Classification of cow diet based on milk Mid Infrared Spectra: A data analysis competition at the “International Workshop on Spectroscopy and Chemometrics 2022”

Classification of cow diet based on milk Mid Infrared Spectra: a data analysis competition at the "International Workshop on Spectroscopy and Chemometrics 2022"

Maria Frizzarin[1,2], Giulio Visentin*[3], Alessandro Ferragina[4], Elena Hayes[5], Antonio Bevilacqua[6], Bhaskar Dhariyal[6], Katarina Domijan[7], Hussain Khan[4], Georgiana Ifrim[6], Thach Le Nguyen[6], Joe Meagher[2,8], Laura Menchetti[9], Ashish Singh[6], Suzy Whoriskey[2,8], Robert Williamson[10], Martina Zappaterra[11], and Alessandro Casa[12]

[1] *Teagasc, Animal & Grassland Research and Innovation Centre, Moorepark, Ireland*
[2] *School of Mathematics and Statistics, University College Dublin, Ireland*
[3] *Department of Veterinary Medical Sciences, University of Bologna, Italy*
[4] *Teagasc Food Research Centre, Ashtown, Ireland*
[5] *Teagasc, Food Research Centre, Moorepark, Ireland*
[6] *School of Computer Science, University College Dublin, Ireland*
[7] *Department of Mathematics and Statistics, National University of Ireland, Maynooth, Ireland*
[8] *Insight Centre for Data Analytics, University College Dublin, Ireland*
[9] *School of Biosciences and Veterinary Medicine, University of Camerino, Italy*
[10] *School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, UK*
[11] *Department of Agricultural and Food Sciences, University of Bologna, Italy*
[12] *Faculty of Economics and Management, Free University of Bozen-Bolzano, Italy*

Abstract

In April 2022, the Vistamilk SFI Research Centre organized the second edition of the "International Workshop on Spectroscopy and Chemometrics – Applications in Food and Agriculture". Within this event, a data challenge was organized among participants of the workshop. Such data competition aimed at developing a prediction model to discriminate dairy cows' diet based on milk spectral information collected in the mid-infrared region. In fact, the development of an accurate and reliable discriminant model for dairy cows' diet can provide important authentication tools for dairy processors to guarantee product origin for dairy food manufacturers from grass-fed animals. Different statistical and machine learning modelling approaches have been employed during the workshop, with different pre-processing steps involved and different degree of complexity. The present paper aims to describe the statistical methods adopted by participants to develop such classification model.

Keywords: Chemometrics, Fourier transform mid-infrared spectroscopy, machine learning, milk quality, food authenticity

1 Introduction

The use of mid-infrared spectroscopy (MIRS) has become a relevant topic in agri-food sciences, due to its capacity to routinely quantify a wide range of important characteristics rapidly and cost-effective. In particular, MIRS is nowadays commonly employed to monitor and quantify

*Corresponding author: address. Email: giulio.visentin@unibo.it

milk quality parameters, such as concentrations of fat, protein, casein, and lactose. These parameters are used for milk quality-based payment schemes, genetic and genomic selection, and as farmers' support tool. Spectral information generated from MIRS analysis have also proven to be effective in predicting fine milk quality parameters, including protein fractions, free amino acids [Bonfatti et al., 2011; McDermott et al., 2016], individual and groups of fatty acids [Soyeurt et al., 2006; Fleming et al., 2017], milk processing traits [Ferragina et al., 2013; Visentin et al., 2015], animal-related characteristics [McParland et al., 2014; Shetty et al., 2017; Ho et al., 2019], and can be used as a tool for the verification of the authenticity of agricultural foods [Cozzolino, 2012]. A more extended list of applications of MIRS in the dairy science framework can be retrieved from the reviews by De Marchi et al. [2014] and Tiplady et al. [2020].

The two-day event *"International Workshop on Spectroscopy and Chemometrics"* was organized by Vistamilk SFI Research Centre in April 2022, following its first edition held in 2021 [Frizzarin et al., 2021a]. The workshop focused on describing the main challenges and applications of near and mid-infrared spectroscopy in food, animal, and agricultural sciences with internationally recognised researchers. Moreover, participants, on a voluntary basis, were provided with a large dataset containing individual cow milk spectra with the sole information on animal's diet for a chemometric data competition. Such data presented many challenges from a methodological and statistical point of view, due to the high dimensionality of the spectral matrices, and strong collinearity between adjacent spectral wavelengths. The chemometric challenge, therefore, encouraged the engagement of participants with different background and skills and required the application of different statistical and machine learning strategies.

The purpose of the data challenge was to develop a model to predict the diet fed to dairy cows by exploiting mid-infrared spectral information. Participants, or groups of participants, were required to apply their developed model to a test set containing only individual milk spectra and to submit their prediction of animals' diet. Since participation was found to be high, six contributions out of twelve were selected, according to criteria based on the accuracy of the predictions and methodological innovativeness, to present their results both at the workshop and in the present manuscript.

2 Data description and challenge

A dataset consisting of 4,364 individual milk spectra from individual cows was collected between May and August in 2015, 2016 and 2017 [O'Callaghan et al., 2016]. The samples were from Holstein Friesian cows with different parity from Irish Dairy Research Herd in Teagasc Moorepark, Fermoy, Co. Cork. Three dietary groups were evaluated with 54 cows being assigned to a dietary group each year. The three diet treatments were grass (GRS) which consisted of perennial ryegrass only, clover (CLV) which consisted of perennial ryegrass with 20% annual clover sward, and total mixed ration (TMR) where cows were fed grass silage, maize silage and concentrates while being maintained indoors for the full season. Milk samples were collected in the morning (AM) and evening (PM) milking session; subsequently AM+PM samples were pooled and analysed weekly using Pro-FOSS FT6000 (FOSS). The output spectrum contained a total of 1060 transmittance data points in the range from 925 cm^{-1} to 5,000 cm^{-1}.

The dataset was divided into training (3275 spectra) and test (1089 spectra) data; for the latter only spectral (i.e., independent variables) information was provided, while diet information, to be used as a classification (i.e., dependent) variable, was available for the training set. The training data included 1094 spectra for GRS, 1120 spectra from CLV and 1061 spectra for TMR. There were no missing values in the training or test set. The specific information about the wavenumbers had not been shared with the participants.

The three dietary groups were carefully selected based on their characteristics. As described by Frizzarin et al. [2021b], pasture-based diets are easily discriminated from TMR diets, while

discriminating between GRS and CLV diets is much more difficult due to the similarities in the sward composition resulting in similar milk composition. However, with the increased pressure to reduce fertilizer use, and the introduction of multi-species swards, the development of a robust discriminant model for classifying milk spectra based on diet is of paramount importance.

After the analysis, the participants submitted their predicted values for the test dataset and a short explanation of the methodology used. The best methods were selected based on the accuracy of the predictions for the test dataset. The accuracy was calculated as the proportion of the correctly classified samples divided by the total number of samples in the test dataset.

3 Modelling approaches and results

3.1 Participant 1

The data were analyzed following different modelling strategies, focusing both on methods that considered spectral proximity of the wavelengths and on methods that do not. All the analyses have been mainly conducted using `Python` libraries `pandas, sklearn, sktime` and `matplotbib` [see Pedregosa et al., 2011, and references therein]. The open source code is available at `https://github.com/mlgig/vistamilk_diet_challenge` and readers can refer to it for the specific details about the implementation of all the methods outlined in this Section.

As a first step, some descriptive statistics were computed, and the outliers have been removed, following both the recommendations given prior to the competition and a visual inspection of the data. In the subsequent step, the labeled dataset was split according to a 3-fold cross-validation (3CV) strategy. Therefore, the best model was selected based on cross-validation accuracy, and then trained on the full training set and used to perform prediction on the provided unlabeled test set.

In order to predict the diet, the following classification strategies were considered:

- **Tabular models:** each sample is considered as a vector of unordered features. In particular, Ridge Classifier, where a penalty shrinking parameters towards zero is imposed on the coefficients of a logistic regression model [see e.g., Hoerl and Kennard, 1970], and Linear Discriminant Analysis (LDA) were tested. In the following, these methods were coupled both with feature selection strategies and with random polynomial feature transformations. The latter approach first used `sklearn` routines to create new features (see, `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html`). For example, for features a, b, c, a polynomial transformation of degree 2 will generate the features $1, a, b, c, ab, ac, bc, a^2, b^2, c^2$. By generating these features, this approach aimed to check if non-linear interactions improved the classification. Finally, a new approach (random polynomial transformation) is presented, which aims to diversify the polynomial features (by random sampling) while keeping low computational requirements.

- **Deep Neural Network Models:** a family of approaches based on deep neural networks, both fully connected and convolutional, were tested. This strategy implicitly generates complex features interactions, as captured by the network architecture.

Note that previously obtained results [Frizzarin et al., 2021a] suggest that tabular methods work quite well with spectroscopy data. Moreover, following the suggestions in Frizzarin et al. [2021b], feature selection strategies were coupled with the information about the presence of water regions in the spectra. In addition, state-of-the-art time series classification algorithms, such as ROCKET [Dempster et al., 2020], MiniROCKET [Dempster et al., 2021], MrSQM [Nguyen and Ifrim, 2021, 2022] and FreshPrince [Middlehurst and Bagnall, 2022], were tested. Lastly, *ensemble methods* were applied, aiming to mix together time series and tabular models, to combine their predictions and strengths. Nonetheless, these approaches have been outperformed

Table 1: Accuracy results, evaluated on the 3-fold cross-validation, for the tabular methods considered, coupled with feature selection strategies.

Method	**Accuracy**
Ridge Classifier	0.760
LDA	0.747
Feature Selection + Ridge Classifier	0.777
Feature Selection + LDA	0.778
No water + Ridge Classifier	0.777
No water + LDA	0.783
Feature Selection + Polynomial Features + LDA	0.844
No water + Feature Selection + Polynomial Features + LDA	0.844

Table 2: Examples of RPolyTransformer features used. Here x_j denote the j-th wavelength.

$(x_{32} * x_{19}) + x_{103} - x_2$
$(x_{102} * (x_{78}) + x_{26})$
$(x_1 - x_{150}) + x_{64} * x_4 * x_5$

by the ones mentioned above, therefore the corresponding results are not shown in the next sections.

