THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

Predicting feed intake in dairy cows from milk mid-infrared (MIR) spectral data using Machine Learning

Student Exam Number: [redacted]

In partial fulfilment of the requirement for the Degree of Master of Science in Quantitative Genetics & Genome Analysis at the University of Edinburgh

Names of Dissertation Supervisors: [redacted]

Word Count: 10,170

ABSTRACT

Feed intake is an economically important characteristic of dairy cows and is known to exhibit genetic variation. Yet it is not explicitly included in national breeding goals. This is largely due to the difficulty associated with measuring individual feed intake in a large population such that a sufficiently large reference population can be created. Repurposing milk mid-infrared spectroscopy data has allowed several phenotypes such as bovine tuberculosis, pregnancy status and methane emissions, to be accurately predicted without the need for direct measurement of the trait. However high accuracies for predicted feed intake have not been achieved using spectral data. Machine learning is a branch of predictive modelling which aims to optimize the performance of a task by learning from "experience" or historical data. In this study, a popular machine learning algorithm known as XGBoost was used alongside the traditional predictive method for milk mid-infrared spectral data, Partial Least Squares (PLS) regression, to derive predictions for individual feed intake in lactating dairy cows. The phenotypic data included 258,885 feed intake recordings from 771 Holstein dairy cows from the Langhill research herd at Crichton Royal Farm in Dumfries, Scotland. Mid-infrared spectral data was aligned to feed intake records to create the datasets for model development. Models were trained using 90% of the dataset and 10-fold cross-validation while 10% of the dataset was removed and retained for use as an external validation set. PLS regression models resulted in feed intake accuracy (i.e., the correlation between predicted and actual feed intake) of 0.65. XGBoost models resulted in a feed intake accuracy of 0.73 after 188 epochs. Genetic analyses generated heritability estimates for actual and MIR-predicted feed intake of 0.03 and 0.05, respectively. Estimates of phenotypic and genetic correlations between actual and MIR-predicted feed intake were 0.94 and 0.62. This study shows that MIR-predicted feed intake values generated from routinely collected milk samples may offer a low-cost, viable alternative to actual individual feed intake for use as a selection criterion in breeding programmes and further validates the use of machine learning in the agriculture sector.

Table of Contents

INTRODUCTION 1

MATERIALS & METHODOLOGY 7

Animals 7

Feed Intake Data 7

Mid-infrared Spectral Data 8

Milk Sampling and MIR Spectral Analysis. 8

Pretreatment and standardisation of MIR Spectral Data. 8

Alignment of Spectral Data to Feed Intake Profile. 9

Data Preparation 10

Training and test datasets 11

Hardware and Software Requirements for Machine Learning 13

Model Development 15

Partial Least Squares Regression. 15

XGBoost. 15

Measures of Accuracy. 17

Application of Prediction Model to Genetic Analysis 17

Pedigree Construction. 17

Feed Intake Heritability Estimates. 17

RESULTS 19

Alignment of Spectral Data to Feed Intake 19

Development of Prediction Tools 19

PLS Regression 19

XGBoost 22
DISCUSSION 27
Predictive performance of XGBoost versus PLS Regression 27
Comparison of Results to Previous Studies 28
Application of ML in Agriculture 30
Choice of ML Model in this study 31
Impact of Genomics and Phenomics on Trait Prediction 32
Difficult-to-measure phenotypes 34
Genetic correlations 36
CONCLUSIONS 39
APPENDICES
REFERENCES

INTRODUCTION

The livestock sector plays a critical role in human nutrition and is a core component of the agriculture economy (Gerber et al., 2010). In many parts of the world, food shortages remain an ever-present struggle with 8.8% of the world's population being undernourished (Johnson, 2009; Roser & Ritchie, 2019). Rapid population growth will continue to increase demand for animal products as the global population is predicted to rise from 7.6 billion to 9.1 billion by 2050. Today, the livestock sector produces over 340 million tonnes of animal products each year – 3x the volume produced fifty years ago (Diouf, 2009; Ritchie & Roser, 2017). It will be a great challenge to increase food supplies at a rate that can reduce the undernourished population and also account for future population growth. In 2012, the United Nations Food and Agriculture Organization forecasted that by 2050 global production of meat is projected to reach 465 million tonnes and production of dairy products to increase to 1,043 million tonnes, a combined total of over 1.5 billion tonnes (Alexandratos, 2012). Although the output associated with livestock has more than tripled, the rise in the population of animals has not been quite so extreme. Global cattle populations saw just a 1.46x increase in the past 50 years, from 988M in 1964 to 1,440M in 2014 (Ritchie & Roser, 2017). The disparity between population growth and volume of animal products demonstrates a significant increase in the output per animal.

One primary contributor to the improvement in quality and quantity of output in dairy cattle, and livestock as a whole, is the use of selective breeding programmes. Breeding programmes can be traced back to the findings of Austrian biologist Gregor Mendel in the 1860s and experiments on the hybridization of pea plants which gave birth to the term Mendelian Inheritance (Fisher, 1919). Selective breeding exploits the influence of genetic variants on certain traits. In modern-day, well-designed breeding programmes have revolutionized the biological efficiency of plant and livestock production through the development of genetically improved stocks (Gjedrem et al., 2012). A selection index allows farmers to manage and

rank their animals by assigning weights to a combination of traits of interest based on their breeding goals. In dairy cattle, breeding goals have traditionally had a narrow focus on increasing milk production; however, we now see national selection indices based on a more holistic approach by also incorporating health and fertility traits (Miglior et al., 2005). Three primary criteria determine the eligibility of a trait for inclusion in a breeding programme - it must be economically important, it must exhibit genetic variation and it should be easily measurable at a low cost in a large population or at least highly correlated with other characteristics which can be measured with such ease (Wallén et al., 2018).

Feed intake is among those attributes explicitly not considered in most breeding programmes (Berry, 2015). Improving feed intake is highly economically important as it represents 50-80% of overall production costs in dairy production systems (Berry, 2015; Shalloo et al., 2004). Economic value is critical as commercial farms must generate a profit. Additionally, feed intake is known to demonstrate genetic variability (Berry & Crowley, 2013; Hurley et al., 2017; Svendsen et al., 1993). Provided genetic variability exists, then breeding for improvement is possible. Even despite antagonistic genetic correlations existing between some phenotypes, for example, milk production and reproductive performance (Berry, Coffey, et al., 2014), provided the genetic correlation is less than perfectly negative (-1), genetic improvement is attainable in all traits (Berry, 2015). The heritability of feed intake in a lactating cow is relatively high (0.10 – 0.54; (Berry et al., 2007; Berry, Coffey, et al., 2014; Veerkamp & Thompson, 1999))and so also satisfies this criterion. The main factor which hampers feed intake's inclusion in breeding programmes is affordable access to routine feed intake measurements for a large population of animals necessary to achieve a high degree of accuracy of selection (Wallén et al., 2018). Average intake measurements can be collected by grouping cows for feeding, filling feed bunks and allowing the group to eat at all feeding bunks, subtracting the amount of feed remaining from the amount offered and dividing this equally by the number of cows

in the feeding group. Motion-activated gates that restrict the dairy cows to a single feeding pen may be used for more accurate measurements (Coffey, 2020) however, this method is expensive, time-consuming and unlikely to be deployed by universally by on every farm.

Despite the difficulty of recording, in some cases, such as research herds and nucleus breeding herds, feed intake is incorporated into their routine measurements (Berry, Coffey, et al., 2014). It may be possible to collate data from across these herds to build predictive models (Banos et al., 2012); however alternative methods for measuring or approximating feed intake remain highly desirable. Mid-infrared spectroscopy is a method of phenotype measurement that relies on the variation of molecular signatures in the milk of dairy cows as a result of different physiological processes (Soyeurt et al., 2006). Mid-infrared spectroscopy of milk samples is an internationally used, noninvasive method of predicting fat, protein and lactose concentration in milk samples during routine milk recording (De Marchi et al., 2014; Denholm et al., 2020; Soyeurt et al., 2006). Mid-infrared (MIR) spectroscopy is the range of the electromagnetic radiation spectrum with wavelengths between 3 and 50 μm. The National Milk Records (NMR) use FOSS spectrometers within the MIR region of wavelengths from 900 to 5,000 cm^{-1} (~1.9-10.8 μm) to generate spectra (FOSS, 2016). Exposure to MIR radiation causes molecules to absorb energy and rotate and vibrate in patterns associated with the molecule's characteristics (Brand et al., 2021). As part of day-to-day dairy herd management, individual animal milk samples are taken and represent a highly cost-effective strategy for predicting economically significant phenotypes. Repurposing this data to develop predictions equations for other economically important phenotypes has been successfully demonstrated in milk fatty acids (Soyeurt et al., 2011), body energy (McParland et al., 2011; Smith et al., 2019), methane emissions (Dehareng et al., 2012), ketone bodies (Grelet et al., 2016), lactoferrin (Soyeurt et al., 2012), pregnancy status (Brand et al., 2021; Lainé et al., 2013; Toledo-Alvarado et al., 2018) and bovine tuberculosis (Denholm et al.,

2020). This prediction method is increasingly used as an efficient, effective, low-cost tool for traditionally expensive and challenging to record phenotypes. Before using milk MIR spectroscopy, simple regression equations had been proposed as predictors on commercial farms – using bodyweight, milk yield and days-in-milk (DIM) as explanatory variables (Fox et al., 2004, 2001). However, poor predictive accuracy has seen MIR Spectroscopy explored as a feasible alternative (Shetty et al., 2017; Wallén et al., 2018). Partial least squares regression has been the primary prediction method for developing phenotype prediction equations in MIR studies (De Marchi et al., 2014; McParland et al., 2011, 2014). Partial least squares (PLS) regression is a form of regression particularly suited to cases of regression where the number of explanatory variables is high and where it is likely that the explanatory variables are correlated. Therefore, the 1,060 wave points of the MIR spectrum are a good fit. A 2018 study developed a set of PLS regression prediction equations for modelling predicted feed intake in lactating Norwegian Red dairy cows using milk mid-infrared spectral data (Wallén et al., 2018). Their best prediction accuracy had a correlation of 0.54 between the actual feed intake measurements and MIR-predicted feed intake when milk yield and live body weight were also included in the explanatory variables; however, just 0.38 when based solely on MIR spectral data.

Although PLS has been used most frequently in milk MIR studies, the method has been criticized for its inability to handle the nonlinearity of relationships between variables and highly correlated wavelengths (Dórea et al., 2018). Another branch of predictive modelling known as machine learning has the potential to improve the quality of prediction due to its ability to model complex relationships between variables (Gianola et al., 2011). Machine learning is a tool that has proved effective in many industries; however, despite a plethora of opportunities, uptake in the agricultural and animal science sectors has been slow (Howard, 2018). These opportunities are present across many parts of the agriculture sector – in crop management, soil management, water management and in our case, livestock

management. There are many different machine learning (ML) algorithms; however, the common aim is to optimize the performance of a task by learning from "experience" or historical data (Benos et al., 2021). The greater the amount of information, the better the ML algorithm will perform - just as a human will execute a task more effectively as they gain more experience (Alpaydin, 2020). The primary result of ML is a measure of generalizability, which is the ability of the algorithm to make the correct prediction when new data is provided using learned rules derived from previous exposure to comparable data (Domingos, 2012). ML algorithms are typically characterized by two distinct processes – training and testing. By evaluating the response variable versus the explanatory variables for many records in the training dataset, the machine "learns" to perform the task incrementally better as it gains experience. Once the learning reaches an optimal point, as expressed by some statistical relationship (e.g., improvement in the mean squared error of prediction), the model completes training. Data previously hidden from the model is evaluated during the testing phase, and the model's performance is assessed based on its prediction accuracy.

Large datasets such as those with milk MIR spectral data are ideal candidates to exploit the power of machine-learning algorithms as predictive tools based on milk spectra and uncover previously unnoticed relationships between the spectra and traits of importance (Denholm et al., 2020). The research group at Scotland's Rural College (SRUC) supporting this study investigates machine learning applications to milk MIR spectral data for various economically important traits. Using a specific ML algorithm from the family of Artificial Neural Networks known as Convolutional Neural Networks (CNNs), they developed predictors for bovine tuberculosis (bTB) (Denholm et al., 2020) with 95% prediction accuracy (i.e., the number of correct predictions divided by the total number of predictions) and pregnancy status (Brand et al., 2021) with 88% prediction accuracy both based solely on the milk MIR spectral data. CNNs are often trained as image classification models. In both cases mentioned previously, the individual spectral

records are converted into greyscale images by resizing the array from 1060 x 1 to 53 x 20 pixels based on a normalized wavelength value. These images serve as the first inputs to the CNN, allowing the model to find new data characteristics important for phenotypic prediction. Another family of machine learning models suitable for this type of analysis is ensemble models. XGBoost is a tree-boosting model which divides the overall predictor into small sub-trees representing simpler, weaker models. The model uses a gradient descent technique to minimise a loss function based on the differences between observed and predicted values to produce the final prediction by repeatedly adding new trees that forecast the mistakes of the previous trees and merging the predictions (Chen & Guestrin, 2016). These models run far faster than CNNs and do not require as much data preprocessing and model tuning.

The objective of this study was to use reference phenotypic data obtained from the dairy research herd at Scotland's Rural College (SRUC) and combine it with milk MIR spectral data to generate prediction models for feed intake using machine learning. The ability to estimate feed intake at high accuracy from a cow's spectral profile would provide a low-cost, noninvasive method for capturing a difficult-to-measure phenotype for significant economic and environmental importance.

MATERIALS & METHODOLOGY

Animals

The Langhill lines of Holstein Friesian dairy cows (n=8,672) from Scotland's Rural College (SRUC) Dairy Research Centre at Crichton Royal Farm in Dumfries, Scotland, were used in this study. Feed intake datasets, mid-infrared spectral datasets, and a 7-generation pedigree for genetic analysis of actual and predicted feed intake amounts were built using the Langhill animal dataset.

Dairy cattle in the Langhill herd have been part of a long-term 2 x 2 factorial experiment evaluating the impact of diet type (High- and Low- concentrate) and genetic lines (Control – the UK average genetic merit for the mass of fat and protein; and Select – cows bred to maximize the performance of these traits) (Roberts & March, 2013; Smith et al., 2019). This approach essentially creates four distinct herds representing the majority of herd types across the UK. Langhill cows are extensively monitored for a variety of kinds of traits which may be helpful for analysis.

Feed Intake Data

Langhill dairy cows have daily feed intake recorded for three consecutive days each week. In this study, feed intake data from exclusively lactating dairy cows were used. Additionally, only cows with experimental feed types and week in milk (WIM) between 1 and 43 were retained. Non-experimental feed types are used to code for animals when they are sick, experiencing a dry period, or are not on trial. These records are removed to preserve data quality. The final dataset included 258,885 unique individual feeding records from February 2012 and July 2021. Each record contained the volume of feed offered (kg/day) and volume of feed refused (kg/day).

Mid-infrared Spectral Data

Milk Sampling and MIR Spectral Analysis.

Mid-infrared analysis of milk samples gathered from the SRUC's Dairy Research Centre was conducted by National Milk Records (NMR) and analysed on their Foss MilkoScan machines (Foss Analytics, Hillerød, Denmark) based at National Milk Laboratories. As part of the NMR's routine milk-recording service, milk sampling of individual cows occurred at consecutive a.m. and p.m. milkings on a fortnightly basis. In contrast, the UK milk recording system takes place monthly (Smith et al., 2019). 43,355 spectral records sampled between February 2012 and June 2021 and 925 unique lactating cows were used. Each spectral record contained 1,060 wave points available for further analysis.

Pretreatment and standardisation of MIR Spectral Data.

A spectrum of 1,060 data points is obtained during MIR analysis of each milk sample. Each point indicates the absorption of mid-infrared light by the milk sample at a specific wavelength within the range of 900 cm^{-1} to 5,000 cm^{-1} (Lainé et al., 2017). The spectral data were subjected to different necessary pretreatments before being used to generate prediction models. Firstly, using the transformation $log10^{-0.5}$, the transmittance values were transformed to a linear absorbance scale. Secondly, spectral data were standardised against a master spectrometer (Grelet et al., 2015) to account for drift caused by spectrum data collection from various MIR machines and across time (Grelet et al., 2014). The Walloon Agricultural Research Centre provides files outlining the standardisation process, used in conjunction with procedures established as part of the InterReg/EU-funded project OptiMIR (Friedrichs et al., 2015). Spectra standardisation has the added benefit of guaranteeing the prediction tools created can be applied to data from other machines that use the same standardisation procedure (Denholm et al., 2020).

Creation of training and test datasets

The Langhill dataset contains measurements for feed offered (the amount provided in the feeding bins) and feed refused (the feed remaining at the end of the feeding period). Both values are measured in kilograms and recorded three times a week on three consecutive days. The Feed Intake (FI) phenotype was created by subtracting the amount refused from the total feed offered.

