., Cullen, P. J., & Sullivan, C. (2017). Developments and Challenges in Online NIR Spectroscopy for Meat Processing. *Comprehensive Reviews in Food Science and Food Safety*, *16*(6), 1172–1187. https://doi.org/10.1111/1541-4337.12295

Ebden, M. (2015). Gaussian Processes for Regression and Classification: A Quick Introduction. (August), 11. http://arxiv.org/abs/1505.02965

Eigenvector Research. (2025). T-Squared Q residuals and Contributions.

Felix Instruments. (2019). F-750 Produce Quality Meter. https://felixinstruments.com/food-science-instruments/nir-spectroscopy/f-750-produce-quality-meter

Feng, J., McGlone, A. V., Currie, M., Clark, C. J., & Jordan, B. R. (2011). Assessment of yellow-fleshed kiwifruit (Actinidia chinensis 'Hort16A') quality in pre- and post-harvest conditions using a portable near-infrared spectrometer. *HortScience*, *46*(1), 57–63. https://doi.org/10.21273/hortsci.46.1.57

Folch-Fortuny, A., Vitale, R., de Noord, O. E., & Ferrer, A. (2017). Calibration transfer between NIR spectrometers: New proposals and a comparative

study. *Journal of Chemometrics*, *31*(3), e2874. https://doi.org/10.1002/cem.2874

Franke, G. R. (2010). Multicollinearity. In *Wiley international encyclopedia of marketing*. American Cancer Society. https://doi.org/10.1002/9781444316568.wiem02066

Goke, A., Serra, S., & Musacchi, S. (2018). Postharvest Dry Matter and Soluble Solids Content Prediction in d'Anjou and Bartlett Pear Using Near-infrared Spectroscopy. *HortScience*, *53*(5), 669–680. https://doi.org/10.21273/HORTSCI12843-17

Google Research. (2024). Google Colaboratory. https://colab.research.google.com/

Gorman, B. (n.d.). Random Forest From Top To Bottom. https://gormanalysis.com/random-forest-from-top-to-bottom/

Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., & Tao, D. (2024). A Survey on Self-supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2024.3415112

Gutiérrez, S., Fernández-Novales, J., Garde-Cerdán, T., Marín-San Román, S., Tardaguila, J., & Diago, M. P. (2023). Multi-sensor spectral fusion to model grape composition using deep learning. *Information Fusion*, *99*, 101865. https://doi.org/10.1016/J.INFFUS.2023.101865

Hapke, B. (1981). Bidirectional reflectance spectroscopy: 1. Theory. *Journal of Geophysical Research: Solid Earth*, *86*(B4), 3039–3054. https://doi.org/10.1029/JB086IB04P03039

Harker, F. R., Carr, B. T., Lenjo, M., MacRae, E. A., Wismer, W. V., Marsh, K. B., Williams, M., White, A., Lund, C. M., Walker, S. B., et al.

(2009). Consumer liking for kiwifruit flavour: A meta-analysis of five studies on fruit quality. *Food Quality and Preference*, *20*(1), 30–41.

Harker, F. R., Gunson, F. A., & Jaeger, S. R. (2003). The case for fruit quality: an interpretive review of consumer attitudes, and preferences for apples. *Postharvest Biology and Technology*, *28*(3), 333–347. https://doi.org/10.1016/S0925-5214(02)00215-6

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, H., Wang, X., Zhang, Y., Chen, Q., & Guan, Q. (2024). A comprehensive survey on contrastive learning. *Neurocomputing*, *610*, 128645. https://doi.org/10.1016/J.NEUCOM.2024.128645

Huang, X., Shi, L., & Suykens, J. A. (2014). Asymmetric least squares support vector machine classifiers. *Computational Statistics and Data Analysis*, *70*(3), 395–405. https://doi.org/10.1016/j.csda.2013.09.015

Jaeger, S. R., Harker, R., Triggs, C. M., Gunson, A., Campbell, R. L., Jackman, R., & Requejo-Jackman, C. (2011). Determining Consumer Purchase Intentions: The Importance of Dry Matter, Size, and Price of Kiwifruit. *Journal of Food Science*, *76*(3), S177–S184. https://doi.org/10.1111/J.1750-3841.2011.02084.X