3.1.1 Tabular models, feature selection and transformation

In Table 1, results for the best tabular methods are presented. Both the ridge classifier, appropriately tuned, and LDA performed quite well, while being extremely fast to train. Nonetheless, the selection of some specific wavelengths seemed to improve the accuracy further. In fact, both the removal of the noisy water regions and the data-driven feature selection (performed using the `SelectFromModel` routine in Python), provides better results.

Nevertheless, all these approaches hover around 80% accuracy, therefore, in order to improve it, the data were augmented considering polynomial features of degree two (using `sklearn` method `PolynomialFeatures(degree = 2)`). This led to an increase of the accuracy to 84.4%. The LDA component visualisation for the model with Feature Selection and Polynomial Features, applied on the unlabeled test dataset, is shown in Figure 1 and a good discrimination between the three classes is clearly visible.

The improvements obtained when considering polynomial features, come at a price in terms of the computational requirements. In fact, starting from the 1060 original wavelengths, the addition of second-degree polynomial features resulted in a total number of variables which made the model estimation task unfeasible. To address this issue, in this work a new *Random Polynomial Features* (`RPolyTransformer` in the following) approach was introduced. The key idea was to implement random sampling in the non-linear feature space. This lead to relevant advantages as the total number of features can be controlled and it can consider both higher-degree (> 2) polynomial features and complex mathematical functions (e.g., cosine, exp).

This strategy firstly generated K random arithmetic expressions (see Table 2 for some examples), which are then used to compute K non-linear features. From the new and the original features, K^* variables are selected using `SelectKBest` from `sklearn`. The hyperparameters K and K^* were optimized via cross-validation in the final model (see the final row of Table 3).

In Table 3 the results obtained with this method, again combined with different classifiers and feature selection approaches and tested with the full data and the data after water region removal, are presented. At first, when combining `RpolyTransformer` with a classifier, a significant drop in the accuracy was observed, if compared with simple tabular models. Ridge was more accurate than LDA but it was still far behind the previous results. However, by carefully filtering the

Figure 1: LDA visualisation for the model Feature Selection + Polynomial Features + LDA, applied to the unlabeled test data to predict class labels.

features either automatically with `SelectFromModel` or manually by removing the water regions, the results improved noticeably. In these experiments, LDA outperforms Ridge consistently. Compared to the `PolynomialFeatures` method, the one proposed here is faster (a few seconds versus a few minutes) and just as accurate. However, the initial results without noise reduction (i.e., feature selection) suggest that this strategy is more sensitive to noise in the data.

3.1.2 Deep Learning Models

When considering deep learning models, the task of exploding the feature space and learning feature interactions is completely deferred to the network, without requiring any feature engineering steps. In turn, deep neural networks require a careful design process, to avoid overfitting and to identify the best model architecture and input modality.

The designed model architectures considered here can be grouped into two main categories, namely, Fully Connected Networks (FCNs) and Convolutional Neural Networks (CNNs). FCNs do not require any manipulation or adaptation of the input data, as each single wavelength is treated as an independent feature and fed to an input unit. In contrast, CNNs require the data to be bi-dimensional, image-like matrices, as they are commonly used to address image classification problems. For this family of networks, the input waves need then to be vertically stacked as 2D arrays and therefore, in order to fit the closest squared dimension, padded with trailing zeros. An example of how the spectroscopy samples can be presented to the CNNs is provided in Figure 2. Additionally, a third group of models is tested for this challenge, namely, CNNs based on dilated kernels (further denoted as CNN_DILATED). Whilst regular CNNs extract features through compact squared filters, or local receptive fields, the CNN_DILATED network utilizes filters that are spatially dilated by a fixed factor [Yu and Koltun, 2015]. Dilated kernels are commonly used in semantic image segmentation.

All the models in this group were trained on both the full training dataset and on the water reduced one. When the CNN models were trained, the full data were shaped into images of shape 33×33 with a padding of 29 values, while the reduced data were shaped into images of shape 23×23 with a padding of 11 values. As already mentioned, all padding values were zeros, and they were appended to the original sequences.

The full list of the implemented architectures is presented in Table S1 in Appendix A.1. The

Table 3: Results for different combinations with `RPolyTransformer`. `SelectFromModel` and `SelectKBest` are feature selection modules to remove noise from data (the former) and select the most discriminative non-linear features (the latter).

Method	**Accuracy**
Region: FULL	
RPolyTransformer + Ridge Classifier	0.717
RPolyTransformer + LDA	0.619
SelectFromModel + RPolyTransformer + SelectKBest + LDA	**0.848**
Region: [925:1585, 1720:2989]	
RPolyTransformer + Ridge Classifier	0.805
RPolyTransformer + LDA	**0.847**
SelectFromModel + RPolyTransformer + SelectKBest + LDA	0.843
Region: [925:1585, 1720:2989, 3738:3807]	
RPolyTransformer + Ridge Classifier	0.811
RPolyTransformer + LDA	0.833
SelectFromModel + RPolyTransformer + SelectKBest + LDA	0.835
Optimized model	
Region: [925:1585, 1720:2989]	
RPolyTransformer($K = 17000$) + SelectKBest($K^* = 7000$) + LDA	**0.864**

Table 4: Training results on the 3CV splits.

Model	**Data**	**Split 1**	**Split 2**	**Split 3**	**Average**
FCN	FULL	0.670	0.677	0.675	0.674
	NO WATER	**0.854**	**0.851**	**0.837**	**0.847**
CNN	FULL	0.686	0.684	0.670	0.680
	NO WATER	0.806	0.836	0.832	0.824
CNN_DILATED	FULL	0.678	0.684	0.652	0.671
	NO WATER	0.824	0.812	0.807	0.814

experiments were conducted on the previously described 3-fold cross-validation splits; note that, for each split, 20% of the training data was held back for validation purposes, to identify network hyperparameters such as number of training epochs, initial learning rate, or regularisation rates. Models were trained for a total of 50,000 epochs, with an early stopping policy used to monitor the validation loss to detect overfitting and save time during the training phase. The final model used to classify the provided unknown data was selected as the overall best performing architecture, and trained over the full training data for a number of epochs set as the average of the epochs reached during the 3CV training.

All models were implemented using `TensorFlow` [Abadi et al., 2016], and trained on a workstation featuring a single GPU, model Nvidia Titan XP. Results are presented in Table 4, which contains the training performances obtained over the 3-folds CV experimental campaign. For all the tested architectures, excluding the water regions from the input waves resulted in a performance increase of roughly 12-13%. The FCN model working on data after water-region removal, achieved the highest accuracy across the 3 splits, with an average of 84.7%. Similar unreported results were obtained also considering a single split validation strategy, which furthermore demonstrated that convolutional models tend to overfit the input data quite fast.

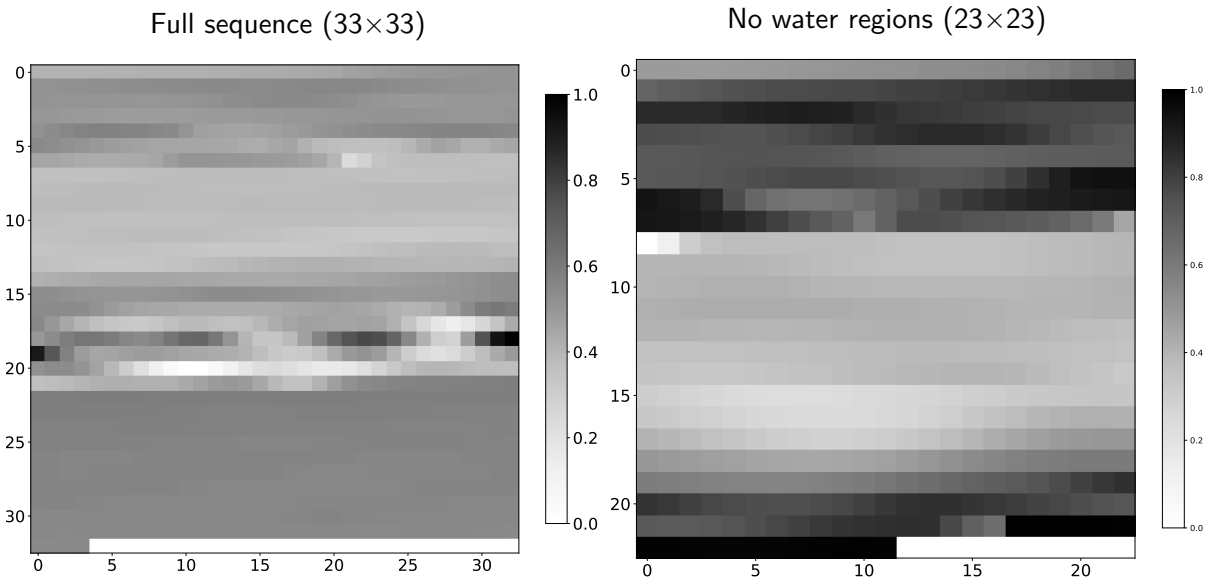

Figure 2: Spectroscopy sequences arranged as image structures. In both examples, the padding values are visible at the bottom of the resulting images. Values are normalised in the 0-1 range for convenience.

Table 5: Confusion matrix obtained by combining LDA and SVM.

		Predicted class		
		CLV	GRS	TMR
True class	CLV	83.5%	17.4%	0.7%
	GRS	15.8%	81.6%	0.8%
	TMR	0.7%	1.1%	98.5%

3.2 Participant 2

All the processing steps and the algorithm implementation was completed using `MATLAB` [MATLAB, 2018]. After having imported the dataset in tabular form, the outliers were identified as those observations with at least one wavelength with more than three scaled median absolute deviation from the wavelength specific median (see `https://uk.mathworks.com/help/matlab/ref/isoutlier.html` for further details). Classification was performed using a set of algorithms such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA). To optimize the number of predicting variables, coefficient's threshold and the regularization parameter was tuned using a 5-fold cross-validation and classification accuracy was evaluated.

The best results were obtained using LDA, which was able to distinguish outdoor grass-feed cow's milk from TMR with an accuracy of 95% while differentiating grass and clover with an accuracy of 68%. Figure 3 allows to visualize class boundaries by plotting the spectra projections in the latent space spanned by the two discriminant functions. From the figure, a clear boundary can be observed between the indoor and outdoor feed classes, while there is a significant overlap between the GRS and CLV classes. Therefore, the extracted components were then considered as an input to a linear SVM model to improve classification between outdoor feed classes. The combination of two classifier (LDA + SVM), resulting in a two-step approach, significantly improved the overall classification accuracy (87.1%) as well as classification accuracy between classes, as shown in Table 5.