An Average Feed Intake (AFI) phenotype was constructed by taking the average of the FI measurements at weekly intervals. The sample date assigned to the average value was the first of the sampling days included in the average. In most cases, three measurements were taken each week; however, occasionally, only one or two recordings were done. Less than three weekly measurements may be due to illness, moving herd, or other inconsistencies in the recording schedule. The AFI phenotype has the advantage of smoothing out day-to-day fluctuations in feed consumed. Additionally, as MIR spectral data is recorded on a fortnightly basis and feed intake daily, the AFI phenotype ensures that the sample dates of the MIR spectral data are more closely fit to the feed eaten in a given week.

Alignment of Spectral Data to Feed Intake Profile.

The MIR spectral data were aligned to the FI phenotype to create three distinct phenotypes, FI-0, FI-1, FI-2 phenotypes were constructed to represent the passage of feed through the digestive tract of the cows, where the number symbolizes the number of days after the feed intake sampling date that the aligned milk MIR spectral data were analysed. These phenotypes will be referred to in this study as the exact-day phenotypes (as opposed to the weekly average phenotype). Table 1 provides a brief illustration of how the exact-day phenotypes differ and their relationship with the aligned milk MIR sampling date. Differences in the predictive model results using the FI-0, FI-1, FI-2 phenotypes will provide insight into the physiological changes incurred because of the volume of feed intake and how long it may take for these changes to be detectable in milk samples.

Table 1. Example breakdown of Feed Intake phenotypes

Milk sample date	FI-0 sample date	FI-1 sample date	FI-2 sample date
July 10th, 2021	July 10th, 2021	July 9th, 2021	July 8th, 2021

In total, eight datasets were created to train and test our prediction models. Each of the four phenotypes was analysed and predicted using the aligned AM and PM milk spectral data. Each record in the FI-0 datasets matched the milk MIR spectral data collected on precisely the same date. The record was discarded if no milk spectral data was aligned with the feeding sample date. This process was repeated for the FI-1 and FI-2 phenotypes; however, the feed intake sample date aligned to the milk sample date was 1 or 2 days before the milk sample date. For the AFI phenotype, a tolerance of ±7 days was applied to find the milk MIR spectral data that most closely matched the week represented by the AFI measurement.

Data Preparation

All datasets were preprocessed and reformatted to ensure they were prepared for the model development phase. Shapiro-Wilk and D'Agostino's K-Squared statistical tests were used to test for normality of the feed phenotypes. Based on failing tests for normality, outliers were handled using the interquartile range instead of the standard method based on standard deviations for normally distributed traits. The interquartile range (75^{th} percentile value - 25^{th} percentile value) was multiplied by a cut-off of ±1.5 to determine the outlier threshold values. Outliers are expected due to human error or faults in the measurement technology; for example, negative feed intake values are impossible and should not be present in the data. Therefore, cleaning the data in advance was necessary to ensure high-quality information was passed to the models. Table 2 presents the summary statistics of each dataset along with the breakdown of the outlier removal process.

Table 2. Summary statistics and outlier removal

Phenotype	Milking	Records	MIR-aligned records	Mean (sd) - kg/d	Min	Max	IQR	Outliers	Final Dataset[1]
AFI	AM	74,086	46,712	44.3 (13.9)	-0.1	233.3	17.7	569	46,143
AFI	PM	74,086	46,257	44.4 (13.9)	-0.1	233.3	17.6	552	45,705
FI (0)	AM	258,885	17,792	42.6 (15.1)	-0.2	301.0	18.4	469	17,323
FI (0)	PM	258,885	17,456	42.8 (15.2)	-0.2	301.0	18.4	468	16,988
FI (-1)	AM	258,885	17,567	43.0 (15.5)	-0.1	271.5	18.0	438	17,129
FI (-1)	PM	258,885	17,217	43.1 (15.5)	-0.1	271.5	18.1	428	16,789
FI (-2)	AM	258,885	17,397	43.1 (17.9)	-0.1	569.8	18.7	514	16,883
FI (-2)	PM	258,885	17,017	43.2 (17.9)	-0.1	569.8	18.7	504	16,513

[1] Final Dataset = total number of MIR-aligned phenotypic records with outliers stripped from the dataset
[2] IQR = Interquartile range

Training and test datasets

Although there are many different machine learning algorithms, they typically all use a similar design that divides the dataset into two groups: training and testing. The training set is passed to the model with the features and labels intact. In our case, the features are the 1,060 MIR spectral wavelengths and the labels are the actual feed intake measurements. The training set is divided into two subsets, one for training and the second for validation during training. The test set is kept entirely separate from the training set and only used when the final model has been trained. The test set aims to replicate a real-world application of the model by providing information based on new input data. This test set is passed through the model with the labels removed, generating a prediction based on the model's learning in the training phase. The true values are then compared versus the predictions to determine the accuracy of the model.

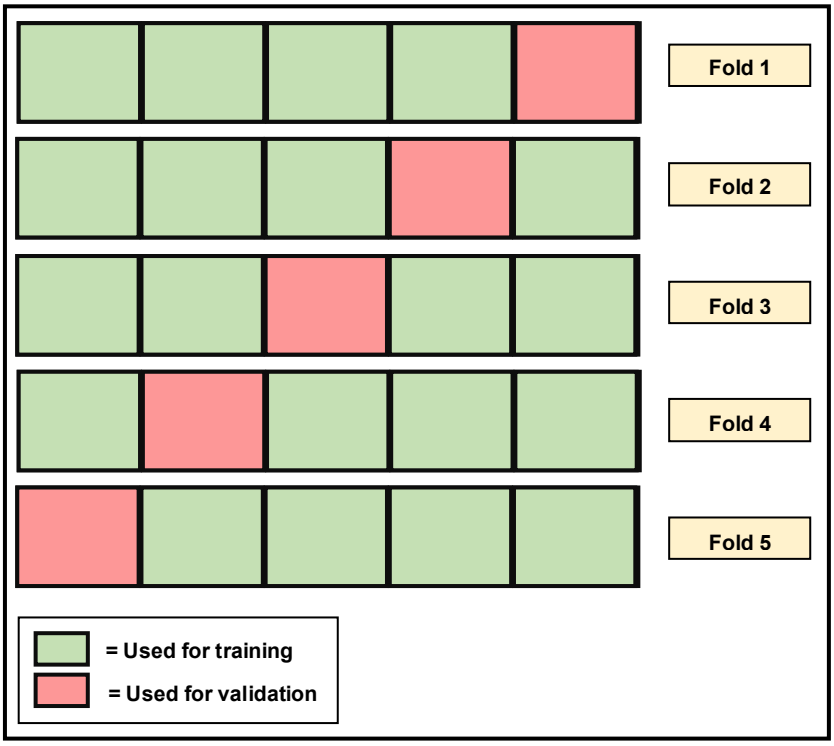

Figure 1: **Illustration of 5-fold cross validation**

In this study, the datasets were split into training and testing sets using a 9:1 ratio and the training set was further subdivided using 10-fold cross-validation. Cross-validation is a technique used to remove bias in our models by dividing the entire dataset into blocks and training the model using all but one block, then testing the model's performance with the omitted block. This process is repeated until every block has been omitted and used as the test set. Figure 1 provides an example of 5-fold cross-validation. The number of folds typically depends on the number of records in the dataset however with large datasets such as those in this study, 10-folds is typically considered sufficient. The performance of the model is then represented by the average performance across all 10 sets. Cross-validation is used to determine the hyperparameters (e.g. number of factors in a PLS regression model) that will be used to fit the final model. Figure 2 provides a schematic overview of the process of data splitting, model training, cross-validation and evaluation of the performance of the final model.

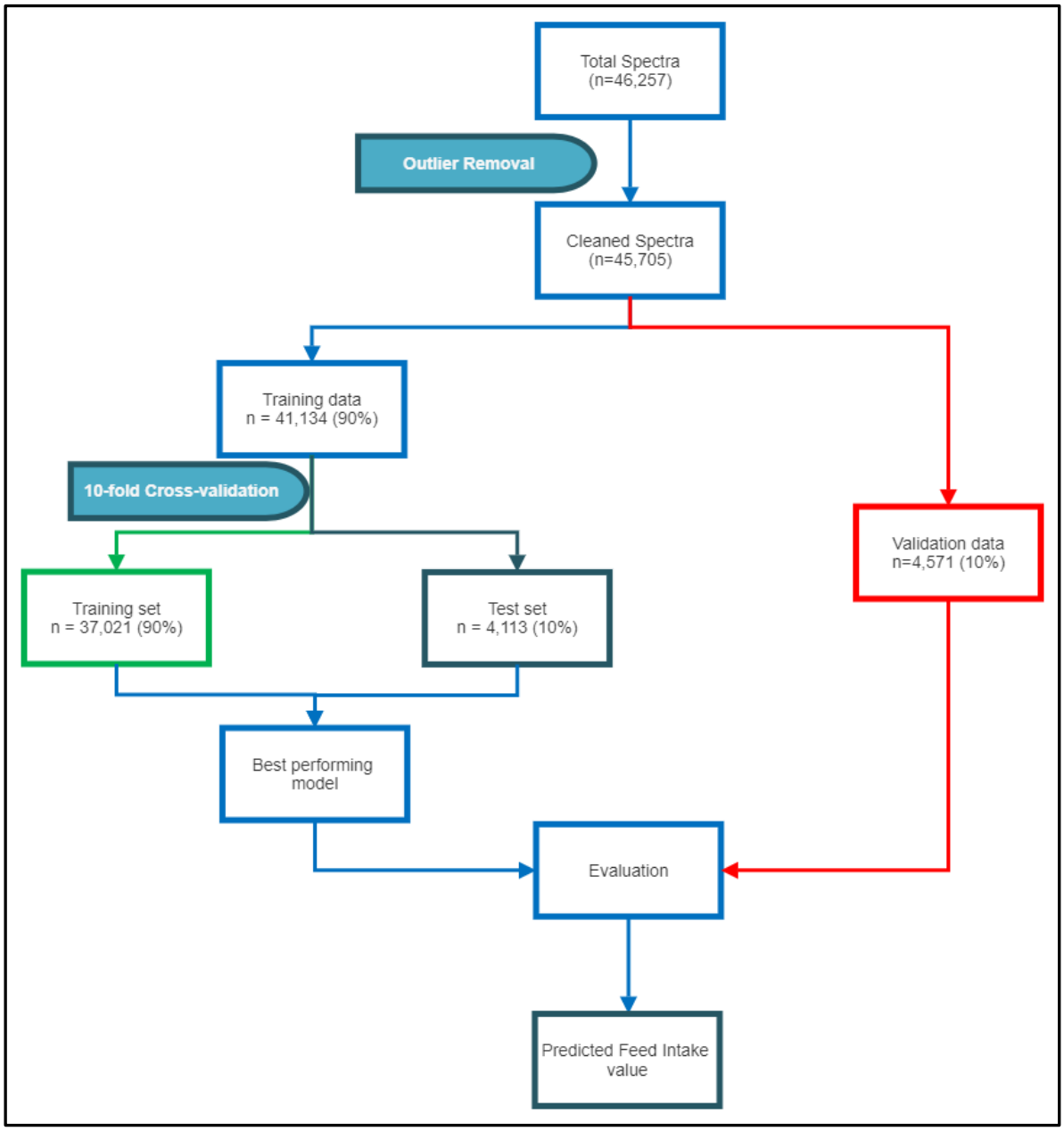

Figure 2: Schematic diagram of the model training and evaluation process (AFI-PM dataset used for illustration)

Hardware and Software Requirements for Machine Learning

Certain hardware and software requirements must be satisfied in order to properly harness the potential of machine learning efficiently. Table 3 shows the complete system specifications utilised in this study which are summarised as follows: NVIDIA DGX Station personal AI supercomputer (NVIDIA Ltd., 2019) with

4 NVIDIA Tesla V100 graphics processing units (GPU), Microsoft Windows, Python 3.6 Virtual Environment, Microsoft SQL Server Management Studio and ASReml-R. When compared to the central processing unit version of XGBoost, the GPU-enabled version offers faster processing rates. The Microsoft SQL Server Management Studio is used to access and manipulate data from the Langhill herd and export datasets for further analysis. All genetic analyses were conducted using ASReml-R, an R-based adaptation of ASReml (Butler et al., 2018).

Table 3. System specifications for machine learning suite (building on an NVIDIA DGX Station[1])

Graphics processing units (GPU)	4x Tesla V100
TFLOPS[2] (mixed precision)	500
GPU Memory	128 GB total system
NVIDIA tensor cores	2,560
NVIDIA CUDA[3] cores	20,480
Central processing unit	Intel Xeon E5-2698 v4 2.2 GHz (20-Core_
System memory	256 GB RDIMM DDR4
Storage	Data: 3X 1.92 TB SSD RAID 0 OS: 1X 1.92 TB SSD
Network	Dual 10GBASE-T (RJ45)
Maximum power requirements	1,500 W
Operation System	Microsoft Windows 10.0 (64 bit) Microsoft Windows Server 2019 10.0 (64 bit)
Software	DGX Recommended GPU Driver CUDA Toolkit Python 3.6.7 XGBoost-GPU Microsoft SQL Server Management Studio 2014 - 12.0.6433.1 (X64) ASReml-R (Version 4.1.0.143)

[1]NVIDIA Ltd. (2019)
[2]TFLOPS = teraflops (i.e., a processor's ability to calculate one trillion floating-point operations per second).
[3]CUDA = Compute Unified Device Architecture. NVIDIA's parallel computing architecture enables increased computing performance by harnessing the power of the GPU to speed up intensive tasks.

Model Development

Partial Least Squares Regression.

Partial Least Squares (PLS) regression is the most common predictive method used for milk MIR spectral data. For each dataset, a PLS regression model was used to predict feed intake from milk spectra data. The PLS model extracts successive linear combinations of the predictors, referred to as factors, that address both response and predictor variance. As is common practice, Savizky-Golay smoothing was applied to the spectral data to remove baseline variation by using the first derivate with filter width 7, polynomial order 2, and mean centering to the raw spectra (McParland et al., 2011; Smith et al., 2019; Soyeurt et al., 2011). The analysis was carried out using Python 3.6 and the Scikit-learn machine learning package (Pedregosa et al., 2011). 10-fold cross-validation was carried out as described previously to monitor performance and determine the number of factors to retain in the optimal model.

XGBoost.

Extreme Gradient Boosting, otherwise known as XGBoost, is a method from a branch of machine learning algorithms known as ensemble methods. Ensemble methods aggregate several ML models to achieve greater predictive performance compared to each component model. Consulting several experts and integrating their perspectives to create a better educated and optimal conclusion is an example of ensemble learning in action (Chakraborty & Elzarka, 2019). XGBoost works on the principle of boosting – where multiple weak learners, such as regression trees, are combined to create a single, more robust learner (Marsland, 2011, p. 154). The main principle behind this method is to learn in a sequential manner, with the current regression tree being fitted to the residuals (errors) from prior trees. The residuals are then updated by adding this new regression tree to the fitted model. Gradient boosting can be expressed as:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} t_k(x_i), \qquad t_k \in T \qquad \textbf{[1]}$$

where y_i are the predicted response, x_i are the inputs and K is the number of functions in the function space T. These functions are included as parameters in XGBoost, allowing to find functions t_k that fit the data incrementally better during training. The functions t_k are learned by minimizing the following objective function:

$$(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(t_k) \qquad \textbf{[2]}$$

where l is a differentiable loss function and Ω is the regularizing function that penalizes overcomplicated models likely to overfit to the training data.

Unlike the PLS models, Savitzky-Golay smoothing is not required for XGBoost, so this preprocessing step was removed. Datasets were loaded into DMatrices, an internal data structure used by XGBoost optimized for memory efficiency and training speed (Punnoose & Ajit, 2016). Early stopping is a machine-learning technique for halting training when the model's performance does not increase, reducing over- and underfitting. By evaluating the root mean square error in the validation set, early stopping occurred when no improvement was attained in 10 consecutive rounds. Similar to the PLS, 10-fold cross-validation was carried out at the training stage. In XGBoost, the original sample is randomly partitioned into 10 equal-sized subsamples. A single subsample is retained as validation data for testing the model, while the remaining nine subsamples are utilised as training data. The cross-validation procedure is then repeated for a specified number of rounds or epochs, with each of the ten subsamples serving as validation data precisely once. In this way, the number of rounds of learning which is optimal for the XGBoost algorithm was obtained for each dataset. XGBoost analysis was carried out using machine learning package XGBoost 0.90 rapidsdev1.

Measures of Accuracy.