Jaiswal, P., Jha, S. N., & Bharadwaj, R. (2012). Non-destructive prediction of quality of intact banana using spectroscopy. *Scientia Horticulturae*, *135*, 14–22. https://doi.org/10.1016/j.scienta.2011.11.021

Kinhal, V. (2024). What Are Kiwifruit Harvest Maturity Indices & Are They Important? https://felixinstruments.com/blog/what-are-kiwifruit-harvest-maturity-indices-and-why-are-they-important/

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-Normalizing Neural Networks. *Advances in Neural Information Processing Systems*, *2017-December*, 972–981. https://arxiv.org/pdf/1706.02515

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*(1), 98–113. https://doi.org/10.1109/72.554195

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 255–258. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.32.9297&rep=rep1&type=pdf

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150–156. https://doi.org/10.1016/J.JESP.2017.09.011

Liu, H., Ong, Y. S., Shen, X., & Cai, J. (2020). When Gaussian Process Meets Big Data: A Review of Scalable GPs. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(11), 4405–4423. https://doi.org/10.1109/TNNLS.2019.2957109

Liu, L., Ji, M., & Buchroithner, M. (2017). Combining partial least squares and the gradient-boosting method for soil property retrieval using Visible Near-Infrared Shortwave Infrared spectra. *Remote Sensing*, *9*(12). https://doi.org/10.3390/rs9121299

Liu, Y., Sun, X., & Ouyang, A. (2010). Nondestructive measurement of soluble solid content of navel orange fruit by visible-NIR spectrometric technique with PLSR and PCA-BPNN. *Lwt*, *43*(4), 602–607. https://doi.org/10.1016/j.lwt.2009.10.008

Madden, H. H. (1978). Comments on the Savitzky-Golay Convolution Method for Least-Squares Fit Smoothing and Differentiation of Digital Data. *Analytical Chemistry*, *50*(9), 1383–1386. https://doi.org/10.1021/ac50031a048

Makalesi, D., Çataltaş, Ö., Tütüncü, K., Kızılötesi, Y., Kullanılan, S., Analizi, V., Bir, T., & Öz, D. (2021). A Review of Data Analysis Techniques Used in Near-Infrared Spectroscopy. *European Journal of Science and Technology*, (25), 475–484. https://doi.org/10.31590/ejosat.882749

Maldonado-Celis, M. E., Yahia, E. M., Bedoya, R., Landázuri, P., Loango, N., Aguillón, J., Restrepo, B., & Guerrero Ospina, J. C. (2019). Chemical Composition of Mango (Mangifera indica L.) Fruit: Nutritional and Phytochemical Compounds. *Frontiers in Plant Science*, *10*, 450160. https://doi.org/10.3389/FPLS.2019.01073/XML

Marsh, K. B., Boldingh, H. L., Shilton, R. S., & Laing, W. A. (2009). Changes in quinic acid metabolism during fruit development in three kiwifruit species. *Functional plant biology : FPB*, *36*(5), 463–470. https://doi.org/10.1071/FP08240

Martens, H., Nielsen, J. P., & Engelsen, S. B. (2003). Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Analytical Chemistry*, *75*(3), 394–404. https://doi.org/10.1021/ac020194w

Martens, H., & Stark, E. (1991). Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, *9*(8), 625–635. https://doi.org/10.1016/0731-7085(91)80188-F

Martins, J. A., Guerra, R., Pires, R., Antunes, M. D., Panagopoulos, T., Brázio, A., Afonso, A. M., Silva, L., Lucas, M. R., & Cavaco, A. M. (2022). SpectraNet–53: A deep residual learning architecture for predicting soluble solids content with VIS–NIR spectroscopy. *Computers*

and Electronics in Agriculture, *197*, 106945. https://doi.org/10.1016/J.COMPAG.2022.106945

Martins, J. A., Rodrigues, D., Cavaco, A. M., Antunes, M. D., & Guerra, R. (2023). Estimation of soluble solids content and fruit temperature in 'Rocha' pear using Vis-NIR spectroscopy and the SpectraNet–32 deep learning architecture. *Postharvest Biology and Technology*, *199*, 112281. https://doi.org/10.1016/J.POSTHARVBIO.2023.112281