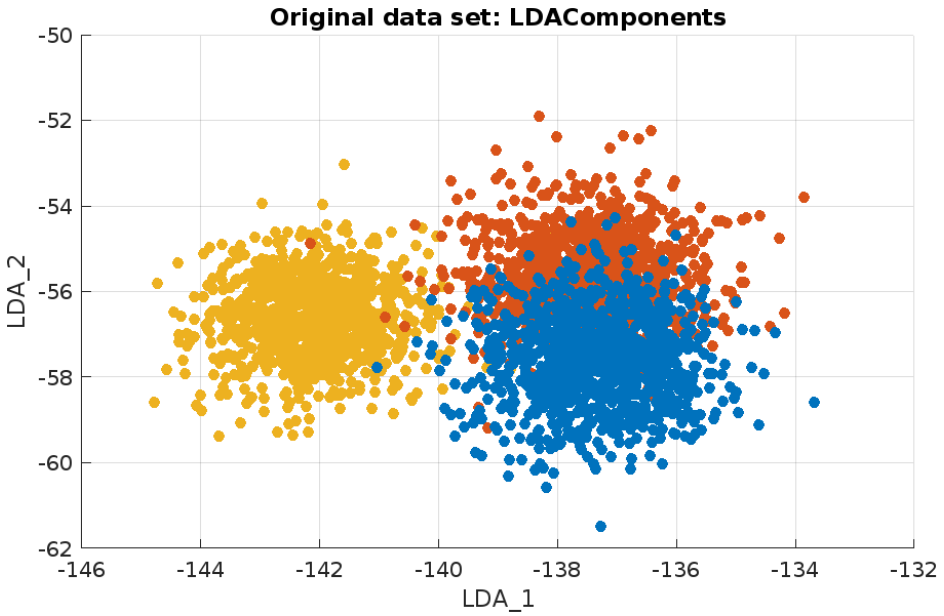

Figure 3: LDA components extracted from the developed model.

3.3 Participant 3

The present work was developed independently by three group members, following a common preliminary analysis of spectral data. Results of the prediction on the test set provided for the chemometric challenge were then compared to assess the agreement between the three different statistical approaches employed.

3.3.1 Preliminary edits on spectral data

These edits were conducted on raw spectral data in both the training and test sets using `Python`. Spectra expressed in transmittance were converted into absorbance by taking the $\log_{10}$ of the reciprocal of the transmittance. Subsequently, spectral wavelengths associated to water absorption, as well as non-informative regions, were deleted. This led to a reduced version of the dataset, that has been used for the subsequent analyses, with 511 remaining wavelengths in the regions between 2,994 and 1,682 cm^{-1} and between 1,578 and 926 cm^{-1}. A graphical representation of this procedure is reported in the supplementary material (Figure S1).

3.3.2 First approach

To explore the multivariate structure of the dataset, Principal Component Analysis (PCA) was exploited on the training dataset, using `prcomp` function in `stats` package and the `factoextra` package [Kassambara and Mundt, 2020] in the `R` environment R Core Team [2020]. The analysis revealed that most of the data variability was explained by the first two Principal Components (PCs), accounting together for the 88% of the total variance (see the scree plot on the left top panel in Figure 4).

Afterwards, possible outliers were detected using the algorithm proposed by Filzmoser et al. [2008] and implemented in the `mvoutlier` package [Filzmoser and Gschwandtner, 2021]; only the observations being both location and scatter outliers were removed from the training dataset. As a results, a total of 63 observations were removed from the training dataset.

After outliers removal, linear discriminant analysis was considered using `lda` function in the `MASS` package [Venables and Ripley, 2002]. To test its accuracy, as a first step the discriminant functions were applied to the training dataset, with the aim of comparing the estimated classification with the actual one. Therefore, LDA was first applied to maximize the differences

Table 6: Summary of the results of the three different approaches.

	Approach 1	Approach 2	Approach 3
Brief description	Two steps DA in `R`	Canonical DA with stepwise method in `SAS`	DA with stepwise methods in `SPSS`
Number of samples (training set)	3180	3116	3153
Number of wavelengths retained	511	88	16
Accuracy (training set)	83.30%	81.32%	71%
Predicted diet for the samples in the test dataset (n cases)			
TMR	344	326	365
CLV	367	342	326
GRS	366	353	386
Agreement between the approaches applied to the test dataset			
Member 1			
Member 2	84.21%		
Member 3	72.90%	70.84%	

between TMR and the CLV+GRS (in the following named PAST group). The LDA returned one Linear Discriminant (LD) function, which was then applied to the training dataset to attribute the TMR diet to observations. Afterwards, LDA was applied again by maintaining in the training set only the observations belonging to the PAST group. The obtained LD function was then applied to the whole training dataset to discriminate between CLV and GRS diets previously categorized as PAST. The vector with the predicted classes was then compared with the vector of actual group classification in the training dataset, thus computing the training accuracy. This appproach resulted in an overall model training accuracy equal to 83.3% (see Table 6); the scatter plot of the first versus second linear dimension scores is depicted in the right top panel in Figure 4. Lastly, the LD functions obtained on the training dataset allowed for the classification of the unknown observations in the test dataset, with the results reported in Table 6.

3.3.3 Second approach

Principal component analysis (`PROC PRINCOMP`, SAS Institute Inc., ver. 9.4) was undertaken on the training set, as in Section 3.3.2. Coherently, outlier removal was then performed by calculating the Mahalanobis distance (MD) as the uncorrected sum of squares of the first four centred and scaled PC scores, explaining up to the 98.21% of the total spectral variance. Outliers were defined as samples whose MD was greater than the 97.5th percentile of a χ^2 distribution with 4 degrees of freedom [Brereton, 2015]. Following this approach, a total of 127 samples were discarded from the training set.

The discriminant model was developed following a multiple-step approach. Firstly, a stepwise discriminant analysis was carried out in order to identify the most significant wavelengths associated with the three different diets using the `PROC STEPDISC`. A total of 88 wavelengths were retained and used for the subsequent canonical discriminant analysis, which was developed through the `PROC DISCRIM`. The proportion of samples correctly classified was 73.38% (CLV), 73.70% (GRS), and 97.62% (TMR), with an overall model accuracy of 81.32%. The scatter plot of the first versus second canonical variables scores is in the bottom left panel of Figure 4. The wavenumbers with the greatest (in absolute value) canonical discriminant function coefficients were between 1,154 and 1,162 cm^{-1}, 2,843 cm^{-1}, 2,874 cm^{-1}, and 2,882 cm^{-1}, thus providing some potentially relevant information to be explored to assess which milk chemical features are

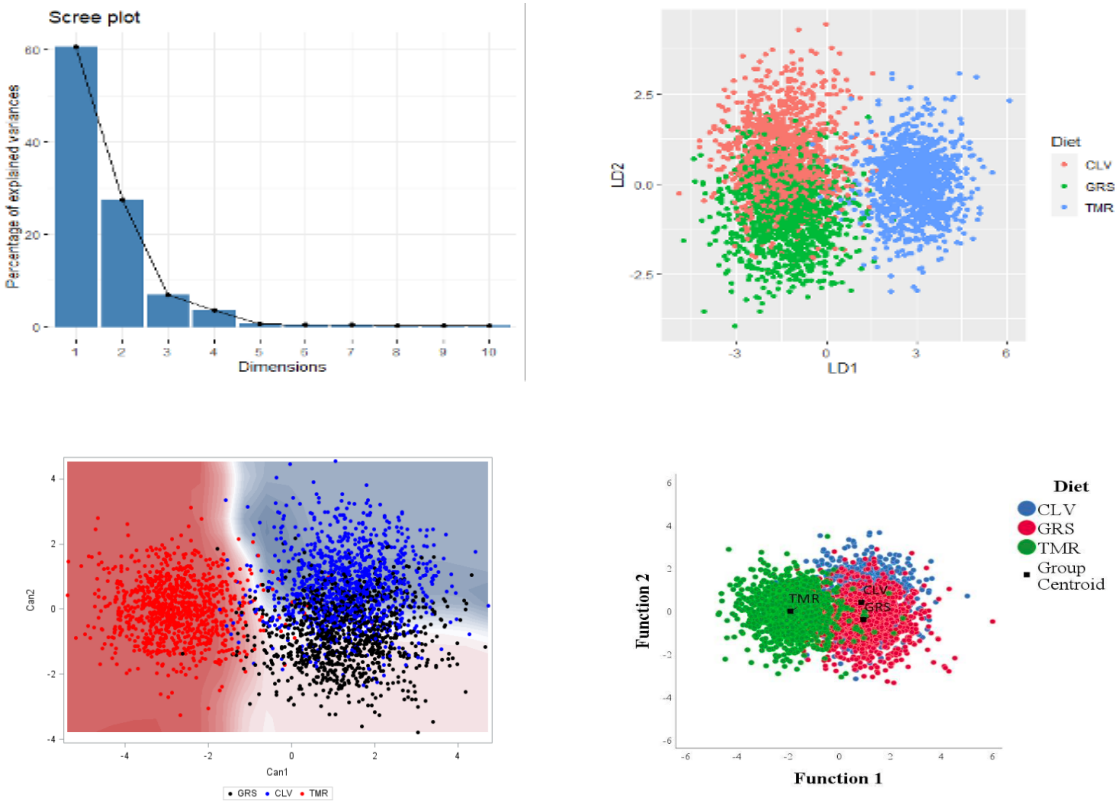

Figure 4: Explained variance by the first 10 principal components (top left), scatter plot of discriminant models developed by member 1 (right top), member 2 (bottom left) and member 3 (bottom right).

more influenced by the dietary regimen. The discriminant model was then applied to the test set to obtain the prediction of cows' diet on unknown milk spectra.

3.3.4 Third approach

Standard assumptions required for multivariate analyses were verified before proceeding to the main analysis. Two diagnostic measures were used to identify the outliers for the predictors and the dependent variables; in the former case Mahalanobis Distance (MD) was used to spot multivariate outliers while, in the latter one, studentized residuals were considered. Samples whose MD was greather than the 97.5th percentile of the MD distribution and studentized residuals greater than 2.5 were removed. During this process, a total of 90 outliers have been identified and excluded. Potential multicollinearity was then verified by Tolerance and Variance Inflation Factors. Moreover, the ratio between the number of cases and predictors was checked as an indicator of the adequacy of the sample size; a ratio of 20 observations for each predictor variable, with the smallest group size exceeding the number of independent variables, is suggested [Meloun and Militkỳ, 2011; Pituch and Stevens, 2015].