Several standard measures in predictive modelling were used to assess how well the models performed. In the cross-validation phase, the root of mean square error (RMSE) and the standard error of RMSE was used to monitor the performance of the model for hyperparameter selection. RMSE is the square root of mean square error (MSE), where MSE is the sum of the squared difference between observed values and model-predicted values divided by the number of data points (n; Bibby and Toutenburg, 1977). In the external validation or testing phase, the square root of the coefficient of determination was used as the primary measure of accuracy (R) of the predictions. The component values, root mean square error of prediction (RMSEP) and the coefficient of determination (R^2) were also recorded. R^2 is the proportion of the variation in the dependent variable that is predictable from the independent variable(s) (Chakraborty & Elzarka, 2019).

Application of Prediction Model to Genetic Analysis

Pedigree Construction.

A seven-generation pedigree was generated from Langhill herd data using SQL Server Management Studio. This required recoding the full 18-generation dataset to ensure descending order of records as are necessary for ingestion for further analysis by ASReml-R. The final pedigree, limited to just seven generations, consisted of 7,322 cows; 1,502 distinct sires, 4,539 distinct dams and 6,158 mating pairs. The number of offspring per sire ranged from 1-62 and per dam from 1-7.

Feed Intake Heritability Estimates.

The utility of milk MIR spectral data as a viable predictive technique can be determined by estimating genetic parameters of predicted feed intake and comparing these parameters to actual feed intake. As the best performing predictive model, predictions from PM milking AFI XGBoost were used for genetic analysis. Variance components for measured AFI and predicted AFI were estimated using repeatability animal linear mixed models (Butler et al., 2018). All

models were adjusted for the fixed effects of genetic line, feed type, lactation number and sampling period (Month x Year). A random additive genetic effect of individuals (cows) and a random permanent environment effect of the cow were also fitted in the models. The permanent environment effect is important to model due to the potential covariance among repeated feed intake observations per individual animal. The general form of the model, in matrix notation, was as follows in Equation 3:

$$\mathbf{y = Xb + Zu + Wpe + e} \quad \textbf{[3]}$$

where **y** is the vector of phenotypic records for feed intake, **b** is the vector of fixed effects of the genetic line (Select or Control), feed type (6 levels), lactation number (1-8) and recording date (year:month), **u** is the vector of solutions for the random additive genetic effect of the animal; **pe** is the vector of solutions for random permanent environmental effects; **e** is the vector of random residuals; and **X, Z, and W** are incidence matrices relating the corresponding effects to the dependent variable. This model was twice; once for predicted AFI and once for actual AFI as the dependent variable. The phenotypic variance was calculated as the sum of additive genetic, permanent environmental and residual variances. Heritability was calculated as the ratio of additive genetic variance to phenotypic variance. Repeatability was expressed as the ratio of the sum of additive genetic variance and permanent environmental variances to the phenotypic variance. Fitting a bivariate adaptation of the model described in Eq.1 allowed phenotypic genetic correlations between predicted and actual feed intake to be calculated as:

$$\frac{cov_{1,2}}{\sqrt{\sigma_1^2 \times \sigma_2^2}} \quad \textbf{[4]}$$

Where $\mathrm{cov}_{1,2}$ is the phenotypic/additive genetic covariance between measured and predicted feed intake and σ_1^2 and σ_2^2 are the phenotypic/additive genetic variances of measured and predicted feed intake respectively (Appendix D).

RESULTS

Alignment of Spectral Data to Feed Intake

Alignment of feed intake phenotypes with associated MIR spectral records produced eight datasets with complete records for analysis. The number of outliers ranged from 428-569 records. This constituted 0.8% of MIR-aligned records for the AFI phenotype, whereas it represented 2.5% of MIR-aligned records for the daily occurrence phenotypes (FI-0, FI-1, FI-2). This is likely owing to the AFI phenotype being constructed as a mean value of three individual feed intake records which has already accounted for part of the variation. Once identified, outliers were removed for subsequent analysis. Table 4 outlines the final datasets for use in training and testing the models. As the AFI phenotype was given a 7-day buffer to the next MIR-spectral record, the number of MIR-aligned records for the AFI phenotype is more than double the amount of the daily occurrence phenotypes.

Table 4. Summary statistics and outlier removal

Phenotype	Milking	Records	MIR-aligned records	Mean (sd) - kg/d	Min	Max	IQR	Outliers	Final Dataset[1]
AFI	AM	74,086	46,712	44.3 (13.9)	-0.1	233.3	17.7	569	46,143
AFI	PM	74,086	46,257	44.4 (13.9)	-0.1	233.3	17.6	552	45,705
FI-0	AM	258,885	17,792	42.6 (15.1)	-0.2	301.0	18.4	469	17,323
FI-0	PM	258,885	17,456	42.8 (15.2)	-0.2	301.0	18.4	468	16,988
FI-1	AM	258,885	17,567	43.0 (15.5)	-0.1	271.5	18.0	438	17,129
FI-1	PM	258,885	17,217	43.1 (15.5)	-0.1	271.5	18.1	428	16,789
FI-2	AM	258,885	17,397	43.1 (17.9)	-0.1	569.8	18.7	514	16,883
FI-2	PM	258,885	17,017	43.2 (17.9)	-0.1	569.8	18.7	504	16,513

[1]Final Dataset = total number of MIR-aligned Phenotypic records with outliers stripped from the dataset

Development of Prediction Tools

PLS Regression

Savitzky-Golay (Sav-Gol) smoothing applied to wavelength values in each dataset had the effect of correcting nonlinearities in the spectral data. Figure 3 depicts the variation of spectra peaks for untreated and first derivates. This

presents the general principle of Sav-Gol smoothing whereby sharp bands are often enhanced at the expense of broad bands and as a result improving the selection of appropriate peaks in the PLS model training (De Marchi et al., 2014). The difference in magnitude of the absorbance values (y-axis) in the first derivative spectrum also demonstrates the smoothing effect of the Sav-Gol filtering.

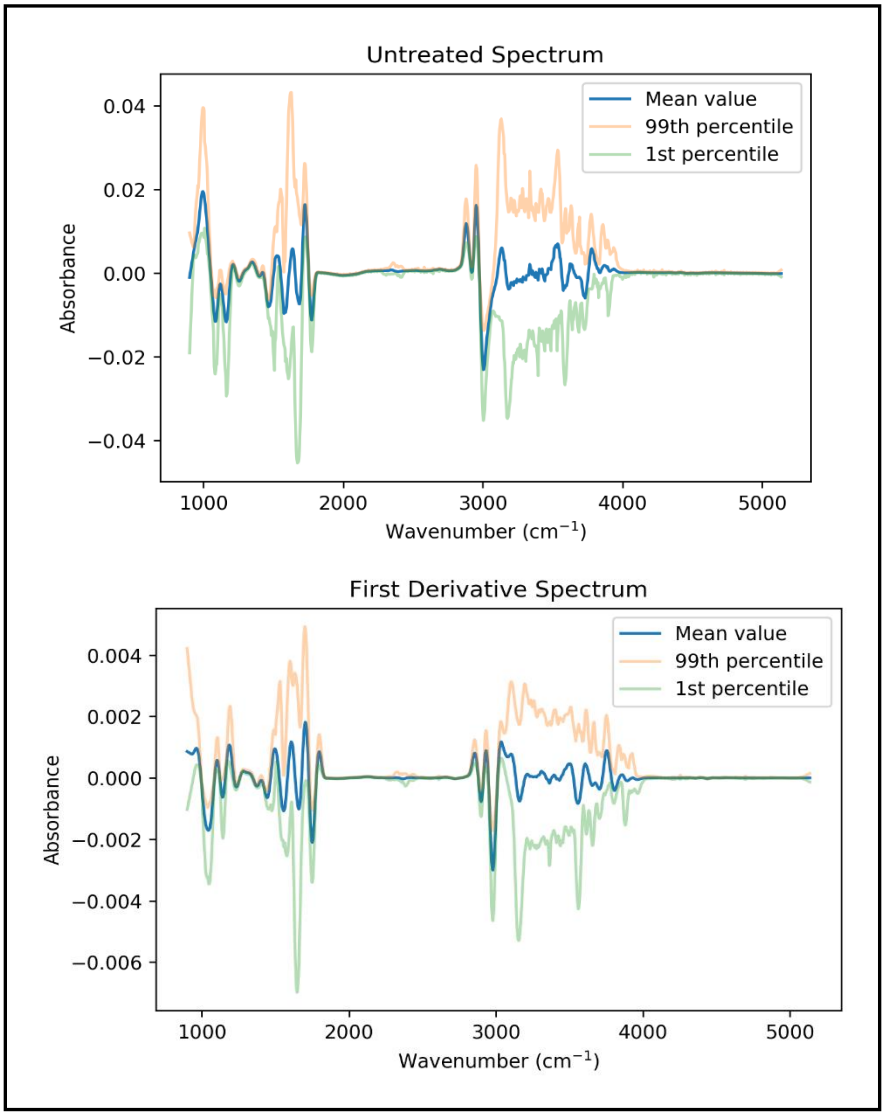

Figure 3: Impact of Savitzky-Golay smoothing on spectral data

Table 5 presents the results of the PLS Regression models developed for each phenotype and milking period. The number of variables maintained in the model was determined using 10-fold cross-validation up to a maximum of 20. The

maximum number of factors were maintained for each dataset as they reduced RMSE and increased the coefficient of determination (R^2). Including a greater number of factors may further reduce RMSE and increase R^2 however in order to reduce the likelihood of overfitting to the training data, 20 factors were chosen as the maximum. This is likely sufficient as the incremental value of an additional factor reduces as the number of factors (or components) increases as seen in Figure 4.

Table 5. Summary of results for PLS Regression

		10-fold cross-validation		External validation		
Phenotype	Milking	Train RMSE (std)	Test RMSE (std)	RMSE_P	R^2	R
AFI	AM	10.077 (0.802)	10.138 (0.807)	10.267	0.376	0.614
	PM	9.813 (0.782)	9.874 (0.785)	9.701	0.418	0.647
FI-0	AM	10.883 (0.895)	11.078 (0.908)	11.038	0.336	0.58
	PM	10.528 (0.865)	10.729 (0.878)	10.757	0.351	0.593
FI-1	AM	10.797 (0.879)	10.957 (0.891)	11.04	0.334	0.578
	PM	10.511 (0.859)	10.735 (0.870)	10.704	0.387	0.622
FI-2	AM	10.753 (0.879)	10.922 (0.892)	10.804	0.38	0.617
	PM	10.486 (0.858)	10.767 (0.874)	10.894	0.367	0.607

[1]All models trained using 20 PLS factors as determined by cross-validation optimization
[2]RMSE = Root Mean Squared Error, mean value for ten-fold validation
[3]std = standard deviation, value for ten-fold validation
[4]RMSE_P = Root Mean Squared Error of Prediction
[5]R2 = Coefficient of determination
[6]R = Accuracy (correlation between predicted and actual feed intake values)

The error of prediction (RMSE_P) was generally greater for external validation than the corresponding Train RMSE for 10-fold cross validation. As expected, increasing accuracy (R) coincided with decreased RMSE values. The accuracy of prediction using external validation ranged from 0.578-0.647. The best performing model was the AFI-PM model with RMSE_P = 9.701 and R = 0.647. Three models (FI-0 AM & PM, FI-1 AM) had an accuracy less than 0.60.

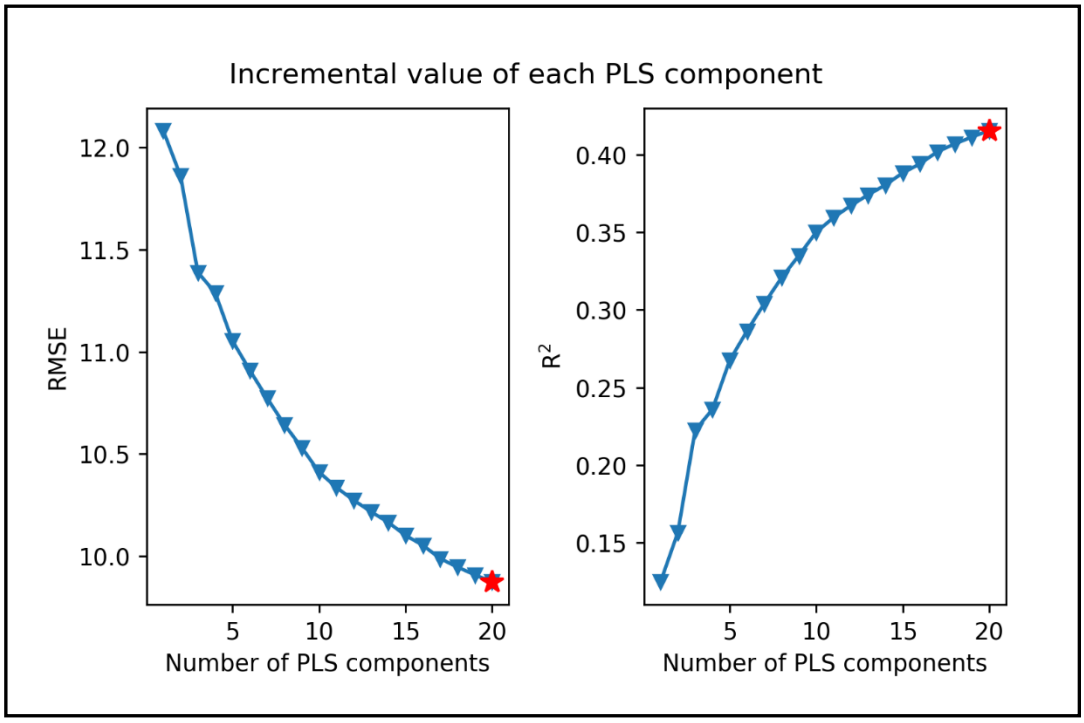

Figure 4: Incremental value of additional PLS components

XGBoost

Table 6 shows the results of the XGBoost models developed for each phenotype and milking period. Cross-validation and an early stopper were used in the training phase of each model to determine the optimal number of epochs. Early stopping is a machine-learning technique for halting training when the model's performance does not increase, thereby reducing over- and underfitting. Validation RMSE was the measure that was tracked with early stopping occurring when no improvement was attained after 10 rounds. The number of epochs optimal for the AFI phenotype was 203 for the AM milking and 188 for the PM milking. The models developed for the exact-day phenotypes (FI-0, FI-1, FI-2) were optimized at far fewer epochs, ranging from 25-33. This may be due to the larger number of MIR-aligned records in the AFI phenotype datasets (Table 4) however it is clear that the XGBoost algorithm had more “to learn” from these datasets.

The Train RMSE were 4.56 and 4.74 for the AFI-AM and AFI-PM models respectively. This is reflective of the high number of epochs these models executed

and the opportunity the model had to learn the features of the training data. This improvement in model training performance was also realized in the external validation datasets with RMSEP = 8.714, R^2 = 0.53 and R = 0.730 in the AFI-PM model. This model narrowly outperformed the AFI-AM model in the external validation. The exact-day phenotypes had larger RMSEs in the 10-fold cross-validation with Train RMSEs ranging from 7.53-7.89 and Test RMSEs ranging from 10.59-10.85. This performance was reflected in the external validation with accuracies (R) ranging from 0.590-0.629. The best performing exact-day phenotypes were the FI-1 PM and FI-2 AM with accuracies of 0.629 and 0.621 respectively.

Table 6. Summary of results for XGBoost

			10-fold cross validation		External Validation		
Phenotype	Milking	Epochs	Train RMSE (std)	Test RMSE (std)	RMSEP	R^2	R
AFI	AM	203	4.56 (0.03)	8.94(0.10)	9.013	0.519	0.722
	PM	188	4.74 (0.03)	8.91 (0.11)	8.714	0.530	0.730
FI-0	AM	26	7.84 (0.05)	10.63 (0.16)	10.901	0.353	0.595
	PM	25	7.89 (0.04)	10.67 (0.18)	10.796	0.346	0.590
FI-1	AM	33	7.61 (0.07)	10.70 (0.14)	10.741	0.369	0.608
	PM	29	7.53 (0.06)	10.59 (0.17)	10.628	0.395	0.629
FI-2	AM	28	7.74 (0.07)	10.85 (0.20)	10.756	0.386	0.621
	PM	27	7.69 (0.08)	10.73 (0.20)	10.900	0.367	0.608

[1]RMSE = Root Mean Squared Error, mean value for ten-fold validation over a variable number of epochs

[2]std = standard deviation, value for ten-fold validation

[3]RMSE_P = Root Mean Squared Error of Prediction

[4]R2 = Coefficient of determination

[5]R = Accuracy (correlation between predicted and actual feed intake values)

The performance of the models can be further visualised as in Figure 5: the distribution of actual vs predicted feed intake (AFI (5a), FI-0 PM (5b)) and Figure 6: the fitted model as a regression line over the scatterplot of actual feed intake versus predicted feed intake (AFI). Figure 4a demonstrates a considerable degree of overlap between the AFI model's predictions. The actual distribution of values has "fatter" tails than the more normal-shaped model-predicted distribution. Greater representation of these values in the training dataset may improve the performance of the model for these values which deviate more severely from the mean. In

contrast to the AFI-PM model, the best performing model, the FI-0 PM model which performed worst in the external validation presents far less overlap in the predicted values over the actual distribution (Figure 4b). This may also be due to an even less normal distribution of actual feed intake values in the FI-0 PM external validation dataset.