Martinsen, P., & Schaare, P. (1998). Measuring soluble solids distribution in kiwifruit using near-infrared imaging spectroscopy. *Postharvest Biology and Technology*, *14*(3), 271–281. https://doi.org/10.1016/S0925-5214(98)00051-9

McGlone, V. A., Clark, C. J., & Jordan, R. B. (2007). Comparing density and VNIR methods for predicting quality parameters of yellow-fleshed kiwifruit (Actinidia chinensis). *Postharvest Biology and Technology*, *46*(1), 1–9. https://doi.org/10.1016/j.postharvbio.2007.04.003

McGlone, V. A., Jordan, R. B., Seelye, R., & Martinsen, P. J. (2002). Comparing density and NIR methods for measurement of Kiwifruit dry matter and soluble solids content. *Postharvest Biology and Technology*, *26*(2), 191–198. https://doi.org/10.1016/S0925-5214(02)00014-5

McGlone, V. A., & Kawano, S. (1998). Firmness, dry-matter and soluble-solids assessment of postharvest kiwifruit by NIR spectroscopy. *Postharvest Biology and Technology*, *13*(2), 131–141. https://doi.org/10.1016/S0925-5214(98)00007-6

McIntire, M., Ratner, D., & Ermon, S. (2016). Sparse Gaussian processes for Bayesian optimization. *32nd Conference on Uncertainty in Artificial Intelligence 2016, UAI 2016*, 517–526.

Mishra, P., Nikzad-Langerodi, R., Marini, F., Roger, J. M., Biancolillo, A., Rutledge, D. N., & Lohumi, S. (2021). Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers? The answer is not always. *TrAC Trends in*

Analytical Chemistry, *143*, 116331. https://doi.org/10.1016/J.TRAC.2021.116331

Mishra, P., & Passos, D. (2021a). A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit. *Chemometrics and Intelligent Laboratory Systems*, *212*(February). https://doi.org/10.1016/j.chemolab.2021.104287

Mishra, P., & Passos, D. (2021b). Deep calibration transfer: Transferring deep learning models between infrared spectroscopy instruments. *Infrared Physics & Technology*, *117*, 103863. https://doi.org/10.1016/J.INFRARED.2021.103863

Mishra, P., Passos, D., Marini, F., Xu, J., Amigo, J. M., Gowen, A. A., Jansen, J. J., Biancolillo, A., Roger, J. M., Rutledge, D. N., & Nordon, A. (2022). Deep learning for near-infrared spectral data modelling: Hypes and benefits. *TrAC Trends in Analytical Chemistry*, *157*, 116804. https://doi.org/10.1016/J.TRAC.2022.116804

Mishra, P., & Woltering, E. (2023). Semi-supervised robust models for predicting dry matter in mango fruit with near-infrared spectroscopy. *Postharvest Biology and Technology*, *200*, 112335. https://doi.org/10.1016/J.POSTHARVBIO.2023.112335

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, *16*, 100258. https://doi.org/10.1016/J.ARRAY.2022.100258

Nawar, S., & Mouazen, A. M. (2017). 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 (Switzerland)*, *17*(10). https://doi.org/10.3390/s17102428

Nicolaï, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and veg-

etable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology*, *46*(2), 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024

Palmer, J. W., Harker, F. R., Tustin, D. S., & Johnston, J. (2010). Fruit dry matter concentration: a new quality metric for apples. *Journal of the Science of Food and Agriculture*, *90*(15), 2586–2594. https://doi.org/10.1002/JSFA.4125

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., & ... (2017). Automatic differentiation in pytorch. https://openreview.net/forum?id=BJJsrmfCZ

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Peiris, K. H., Dull, G. G., Leffler, R. G., & Kays, S. J. (1998). Near-infrared spectrometric method for nondestructive determination of soluble solids content of peaches. *Journal of the American Society for Horticultural Science*, *123*(5), 898–905. https://doi.org/10.21273/jashs.123.5.898