LDA was then chosen as the main discriminative approach. The stepwise method, using Wilks' lambda Λ as criterion, was adopted to reduce multicollinearity and increase the case/predictors ratio, improving the adequacy of the sample size. Box's test and log determinants were considered to verify the equality of covariance matrices. The canonical correlation and the proportion of between-group variance that is due to each variate were used as measures of effect size [Pituch and Stevens, 2015], while the performance of the LDA was evaluated by classification-related statistics and leave-one-out CV [Hahs-Vaughn, 2016]. The `Scoring Wizard` command was finally used to apply the discriminant functions (DF) to the test dataset,

and the predicted probability was calculated to assess its performance. Analyses were performed with SPSS software [IBM Corp., 2017].

Standardized canonical DF coefficients of the variables selected by DA and measures of effect size are shown in Table S2 in the Supplementary Material. More than 90% of the total difference between the groups was attributable to the first DF, with the Wilks' Λ (0.330) indicating that it has a significant discriminating capacity (p-value < 0.001). Wavenumber 2,851 cm^{-1} and 2,890 cm^{-1} mostly contributed to the discrimination of cows' diet. The second DF only explained 6% of the total variance, being nonetheless still significant (Wilks' $\Lambda = 0.902$; p-value < 0.001). Centroids (Table S3) and the plot of DF scores (bottom right panel in Figure 4) indicated that the first DF appropriately discriminate the TMR group from the others (i.e., CLV and GRS). On the other hand, group separation on the second DF was poor; in particular, CLV and GRS clusters were not clearly distinguished. The cross-validation procedure indicated an overall model accuracy of 71% (see Table 6), with different sensitivity between groups: over 90% for TMR samples, and below 65% for CLV and GRS samples. The application of DFs to predict the diet of cows in the test data set showed a similar trend, with an expected sensitivity of 64%, 63%, and 87% for CLV, GRS, and TMR diets, respectively (Table S4).

Lastly note that, all the three approaches were applied and the results were compared at the end of the competition. Despite the better prediction performance shown by the first approach on the training set, the second approach proved to be the best for the prediction of the test set.

3.4 Participant 4

A conventional machine learning pipeline was used, composed of feature (i.e., wavelength) selection and classification, with no outliers being removed from the original dataset. Dimensionality reduction techniques such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA), as well as Extended Multiplicative Scatter Correction (EMSC) and a data augmentation approach were tested to improve the classification results [Bjerrum et al., 2017]. EMSC represents a preprocessing technique which removes multiplicative effects potentially caused by physical phenomena such as light scattering, which is commonly seen in reflectance spectroscopy, thus allowing for easier modelization of chemical effects. On the other hand, the data augmentation scheme increases the data set ten fold by adding random variations in offset, multiplication, and slope, nine times to each sample. The variations were $\pm$ 0.1 times the standard deviation of the training set for the offset, multiplication was 1 ± 0.1 times the standard deviations, and the slope adjustement was between 0.95 and 1.05 [Bjerrum et al., 2017].

Subsequently a range of different classifiers, which have successfully been adopted before on infrared spectroscopy data, were used. In particular, the considered models were K-nearest Neighbour [K-NN; Balabin and Safieva, 2011], Random Forest [RF; Chen et al., 2021], Support Vector Classification [SVC; Ji-yong et al., 2013], Multilayer-perceptron [MLP; Balabin and Safieva, 2008], Linear Discriminant Analysis [LDA; Khuwijitjaru et al., 2020], Decision Tree Classification [Geronimo et al., 2019], Nu-Support Vector Classification [NuSVC; Terouzi et al., 2013], AdaBoost Classification [Wu et al., 2017], Gradient Boosting Classification [Munera et al., 2021], Gaussian Naive Bayes [Bhati and Bhattacharya, 2020] and Quadratic Discriminant Analysis [QDA; Oravec et al., 2019]. Other investigated predictive methods belonged to the group of deep Learning (DL) techniques, and in particular one-dimensional (1D) Convolutional Neural Network (CNN). This 1D CNN makes use of six one dimensional convolutional layers, and a number of max pooling, batch normalization and dropout layers. Each 1D CNN layer is followed by a max-pooling and batch normalization layer. The one-dimensional CNN only used the raw spectra, as the use of PCA and FastICA would be detrimental due to the transformation of the sequence of the data.

Prior to the analyses, the dataset was split into a training set (80% of the data), to train different models, and a validation set (remaining 20% of the data), used to optimise the hyperparameters and to identify the best methods to be used for the final testing. This split was

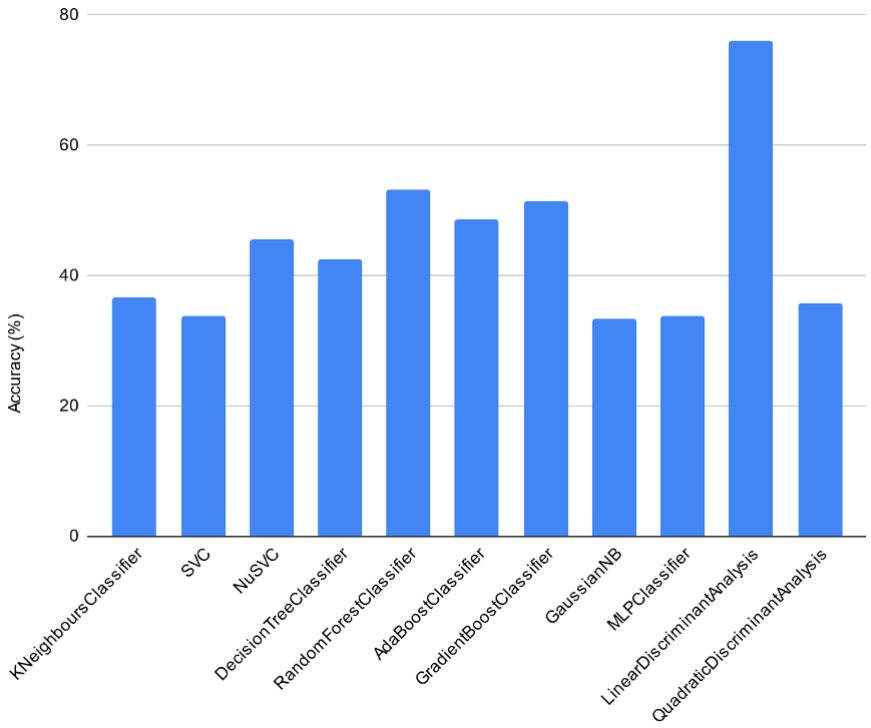

Figure 5: Results of classifiers on a 80/20 test-train split.

made by utilizing the `train_test_split` function provided through `scikit-learn` [Pedregosa et al., 2011].

An initial experiment was performed on all classifiers without the use of data augmentation or feature selection. This was carried out to explore which classification method was performing better with the raw spectral data. Figure 5 shows the results obtained from the initial step with the 80/20 train/validation for different classifiers. All results gathered were averages taken from three training and validation predictions for each model. LDA gave the best results with an accuracy of 76%, whereas the MLP and SVC produce some of the worst performances with accuracies around 33%.

In the second stage, the classifiers were tested in conjunction with PCA, ICA or data augmentation: for PCA and FastICA `scikit-learn` methods were used, with the parameters being setted as `FastICA(tol = 0.02, max_iter = 4000)` and `PCA(n_components = 800)`. The use of PCA and ICA altered the data by reducing the dimensionality, while on the other hand data augmentation increased the number of samples. For data augmentation, the data augment function from Bjerrum et al. [2017] was used. This increased the number of training samples from 3,244 to 19,464. At this stage, only a subset of the previously tested model were considered, based on their performances in the previous step. Figure 6 shows the results of each classifier with each pre-processing method (base, ICA, PCA, data augmentation (Aug)). From these results, it was noted that LDA following data augmentation achieved the highest accuracy with 82.7%. The greatest improvement in the predictions was observed using MLP after ICA (improvement of 41%). An additional experiment was then carried out with just the use of the LDA model. This was to show the importance of regions within the spectra, and a number of different wavelength region were tested. Therefore, figure 7 shows the results of the LDA when removing different spectral regions.

There was a general increase in accuracy over the base approach when data augmentation was used, with the only exception of CNN. With regard to wavelengths selection, there was no

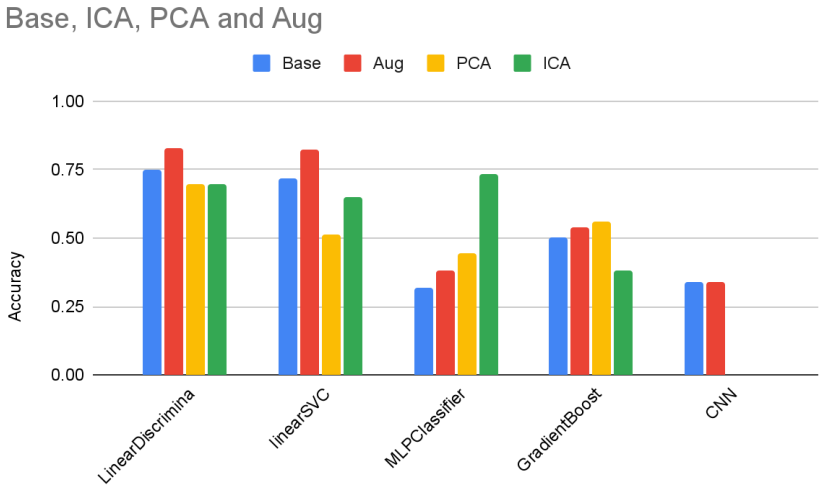

Figure 6: Results of classifiers on with different pre-processing methods.