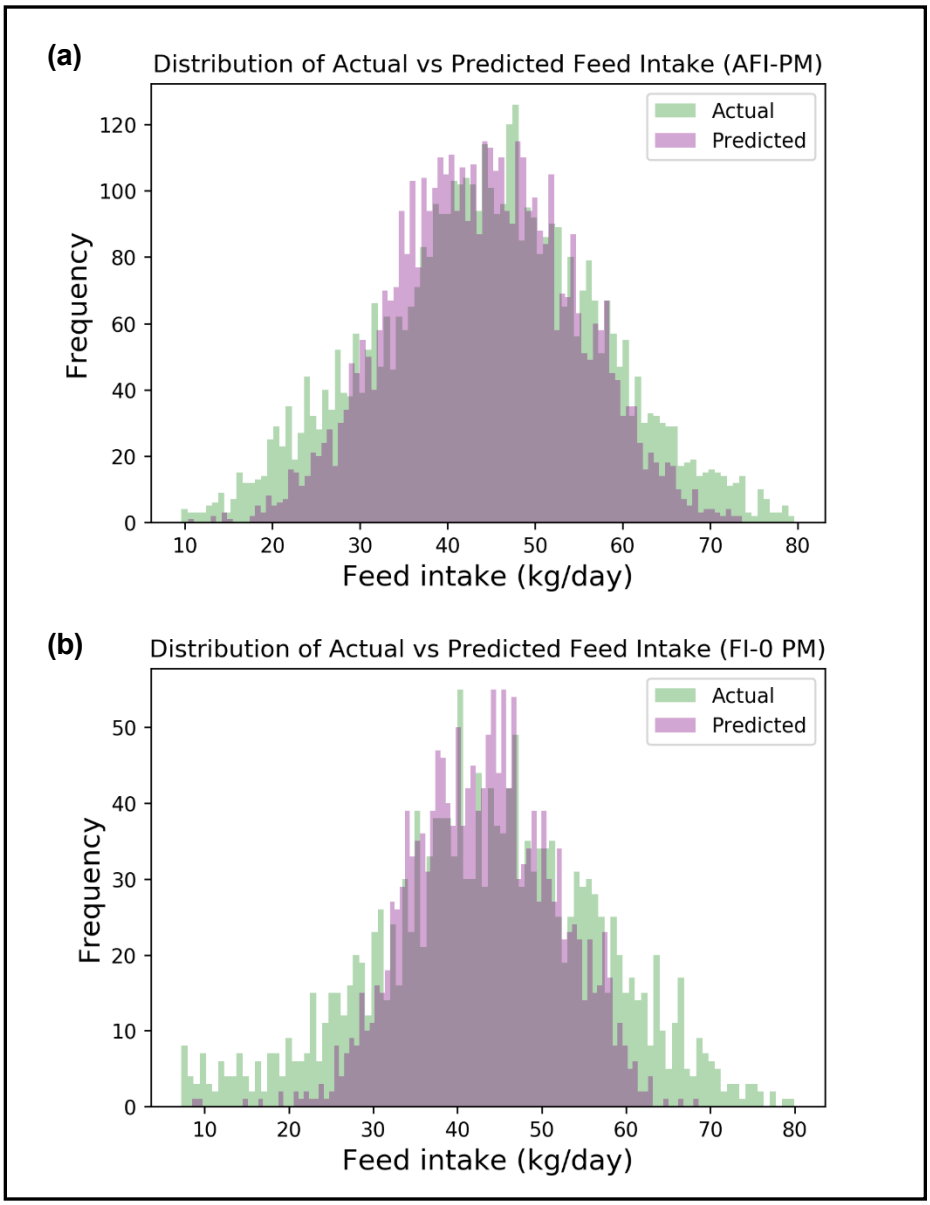

Figure 5: Distributions of actual and predicted feed intake for best (a) and worst (b) performing models

The optimal fitted model (Figure 6) demonstrates the strong performance of the AFI-PM model with a Pearson's correlation (R) between actual feed intake and predicted intake of 0.73. Given a large number of datapoints in the AFI-PM external validation dataset (n = 4,570, Figure 2) transparency was applied to the datapoints to present the volume of points overlapping. The model fits the bulk of datapoints accurately however there are a number of outliers where the actual feed intake is large, eg 78 kg/day and the model-predicted feed intake is much lower at 32 kg/day. This illustrates that although the model performs well overall, large errors can still be encountered for some records when feed intake is predicted based purely on MIR-spectral data.

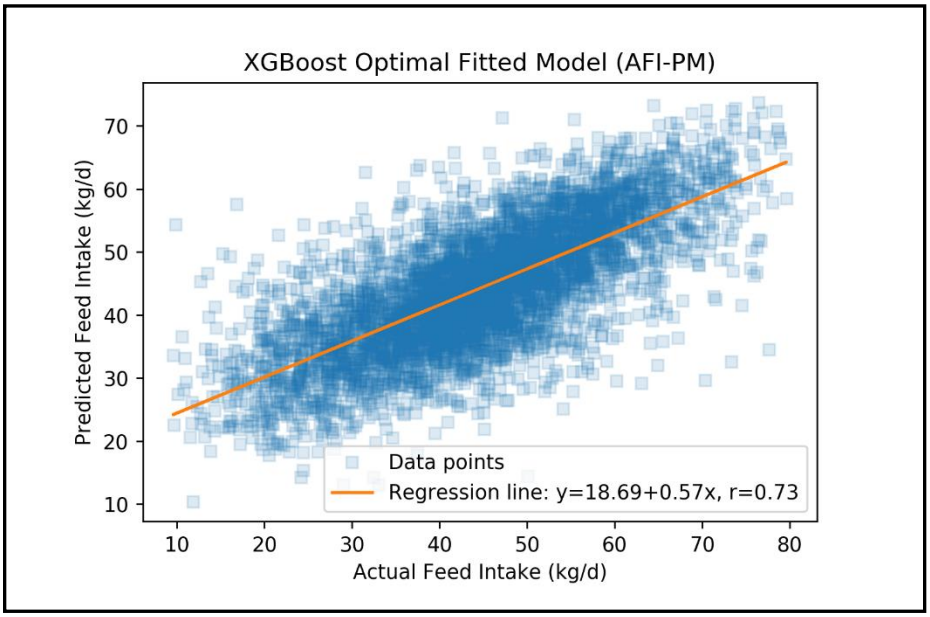

Figure 6: Scatterplot of actual vs predicted feed intake with fitted regression line

Application of Prediction Tool to Genetic Analysis

Estimates of additive genetic and permanent environmental variance, heritability, and repeatability of actual and predicted AFI based on the XGBoost generated predictions for the AFI-PM dataset are reported in Table 7. As illustrated previously by Figure 5, the distribution of predicted AFI values is much narrower than the distribution of actual AFI. This explains the disparity between the

phenotypic variances of actual AFI (117.343) and predicted AFI (57.229). The heritability of AFI in lactating dairy cows was 0.05 (0.02) in the MIR-predicted AFI phenotype and 0.03 (0.02) in the measured AFI phenotype. The heritability of AFI in the entire AFI dataset was 0.13 (0.02). This estimate is consistent with the most recent international estimates for feed intake in dairy cows (0.10-0.60). The estimates for the smaller AFI and MIR-predicted AFI are slightly outside this range. The bivariate repeatability model estimated phenotypic correlation 0.937 (0.009) and genetic correlation 0.618 (0.165) between AFI and MIR-predicted AFI.

Table 6. Genetic parameters of actual and MIR-predicted average feed intake (AFI)

Phenotype	Cows	n	σ^2_p	σ^2_a	σ^2_{pe}	h^2	rep
AFI	744	4571	117.343 (2.67)	3.564 (2.31)	12.918 (2.46)	0.030 (0.020)	0.140 (0.014)
Pred AFI	744	4571	57.229 (1.25)	2.877 (1.027)	1.404 (0.921)	0.051 (0.017)	0.075 (0.012)
AFI[2]	841	72960	131.643 (2.05)	17.221 (3.018)	12.052 (1.862)	0.131 (0.021)	0.222 (0.012)

[1] Standard deviation of each metric displayed in parentheses
[2] Genetic parameters for the full dataset of AFI phenotypes (both MIR-aligned and non-aligned records included)
[3] n = number of AFI records, σ^2_p = phenotypic variance; σ^2_a = additive genetic variance; σ^2_{pe} = permanent environmental variance; h^2 = heritability; rep = repeatability

DISCUSSION

Predictive performance of XGBoost versus PLS Regression

This study set out to determine the suitability of machine learning (ML) techniques to generate prediction models for feed intake in dairy cows using milk MIR spectral data. Comparison of the results of the XGBoost models offers insight into the capabilities of ML techniques over the standard PLS regression models which have been deployed in the past decade. The predictive performance of the models for the exact-day phenotypes was very similar for the PLS and XGBoost models with only small deviations in RMSE and R in the external validation set. However, for the AFI phenotype, the XGBoost models outperformed the PLS regression models with accuracies of 0.72 vs 0.61 in the AM milking dataset and 0.73 vs 0.65 in the PM milking. The disparity in performance between PLS and XGBoost may be greater in the AFI phenotype datasets due to the larger number of MIR-aligned samples in these datasets (Table 4). This offers the ML algorithm a greater number of learning rounds (epochs) in training before overfitting to the training data occurs.

Although the performance of both the PLS and XGBoost models for the exact-day phenotypes is similar, an interesting trend arises in that the accuracy of the algorithms increases as the time from the feed sample recording date increases and then decreases at the FI-2 PM dataset. This is evident as the FI-1 PM and FI-2 AM models are the best performing for the XGBoost and the PLS regression models followed by a slight decline in accuracy of the FI-2 PM dataset. This may suggest that the passage of food through the rumen and biological processes which take place, as a result, are most evident in milk samples approximately 36-48 hours after food has been ingested. Despite this, the results are only marginally different and given the small sample size of the external validation sets for the exact-day datasets further study would be required to determine any statistically significant result.

It is worth noting that the time of development, as well as the model runtime for the XGBoost models, was significantly less than the PLS models. Datasets require significant pre-processing for use in PLS models, e.g. Savizky-Golay smoothing. The raw untreated data can be passed to the XGBoost models, and the algorithm will be able to manage the nonlinearity of relationships between variables and highly correlated wavelengths. XGBoost models were trained in a matter of minutes on the GPU-optimized algorithm whereas the equivalent PLS models required hours of training. This makes XGBoost a very attractive predictive model to quickly evaluate the potential usability of a set of independent variables (e.g., wavelengths) on a given dependent variable (e.g., feed intake). Once the suitability is established, further tuning of hyperparameters such as number of training epochs, tree depth, number of trees can be carried out to improve model performance.

Comparison of Results to Previous Studies

The significant proportion of variable costs attributable to feed (50-60%) in dairy farms (Meyer et al., 2017) and the environmental impact due to correlations with greenhouse gas emissions makes feed intake a highly studied trait in dairy cows. The difficulty attaining accurate, routine measurements of feed intake in large numbers has led to many studies seeking to validate the accuracy of predictive models for feed intake in different countries and different breeds of dairy cows.

Table 7 presents a comparison of results from recent studies focused on the application of MIR spectroscopy to feed intake prediction in dairy cows. In each of the studies, dry matter intake (DMI) was used as the phenotype of interest. This measure of feed intake is highly correlated with total feed intake and therefore this paper focused on the use of total feed intake primarily due to time limitations and additionally because a predictive tool built on total feed intake would be more accessible and manageable for farmers. Each of the comparable studies applied some form of dimensionality reduction by removing wavelengths prior to training

their model. Dórea et al. (2018) eliminated wavelengths with less than 1% coefficient of variation among animals. Shetty et al. (2017) removed wavelength regions associated with the high levels of water absorption which may interfere with the results. This is a common step in training a PLS regression model to remove potential sources of noise in the set of independent variables. However, a recent study by Tiplady et al (2021) which conducted a genome-wide association study of individual wavenumbers in milk MIR spectra of dairy cattle demonstrated that many of the QTLs found were for wavenumbers in the mid-infrared range which were characterised by low signal-to-noise ratios. Normally, spectral data in these low-signal-to-noise areas are excluded from analyses however these findings suggest that the regions typically discarded potentially may be informative.

Each of the studies listed in Table 7 built several other PLS regression models with additional independent variables included which were aligned with the MIR-spectral and feed intake records. For example, in Wallén et al.'s paper (2018) the best performing model had an accuracy of 0.54 when fitted with both live body weight and daily milk yield as predictor variables. However, this approach increases the technology and investment required by farmers to implement such a predictive tool. One of the major advantages of milk MIR spectroscopy for dairy cattle is the repurposing of existing data collection. Therefore, models built using solely milk MIR spectra were evaluated in this study.

The accuracy of the models built using PLS regression range from 0.40-0.70. This range is similar to the accuracies found by the PLS regression models built in this study (Table 5) which range from 0.578-0.647. Comparison of the external validation results of the PLS regression and XGBoost models in this study (Table 5&6) suggested that XGBoost may be capable of building a stronger predictive model than the traditional PLS regression method. This is further reinforced by evaluating the result of the XGBoost model versus the best performing PLS models from comparable studies (Table 7). However given the presence of repeated feed intake observations per individual animal in our external validation dataset, as well

as the crossover between the animals (but importantly not records) in the training and testing phases, bias may have been introduced to the accuracy results. Therefore, further validation of the technique with a separate population of dairy cows would illustrate the true predictive value of our model. That said, the results do serve to further validate a growing body of work (Brand et al., 2021; Denholm et al., 2020; Dórea et al., 2018; Mota et al., 2021) investigating the use of machine learning techniques in agriculture and applied to phenotype prediction in dairy cattle using milk MIR spectral data.

Table 7. Comparison of recent milk MIR-spectral studies on feed intake prediction

Study	Trait	Cows	n	Dairy breed	Spectra pre-treatments	Validation procedure	Model	Accuracy (R)
(Dórea et al., 2018)	DMI	308[1]	1279	Holstein	36/361 WLs	10-fold X val	PLS	0.62/0.70
(Shetty et al., 2017)	DMI	140	1044	Holstein/Jersey	1D, 511 WLs	5-fold X val	PLS	0.55
(Wallén et al., 2018)	DMI	160	857	Norwegian Red	1D, variable WLs	LOO X val	PLS	0.4
This paper	AFI	744	41135	Holstein	1060 WLs	10-fold X val	XGBoost	0.73

[1] Animals in this study were exclusively mid-lactation (127 ± 27 DIM)

[2] n = number of samples used for calibration model

[3] DMI = Dry Matter Intake; AFI = Average (total) Feed Intake 1D = first derivate; 2D = second derivative; WLs = wavelengths; X val = cross-validation; PLS = Partial Least Squares Regression

Application of ML in Agriculture

This study aimed to further validate the use of machine learning techniques in the agriculture sector. With the rise of big data technologies and high-performance computers, machine learning has opened up new possibilities for data-intensive tasks in the multi-disciplinary “agritech“ industry. Farm management systems are turning into real-time artificial intelligence-enabled applications that give comprehensive recommendations and insights to farmers, providing decision support and action by using machine learning to sensor data. Predictive tools built using milk MIR spectral data is an efficient, cost-effective repurposing of large datasets that harness the power of machine learning to improve overall livestock management. In the future, the integration of automated data collection, data analysis, machine learning implementation and decision-making with knowledge-

based agriculture will serve to boost output levels and improve the quality of bio-products (Liakos et al., 2018).

The threshold of predictive accuracy required to make a predictive ML model effective in commercial herds varies depending on the phenotype in question. In a categorical variable, such as pregnancy status (Brand et al., 2021) or bovine tuberculosis (Denholm et al., 2020) where the model returns pregnant/not pregnant or infected/not infected, the certainty of the outcome is critical for the next steps and the actions which are required to be taken. In the previously referenced studies, the accuracies of these models were 95% and 97% respectively. For a continuous phenotype, such as feed intake, the required accuracy for a commercially viable predictive tool is less. For example, if a model allows farmers to rank their cows based on feed intake with greater accuracy 50% of the time, in the long run, improved breeding programmes could lead to an increase in annual profit of tens or hundreds of millions on a national scale. Therefore, the accuracy result of 0.73 (Table 6) in the XGBoost AFI-PM model may be viable to provide true economic value in the future. Further study into modelling the incremental impact of improved genetic evaluations as a result of more accurate feed intake phenotypes and breeding values would be highly insightful and demonstrate a concrete application of machine learning techniques applied in agriculture.