Peiris, K. H., Dull, G. G., Leffler, R. G., & Kays, S. J. (1999). Spatial variability of soluble solids or dry-matter content within individual fruits, bulbs, or tubers: Implications for the development and use of NIR spectrometric techniques. *HortScience*, *34*(1), 114–118. https://doi.org/10.21273/hortsci.34.1.114

Pelliccia, D. (2018). Outliers Detection with {PLS} Regression for {NIR} Spectroscopy in {Python}. https://nirpyresearch.com/outliers-detection-pls-regression-nir-spectroscopy-python/

Peng, C., Zhang, S., Nie, L., & Zang, H. (2025). AI-enhanced multimodal spectroscopy for monitoring volatile organic compounds in pharmaceutical

wastewater. *Water Research*, *287*, 124476. https://doi.org/10.1016/j.watres.2025.124476

Puneet Mishra & Passos, D. (2021). Deep multiblock predictive modelling using parallel input convolutional neural networks. *Analytica Chimica Acta*, *1163*, 338520. https://doi.org/10.1016/J.ACA.2021.338520

Qi, P., Zhou, W., & Han, J. (2017). A method for stochastic L-BFGS optimization. *2017 2nd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2017*, 156–160. https://doi.org/10.1109/ICCCBDA.2017.7951902

Ramadan, A., Robert, G., Kersaudy, R., Rouabah, M., Abatzoglou, N., & Gosselin, R. (2025). Calibration transfer and maintenance in the pharmaceutical industry: a systematic review. *European Journal of Pharmaceutical Sciences*, *209*, 107114. https://doi.org/10.1016/J.EJPS.2025.107114

Reich, G. (2005). Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. *Advanced Drug Delivery Reviews*, *57*(8), 1109–1143. https://doi.org/10.1016/j.addr.2005.01.020

Richardson, D. P., Ansell, J., & Drummond, L. N. (2018). The nutritional and health attributes of kiwifruit: a review. *European Journal of Nutrition 2018 57:8*, *57*(8), 2659–2676. https://doi.org/10.1007/S00394-018-1627-Z

Rinnan, A., Berg, F. v. d., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, *28*(10), 1201–1222. https://doi.org/10.1016/J.TRAC.2009.07.007

Rogozhnikov, A. (2016). Gradient Boosting explained [demonstration]. https://gormanalysis.com/gradient-boosting-explained/

Rosipal, R., & Trejo, L. J. (2000). 10.1162/15324430260185556. *CrossRef Listing of Deleted DOIs*, *1*, 27. https://doi.org/10.1162/15324430260185556

Said, M., Wahba, A., & Khalil, D. (2022). Semi-supervised deep learning framework for milk analysis using NIR spectrometers. *Chemometrics and Intelligent Laboratory Systems*, *228*, 104619. https://doi.org/10.1016/J.CHEMOLAB.2022.104619

Scalisi, A., & O'Connell, M. G. (2021). Relationships between Soluble Solids and Dry Matter in the Flesh of Stone Fruit at Harvest. *Analytica 2021, Vol. 2, Pages 14-24*, *2*(1), 14–24. https://doi.org/10.3390/ANALYTICA2010002

Schotsmans, W. C., Mawson, A. J., & MacKay, B. (2007). Comparison of destructive and non-destructive (NIR) dry matter determination for 'Hort16A' (ZESPRI™ GOLD) kiwifruit. *Acta Horticulturae*, *753*, 283–288. https://doi.org/10.17660/ACTAHORTIC.2007.753.35

Serra, S., Goke, A., Diako, C., Vixie, B., Ross, C., & Musacchi, S. (2019). Consumer perception of d'Anjou pear classified by dry matter at harvest using near-infrared spectroscopy. *International Journal of Food Science and Technology*, *54*(6), 2256–2265. https://doi.org/10.1111/IJFS.14140

Singh, N., Kaur, S., Mithraa, T., Verma, V. K., Kumar, A., Choudhary, V., & Bhardwaj, R. (2024). ProTformer: Transformer-based model for superior prediction of protein content in lablab bean (Lablab purpureus L.) using Near-Infrared Reflectance spectroscopy. *Food Research International*, *197*, 115161. https://doi.org/10.1016/J.FOODRES.2024.115161

Snelson, E., & Ghahramani, Z. (2005). Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 1257–1264.