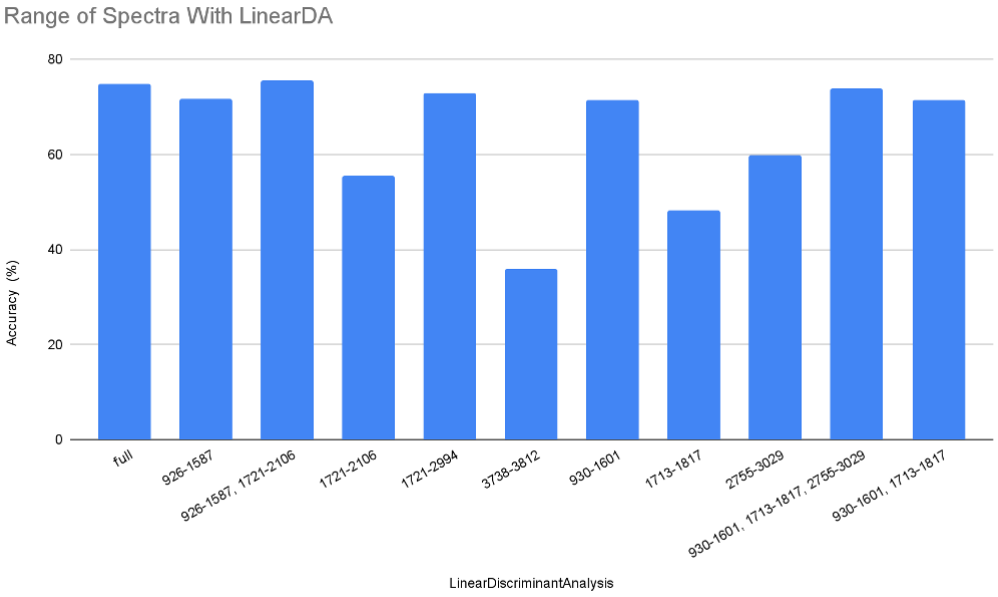

Figure 7: Results of Linear Discriminant Analysis for different feature selection.

noticeable increase in accuracy when focusing on a specific region in the spectra. Nonetheless, the majority of the relevant information lied within the region from 925 cm^{-1} and 1597 cm^{-1}, and there was a slight increase in the accuracy of prediction of around 1% when using the range of 925 to 1585 cm^{-1} and 1717 to 2103 cm^{-1} compared to the full set of wavelengths.

3.5 Participant 5

In order to prepare the data set for predictive analysis, some pre-processing was considered. As directed by the challenge organisers, outlying spectra were removed such that the data set consisted of 3243 transmittance spectra covering 1060 wavelengths. Spectra were transformed to absorbance values by taking $\log_{10}$ of the reciprocal of the transmittance values. In addition, following Frizzarin et al. [2021b], a subset of 534 wavelengths that lay outside the water-related

high-noise-level regions were identified as relevant for predicting a cow's diet, although the water-regions were not excluded at this point in the analysis.

To ensure a robust assessment, the dataset was split into training and validation sets. In this case, the validation set was constructed to control for batch effect confounding, which may bias estimates for out-of-sample prediction [Soneson et al., 2014]. Inspection of the data set revealed that rows were ordered to have several consecutive observations of each diet. Therefore, it was assumed that each set of consecutive diet observations belonged to a single batch. In this manner, 90 batches, 30 for each diet, were identified. In addition, the data was collected over three years [Frizzarin et al., 2021b], and so it was assumed that the first 30 batches were collected in the first year of the study, the next 30 in the second year, and the final 30 in the third. Based on these assumptions, the validation set consisted of 996 spectra from 30 batches collected in the study's third year, which included ten batches for each diet, while models have been trained on the 2247 remaining spectra. Training data was randomly split into $V = 10$ folds, with each fold including two batches from each diet. Possible batch effect of repeated measurements for a single cow were ignored.

In order to describe the predictive model used in this analysis, let $\mathcal{D} = \{y_i, \mathbf{x}_i\}_{i=1}^{N}$ denote the observed data, where the response variable $y_i \in \{1, \ldots, M\}$ represents the diet of the i-th cow and covariates $\mathbf{x}_i \in \mathbb{R}^D$ represent the corresponding milk absorbance spectrum. Note that this analysis considers $M = 3$ diets, $D = 1060$ wavelengths, and $N = 3243$ training observations. The objective of the proposed predictive models is to learn $\mathbb{P}(y \mid \mathbf{x})$, that is, the probability that a given milk sample comes from a grass, clover or TMR-fed cow, given the spectrum for that sample.

The first step in constructing a predictive model is to define a deterministic mapping function $g : \mathbf{x}_i \to \mathbf{z}_i$, for $\mathbf{z}_i \in \mathbb{R}^{D'}$, with $D' < D$, which describes a feature extraction procedure. Two approaches to feature extraction were considered here. The first simply selected the $D' = 534$ relevant wavelengths identified by Frizzarin et al. [2021b] such that $\mathbf{z}_i$ is the i-th absorbance spectrum after removing the high-noise-level water regions and standardises each wavelength. The second was based on the wavelet transform, a popular technique for signal processing which can be applied for data compression, smoothing, and multi-resolution analysis [Nason, 2008], and proceeds in three steps. After setting high-noise-level regions of each spectrum to 0, a thresholded wavelet transform provides a set of wavelet coefficients. The feature vector $\mathbf{z}_i$ is then the vector of wavelet coefficients that are non-zero for at least one of the N spectra, in this case $D' = 594$. The thresholded wavelet transform is available with the `wavethresh` R package [Nason, 2016], using Daubechies least symmetric wavelet as the mother wavelet and Bayesian approach to thresholding wavelet coefficients [Abramovich et al., 1998]. Note that setting wavelengths in the high-noise-level regions to 0 means the wavelet transform preserves the spectral distance between wavelengths while ensuring that the corresponding wavelet coefficients are 0.

Given the feature vector $\mathbf{z}_i = g(\mathbf{x}_i)$, a multinomial regression model for diet was assumed, such that

$$\mathbb{P}(y_i = m \mid \mathbf{z}_i) = \frac{\exp\left(\boldsymbol{\beta}_m^\top \mathbf{z}_i\right)}{\sum_{l=1}^{M} \exp\left(\boldsymbol{\beta}_l^\top \mathbf{z}_i\right)}, \tag{1}$$

for $m = 1, \ldots M$ where $\boldsymbol{\beta}_m \in \mathbb{R}^{D'}$, implicitly assuming that $\mathbf{z}_i$ includes an intercept term. The `glmnet` package [Friedman et al., 2010] fits this model to data efficiently. For simplicity, a LASSO model was fitted, where 10-fold cross-validation on the training data informs the penalty hyperparameter.

Finally, the predictive performance of the proposed models was compared by analysing their log-loss on the validation data set. That is, for a validation data set and a model $\mathcal{M}_j$ for $\mathbf{z}_i = g(\mathbf{x}_i)$, the log-loss is defined as

$$\ell_j = -\frac{1}{N'} \sum_{i=1}^{N'} \sum_{m=1}^{M} \mathbb{I}(y_i = m) \ln \mathbb{P}(y_i = m \mid \mathbf{z}_i, \mathcal{M}_j), \tag{2}$$

Table 7: Predictive model assessment.

Model	In-sample log-loss	Validation log-loss
Raw Spectra	0.57	0.82
Wavelet Coefficents	0.74	0.88

where N' is the number of observations in the validation set, $\mathbb{I}(\mathcal{A})$ is the usual indicator function that is equal to 1 when $\mathcal{A}$ is true and 0 otherwise and $\mathbb{P}(y_i = m \mid \mathbf{z}_i, \mathcal{M}_j)$ is the probability under $\mathcal{M}_j$ that $y_i = m$ given $\mathbf{z}_i$. The log-loss is a proper scoring rule for evaluating predictive models [Gneiting and Raftery, 2007], where smaller scores are better, and so encourages the analysts to express their true belief about the data. It is also straightforward to set benchmarks for assessing the quality of predictions a priori. For example, for any M a mean log loss of 0 represents perfect predictive performance, while when $M = 3$ as in the considered case, a mean log loss of $-\ln(1/3) \approx 1.1$ represents "guessing", where we predict each category uniformly at random. For completeness, the classification accuracy of $\mathcal{M}_j$ was also assessed.

The results of this analysis are presented in Table 7. The first model considered was a LASSO-penalized multinomial regression of the raw milk spectra on the diet, where high-noise-level regions of the spectrum was excluded and the wavelengths standardised. The tuning parameter λ, controlling the strength of the penalization, was selected to minimise the multinomial deviance (a statistic proportional to the mean log-loss) via 10-fold cross-validation. The log-loss of this model on the training set was 0.57, which corresponds to a diet classification accuracy of 77%. A closer examination of the predictions revealed that when `CLV` and `GRS` were treated as a single category (pasture-fed), it was possible to predict `TMR` with an accuracy of 94%. When trying to predict whether the cow was fed `CLV`, given that it was pasture-fed, an accuracy of 72% was achieved. Predictive performance was much poorer on the validation set, with an overall log-loss of 0.82, corresponding to an accuracy of 58%. The model predicted `TMR` with an accuracy of 88%. However, for cows known to be pasture-fed, it predicted `CLV` with an accuracy of 49%.

The second model considered a multinomial regression of the non-zero thresholded wavelet transform coefficients of the milk spectra on diet. As above, the model was fitted by maximising a penalised log-likelihood and by using 10-fold cross-validation to tune λ. For this model, the log-loss on the training set was equal to 0.74, corresponding to an accuracy of 69%, although it predicted `TMR` with an accuracy of 88%. For pasture-fed cows, it predicted `CLV` with an accuracy of 68%. As with the first model, performance dropped for the validation set. The log-loss was 0.88 and `TMR` accuracy was 79%. Given that a cow was pasture-fed, the `CLV` accuracy was 47%. These results are summarised in Table 7.

The obtained results clearly showed that milk spectra carry a signal distinguishing pasture-fed cows from `TMR`, but that it was difficult to distinguish between `CLV` and `GRS`. However, the predictive performance was much poorer on the validation dataset than for the training one, indicating that the adopted models did not offer a robust out-of-sample predictions. Without careful consideration of potential batch effect confounders within the sampled spectra, we are likely to overestimate the out-of-sample performance of our models. Collecting data from more cows over a more extended period should alleviate this issue and allow more robust models to be developed.

Lastly, no evidence was found to suggest that wavelet transformed spectra provided helpful insight into the cows' diet. However, that is not to say that some alternative basis expansion could improve the current predictive models. In fact, given more data on the relationship between milk spectra and diet, the development of models which allow for non-linear relationships between wavelengths may prove a fruitful avenue for future research.

3.6 Participant 6

As a first step, the training set was centered and scaled and the same transformation was applied to the test set. In the following analyses, no outliers were removed while all the spectra were transformed from transmittance to absorbance. Wavelengths from high-noise level spectral regions between 1720 and 1592 cm^{-1}, between 3698 and 2996 cm^{-1}, and greater than 3,818 cm^{-1} were removed from the analysis following Frizzarin et al. [2021b].