Choice of ML Model in this study

Many different types of machine learning models can be applied to different scenarios such as regression, clustering, Bayesian models, instance-based models, decision trees, artificial neural networks, support vector machines and ensemble learning. For example, K-Nearest Neighbours (KNN) is a commonly used classification instance based model which is used to classify data points or records based on the class of their nearest “neighbours” as calculated by some function (Punnoose & Ajit, 2016). XGBoost was chosen for this study based on its processing speed, relatively few pre-processing requirements and its predictive performance versus other models. XGBoost is an ensemble technique that scales

beyond billions of examples using far fewer memory or computational resources than other ML algorithms. When tested versus other popular predictive models, XGBoost is 10 times quicker and scales to billions of records in distributed or memory-limited environments (Chakraborty & Elzarka, 2019). XGBoost's scalability is due to a number of key systems and algorithmic improvements. These include a unique tree learning algorithm (Eq. 1) and a weighted quantile procedure that enables handling instance weights in approximate tree learning (Eq. 2). Parallel and distributed computing make learning more efficient, allowing for more rapid model exploration (Chen & Guestrin, 2016). The impact of this technique has been acknowledged in many instances and has become commonplace in machine learning and data mining challenges. In 2015, XGBoost was utilised in 17 of the 29 challenge-winning solutions published on Kaggle. Deep neural networks, the second most common approach, were employed in just 11 of the answers. Based on the superior performance and particularly given the time limitations of this paper, XGBoost became the ML model of choice. Previous studies using machine learning to analyse milk MIR spectra implemented a form of neural network known as convolutional neural networks (CNNs) (Brand et al., 2021; Denholm et al., 2020; Smith et al., 2019). This method requires the milk MIR spectra to be converted into 53x20 pixel images as inputs to the system and more rigorous model parameter tuning. Although studies comparing the methods show that XGBoost outperforms CNNs, the application of alternative ML algorithms such as CNNs may be an area of further investigation in the pursuit of improved predictive accuracy of feed intake from milk MIR spectral data.

Impact of Genomics and Phenomics on Trait Prediction

The progression of dairy cattle breeding goals towards being more holistic and including more functional traits (Miglior et al., 2005) has invited discussion on the inclusion and possible weighting of a variety of traits not explicitly included in breeding goals such as feed intake, output quality and the environmental footprint (Berry, 2015). Genomic selection is the process of establishing prediction

equations from the combination of phenotypic information and SNP genetic markers from a group of animals termed the reference population (Meuwissen et al., 2001). This enables the evaluation of genotyped animals without the need for phenotypic data. The use of genomic selection and a reference population offers the opportunity to include traits that are expensive or difficult to record, such as feed intake, as a breeding goal even with a limited number of records. As the cost of genotyping using high-density SNP chips decreases, sequence data may be worth considering as an alternative to the existing imputation-based approach used for low-density SNP chips. Sequence data are thought to facilitate cross-breed evaluation, which would allow for larger reference populations by including data from multiple breeds and lead to improved prediction accuracy (de Haas et al., 2017). Despite the relaxation of the requirement for very large populations of phenotyped animals, a reference population of several thousand animals is still required to properly assess the contribution of each genomic region to the expression of the phenotype in question (Hayes & Goddard, 2010). Even if high-quality genomic selection is achieved for a given trait, phenotyping strategies will remain necessary to achieve animal breeding goals and monitor the performance of animal breeding strategies (Berry, 2015).

Prior to the widespread use of genomic selection and the rapid decline in price per genotype, global breeding companies prioritised collecting enough genotypes on older bulls to produce an informative reference population. As the phenotypes were accessible through national or international genetic evaluation centers, such as Interbull, the focus was placed on genotyping a large enough population capable of providing accurate genomic breeding values (gEBVs) (Coffey, 2020). This is no longer the case as many young female cows and nearly all candidate bulls are routinely genotyped. The limiting factor for future evaluations of novel traits of economic importance, such as feed intake, is now the phenotypes. Phenomics is the science concerned with the acquisition of phenotypic data on a large scale (Houle et al., 2010). Phenomics is a natural complement to genomics

as the accessibility of accurate and efficient phenotyping tools are essential pre-requisites to the effective genomic selection of livestock animals designed to increase genetic gain of selected traits (Lillehammer et al., 2011). Although feed intake is economically important and exhibits genetic variation, the expense associated with the measurement of individual animal feed intake makes it unsuitable for routine measurement in commercial and smaller herds. However, MIR-predicted feed intake may offer a low-cost alternative should it demonstrate a high genetic correlation with the actual measurement (Berry, 2015). Therefore, the results of our genetic analysis which demonstrate a genetic correlation of 0.618 (0.165) are an indication that MIR-predicted feed intake may be a viable alternative. However, performing genetic studies on animal phenotypes requires accurate data collection (or prediction) on a sufficient number of animals in order to unravel the genetics of that characteristic (Banos et al., 2012). Therefore, the results of the genetic analysis of MIR-predicted feed intake in this study may be limited due to the number of cows in the external validation dataset of the model (n = 4571).

Difficult-to-measure phenotypes

The omission of feed intake, like many difficult-to-measure traits, from dairy cow breeding objectives, is predominantly due to a lack of information from which to make selection decisions. Several of these difficult-to-measure phenotypes are coincidentally traits that are currently of global importance in terms of climate change and are necessary today to allow farmers to make socially important selection decisions (i.e., cattle that produce food with less resource use and have a lower environmental impact) (Coffey, 2020). Therefore, significant focus is being placed on the creation of larger reference populations of these phenotypes. Limited phenotypic records are available in different countries from research or nucleus herds such as at Langhill. Banos et. al (2012) previously described an approach for the collation of research data from 4 European countries. De Haas et al. (2012) also demonstrated that the accuracy of genomic selection for dry matter intake in

dairy cows could be improved by collating data from research populations in Australia, the Netherlands and the United Kingdom. This solution may be viable however it is time-consuming, costly and requires a large degree of collaboration between research facilities.

The potential of milk MIR spectroscopy as a phenotyping technique has been thoroughly examined and proven successful for many traits previously discussed (Berry & Crowley, 2013; Denholm et al., 2020). A particular benefit of milk MIR spectroscopy as a phenotyping tool is that all milk samples are already subjected to MIR spectroscopy analysis to assess the fat, protein, and lactose concentrations (as well as other components) as part of routine evaluation of national herds. As a result, the marginal cost of obtaining an additionally predicted phenotype from the milk MIR spectra is negligible if an adequate prediction equation is developed. Milk MIR spectra have been shown to have a high predictive ability for characteristics that are difficult to measure, such as milk fatty acids (Maurice-Van Eijndhoven et al., 2013), milk coagulation properties (Cecchinato et al., 2009) and a moderate predictive ability for milk protein fractions (Bonfatti et al., 2011). MIR has also been investigated as a technique for predicting phenotypes relating to animal health (Grelet et al., 2016), fertility (Toledo-Alvarado et al., 2018), body energy status (McParland et al., 2012), feed efficiency (McParland et al., 2014), and methane emission (Wang & Bovenhuis, 2019). As indicated by this study and other recent studies, some machine learning techniques can provide similar or even better predictive performance than the standard approach applied to MIR spectral data (Brand et al., 2021; Denholm et al., 2020; Dórea et al., 2018; Mota et al., 2021). These studies illustrate the promise of MIR spectroscopy as a predictive method for difficult-to-measure phenotypes. Successful utilisation of any predicted phenotype in breeding programmes is dependent on the strength of phenotypic and genetic correlation between the predicted trait and the measured trait as well as the heritability of the MIR-predicted phenotype (Tiplady et al., 2021). Therefore,

future emphasis should be placed on extending studies on the predictive accuracy of MIR-predicted traits to also include genetic analysis of the predictions.

Genetic correlations

Genetic correlations are a manifestation of either the same genomic mutation influencing both traits (a pleiotropic effect) or distinct genomic mutations affecting both traits, but which are likely to be inherited together (i.e., linkage). Selection alters genetic correlations, and if selection has targeted improvement in both traits (e.g., milk production and reproduction success) the correlation is anticipated to become unfavourable (Falconer & Mackay, 1996), as currently observed in dairy cattle (Berry, Wall, et al., 2014). This is due to the fact that pleiotropic alleles (alleles that impact more than one trait) that favour both traits will become quickly fixed under selection, contributing little to the variation or covariance between the two traits. Alleles that impact both animal traits in opposite directions will remain at intermediate frequencies and so contribute more to the covariance between the traits; however, this also means that there will be minimal selection response. As rapid selection for greater milk production and reproductive performance is effective, this implies that there is still a significant amount of exploitable covariance, which might be attributable to pleiotropy or linkage (Berry, Wall, et al., 2014). The use of genomic data can assist in explaining the genomic architecture behind estimated genetic correlations; the component of the antagonistic correlation due to linkage can be resolved with the right genomic data. This might lead to a weakening of the genetic link between favourable performance qualities and poor reproductive outcomes. For characteristics like feed intake and milk production, where the objective is to modify the positively associated attributes (Berry & Crowley, 2013) in opposite directions, such an approach is particularly important.

Therefore, although including feed intake as a selection criterion in breeding schemes is both economically and environmentally desirable, the complexity of the correlated effects must also be considered. Rebuilding the present genetic

evaluation system and breeding goal to incorporate feed intake breeding values could create disruption within the industry's current genetic evaluation structures. One proposed solution to this problem is the use of residual feed intake (RFI) instead of total feed intake. RFI is defined as the difference between actual feed intake and predicted feed intake (Archer et al., 1999), accounting for known energy sinks such as milk production, fertility and adequate body reserves (Meyer et al., 2017). In this way feed efficiency may be incorporated into selection indices without impacting existing selection criteria. However, RFI is not universally accepted as the correct handling of feed intake records and further validation of its viability would be necessary.

One particularly topical genetic correlation of feed intake in livestock is methane emissions. Climate change is a major international concern, and the production of greenhouse gases (GHG) has long been recognised as a significant component (Gerber et al., 2010). According to estimations from the Intergovernmental Panel on Climate Change (IPCC), animal agriculture is responsible for 8.0 to 10.8% of global GHG emissions(O'Mara, 2011). However, this proportion can climb to 18% if a complete lifecycle analysis is conducted (i.e., accounting for the production of animal agriculture inputs as well as changes in land use such as deforestation) (Berry, 2015). O'Mara (2011) highlighted that cattle are the world's largest source of GHG emissions. Enteric methane (CH_4) is produced by ruminants during the microbial digestion of feed in the rumen and so unsurprisingly feed intake accounts for the most variation in daily CH_4 emissions (de Haas et al., 2017; Dehareng et al., 2012). Like feed intake, methane emissions are not included in national breeding goals largely due to the lack of easily accessible phenotypic data and no direct economic value in recording for producers. Studies have shown a positive phenotypic (0.72) and genetic (0.32) correlation between CH_4 emission and RFI in dairy cattle which implies that RFI selection may be a suitable method for decreasing CH_4 emissions (de Haas et al., 2011). CH_4 emissions have also been successfully predicted through the use of

milk MIR spectral data via PLS regression models with a cross-validation coefficient of determination equal to 0.69-0.77 (Dehareng et al., 2012). Accurate predictions of feed intake may offer the opportunity to indirectly select for CH_4 emissions through MIR-predicted feed intake values. In this way, selection for animals with both decreased environmental footprints and decreased feed intake may be achievable.

CONCLUSIONS

The use of XGBoost improved the accuracy of feed intake predictions in lactating dairy cows. The improvement of XGBoost's predictive ability over PLS indicates possible nonlinear relationships between feed intake and the milk MIR spectral data. Using all 1,060 untreated wavelength values appears to be beneficial in analyzing the relationship between the spectral data and feed intake. Future research should begin by first making use of all wave points before concluding that certain regions may not be useful. Moreover, genetic analyses of MIR-predicted feed intake and its genetic correlation with actual feed intake present the possibility of using MIR-predicted feed intake as a substitute for actual feed intake in breeding schemes.

Ultimately, everyone will benefit from healthier, more efficient cows. Selecting for cows that consume less and emit less harmful gases has the potential to simultaneously create more profitable farming and reduced global warming as a result of greenhouse gas emissions. As the climate crisis continues to accelerate, paired with the nutritional requirements of a rapidly expanding human population, interest in feed intake in livestock is likely to increase. The use of MIR spectroscopy and machine learning techniques such as XGBoost can be further validated by this study as valuable phenotyping tools and will play a critical role in developing tools that can be deployed in national and commercial herds.

APPENDICES

Appendix A – Data Pre-processing script

```
import numpy as np; print('numpy Version:', np.__version__)
import pandas as pd; print('pandas Version:', pd.__version__)
```

```
numpy Version: 1.16.4
pandas Version: 0.24.2
```

Reading Dataset

```
file = 'SQL_extracted
datasets/FINALlanghill_mir_extractready_revised_7pm_mir_aligned_20210711.
csv'
```

```
data_2 = pd.read_csv('SQL_extracted
datasets/Formatted_data/ZERODAYS_langhill_mir_extractready2_0am_mir_align
ed_20210727.csv', sep=',')
#data_3 =
#data_4 =
```

```
data_ = pd.read_csv(file, sep=',')
print(data_.shape) #Shape is 417657 rows and 18 columns

#data_.head
```

```
(74086, 15)
```

```
#Remove all rows for which there was no aligned spectral data
data_.dropna(subset=['spectra'], inplace=True)
print(data_.shape)
```

```
(17017, 15)
```

```
data_['weeklyave_FI'].describe()
```

```
count    74086.000000
mean        46.617685
std         15.041247
min         -0.136667
25%         37.467625
50%         46.710000
75%         55.957292
max        441.660000
Name: weeklyave_FI, dtype: float64
```

```
#Break spectra from long format into individual columns
spectra_ = data_['spectra'].str.split(',', n=1060, expand=True)
print(spectra_.shape)
```

```
(17017, 1060)
```

```
#Rename each spectral wavepoint and append to the data table
for i in range(1060):

      data_['wp%s' % (i+1)] = spectra_[i]

print(data_.shape)

#Drop the spectra column as it is now represented in the other columns
data_.drop(columns=['spectra'], inplace=True)

print(data_.shape)

#Data is now read in and correctly aligned
data_.head()
```

```
(17017, 1075)
(17017, 1074)
```

Data Formatting

```
#Reformatting all wavepoint columns
#Formatting as float32 - 32 is preferred for GPU

data_.iloc[:, 14:1704] = data_.iloc[:, 14:1074].apply(lambda x:
x.astype('float32')) #Takes a while to run, but seems to work
```

Investigating data

Outlier Removal

Causes: - Measurement or input error - Data corruption - True outlier observation

Methods - Standard Deviation (For normally distributed traits) - IQR

```
from statsmodels.graphics.gofplots import qqplot

qqplot(data_['FI'], line ='s')

from scipy.stats import shapiro
# normality test
```

```
stat, p = shapiro(data_['FI'])
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
    print('Sample looks Gaussian (fail to reject H0)')
else:
    print('Sample does not look Gaussian (reject H0)')

# D'Agostino and Pearson's Test
from scipy.stats import normaltest
#normality test
stat, p = normaltest(data_['FI'])
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
    print('Sample looks Gaussian (fail to reject H0)')
else:
    print('Sample does not look Gaussian (reject H0)')

# Anderson-Darling Test
from scipy.stats import anderson
# normality test
result = anderson(data_['FI'])
print('Statistic: %.3f' % result.statistic)
p = 0
for i in range(len(result.critical_values)):
    sl, cv = result.significance_level[i], result.critical_values[i]
    if result.statistic < result.critical_values[i]:
        print('%.3f: %.3f, data looks normal (fail to reject H0)' % (sl,
cv))
    else:
        print('%.3f: %.3f, data does not look normal (reject H0)' % (sl,
cv))

# IQR Method
# Identify outliers that are a factor of k of the IQR below the 25th perc
or above 75th perc
# A factor of k of 3 or more can be used to identify values that are
extreme outliers
# On a boxplot these limits are the whiskers at the end of the boxplot
lines. (Outliers are dots)

#Calculate IQR
q25, q75 = np.percentile(data_['weeklyave_FI'], 25) ,
np.percentile(data_['weeklyave_FI'], 75)
iqr = q75 - q25
print('Q25 = %.3f' % (q25))
```

```
print('Q75 = %.3f' % (q75))
print('IQR = %.3f' % (iqr))
```

```
Q25 = 37.468
Q75 = 55.957
IQR = 18.490
```

```
#Calculate the outlier cutoff
cut_off = iqr * 1.5
lower, upper = q25 - cut_off, q75 + cut_off

#Identify outliers
outliers = data_[(data_.weeklyave_FI > upper) | (data_.weeklyave_FI <
lower)]
#Remove outliers
data_cleaned = data_[(data_.weeklyave_FI <= upper) & (data_.weeklyave_FI
>= lower)]
print('Total observations: %d' % len(data_))

print('Identified outliers: %d' % len(outliers))
print('Non-outlier observations: %d' % len(data_cleaned))
```

```
Total observations: 74086
Identified outliers: 1126
Non-outlier observations: 72960
```