Snelson, E. (2007). Flexible and efficient Gaussian process models for machine learning. *ACM SIGKDD Explorations Newsletter*, *7*(2001), 1–135. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.4041&rep=rep1&type=pdf%5Cnhttp://portal.acm.org/citation.cfm?id=1117456

Sow, A., Traore, I., Diallo, T., Traore, M., & Ba, A. (2022). Comparison of Gaussian process regression, partial least squares, random forest and

support vector machines for a near infrared calibration of paracetamol samples. *Results in Chemistry*, *4*. https://doi.org/10.1016/j.rechem.2022.100508

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

United Fresh New Zealand. (2024). *Fresh Facts 2024: New Zealand's Fresh Fruit \& Vegetable Industry* (tech. rep.). United Fresh New Zealand. https://unitedfresh.co.nz/assets/site/Fresh-Facts-2024-%5C%E2%5C%80%5C%93-Online-Version.pdf%20www.unitedfresh.co.nz

van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. https://doi.org/10.1007/S10994-019-05855-6/FIGURES/5

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need, 1. https://arxiv.org/pdf/1706.03762

Vrasti, R., Grant, B. F., Chatterji, S., Üstün, B. T., Mager, D., Olteanu, I., & Badoi, M. (1998). An Introduction to Partial Least Squares Regression. *Proceedings of the twentieth annual {SAS} users group international conference*, *4*(4), 8.

Walsh, J. (2024). *Deep Learning in Estimation of Fruit Attributes Using Near Infrared Spectroscopy* [Doctoral dissertation, CQUniversity].

Walsh, J., Neupane, A., Koirala, A., Li, M., & Anderson, N. (2023). Review: The evolution of chemometrics coupled with near infrared spectroscopy for fruit quality evaluation. II. The rise of convolutional neural networks. *Journal of Near Infrared Spectroscopy*, *31*(3), 109–125. https://doi.org/10.1177/09670335231173140

Walsh, K. B., McGlone, V. A., & Han, D. H. (2020). The uses of near infrared spectroscopy in postharvest decision support: A review. *Postharvest*

Biology and Technology, *163*, 111139. https://doi.org/10.1016/J.POSTHARVBIO.2020.111139

Wang, H., & Hu, D. (2005). Comparison of SVM and LS-SVM for regression. *Proceedings of 2005 International Conference on Neural Networks and Brain Proceedings, ICNNB'05*, *1*, 279–283. https://doi.org/10.1109/icnnb.2005.1614615

Wang, H., Peng, J., Xie, C., Bao, Y., & He, Y. (2015). Fruit quality evaluation using spectroscopy technology: A review. *Sensors (Switzerland)*, *15*(5), 11889–11927. https://doi.org/10.3390/s150511889

Wohlers, M., McGlone, A., Frank, E., & Holmes, G. (2023). Augmenting NIR Spectra in deep regression to improve calibration. *Chemometrics and Intelligent Laboratory Systems*, *240*, 104924. https://doi.org/10.1016/J.CHEMOLAB.2023.104924

Wohlers, M., McGlone, A., Frank, E., & Holmes, G. (2025). Assessing machine learning models for near-infrared regression by measuring stability towards diffeomorphisms. *Chemometrics and Intelligent Laboratory Systems*, *264*, 105449. https://doi.org/10.1016/J.CHEMOLAB.2025.105449

Wold, S., Sjöström, M., & Eriksson, L. (2001a). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Wold, S., Sjöström, M., & Eriksson, L. (2001b). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Woodward, T. (2007). *KiwiTech Bulletin No. N45: Taste, Dry Matter and Brix*. https://www.agribusiness.school.nz/pluginfile.php/2247/mod_folder/content/0/Consumer%20Preferences/Taste%2C%20Dry%20Matter%20and%20Brix.pdf?forcedownload=1

Workman, J. J. (2018). A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy. *Applied Spectroscopy*, *72*(3), 340–365. https://doi.org/10.1177/0003702817736064