The Fisher score, being the ratio of between to within diet group variance, was calculated for all the wavelengths in the training set. For wavelength j, the Fisher score is given by:

$$\text{Fisher score}_j = \frac{\sum_{m=1}^{M}\sum_{i=1}^{n}\mathbb{I}(y_i = m)(\bar{x}_{\cdot j}^{(m)} - \bar{x}_{\cdot j})^2}{\sum_{m=1}^{M}\sum_{i=1}^{n}\mathbb{I}(y_i = m)(x_{ik} - \bar{x}_{\cdot j}^{(m)})^2}$$

where j denotes the wavelength index, $i = 1, \ldots, n$ denotes the spectra with n being the number of spectra in the training set, m denotes the diet group with $M = 3$, $\mathbb{I}(y_i = m)$ is an indicator of diet group spectra i, $\bar{x}_{\cdot j}$ is the average of wavelength j for all spectra ($i = 1, \ldots, n$), $\bar{x}_{\cdot j}^{(m)}$ is the average of wavelength j in diet group m. A wavelength with the highest Fisher score in each of the discarded regions was kept in the analysis. Wavelengths with Fisher score lower than 0.002 were removed from further analysis, thus leaving 380 wavelengths. In order to compare algorithms and carry out further feature selection, the training set was itself randomly split 75/25 into training and testing sets stratified by diet. A genetic algorithm [Holland, 1992], implemented in library `genalg` [Willighagen and Ballings, 2022] was used as a stochastic search method to find an optimal subset of input wavelengths for classification. Individuals in the GA population were represented by binary strings denoting wavelengths to be included or excluded for prediction. Objective function was set to be the average accuracy from ten cross-validated fits of linear discriminant analysis (LDA) of the training subset. GA was run for 200 iterations with population size set at 200. Figure 8 shows the spectra absorbance and the corresponding Fisher scores, with points denoting the wavelengths selected by the GA.

The best configuration from the final GA population had 70 wavelengths included. These wavelengths were used as inputs to the following classification algorithms:

- Linear discriminant analysis (LDA), library `MASS` [Venables and Ripley, 2002];
- Partial least squares discriminant analysis (PLS-DA) [Mevik et al., 2020];
- Least absolute shrinkage and selection operator [LASSO; Tibshirani, 1996], library `glmnet` [Friedman et al., 2010];
- Elastic net [EN; Zou and Hastie, 2005], library `glmnet`;
- Random Forest [RF; Breiman, 2001], library `ranger` [Wright and Ziegler, 2017];
- Support vector machines [Vapnik, 1998], library `kernlab` [Karatzoglou et al., 2004];
- Bayesian kernel projection classifier [BKPC Domijan and Wilson, 2011], library `BKPC` [Domijan, 2018].

All analyses were done using `R` [R Core Team, 2020], the code is available in the github repository `https://github.com/domijan/KD_Vistamilk2022`.

The training set was randomly split into ten further training/testing sets of equal size, stratified on diet. The average accuracy and standard deviation over the ten random splits for all the classification algorithms are given in Table 8. LDA performed best with average accuracy of 77.4%. PLS-DA and EN overall accuracy was of 76.9%, 76.5% respectively. The algorithms were tuned using further cross-validation of the training sets. For BKPC and SVM, the best results were obtained with a linear kernel. The predictions of the LDA were submitted to the competition. Moreover, genetic algorithm was able to select a much smaller subset of wavelengths without loss of classification performance.

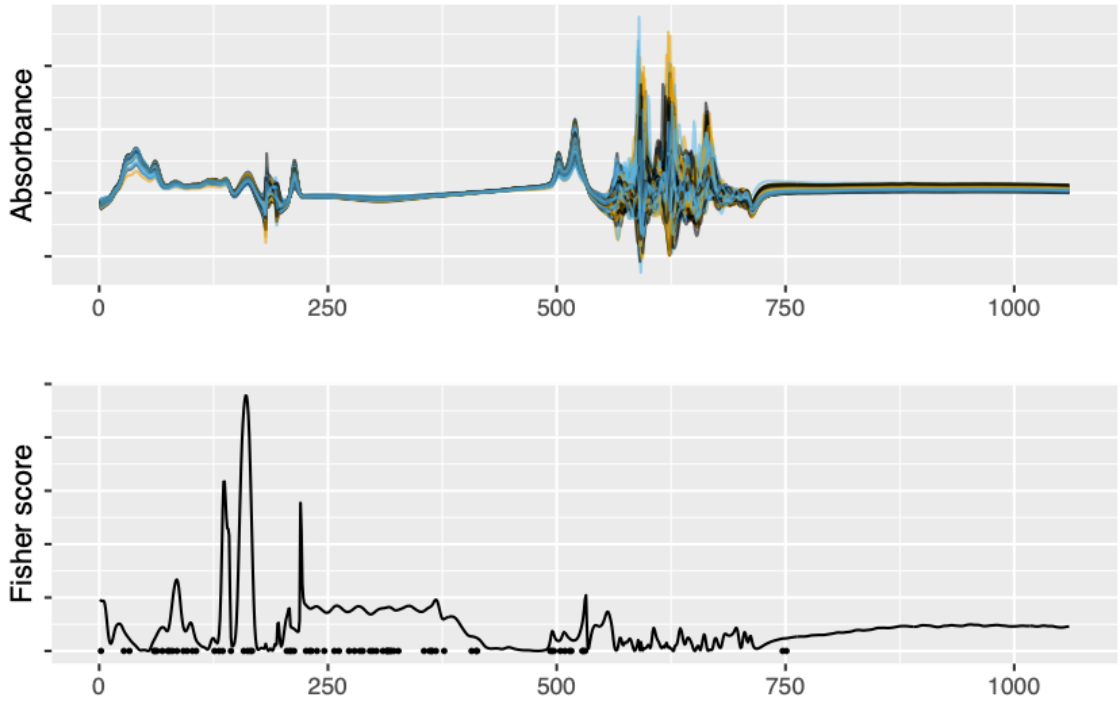

Figure 8: Spectra absorbance and the corresponding Fisher score with points on the x-axis denoting the wavelengths selected by the GA.

Table 8: Average accuracy for over ten random splits of the training set for classifiers. LDA: linear discriminant analysis; PLS: partial least squares regression; EN: elastic net; BKPC: Bayesian kernel projection classifier; SVM: support vector machine; LASSO: Least absolute shrinkage and selection operator; RF: random forest.

Accuracy	LDA	PLS	EN	BKPC	SVM	LASSO	RF
Mean	0.774	0.769	0.765	0.759	0.738	0.736	0.509
SD	0.008	0.009	0.007	0.008	0.007	0.006	0.014

4 Discussion

While the dataset provided for the data competition included three different classes to discriminate (i.e. TRM, GRS, and CLV), the main difficulty of the present data competition was concerned with the discrimination between GRS and CLV diets. In fact, the ability of distinguishing pasture and TMR dietary regimens has been already documented [Frizzarin et al., 2021b], with the discrimination being driven mainly by the different content of fatty acids (FA) in milk [Agradi et al., 2020]. In particular, milk from pasture based diet is generally richer in saturated FA such as linoleic acid, poorer in saturate FA, and have a lower omega6/omega3 ratio [see e.g. Chilliard et al., 2007; Dewhurst et al., 2006; Ferlay et al., 2013, 2017]. As MIR is known to be able to predict, with a certain degree of accuracy, the different FA in milk [Soyeurt et al., 2011], spectral data are therefore capable to discriminate also TMR and pasture diets.

On the other hand, since GRS and CLV dietary regimens differed only for the inclusion of 20% annual clover in perennial ryegrass sward for the CLV diet, induced differences in the FA might be less clear. As a consequence, to discriminate GRS and CLV exploiting spectral information only, a careful and accurate tuning of the modelling choices was required. In this regard, interestingly, some participants proposed two-steps classification approaches, with the first step focusing on TMR and pasture based diets, while the second one aimed at distinguishing CLV from GRS samples. As an example, participant 2 highlighted a potentially significant gain

Table 9: Accuracy computed on the test dataset for all the participants.

Participant	Sect 3.1.1	Sect 3.1.2	Sect 3.2	Sect 3.3.2	Sect 3.3.3
Test accuracy	0.871	0.837	0.798	0.711	0.783
Participant	Sect 3.3.4	Sect 3.4	Sect 3.5	Sect 3.6	
Test accuracy	0.796	0.786	0.724	0.766	

in terms of accuracy when considering an ensemble approach, where components extracted from LDA was used to train a linear SVM, better discriminating between GRS and CLV. Again, in Section 3.3.2 two consecutive LDA models have been fitted, with the first one being used to discriminate TMR from pasture while the second, exploiting the discriminant function on the pasture samples only, was trained to classify GRS and CLV.

Generally speaking, linear approaches introduce a gain in interpretability of the results, while paying a price in terms of accuracy. Nonetheless, the review of the different approaches presented in this paper showed that strong performances were achieved resorting to linear classifiers. In fact, remarkable results were obtained when adopting LDA-based approaches (see, e.g., participants 1, 2, 4 and 6), which were certainly proven effective in discriminating TMR and pasture diets and, as highlighted above, were also used as a building block for promising two-steps procedures. Nevertheless, the approaches presented in Sections 3.1.1 and 3.1.2, which attained the best test set prediction accuracies as it is displayed in Table 9, pointed towards the need of considering non-linearities, especially when the aim is to discriminate between GRS and CLV. This is confirmed by the confusion matrix displayed in Table 10, where it is shown that these two different dietary regimens are discriminated remarkably well, especially if considering their similarities from a compositional standpoint. Note that, while with FCN interpretation of the results and exploration of the most informative wavelengths are compromised, the approach in Section 3.1.1, which is considering again LDA as the final classifier, tends to be more transparent. However, the clever random polynomial variables generation proposed tends to produce new features which are difficult to interpret from a chemical standpoint. Therefore, as it often happens in modern data analysis routine, the adopted approaches have to be tailored on the specific aim to pursue, often dealing with the standard trade-off between accuracy and interpretability.