Export into folder for model use

```
fileName = file.split("/")[1]
fileName = fileName.split(".")[0]

data_cleaned.to_csv('SQL_extracted datasets/Formatted_data/' +fileName +
'GENETIC.csv', index = False)
```

Appendix B – XGBoost Model Development and Testing Script

Setup

```
from platform import python_version

print(python_version())
```

```
!nvidia-smi
```

```
Tue Sep 14 22:48:13 2021
+-----------------------------------------------------------------------------
-----+
| NVIDIA-SMI 418.126.02   Driver Version: 418.126.02   CUDA Version: 10.1
|
|-------------------------------+----------------------+----------------
-----+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr.
ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Comput
e M. |
|===============================+======================+================
=====|
|   0  Tesla V100-DGXS...  On   | 00000000:07:00.0 Off |
0 |
| N/A   37C    P0    50W / 300W |  22153MiB / 32478MiB |      0%      Def
ault |
+-------------------------------+----------------------+----------------
-----+
|   1  Tesla V100-DGXS...  On   | 00000000:08:00.0 Off |
0 |
| N/A   38C    P0    52W / 300W |  14880MiB / 32478MiB |      0%      Def
ault |
+-------------------------------+----------------------+----------------
-----+
|   2  Tesla V100-DGXS...  On   | 00000000:0E:00.0 Off |
0 |
| N/A   37C    P0    51W / 300W |  14880MiB / 32478MiB |      0%      Def
ault |
+-------------------------------+----------------------+----------------
-----+
|   3  Tesla V100-DGXS...  On   | 00000000:0F:00.0 Off |
0 |
| N/A   38C    P0    50W / 300W |  19912MiB / 32478MiB |      0%      Def
ault |
+-------------------------------+----------------------+----------------
-----+
```

```
+-----------------------------------------------------------------------------
-----+
| Processes:                                                       GPU Me
mory |
|  GPU       PID   Type   Process name                             Usage
|
|=============================================================================
=====|
+-----------------------------------------------------------------------------
-----+
```

```
import numpy as np; print('numpy Version:', np.__version__)
import pandas as pd; print('pandas Version:', pd.__version__)
import matplotlib.pyplot as plt
from matplotlib.ticker import StrMethodFormatter;
import xgboost as xgb; print('XGBoost Version:', xgb.__version__)
```

```
numpy Version: 1.16.4
pandas Version: 0.24.2
XGBoost Version: 0.90.rapidsdev1
```

Reading Dataset

```
data_cleaned = pd.read_csv('SQL_extracted datasets/Formatted_data/FINALla
nghill_mir_extractready_revised_7pm_mir_aligned_20210711.csv', sep=',')
print(data_cleaned.shape) #Shape is 417657 rows and 18 columns

#data_.head
```

```
(45705, 1074)
```

Split Data - Basic split

Training - 90% Testing - 10%

```
#Split data into 3 - train validation and test
#Once model trained, test the model using best interation of predictions
#For comparison of predicted vs observed
#Plots of accuracy

#Alternative split method which will randomly split the data
from sklearn.model_selection import train_test_split

test_size = 0.1
X = data_cleaned.iloc[:, 14:1074]
y = data_cleaned['weeklyave_FI']

X_test, X_train, y_test, y_train =train_test_split(X,y, train_size=0.1, r
andom_state = 11)
```

```
#_train, X_valid, y_train, y_valid = train_test_split(X_rem,y_rem, train_
size=0.8)

# check dimensions
print('X_train: ', X_train.shape,  'y_train: ', y_train.shape)
#print('X_validation', X_valid.shape, 'y_validation: ', y_valid.shape)
print('X_test', X_test.shape, 'y_test: ', y_test.shape)


# check the proportions
total = X_train.shape[0]  + X_test.shape[0]
print('X_train proportion:', X_train.shape[0] / total)
#print('X_validation proportion:', X_valid.shape[0] / total)
print('X_test proportion:', X_test.shape[0] / total)
```

```
X_train:  (41135, 1060) y_train:  (41135,)
X_test (4570, 1060) y_test:  (4570,)
X_train proportion: 0.9000109397221311
X_test proportion: 0.09998906027786894
```

Model running

```
dtrain = xgb.DMatrix(X_train, label=y_train)
#dvalidation = xgb.DMatrix(X_valid, label=y_valid)
dtest = xgb.DMatrix(X_test)
```

```
/opt/conda/envs/rapids/lib/python3.6/site-packages/xgboost/core.py:604: F
utureWarning: Series.base is deprecated and will be removed in a future v
ersion
  if getattr(data, 'base', None) is not None and \
```

Must set three types of parameters: general parameters, booster parameters and task parameters. - General parameters relate to which booster we are using to do boosting, commonly tree or linear model. FOR REGRESSION OUTPUT VARS CAN USE EITHER TREE OR LINEAR! - Booster parameters depend on which booster you have chosen - Learning task parameters decide on the learning scenario. For example, regression tasks may use different parameters with ranking tasks.

gbtree

```
# instantiate params
params = {}
```

```
# general params
general_params = {'silent': 1, 'booster' : 'gbtree'}
params.update(general_params)

# booster params
n_gpus = 3
booster_params = {}

if n_gpus != 0:
    booster_params['tree_method'] = 'gpu_hist'
    booster_params['n_gpus'] = n_gpus
params.update(booster_params)

# learning task params
#Check XGBoost manual for full parameter descriptions
learning_task_params = {'eval_metric': 'rmse', 'objective': 'reg:linear'}
params.update(learning_task_params)
print(params)
```

```
{'silent': 1, 'booster': 'gbtree', 'tree_method': 'gpu_hist', 'n_gpus': 3
, 'eval_metric': 'rmse', 'objective': 'reg:linear'}
```

```
from xgboost import cv

xgb_cv = cv(dtrain = dtrain, params = params, nfold = 10, early_stopping_
rounds = 10, metrics = "rmse")

xgb_cv

# model training settings
#Could set rounds to 10,000 with an early stopper. If no change in 25 loo
ps can stop and return to best iteration
#evallist = [(dvalidation, 'validation'), (dtrain, 'train')]
#watchlist = [(dvalidation,'valid'), (dtrain, 'train')]
watchlist = [ (dtrain, 'train')]

#evals_result = {'valid':{}, 'train':{}}
evals_result = {'train':{}}

num_round = 208

import os
os.environ['KMP_DUPLICATE_LIB_OK']='True'
from xgboost import XGBClassifier


#Train model
bst = xgb.train(params,
                dtrain,
```

```
                num_round,
                watchlist,
                evals_result = evals_result,
                early_stopping_rounds = 15,
                verbose_eval = 100
                )
```

```
[0] train-rmse:32.778
Will train until train-rmse hasn't improved in 15 rounds.
[100]   train-rmse:6.20902
[200]   train-rmse:4.76593
[207]   train-rmse:4.6883
```

```
preds2 = xgb_cv.predict(dtest)
```

Testing Model Performance

```
preds = bst.predict(dtest, ntree_limit =  bst.best_ntree_limit)
#preds
```

```
import scipy as scipy
slope, intercept, r, p, stderr = scipy.stats.linregress(y_test, preds)
```

```
line = f'Regression line: y={intercept:.2f}+{slope:.2f}x, r={r:.2f}'
line
```

```
'Regression line: y=18.78+0.58x, r=0.73'
```

```
fig, ax = plt.subplots()
ax.plot(y_test,preds, linewidth=0, marker='s', label='Data points', alpha
= 0.15)
ax.plot(y_test, intercept + slope * y_test, label=line)

ax.set_xlabel('Actual')
ax.set_ylabel('Predicted')
ax.legend(facecolor='white')
plt.show()
```

```
#plt.figure(figsize=(8,6))
fig, ax = plt.subplots()


ax.hist(y_test, bins=100, alpha=0.3, label="Actual", color = "green")
ax.hist(preds, bins=100, alpha=0.35, label="Predicted", color = "purple")

ax.set_xlabel("Feed intake (kg/day)", size=14)
ax.set_ylabel("Frequency", size=14)
ax.legend(facecolor='white')

plt.title("Distribution of Actual vs Predicted Feed Intake (FI-0 PM)")
```

```
plt.savefig('XGBDist2.png', dpi=250)

#plt.savefig("overlapping_histograms_with_matplotlib_Python.png")

plt.figure(figsize=(8,6))
plt.hist(y_test, bins=100, alpha=0.3, label="Actual", color = "green")
plt.hist(preds, bins=100, alpha=0.35, label="Predicted", color = "purple"
)

plt.xlabel("Feed intake (kg/day)", size=14)
plt.ylabel("Frequency", size=14)
plt.title("Distribution of Actual vs Predicted ")
plt.legend(loc='upper right')
#plt.savefig("overlapping_histograms_with_matplotlib_Python.png")

y_test.shape

preds.shape

from scipy.stats import pearsonr
# calculate Pearson's correlation
corr, p = pearsonr(y_test, preds)
print('Pearsons correlation: %.3f' % corr)

from sklearn.metrics import mean_squared_error, r2_score
# calculate coeff of determination
r2 = r2_score(y_test, preds)
print('Coeff of determination: %.3f' % r2)

from sklearn.metrics import mean_squared_error
rmse = np.sqrt(mean_squared_error(y_test, preds))
print('RMSE of prediction: %.3f' % rmse)

from xgboost import cv


r,p = pearsonr(preds, y_test)

import scipy as sp
m_slope, c_intercept, r_value, p_value, std_err = sp.stats.linregress(y_t
est, preds)

m, c = np.polyfit(y_test, preds, 1
                  )

std_err

xgb.plot_importance("bst")

xgb.plot_importance("bst", importance_type = "cover")

xgb.plot_importance("bst", importance_type = "gain")
```

Data Export

- Create dataframe with just actual values, predicted values and EAR_TAG

```
export_data = {'Actual': y_test,
               'Preds': preds
               }

df = pd.DataFrame(export_data, columns = ['Actual', 'Preds'])
print(df)

df = pd.merge(df, data_cleaned.iloc[:,1:14], left_index=True, right_index
=True)

df.shape

(4570, 15)

df.to_csv('exported_dataframe_withpredsNEW.csv')
```

Appendix C – PLS Model Development and Testing Script

Setup

```
!nvidia-smi
```

```
#PLS imports
from sys import stdout
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from scipy.signal import savgol_filter

from sklearn.cross_decomposition import PLSRegression
from sklearn.model_selection import cross_val_predict
from sklearn.metrics import mean_squared_error, r2_score
```

```
import numpy as np; print('numpy Version:', np.__version__)
import pandas as pd; print('pandas Version:', pd.__version__)
import matplotlib.pyplot as plt
from matplotlib.ticker import StrMethodFormatter;
import xgboost as xgb; print('XGBoost Version:', xgb.__version__)
```

```
numpy Version: 1.16.4
pandas Version: 0.24.2
XGBoost Version: 0.90.rapidsdev1
```

Reading Dataset

```
data_cleaned = pd.read_csv('SQL_extracted datasets/Formatted_data/FINALla
nghill_mir_extractready_revised_7pm_mir_aligned_20210711.csv', sep=',')
print(data_cleaned.shape) #Shape is 417657 rows and 18 columns

#data_.head
```

```
(45705, 1074)
```

External Test set

```
X = data_cleaned.iloc[:, 14:1074]
y = data_cleaned['weeklyave_FI']
```

```
X2 = savgol_filter(X, 17, polyorder=2, deriv=2)
X1 = savgol_filter(X, 17, polyorder=2, deriv=1)
```

```
#Alternative split method which will randomly split the data
from sklearn.model_selection import train_test_split
```

```
test_size = 0.1


X_test, X_train, y_test, y_train =train_test_split(X1,y, train_size=0.1,
random_state = 11)
#_train, X_valid, y_train, y_valid = train_test_split(X_rem,y_rem, train_
size=0.8)

# check dimensions
print('X_train: ', X_train.shape,  'y_train: ', y_train.shape)
#print('X_validation', X_valid.shape, 'y_validation: ', y_valid.shape)
print('X_test', X_test.shape, 'y_test: ', y_test.shape)


# check the proportions
total = X_train.shape[0]  + X_test.shape[0]
print('X_train proportion:', X_train.shape[0] / total)
#print('X_validation proportion:', X_valid.shape[0] / total)
print('X_test proportion:', X_test.shape[0] / total)
```

```
X_train:  (41135, 1060) y_train:  (41135,)
X_test (4570, 1060) y_test:  (4570,)
X_train proportion: 0.9000109397221311
X_test proportion: 0.09998906027786894
```

PLS Prep

```
y = y_train
X = X_train
```

```
y.shape
```

```
X.shape
```

```
wl = np.arange(900, 5140, 4)
```

```
print(len(wl))
```

```
1060
```

```
fig, ax = plt.subplots()
ax.plot(wl, np.mean(X, axis = 0),   label = "Mean value", alpha = 1)
ax.plot(wl, np.percentile(X, 99, axis = 0),  label = "99th percentile", a
lpha = .35)
ax.plot(wl, np.percentile(X, 1, axis = 0), label = "1st percentile",  alp
ha = .35)
#ax.plot(y_test, intercept + slope * y_test, label=line)

plt.title('Untreated Spectrum')
```

```
ax.set_xlabel('Wavenumber (cm$^{-1}$)')
ax.set_ylabel('Absorbance')
ax.legend(facecolor='white')
plt.savefig('output3.png', dpi=250)
plt.show()
#fig.tight_layout()



fig, ax = plt.subplots()
ax.plot(wl, np.mean(X1, axis = 0), label = "Mean value",  alpha = 1)
ax.plot(wl, np.percentile(X1, 99, axis = 0), label = "99th percentile", a
lpha = .35)
ax.plot(wl, np.percentile(X1, 1, axis = 0), label = "1st percentile",  al
pha = .35)

#ax.plot(y_test, intercept + slope * y_test, label=line)

plt.title('First Derivative Spectrum')
ax.set_xlabel('Wavenumber (cm$^{-1}$)')
ax.set_ylabel('Absorbance')
ax.legend(facecolor='white')
fig.tight_layout()
plt.savefig('output6.png', dpi=250)

plt.show()

# Plot the data absorption at each datapoint - need to clarify the range
& unit of wavelengths


with plt.style.context('ggplot'):
    plt.plot(wl, X.T)
    plt.xlabel("Datapoint - some wavelength value")
    plt.ylabel("Absorbance")
```