Yamashita, G. H., Anzanello, M. J., Soares, F., Rocha, M. K., & Fogliatto, F. S. (2022). Selecting relevant wavelength intervals for PLS calibration based on absorbance interquartile ranges. *Chemometrics and Intelligent Laboratory Systems*, *231*, 104689. https://doi.org/10.1016/J.CHEMOLAB.2022.104689

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., & Hsieh, C. J. (2019). Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. *8th International Conference on Learning Representations, ICLR 2020*. https://arxiv.org/pdf/1904.00962

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *Proceedings of Machine Learning Research*, *139*, 12310–12320. https://arxiv.org/pdf/2103.03230

Zespri Group Limited. (2024). The Kiwifruit Harvesting Guide for Growers. https://canopy.zespri.com/content/dam/new-canopy/nz/en/documents/public/harvesting/Harvesting-Best-Practice-Guide-Growers.pdf

Zhang, L., Liu, J., Wei, Y., An, D., & Ning, X. (2025). Self-supervised learning-based multi-source spectral fusion for fruit quality evaluation: A case study in mango fruit ripeness prediction. *Information Fusion*, *117*, 102814. https://doi.org/10.1016/J.INFFUS.2024.102814

Zhang, X., Lin, T., Xu, J., Luo, X., & Ying, Y. (2019). DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis. *Analytica Chimica Acta*, *1058*, 48–57. https://doi.org/10.1016/J.ACA.2019.01.002

Appendix A

Co-authership Forms

Co-Authorship Form

School of Graduate Research
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 5096
Email: SGR@waikato.ac.nz
Website: http://www.waikato.ac.nz/students/research-degree

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work**. Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 4: Augmenting NIR Spectra in deep regression to improve calibration
Published in Chemometrics and Intelligent Laboratory Systems 240 (2023) 104924

Nature of contribution by PhD candidate	Conceived of ideas, performed experiments, wrote the paper
Extent of contribution by PhD candidate (%)	70

CO-AUTHORS

Name	Nature of Contribution
Geoff Holmes	Supervision, discussion, paper revision
Eibe Frank	Supervision, discussion, paper revision
Andrew McGlone	Supervision, discussion, paper revision

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

Name	Signature	Date
Geoff Holmes		24/10/2025
Eibe Frank		24/10/2025
Andrew McGlone		29/10/2025

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

School of Graduate Research
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 5096
Email: SGR@waikato.ac.nz
Website: http://www.waikato.ac.nz/students/research-degree

Co-Authorship Form

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work**. Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 5: Assessing machine learning models for near-infrared regression by measuring stability towards diffeomorphisms
Published in Chemometrics and Intelligent Laboratory Systems 264 (2025) 105449

Nature of contribution by PhD candidate	Conceived of ideas, performed experiments, wrote the paper
Extent of contribution by PhD candidate (%)	75

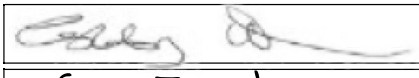
CO-AUTHORS

Name	Nature of Contribution
Geoff Holmes	Supervision, discussion, paper revision
Eibe Frank	Supervision, discussion, paper revision
Andrew McGlone	Supervision, discussion, paper revision

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

Name	Signature	Date
Geoff Holmes		24/10/2025
Eibe Frank		24/10/2025
Andrew McGlone		29/10/25

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Co-Authorship Form

School of Graduate Research
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 5096
Email: SGR@waikato.ac.nz
Website: http://www.waikato.ac.nz/students/research-degree

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work**. Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 6 Barlow Twins for Semi-Supervised Learning in NIR Spectroscopy
Submitted to Chemometrics and Intelligent Laboratory Systems 23/10/2025

Nature of contribution by PhD candidate	Conceived of ideas, performed experiments, wrote the paper
Extent of contribution by PhD candidate (%)	80

CO-AUTHORS

Name	Nature of Contribution
Geoff Holmes	Supervision, discussion, paper revision
Eibe Frank	Supervision, discussion, paper revision
Andrew McGlone	Supervision, discussion, paper revision

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

Name	Signature	Date
Geoff Holmes		24/10/2025
Eibe Frank		24/10/2025
Andrew McGlone		29/10/2025