Data transformations, such as first and second derivative, are extensively used in near infrared spectroscopy. In the current study with MIRS data, as widely undertaken, the only transformation applied to the spectral data was their conversion from transmittance to absorbance, since the other tested transformations did not show a strong impact on the quality of the predictions. On the other hand the removal of noisy and non-informative spectral regions seemed to be of fundamental importance, as reported by the participants which tested their prediction methods before and after their removal. For example, results from Section 3.1 showed an improvement of 11.6% and of 25.7% when ridge regression and LDA were respectively used in combination of new polynomial variables generation after water regions removal. Again, in Section 3.1.2 an improvement of the prediction performance, from 17.5% (CNN) to 20.5% (FCN), after removing the water regions also when using deep learning methods is shown. Participant 1 also demonstrated the possibility to select the important variables directly from the spectra, in fact they achieved the best prediction results using a variables selection approach starting from all the spectral information (see Table 1). Variable selection was also tested in Section 3.6, where a genetic algorithm was used to select a smaller subset of wavelengths without substantial loss in classification performance.

In Section 3.3, the participants investigated the pairwise agreement among the three different approaches, to calculate by comparing the observations and quantifying the percentage of

Table 10: Final confusion matrix obtained with the approach outlined in Section Sect 3.1.1.

		Actual		
		CLV	GRS	TMR
Predicted	CLV	312	55	5
	GRS	61	300	5
	TMR	6	7	326

classifications in agreement on the total number of observations (Table 6). Methods applied by members 1 and 2 gave similar predictions (agreement of 84.21%), whereby agreement between predictions from member 3 was between 70.84% (with member 2) and 72.90% (with member 1). Although strong, the discrepancies among the three predictions could be due to: i) the different number of samples retained for model development, and ii) the different number of predictors (i.e., wavelengths) used for training, considering that the first member used the entire edited spectra, whereby the second and third applied different algorithms for wavelengths selection. This investigation from the third participant permits to understand that differences in data editing and different methodologies selected for the predictions, even if similar, brought to consistently different class predictions.

A final discussion point was related to the creation of the test dataset. The dataset was created by the organizers, who splitted the original dataset in 75% training and 25% test dataset, considering a correct division of the classes across years into the 2 datasets. The discussion revolved around whether or not divide the dataset into 75% training and 25% testing, or dividing the dataset according to time components, like keeping the samples recorded in 2015 and 2016 into the training dataset, and the samples recorded in 2017 in the test dataset. Such temporal division would permit to understand if samples recorded in previous years can predict future information.

5 Conclusion

Thanks to the high number of participants, with different backgrounds, who provided their prediction results, the data competition was a thought-provoking occasion to discuss some of the challenges arising when analyzing spectral data and provided insightful indications.

As mentioned in the paper and as it was previously shown in Frizzarin et al. [2021b], the stronger compositional dissimilarities between pasture-based diet and TMR-based ones induced an easier discrimination between the corresponding classes. This generally led to overall good performances, in terms of accuracy, for the adopted methods (see Table 9). On the other hand, the distinction between milk samples originated from GRS and CLV was more challenging. Nonetheless, as it is shown in Table 10, some hand-crafted strategies specifically proposed for this competition showed more than promising results also when employed to detect differences in the composition between distinct pasture-based feeding regimens. In particular, non-linear transformations of the original wavelengths and two-steps classification approaches, outlined in Section 3.1 and 3.3, seemed to be effective in solving this problem.

Pre-treatments were generally not beneficial for the improvement of the prediction equations, while the deletion of the spectral regions related to water (with manual selection of these regions or by means of automatic variable selection procedures) improved the prediction results. The utilization of linear models, in particular LDA, provided some of the best results, and the overall best prediction was achieved using LDA applied after wavelengths selection and random polynomial generation, as it was shown in Table 9. When spectral analyses are undertaken it is important to know not only the best possible statistical methods to use for the analyses, but also what is the best data editing for such data.

A Supplementary material

A.1 Deep neural network architecture

Table S1: List of the deep model architectures considered in Section 3.1.2, including the number of trainable parameters for each model and the type of input data they accept.

Model Architecture	Parameters	Input Data and Shape
FCN - Dense layers of 1024, 512, 128, 64 and 32 units - Output layer of 3 units - Dropout for dense layers, drop rate of 0.2 - ELU activation for hidden layers - softmax activation for output layer - Adam optimiser, initial learning rate of 0.0001 - Categorical cross entropy as loss function	1,785,923	- Linear, full (1060) - Linear, reduced (518)
CNN - Convolutional layers with 32, 64 and 128 filters - Filters of shape (3, 3), (2, 2) and (2, 2) - Flattening layer - Dense layers of 512, 256, 128, 64, and 32 units - Output layer of 3 units - ELU activation for hidden layers - softmax activation for output layer - Adam optimiser, initial learning rate of 0.0001 - Categorical cross entropy as loss function	55,332,419	- Squared, full (33×33) - Squared, reduced (23×23)
CNN_DILATED - Same architecture as CNN - Kernels built with a dilation rate of (2, 2)	41,176,643	- Squared, full (33×33) - Squared, reduced (23×23)

A.2 Participant 3

Table S2: Standardized canonical discriminant function coefficients of the variables selected by DA and effective size measures.

Wavenumber, cm^{-1}	**Function**	
	1	2
1069	2.899	0.298
1130	-3.790	0.416
1181	-2.003	5.371
1269	-7.321	-2.495
1292	10.544	-3.045
1377	-5.860	-0.482
1416	-5.885	1.267
1439	12.710	1.112
1474	-4.689	3.714
1539	-3.816	-2.385
1577	4.442	1.247
1752	11.958	6.035
2782	-1.459	0.875
2851	-15.686	-13.612
2890	16.085	3.459
2932	-4.166	0.916
Eigenvalue	1.732	0.109
& of variance	94.1%	5.9%
Canonical correlation	0.796	0.313

Table S3: Group means (centroids) for the Discriminant Functions

Diet	**Function**	
	1	2
CLV	0.872	0.403
GRS	0.954	-0.400
TMR	-1.895	-0.012

Table S4: Classification related statistics and leave-one-out cross-validation. [a] 71% of original grouped cases correctly classified. [b] Cross-validation is done only for those cases in the analysis. In cross-validation, each case is classified by the functions derived from all cases other than that case. 70.5% of cross-validatd grouped cases correctly classified.

		Diet	Predicted Group Membership CLV	GRS	TMR	Total
Original[a]	**Count**	**CLV**	629	363	83	1075
		GRS	323	668	62	1053
		TMR	39	44	942	1025
	%	**CLV**	58.5	33.8	7.7	100.0
		GRS	30.7	63.4	5.9	100.0
		TMR	3.8	4.3	91.9	100.0
Cross-validated[b]	**Count**	**CLV**	620	369	86	1075
		GRS	326	663	64	1053
		TMR	39	47	939	1025
	%	**CLV**	57.7	34.3	8.0	100.0
		GRS	31.0	63.0	6.1	100.0
		TMR	3.8	4.6	91.6	100.0

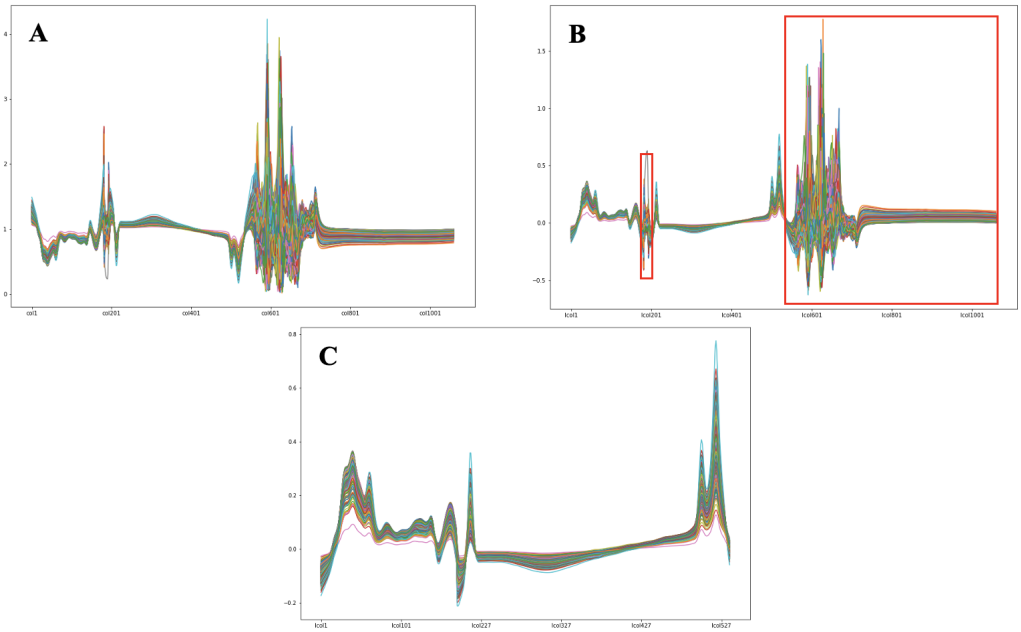

Figure S1: Line plot of raw spectra expressed in transmittance (A), conversion of raw spectra from transmittance fo absorbance (B; red boxes indicate low signal-to-noise regions), and raw spectra in absorbance after noisy area removal (C).

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) and the Department of Agriculture, Food and Marine on behalf of the Government of Ireland under grant number (16/RC/3835), the SFI Insight Research Centre under grant number (SFI/12/RC/2289_P2) and the SFI Starting Investigator Research Grant "Infrared spectroscopy analysis of milk as a low-cost solution to identify efficient and profitable dairy cows" (18/SIRG/5562).

Declaration of interests

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

Data used in the present paper are available upon request from the corresponding author.

References

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. J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D. G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P. A., Vanhoucke, V., Vasudevan, V., Viégas, F. B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467.

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society: Series B*, 60(4):725–749.

Agradi, S., Curone, G., Negroni, D., Vigo, D., Brecchia, G., Bronzo, V., Panseri, S., Chiesa, L. M., Peric, T., Danes, D., and Menchetti, L. (2020). Determination of fatty acids profile in original brown cows dairy products and relationship with alpine pasture farming system. *Animals*, 10(7):1231.

Balabin, R. M. and Safieva, R. Z. (2008). Gasoline classification by source and type based on near infrared (NIR) spectroscopy data. *Fuel*, 87(7):1096–1101.

Balabin, R. M. and Safieva, R. Z. (2011). Biodiesel classification by base stock type (vegetable oil) using near infrared spectroscopy data. *Analytica Chimica Acta*, 689(2):190–197.

Bhati, I. and Bhattacharya, M. (2020). An IOT-based system for classification and identification of plastic waste using near infrared spectroscopy. In *Proceedings of the 2nd International Conference on Communication, Devices and Computing*, pages 697–703. Springer.