https://www.kaggle.com/phamvanvung/partial-least-squares-regression-in-python

```
#X2 = savgol_filter(X, 17, polyorder=2, deriv=2)
X1 = savgol_filter(X, 17, polyorder=2, deriv=1)


X1.shape

plt.figure(figsize=(8, 4.5))
with plt.style.context('ggplot'):
    plt.plot(wl, X1.T)
```

```
    plt.xlabel("Wavelengths (nm)")
    plt.ylabel("D1 Absorbance")
    plt.show()

plt.figure(figsize=(8, 4.5))
with plt.style.context('ggplot'):
    plt.plot(wl, X2.T)
    plt.xlabel("Wavelengths (nm)")
    plt.ylabel("D2 Absorbance")
    plt.show()

def optimise_pls_cv(X1, y, n_comp, plot_components = True):
    # Define PLS object
    pls = PLSRegression(n_components=n_comp)

    # Cross-validation
    y_cv = cross_val_predict(pls, X1, y, cv=10)

    # Calculate scores
    r2 = r2_score(y, y_cv)
    mse = mean_squared_error(y, y_cv)
    rpd = y.std()/np.sqrt(mse)

    return (y_cv, r2, mse, rpd)

%%time

r2s = []
mses = []
rpds = []
xticks = np.arange(1, 21)
for n_comp in xticks:
    y_cv, r2, mse, rpd = optimise_pls_cv(X, y, n_comp)
    r2s.append(r2)
    mses.append(mse)
    rpds.append(rpd)

CPU times: user 3h 40min 32s, sys: 3h 28min 54s, total: 7h 9min 27s
Wall time: 16min 4s



 # Calculate and print the position of minimum in MSE
    msemin = np.argmin(mses)
    print("Suggested number of components: ", msemin+1)
    stdout.write("\n")

Suggested number of components:  20

import matplotlib.ticker as mticker
```

```
# Plot the mses
def plot_metrics(vals, ylabel, objective):
    #with plt.style.context('ggplot'):
    plt.plot(range(1, 21), np.array(vals), '-v')
    if objective=='min':
        idx = np.argmin(vals)
    else:
        idx = np.argmax(vals)
    plt.plot(xticks[idx], np.array(vals)[idx], '*', ms=10, color = "red")

    plt.ticklabel_format(style='plain', axis='x', useOffset=False)
    #plt.plot(range(1, 20), logger.acc)
    plt.gca().xaxis.set_major_locator(mticker.MultipleLocator(1))
    plt.xlabel('Number of PLS components')
    #plt.xticks(np.arange(1, 21,1.0))
    #plt.xticks(range(1,20))
    plt.ylabel(ylabel)
    plt.title('Incremental value of each PLS component')

    plt.show()

fig, ax = plt.subplots()
ax.plot(wl, np.mean(X1, axis = 0), label = "Mean value",  alpha = 1)
ax.plot(wl, np.percentile(X1, 99, axis = 0), label = "99th percentile", a
lpha = .35)
ax.plot(wl, np.percentile(X1, 1, axis = 0), label = "1st percentile",  al
pha = .35)

#ax.plot(y_test, intercept + slope * y_test, label=line)

plt.title('First Derivative Spectrum')
ax.set_xlabel('Wavenumber (cm$^{-1}$)')
ax.set_ylabel('Absorbance')
ax.legend(facecolor='white')


fig.tight_layout()
plt.savefig('output6.png', dpi=250)

plt.show()

fig, (ax1, ax2) = plt.subplots(1, 2)
fig.suptitle('Incremental value of each PLS component')
objective = 'min'
ylabel = 'RMSE'
vals = np.sqrt(mses)

ax1.plot(range(1, 21), np.array(vals), '-v')
```

```
if objective=='min':
    idx = np.argmin(vals)
else:
    idx = np.argmax(vals)
ax1.plot(xticks[idx], np.array(vals)[idx], '*', ms=10, color = "red")

ax1.ticklabel_format(style='plain', axis='x', useOffset=False)
#plt.plot(range(1, 20), logger.acc)
#plt.gca().xaxis.set_major_locator(mticker.MultipleLocator(1))
ax1.set_xlabel('Number of PLS components')
#plt.xticks(np.arange(1, 21,1.0))
#plt.xticks(range(1,20))
ax1.set_ylabel(ylabel)


objective = 'max'
ylabel = 'R$^{2}$'
vals =r2s
ax2.plot(range(1, 21), np.array(vals), '-v')
if objective=='min':
    idx = np.argmin(vals)
else:
    idx = np.argmax(vals)
ax2.plot(xticks[idx], np.array(vals)[idx], '*', ms=10, color = "red")

ax2.ticklabel_format(style='plain', axis='x', useOffset=False)
#plt.plot(range(1, 20), logger.acc)
#ax2.gca().xaxis.set_major_locator(mticker.MultipleLocator(1))
ax2.set_xlabel('Number of PLS components')
#plt.xticks(np.arange(1, 21,1.0))
#plt.xticks(range(1,20))
ax2.set_ylabel(ylabel)
#fig.tight_layout()
fig.tight_layout(rect=[0, 0.03, 1, 0.95])


plt.savefig('outputPLS.png', dpi=250)



plot_metrics(np.sqrt(mses), 'RMSE', 'min')

plot_metrics(r2s, 'R$^{2}$', 'max')

x

[0, 5, 9, 10, 15]
```

```
# Define PLS object with optimal number of components
pls_opt = PLSRegression(n_components=msemin+1)
# Fit to the entire dataset
pls_opt.fit(X, y)
y_c = pls_opt.predict(X)


np.std(y)
```

```
%%time

# Cross-validation
y_cv = cross_val_predict(pls_opt, X, y, cv=10)
# Calculate scores for calibration and cross-validation
score_c = r2_score(y, y_c)
score_cv = r2_score(y, y_cv)
# Calculate mean squared error for calibration and cross validation
mse_c = mean_squared_error(y, y_c)
rmse_c = np.sqrt(mse_c)
mse_cv = mean_squared_error(y, y_cv)
rmse_cv = np.sqrt(mse_cv)
print('R2 calib: %5.3f'  % score_c)
print('R2 CV: %5.3f'  % score_cv)
print('R calib: %5.3f'  % np.sqrt(score_c)      )
print('R CV: %5.3f'  % np.sqrt(score_cv)  )
print('RMSE calib: %5.3f' % rmse_c)
print('RMSE CV: %5.3f' % rmse_cv)

# Plot regression and figures of merit
rangey = max(y) - min(y)
rangex = max(y_c) - min(y_c)
# Fit a line to the CV vs response
z = np.polyfit(y, y_c, 1)
with plt.style.context(('ggplot')):
    fig, ax = plt.subplots(figsize=(9, 5))
    ax.scatter(y_c, y, c='red', edgecolors='k')
    #Plot the best fit line
    ax.plot(np.polyval(z,y), y, c='blue', linewidth=1)
    #Plot the ideal 1:1 line
    ax.plot(y, y, color='green', linewidth=1)
    plt.title('$R^{2}$ (CV): '+str(score_cv))
    plt.xlabel('Predicted $^{\circ}$Feed Intake')
    plt.ylabel('Measured $^{\circ}$Feed Intake')
    plt.show()

#Get std of RMSE
ysq = np.sqrt(y)
y_csq = np.sqrt(y_c[:,0])
e = (ysq-y_csq)
```

```
sd = np.std(e)

print("Std of RMSE calib: ", sd)

y_cvsq = np.sqrt(y_cv[:,0])
e = (ysq-y_cvsq)
sd = np.std(e)

print("Std of RMSE CV: ", sd)
```

```
y_cv, r2, mse, rpd = optimise_pls_cv(X1, y, 20)
rmse = np.sqrt(mse)
corr = np.sqrt(r2)

print('Corr: %0.4f, R2: %0.4f, RMSE: %0.4f, RPD: %0.4f' %(corr,r2, rmse,
rpd))

preds = pls_opt.predict(X_test)
preds = preds[:,0]

from scipy.stats import pearsonr
# calculate Pearson's correlation
corr, p = pearsonr(y_test, preds)
print('Pearsons correlation: %.3f' % corr)
```

```
Pearsons correlation: 0.647
```

```
from sklearn.metrics import mean_squared_error, r2_score
# calculate coeff of determination
r2 = r2_score(y_test, preds)
print('Coeff of determination: %.3f' % r2)
```

```
Coeff of determination: 0.418
```

```
from sklearn.metrics import mean_squared_error
rmse = np.sqrt(mean_squared_error(y_test, preds))
print('RMSE of prediction: %.3f' % rmse)
```

```
RMSE of prediction: 9.700
```

Appendix D – Genetic Analyses

---

title: "Analysis of MIR-predicted FI"

author: "Ross Finnegan"

date: "15 June 2021"

output:

  word_document: default

  html_document: default

```{r setup, include=FALSE}

knitr::opts_chunk$set(echo = TRUE)

```

R Markdown

```{r results='hide'}

library("asreml")

source("qgm_helpers_t2.R")

```

Reading phenotypic data and pedigree

```{r}

ped <- read.csv(file= "ped_gen0to6.csv", header= TRUE, na.strings=c("NA"))

head(ped)

tail(ped)
```

```

```{r}
predAFI <- read.csv(file="full_df_with_preds_XGBPM.csv",header=TRUE,na.strings=c("NA"))
head(predAFI)
tail(predAFI)
```

Preparing for the analysis

```{r}
predAFI$icowi <- predAFI$icow
predAFI$sample_date_formatted <- as.Date(predAFI$sample_date, format = "%d/%m/%Y %H:%M")
predAFI$YEAR <- format(predAFI$sample_date_formatted, format="%Y")
predAFI$MONTH <- format(predAFI$sample_date_formatted, format="%M")
```

Declaring factors

```{r}
predAFI$GENETIC_GROUP <- factor(predAFI$GENETIC_GROUP)
predAFI$FEED_TYPE <- factor(predAFI$FEED_TYPE)
predAFI$icow <- factor(predAFI$icow)
predAFI$icowi <- factor(predAFI$icowi)
predAFI$YEAR <- factor(predAFI$YEAR)
```
```

predAFI$MONTH <- factor(predAFI$MONTH)

```

```{r}

ainv <- ainverse(ped[,1:3])

```

Summary of files

```{r}

qgm_ped_summary(ped[,1:3])

```

```{r}

qgm_ped_summary(predAFI[,10:12])

```

Fitting the Individual (Animal) Model

```{r}

# Predicted AFI

predAFI2.asr <- asreml(fixed= Preds ~ 1 + GENETIC_GROUP + FEED_TYPE + LACT_NO +  YEAR:MONTH, random= ~ vm(icow,ainv) +idv(icowi) , residual= ~ idv(units), data=predAFI)

```

```{r}

# Actual AFI of predicted set

actual_p.asr <- asreml(fixed= Actual ~ 1 + GENETIC_GROUP + FEED_TYPE + LACT_NO + YEAR:MONTH, random= ~ vm(icow,ainv) +idv(icowi) , residual= ~ idv(units), data=predAFI)

```

```{r}

qgm_summary(predAFI2.asr)

```

Calculating variance components

```{r}

# Predicted AFI

vpredict(predAFI2.asr, vp ~ V1+V2+V3)

vpredict(predAFI2.asr, h ~ V2/(V1+V2+V3))

vpredict(predAFI2.asr, rep ~ (V1+ V2)/(V1+V2+V3) )

vpredict(predAFI2.asr, vpe ~ V1)

vpredict(predAFI2.asr, va ~ V2)

```

```{r}

qgm_summary(actual_p.asr)

```

```{r}

# Actual AFI of predicted set

vpredict(actual_p.asr, vp ~ V1+V2+V3)

```
vpredict(actual_p.asr, h ~ V2/(V1+V2+V3))

vpredict(actual_p.asr, rep ~ (V1+ V2)/(V1+V2+V3))

vpredict(actual_p.asr, vpe ~ V1)

vpredict(actual_p.asr, va ~ V2)
```

```{r}
# Actual AFI of full set

fullset <- read.csv(file="full3bin.csv",header=TRUE,na.strings=c("NA"), quote = "")
```

```{r}
# Preparing for the analysis

fullset$icowi <- fullset$icow

fullset$sample_date_formatted <- as.Date(fullset$sample_date, format = "%d/%m/%Y %H:%M")

fullset$YEAR <- format(fullset$sample_date_formatted, format="%Y")

fullset$MONTH <- format(fullset$sample_date_formatted, format="%M")
```

Declaring factors

```{r}
fullset$GENETIC_GROUP <- factor(fullset$GENETIC_GROUP)

fullset$FEED_TYPE <- factor(fullset$FEED_TYPE)

fullset$icow <- factor(fullset$icow)
```

```
fullset$icowi <- factor(fullset$icowi)

fullset$YEAR <- factor(fullset$YEAR)

fullset$MONTH <- factor(fullset$MONTH)
```

```{r}
actual_f.asr <- asreml(fixed= weeklyave_FI ~ 1 + GENETIC_GROUP + FEED_TYPE + LACT_NO + YEAR:MONTH, random= ~ vm(icow,ainv) +idv(icowi) , residual= ~ idv(units), data=fullset)
```

```{r}
# Actual AFI of predicted set

vpredict(actual_f.asr, vp ~ V1+V2+V3)

vpredict(actual_f.asr, h ~ V2/(V1+V2+V3))

vpredict(actual_f.asr, rep ~ (V1+ V2)/(V1+V2+V3) )

vpredict(actual_f.asr, vpe ~ V1)

vpredict(actual_f.asr, va ~ V2)
```

```{r}
# Actual AFI of full set

fullset2 <- read.csv(file="full3bin2.csv",header=TRUE,na.strings=c("NA"), quote = "")
```

```{r}

# Preparing for the analysis

fullset2$icowi <- fullset2$icow

fullset2$sample_date_formatted <- as.Date(fullset2$sample_date, format = "%d/%m/%Y %H:%M")

fullset2$YEAR <- format(fullset2$sample_date_formatted, format="%Y")

fullset2$MONTH <- format(fullset2$sample_date_formatted, format="%M")

```

Declaring factors

```{r}

fullset2$GENETIC_GROUP <- factor(fullset2$GENETIC_GROUP)

fullset2$FEED_TYPE <- factor(fullset2$FEED_TYPE)

fullset2$icow <- factor(fullset2$icow)

fullset2$icowi <- factor(fullset2$icowi)

fullset2$YEAR <- factor(fullset2$YEAR)

fullset2$MONTH <- factor(fullset2$MONTH)

```

```{r}

qgm_ped_summary(fullset2[,8:10])

```

```{r}

actual_f2.asr <- asreml(fixed= weeklyave_FI ~ 1 + GENETIC_GROUP + FEED_TYPE + LACT_NO + YEAR:MONTH, random= ~ vm(icow,ainv) +idv(icowi) , residual= ~ idv(units), data=fullset2)

```

```{r}

# Actual AFI of predicted set

vpredict(actual_f2.asr, vp ~ V1+V2+V3)

vpredict(actual_f2.asr, h ~ V2/(V1+V2+V3))

vpredict(actual_f2.asr, rep ~ (V1+ V2)/(V1+V2+V3) )

```

Genetic correlation

```{r}

library("ggplot2")

library("GGally")

```

```{r}

ggpairs(data = predAFI[,c("Actual","Preds")], diag= list(continuous= "bar", binwidth= 30), lower= list(continuous= wrap(ggally_points, size= 0.01, alpha= 0.1)))

```

```{r}

corr.asr <- asreml(fixed = cbind(Preds, Actual) ~ trait + trait:GENETIC_GROUP + trait:FEED_TYPE + trait:LACT_NO +trait:YEAR:MONTH , random
```

=~us(trait):vm(icow,ainv) + id(icowi):us(trait), residual = ~id(units):us(trait), data = predAFI)

```

```{r}

qgm_summary(corr.asr)

```

```{r}

#Phenotypic correlation & genotypic correlation

vpredict(corr.asr, vp11 ~ (V1+V4+V8) ) # var(1)

vpredict(corr.asr, vp12 ~ (V2+V5+V9) ) # cov(1,2)

vpredict(corr.asr, vp22 ~ (V3+V6+V10)) # var(2)

vpredict(corr.asr, rp12 ~ (V2+V5+V9)/sqrt((V1+V4+V8)*(V2+V5+V9))) # cor(1,2)

vpredict(corr.asr, gencor ~ V2/sqrt(V1*V3))

```
```

REFERENCES

Alexandratos, N. (2012). *World Agriculture towards 2030/2050: The 2012 revision*. 154.

Alpaydin, E. (2020). Introduction to Machine Learning. In *Introduction to Machine Learning* (pp. 1–20). MIT Press. https://books.google.co.uk/books?hl=en&lr=&id=tZnSDwAAQBAJ&oi=fnd&pg=PR7&dq=Vieira,+S.%3B+Lopez+Pinaya,+W.H.%3B+Mechelli,+A.+Introduction+to+Machine+Learning%3B+Mechelli,+A.,+Vieira,+S.B.T.-M.L.,+Eds.%3B+Academic+Press:+Cambridge,+MA,+USA,+2020%3B+Chapter+1%3B+pp.+1%E2%80%9320.+ISBN+978-0-12-815739-8.&ots=F3TY8-9tvl&sig=bUxJLOigPw8j-cyvYMKtMdkSmGg&redir_esc=y#v=onepage&q&f=false

Archer, J., Arthur, P., Richardson, E. C., & Herd, R. M. (1999). Potential for selection to improve efficiency of feed use in beef cattle: A review. *Australian Journal of Agricultural Research - AUST J AGR RES*, *50*. https://doi.org/10.1071/A98075

Banos, G., Coffey, M. P., Veerkamp, R. F., Berry, D. P., & Wall, E. (2012). Merging and characterising phenotypic data on conventional and rare traits from dairy cattle experimental resources in three countries. *Animal*, *6*(7), 1040–1048. https://doi.org/10.1017/S1751731111002655

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine Learning in Agriculture: A Comprehensive Updated Review. *Sensors*, *21*(11), 3758. https://doi.org/10.3390/s21113758

Berry, D. P. (2015). Breeding the dairy cow of the future: What do we need? *Animal Production Science*, *55*(7), 823–837. https://doi.org/10.1071/AN14835

Berry, D. P., Coffey, M. P., Pryce, J. E., de Haas, Y., Løvendahl, P., Krattenmacher, N., Crowley, J. J., Wang, Z., Spurlock, D., Weigel, K., Macdonald, K., & Veerkamp, R. F. (2014). International genetic evaluations for feed intake in dairy cattle through the collation of data from multiple sources. *Journal of Dairy Science*, *97*(6), 3894–3905. https://doi.org/10.3168/jds.2013-7548

Berry, D. P., & Crowley, J. J. (2013). CELL BIOLOGY SYMPOSIUM: Genetics of feed efficiency in dairy and beef cattle1. *Journal of Animal Science*, *91*(4), 1594–1613. https://doi.org/10.2527/jas.2012-5862

Berry, D. P., Horan, B., O'Donovan, M., Buckley, F., Kennedy, E., McEvoy, M., & Dillon, P. (2007). Genetics of Grass Dry Matter Intake, Energy Balance, and Digestibility in Grazing Irish Dairy Cows. *Journal of Dairy Science*, *90*(10), 4835–4845. https://doi.org/10.3168/jds.2007-0116

Berry, D. P., Wall, E., & Pryce, J. E. (2014). Genetics and genomics of reproductive performance in dairy and beef cattle. *Animal*, *8*(s1), 105–121. https://doi.org/10.1017/S1751731114000743

Bonfatti, V., Cecchinato, A., Gallo, L., Blasco, A., & Carnier, P. (2011). Genetic analysis of detailed milk protein composition and coagulation properties in

Simmental cattle. *Journal of Dairy Science*, *94*(10), 5183–5193. https://doi.org/10.3168/jds.2011-4297

Brand, W., Wells, A. T., Smith, S. L., Denholm, S. J., Wall, E., & Coffey, M. P. (2021). Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. *Journal of Dairy Science*, S0022030221000692. https://doi.org/10.3168/jds.2020-18367

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., & Thompson, R. (2018). *ASReml-R Reference Manual Version 4*. 188.