Bjerrum, E. J., Glahder, M., and Skov, T. (2017). Data augmentation of spectral data for convolutional neural network (CNN) based deep chemometrics. *arXiv preprint arXiv:1710.01927*.

Bonfatti, V., Di Martino, G., and Carnier, P. (2011). Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of simmental cows. *Journal of Dairy Science*, 94(12):5776–5785.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brereton, R. G. (2015). The mahalanobis distance and its relationship to principal component scores. *Journal of Chemometrics*, 29(3):143–145.

Chen, G., Zhang, X., Wu, Z., Su, J., and Cai, G. (2021). An efficient tea quality classification algorithm based on near infrared spectroscopy and random forest. *Journal of Food Process Engineering*, 44(1):e13604.

Chilliard, Y., Glasser, F., Ferlay, A., Bernard, L., Rouel, J., and Doreau, M. (2007). Diet, rumen biohydrogenation and nutritional quality of cow and goat milk fat. *European Journal of Lipid Science and Technology*, 109(8):828–855.

Cozzolino, D. (2012). Recent trends on the use of infrared spectroscopy to trace and authenticate natural and agricultural food products. *Applied Spectroscopy Reviews*, 47(7):518–530.

De Marchi, M., Toffanin, V., Cassandro, M., and Penasa, M. (2014). Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. *Journal of Dairy Science*, 97(3):1171–1186.

Dempster, A., Petitjean, F., and Webb, G. I. (2020). ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495.

Dempster, A., Schmidt, D. F., and Webb, G. I. (2021). Minirocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 248–257. Association for Computing Machinery.

Dewhurst, R., Shingfield, K., Lee, M., , and Scollan, N. (2006). Increasing the concentrations of beneficial polyunsaturated fatty acids in milk produced by dairy cows in high-forage systems. *Animal Feed Science and Technology*, 131(3-4):168–206.

Domijan, K. (2018). *BKPC: Bayesian Kernel Projection Classifier*. R package version 1.0.1.

Domijan, K. and Wilson, S. P. (2011). Bayesian kernel projections for classification of high dimensional data. *Statistics and Computing*, 21(2):203–216.

Ferlay, A., B, G., and Y, C. (2013). Maitrise par l'alimentation des teneurs en acides gras et en composes vitaminiques du lait de vache. *INRAE Productions Animales*, 26(2):177—-192.

Ferlay, A., Bernard, L., Meynadier, A., and Malpuech-Brugère, C. (2017). Production of trans and conjugated fatty acids in dairy ruminants and their putative effects on human health: A review. *Biochimie*, 141:107–120.

Ferragina, A., Cipolat-Gotet, C., Cecchinato, A., and Bittante, G. (2013). The use of fourier-transform infrared spectroscopy to predict cheese yield and nutrient recovery or whey loss traits from unprocessed bovine milk samples. *Journal of Dairy Science*, 96(12):7980–7990.

Filzmoser, P. and Gschwandtner, M. (2021). *mvoutlier: Multivariate Outlier Detection Based on Robust Methods*. R package version 2.1.1.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711.

Fleming, A., Schenkel, F., Chen, J., Malchiodi, F., Bonfatti, V., Ali, R., Mallard, B., Corredig, M., and Miglior, F. (2017). Prediction of milk fatty acid content with mid-infrared spectroscopy in canadian dairy cattle using differently distributed model development sets. *Journal of Dairy Science*, 100(6):5073–5081.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Frizzarin, M., Bevilacqua, A., Dhariyal, B., Domijan, K., Ferraccioli, F., Hayes, E., Ifrim, G., Konkolewska, A., Nguyen, T. L., Mbaka, U., Ranzato, G., Singh, A., Stefanucci, M., and Casa, A. (2021a). Mid infrared spectroscopy and milk quality traits: a data analysis competition at the international workshop on spectroscopy and chemometrics 2021". *Chemometrics and Intelligent Laboratory Systems*.

Frizzarin, M., O'Callaghan, T., Murphy, T., Hennessy, D., and Casa, A. (2021b). Application of machine-learning methods to milk mid-infrared spectra for discrimination of cow milk from pasture or total mixed ration diets. *Journal of Dairy Science*, 104(12):12394–12402.

Geronimo, B. C., Mastelini, S. M., Carvalho, R. H., Júnior, S. B., Barbin, D. F., Shimokomaki, M., and Ida, E. I. (2019). Computer vision system and near-infrared spectroscopy for identification and classification of chicken with wooden breast, and physicochemical and technological characterization. *Infrared Physics & Technology*, 96:303–310.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Hahs-Vaughn, D. L. (2016). *Applied multivariate statistical concepts*. Routledge.

Ho, P., Bonfatti, V., Luke, T., and Pryce, J. (2019). Classifying the fertility of dairy cows using milk mid-infrared spectroscopy. *Journal of Dairy Science*, 102(11):10460–10470.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

IBM Corp. (2017). *IBM SPSS Statistics for Windows, Version 25.0*. R Foundation for Statistical Computing, Armonk, NY.

Ji-yong, S., Xiao-bo, Z., Xiao-wei, H., Jie-wen, Z., Yanxiao, L., Limin, H., and Jianchun, Z. (2013). Rapid detecting total acid content and classifying different types of vinegar based on near infrared spectroscopy and least-squares support vector machine. *Food Chemistry*, 138(1):192–199.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Kassambara, A. and Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7.

Khuwijitjaru, P., Boonyapisomparn, K., and Huck, C. (2020). Near-infrared spectroscopy with linear discriminant analysis for green 'robusta' coffee bean sorting. *International Food Research Journal*, 27(2):287–294.

MATLAB (2018). *version 9.4 (R2018a)*. The MathWorks Inc., Natick, Massachusetts.

McDermott, A., Visentin, G., De Marchi, M., Berry, D., Fenelon, M., O'Connor, P., Kenny, O., and McParland, S. (2016). Prediction of individual milk proteins including free amino acids in bovine milk using mid-infrared spectroscopy and their correlations with milk processing characteristics. *Journal of Dairy Science*, 99(4):3171–3182.

McParland, S., Lewis, E., Kennedy, E., Moore, S., McCarthy, B., O'Donovan, M., Butler, S. T., Pryce, J., and Berry, D. (2014). Mid-infrared spectrometry of milk as a predictor of energy intake and efficiency in lactating dairy cows. *Journal of Dairy Science*, 97(9):5863–5871.

Meloun, M. and Militkỳ, J. (2011). *Statistical data analysis: A practical guide*. Woodhead Publishing Limited.

Mevik, B.-H., Wehrens, R., and Liland, K. H. (2020). *pls: Partial Least Squares and Principal Component Regression*. R package version 2.7-3.

Middlehurst, M. and Bagnall, A. (2022). The freshprince: A simple transformation based pipeline time series classifier. *CoRR*, abs/2201.12048.

Munera, S., Gómez-Sanchís, J., Aleixos, N., Vila-Francés, J., Colelli, G., Cubero, S., Soler, E., and Blasco, J. (2021). Discrimination of common defects in loquat fruit cv. 'Algerie' using hyperspectral imaging and machine learning techniques. *Postharvest Biology and Technology*, 171:111356.

Nason, G. (2016). *wavethresh: Wavelets Statistics and Transforms*. R package version 4.6.8.

Nason, G. P. (2008). *Wavelet methods in statistics with R*. Springer.

Nguyen, T. L. and Ifrim, G. (2021). Mrsqm: Fast time series classification with symbolic representations. *arXiv preprint arXiv:2109.01036*.

Nguyen, T. L. and Ifrim, G. (2022). A short tutorial for time series classification and explanation with mrsqm. *Software Impacts*, 11:100197.

Oravec, M., Beganović, A., Gál, L., Čeppan, M., and Huck, C. W. (2019). Forensic classification of black inkjet prints using fourier transform near-infrared spectroscopy and linear discriminant analysis. *Forensic Science International*, 299:128–134.

O'Callaghan, T. F., Hennessy, D., McAuliffe, S., Kilcawley, K. N., O'Donovan, M., Dillon, P., Ross, R. P., and Stanton, C. (2016). Effect of pasture versus indoor feeding systems on raw milk composition and quality over an entire lactation. *Journal of Dairy Science*, 99(12):9424–9440.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Pituch, K. A. and Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS*. Routledge.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shetty, N., Difford, G., Lassen, J., Løvendahl, P., and Buitenhuis, A. (2017). Predicting methane emissions of lactating danish holstein cows using fourier transform mid-infrared spectroscopy of milk. *Journal of Dairy Science*, 100(11):9052–9060.

Soneson, C., Gerster, S., and Delorenzi, M. (2014). Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PloS one*, 9(6):e100335.

Soyeurt, H., Dardenne, P., Dehareng, F., Lognay, G., Veselko, D., Marlier, M., Bertozzi, C., Mayeres, P., and Gengler, N. (2006). Estimating fatty acid content in cow milk using mid-infrared spectrometry. *Journal of Dairy Science*, 89(9):3690–3695.

Soyeurt, H., Dehareng, F., Gengler, N., McParland, S., Wall, E., Berry, D., Coffey, M., and Dardenne, P. (2011). Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. *Journal of Dairy Science*, 94(4):1657–1667.

Terouzi, W., Platikanov, S., de Juan Capdevila, A., and Oussama, A. (2013). Classification of olives from moroccan regions by using direct ft-ir analysis: Application of support vector machines (svm). *International Journal of Innovation and Applied Studies*, 3(2):493–503.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

Tiplady, K., Lopdell, T., Littlejohn, M., and Garrick, D. (2020). The evolving role of fourier-transform mid-infrared spectroscopy in genetic improvement of dairy cattle. *Journal of Animal Science and Biotechnology*, 11(1):1–13.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Visentin, G., McDermott, A., McParland, S., Berry, D., Kenny, O., Brodkorb, A., Fenelon, M., and De Marchi, M. (2015). Prediction of bovine milk technological traits from mid-infrared spectroscopy analysis in dairy cows. *Journal of Dairy Science*, 98(9):6620–6629.

Willighagen, E. and Ballings, M. (2022). *genalg: R Based Genetic Algorithm*. R package version 0.2.1.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Wu, X., Fu, H., Tian, X., Wu, B., and Sun, J. (2017). Prediction of pork storage time using fourier transform near infrared spectroscopy and Adaboost-ULDA. *Journal of Food Process Engineering*, 40(6):e12566.

Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.