Cecchinato, A., De Marchi, M., Gallo, L., Bittante, G., & Carnier, P. (2009). Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. *Journal of Dairy Science*, *92*(10), 5304–5313. https://doi.org/10.3168/jds.2009-2246

Chakraborty, D., & Elzarka, H. (2019). Advanced machine learning techniques for building performance simulation: A comparative analysis. *Journal of Building Performance Simulation*, *12*(2), 193–207. https://doi.org/10.1080/19401493.2018.1498538

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Coffey, M. (2020). Dairy cows: In the age of the genotype, #phenotypeisking. *Animal Frontiers*, *10*(2), 19–22. https://doi.org/10.1093/af/vfaa004

de Haas, Y., Calus, M. P. L., Veerkamp, R. F., Wall, E., Coffey, M. P., Daetwyler, H. D., Hayes, B. J., & Pryce, J. E. (2012). Improved accuracy of genomic prediction for dry matter intake of dairy cattle from combined European and Australian data sets. *Journal of Dairy Science*, *95*(10), 6103–6112. https://doi.org/10.3168/jds.2011-5280

de Haas, Y., Pszczola, M., Soyeurt, H., Wall, E., & Lassen, J. (2017). Invited review: Phenotypes to genetically reduce greenhouse gas emissions in dairying. *Journal of Dairy Science*, *100*(2), 855–870. https://doi.org/10.3168/jds.2016-11246

de Haas, Y., Windig, J. J., Calus, M. P. L., Dijkstra, J., de Haan, M., Bannink, A., & Veerkamp, R. F. (2011). Genetic parameters for predicted methane production and potential for reducing enteric emissions through genomic selection. *Journal of Dairy Science*, *94*(12), 6122–6134. https://doi.org/10.3168/jds.2011-4439

De Marchi, M., Toffanin, V., Cassandro, M., & Penasa, M. (2014). Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. *Journal of Dairy Science*, *97*(3), 1171–1186. https://doi.org/10.3168/jds.2013-6799

Dehareng, F., Delfosse, C., Froidmont, E., Soyeurt, H., Martin, C., Gengler, N., Vanlierde, A., & Dardenne, P. (2012). Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. *Animal*, *6*(10), 1694–1701. https://doi.org/10.1017/S1751731112000456

Denholm, S. J., Brand, W., Mitchell, A. P., Wells, A. T., Krzyzelewski, T., Smith, S. L., Wall, E., & Coffey, M. P. (2020). Predicting bovine tuberculosis status of

dairy cows from mid-infrared spectral data of milk using deep learning. *Journal of Dairy Science*, *103*(10), 9355–9367. https://doi.org/10.3168/jds.2020-18328

Diouf, J. (2009). FAOs Director-General on How to Feed the World in 2050. *Popul. Dev. Rev 35 (2009)*.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Dórea, J. R. R., Rosa, G. J. M., Weld, K. A., & Armentano, L. E. (2018). Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. *Journal of Dairy Science*, *101*(7), 5878–5889. https://doi.org/10.3168/jds.2017-13997

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Benjamin-Cummings Pub Co.

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, *52*(2), 399–433. https://doi.org/10.1017/S0080456800012163

FOSS. (2016). *FTIR Analysis of Food and Agric. Products.* https://www.fossanalytics.com/en

Fox, D. G., Tedeschi, L. O., Tylutki, T. P., Russell, J. B., Van Amburgh, M. E., Chase, L. E., Pell, A. N., & Overton, T. R. (2004). The Cornell Net Carbohydrate and Protein System model for evaluating herd nutrition and nutrient excretion.

Animal Feed Science and Technology, *112*(1), 29–78. https://doi.org/10.1016/j.anifeedsci.2003.10.006

Friedrichs, P., Bastin, C., Dehareng, F., Wickham, B., & Massart, X. (2015). Final OptiMIR Scientific and Expert Meeting: From milk analysis to advisory tools. Palais des Congrès, Namur (Belgium), 16-17 April 2015. *BASE*. https://popups.uliege.be/1780-4507/index.php?id=11963

Gerber, P., Key, N., Portet, F., & Steinfeld, H. (2010). Policy options in addressing livestock's contribution to climate change. *Animal*, *4*(3), 393–406. https://doi.org/10.1017/S1751731110000133

Gianola, D., Okut, H., Weigel, K. A., & Rosa, G. J. (2011). Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows and wheat. *BMC Genetics*, *12*(1), 87. https://doi.org/10.1186/1471-2156-12-87

Gjedrem, T., Robinson, N., & Rye, M. (2012). The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. *Aquaculture*, *350–353*, 117–129. https://doi.org/10.1016/j.aquaculture.2012.04.008

Grelet, C., Bastin, C., Gelé, M., Davière, J.-B., Johan, M., Werner, A., Reding, R., Fernandez Pierna, J. A., Colinet, F. G., Dardenne, P., Gengler, N., Soyeurt, H., & Dehareng, F. (2016). Development of Fourier transform mid-infrared calibrations to predict acetone, β-hydroxybutyrate, and citrate contents in bovine milk through a European dairy network. *Journal of Dairy Science*, *99*(6), 4816–4825. https://doi.org/10.3168/jds.2015-10477

Grelet, C., Fernández Pierna, J. A., Dardenne, P., Baeten, V., & Dehareng, F. (2015). Standardization of milk mid-infrared spectra from a European dairy network. *Journal of Dairy Science*, *98*(4), 2150–2160. https://doi.org/10.3168/jds.2014-8764

Grelet, C., Pierna, J. A. F., Soyeurt, H., Dehareng, F., Gengler, N., & Dardenne, P. (2014). *Creation of universal MIR calibrations by standardization of milk spectra: Example of fatty acids*. 1.

Hayes, B., & Goddard, M. (2010). Genome-wide association and genomic selection in animal breedingThis article is one of a selection of papers from the conference “Exploiting Genome-wide Association in Oilseed Brassicas: A model for genetic improvement of major OECD crops for sustainable farming”. *Genome*, *53*(11), 876–883. https://doi.org/10.1139/G10-076

Houle, D., Govindaraju, D. R., & Omholt, S. (2010). Phenomics: The next challenge. *Nature Reviews Genetics*, *11*(12), 855–866. https://doi.org/10.1038/nrg2897

Howard, J. (2018). *Deep Learning: The tech that’s changing everything, except animal breeding and genetics* [Plenary address]. 11th World Congress on Genetics Applied to Livestock Production, Auckland, New Zealand. https://icarinterbullwcgalp.zerista.com/event/member/453201

Hurley, A. M., López-Villalobos, N., McParland, S., Lewis, E., Kennedy, E., O’Donovan, M., Burke, J. L., & Berry, D. P. (2017). Genetics of alternative definitions of feed efficiency in grazing lactating dairy cows. *Journal of Dairy Science*, *100*(7), 5501–5514. https://doi.org/10.3168/jds.2016-12314

Johnson, H. F. (2009). Food and food security—A global perspective. *Food 2009 Anniversary Conference–Keynote Speech (p. 14)*.

Lainé, A., Bastin, C., Grelet, C., Hammami, H., Colinet, F. G., Dale, L. M., Gillon, A., Vandenplas, J., Dehareng, F., & Gengler, N. (2017). Assessing the effect of pregnancy stage on milk composition of dairy cows using mid-infrared spectra. *Journal of Dairy Science*, *100*(4), 2863–2876. https://doi.org/10.3168/jds.2016-11736

Lainé, A., Mabrouk, H. B., Dale, L.-M., Bastin, C., & Gengler, N. (2013). *How to use mid-infrared spectral information from milk recording system to detect the pregnancy status of dairy cows*. 6.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors*, *18*(8), 2674. https://doi.org/10.3390/s18082674

Lillehammer, M., Meuwissen, T. H. E., & Sonesson, A. K. (2011). A comparison of dairy cattle breeding designs that use genomic selection. *Journal of Dairy Science*, *94*(1), 493–500. https://doi.org/10.3168/jds.2010-3518

Marsland, S. (2011). *Machine Learning: An Algorithmic Perspective*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420067194

Maurice-Van Eijndhoven, M. H. T., Bovenhuis, H., Soyeurt, H., & Calus, M. P. L. (2013). Differences in milk fat composition predicted by mid-infrared spectrometry among dairy cattle breeds in the Netherlands. *Journal of Dairy Science*, *96*(4), 2570–2582. https://doi.org/10.3168/jds.2012-5793

McParland, S., Banos, G., McCarthy, B., Lewis, E., Coffey, M. P., O'Neill, B., O'Donovan, M., Wall, E., & Berry, D. P. (2012). Validation of mid-infrared spectrometry in milk for predicting body energy status in Holstein-Friesian cows. *Journal of Dairy Science*, *95*(12), 7225–7235. https://doi.org/10.3168/jds.2012-5406

McParland, S., Banos, G., Wall, E., Coffey, M. P., Soyeurt, H., Veerkamp, R. F., & Berry, D. P. (2011). The use of mid-infrared spectrometry to predict body energy status of Holstein cows1. *Journal of Dairy Science*, *94*(7), 3651–3661. https://doi.org/10.3168/jds.2010-3965

McParland, S., Lewis, E., Kennedy, E., Moore, S. G., McCarthy, B., O'Donovan, M., Butler, S. T., Pryce, J. E., & Berry, D. P. (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. https://doi.org/10.3168/jds.2014-8214

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, *157*(4), 1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Meyer, S., Amer, P., Baes, C., Miglior, F., Richardson, C., Wall, E., & Coffey, M. (2017). Options for incorporating feed intake into national selection indexes. *Interbull Bulletin*, *51*, Article 51. https://journal.interbull.org/index.php/ib/article/view/1429

Miglior, F., Muir, B. L., & Van Doormaal, B. J. (2005). Selection Indices in Holstein Cattle of Various Countries. *Journal of Dairy Science*, *88*(3), 1255–1263. https://doi.org/10.3168/jds.S0022-0302(05)72792-2

Mota, L. F. M., Pegolo, S., Baba, T., Peñagaricano, F., Morota, G., Bittante, G., & Cecchinato, A. (2021). Evaluating the performance of machine learning methods and variable selection methods for predicting difficult-to-measure traits in Holstein dairy cattle using milk infrared spectral data. *Journal of Dairy Science*, *104*(7), 8107–8121. https://doi.org/10.3168/jds.2020-19861

O'Mara, F. P. (2011). The significance of livestock as a contributor to global greenhouse gas emissions today and in the near future. *Animal Feed Science and Technology*, *166–167*, 7–15. https://doi.org/10.1016/j.anifeedsci.2011.04.074

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, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Punnoose, R., & Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. *International Journal of Advanced Research in Artificial Intelligence*, *5*(9). https://doi.org/10.14569/IJARAI.2016.050904

Ritchie, H., & Roser, M. (2017). Meat and Dairy Production. *Our World in Data*. https://ourworldindata.org/meat-production

Roberts, D. J., & March, M. (2013). Feeding systems for dairy cows: Homegrown versus by-product feeds. In *Recent Advances in Animal Nutrition.* (pp. 61–70). PC Garnsworthy and J. Wiseman, ed. Context Products Ltd.

Roser, M., & Ritchie, H. (2019). Hunger and Undernourishment. *OurWorldInData.Org*. 'https://ourworldindata.org/hunger-and-undernourishment'

Shalloo, L., Dillon, P., Rath, M., & Wallace, M. (2004). Description and Validation of the Moorepark Dairy System Model. *Journal of Dairy Science*, *87*(6), 1945–1959. https://doi.org/10.3168/jds.S0022-0302(04)73353-6

Shetty, N., Løvendahl, P., Lund, M. S., & Buitenhuis, A. J. (2017). Prediction and validation of residual feed intake and dry matter intake in Danish lactating dairy cows using mid-infrared spectroscopy of milk. *Journal of Dairy Science*, *100*(1), 253–264. https://doi.org/10.3168/jds.2016-11609

Smith, S. L., Denholm, S. J., Coffey, M. P., & Wall, E. (2019). Energy profiling of dairy cows from routine milk mid-infrared analysis. *Journal of Dairy Science*, *102*(12), 11169–11179. https://doi.org/10.3168/jds.2018-16112

Soyeurt, H., Bastin, C., Colinet, F. G., Arnould, V. M.-R., Berry, D. P., Wall, E., Dehareng, F., Nguyen, H. N., Dardenne, P., Schefers, J., Vandenplas, J., Weigel, K., Coffey, M., Théron, L., Detilleux, J., Reding, E., Gengler, N., & McParland, S. (2012). Mid-infrared prediction of lactoferrin content in bovine milk: Potential indicator of mastitis. *Animal*, *6*(11), 1830–1838. https://doi.org/10.1017/S1751731112000791

Soyeurt, H., Dardenne, P., Dehareng, F., Lognay, G., Veselko, D., Marlier, M., Bertozzi, C., Mayeres, P., & Gengler, N. (2006). Estimating Fatty Acid Content in Cow Milk Using Mid-Infrared Spectrometry. *Journal of Dairy Science*, *89*(9), 3690–3695. https://doi.org/10.3168/jds.S0022-0302(06)72409-2

Soyeurt, H., Dehareng, F., Gengler, N., McParland, S., Wall, E., Berry, D. P., Coffey, M., & 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. https://doi.org/10.3168/jds.2010-3408

Subcommittee on Dairy Cattle Nutrition. (2001). Nutrient requirements of dairy cattle. *National Academies Press*.

Svendsen, M., Skipenes, P., & Mao, I. L. (1993). Genetic parameters in the feed conversion complex of primiparous cows in the first two trimesters. *Journal of Animal Science*, *71*(7), 1721–1729. https://doi.org/10.2527/1993.7171721x

Tiplady, K. M., Lopdell, T. J., Reynolds, E., Sherlock, R. G., Keehan, M., Johnson, T. JJ., Pryce, J. E., Davis, S. R., Spelman, R. J., Harris, B. L., Garrick, D. J., & Littlejohn, M. D. (2021). Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle. *Genetics Selection Evolution*, *53*(1), 62. https://doi.org/10.1186/s12711-021-00648-9

Toledo-Alvarado, H., Vazquez, A. I., de los Campos, G., Tempelman, R. J., Bittante, G., & Cecchinato, A. (2018). Diagnosing pregnancy status using infrared spectra and milk composition in dairy cows. *Journal of Dairy Science*, *101*(3), 2496–2505. https://doi.org/10.3168/jds.2017-13647

Veerkamp, R. F., & Thompson, R. (1999). A Covariance Function for Feed Intake, Live weight, and Milk Yield Estimated Using a Random Regression Model. *Journal of Dairy Science*, *82*(7), 1565–1573. https://doi.org/10.3168/jds.S0022-0302(99)75384-1

Wallén, S. E., Prestløkken, E., Meuwissen, T. H. E., McParland, S., & Berry, D. P. (2018). Milk mid-infrared spectral data as a tool to predict feed intake in lactating Norwegian Red dairy cows. *Journal of Dairy Science*, *101*(7), 6232–6243. https://doi.org/10.3168/jds.2017-13874

Wang, Q., & Bovenhuis, H. (2019). Validation strategy can result in an overoptimistic view of the ability of milk infrared spectra to predict methane emission of dairy cattle. *Journal of Dairy Science*, *102*(7), 6288–6295. https://doi.org/10.3168/jds.2018-15684

.