Journal of Dairy Research

cambridge.org/dar

Using machine learning methods to predict dry matter intake from milk mid-infrared spectroscopy data on Swedish dairy cattle

Suraya Mohamad Salleh[1,2], Rebecca Danielsson[1] and Cecilia Kronqvist[1]

[1]Department of Animal Nutrition and Management, Swedish University of Agricultural Science, SE-750 07 Uppsala, Sweden and [2]Department of Animal Science, Faculty of Agriculture, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Research Article

Cite this article: Salleh SM, Danielsson R and Kronqvist C (2023). Using machine learning methods to predict dry matter intake from milk mid-infrared spectroscopy data on Swedish dairy cattle. *Journal of Dairy Research* **90**, 5–8. https://doi.org/10.1017/S0022029923000171

Received: 7 July 2022
Revised: 25 January 2023
Accepted: 25 January 2023
First published online: 1 March 2023

Keywords:
Milk MIRS; partial least-squares regression; random forest regression; support vector machine regression

Author for correspondence:
Suraya Mohamad Salleh:
Email: suraya.mohamad.salleh@slu.se, surayams@upm.edu.my

Abstract

In this research communication we compare three different approaches for developing dry matter intake (DMI) prediction models based on milk mid-infrared spectra (MIRS), using data collected from a research herd over five years. In dairy production, knowledge of individual DMI could be important and useful, but DMI can be difficult and expensive to measure on most commercial farms as cows are commonly group-fed. Instead, this parameter is often estimated based on the age, body weight, stage of lactation and body condition score of the cow. Recently, milk MIRS have also been used as a tool to estimate DMI. There are different methods available to create prediction models from large datasets. The main data used were total DMI calculated as a 3-d average, coupled with milk MIRS data available fortnightly. Data on milk yield and lactation stage parameters were also available for each animal. We compared the performance of three prediction approaches: partial least-squares regression, support vector machine regression and random forest regression. The full milk MIRS alone gave low to moderate prediction accuracy ($R^2 = 0.07–0.40$), regardless of prediction modelling approach. Adding more variables to the model improved R^2 and decreased the prediction error. Overall, partial least-squares regression proved to be the best method for predicting DMI from milk MIRS data, while MIRS data together with milk yield and concentrate DMI at 3–30 d in milk provided good prediction accuracy ($R^2 = 0.52–0.65$) regardless of the prediction tool used.

Dry matter intake (DMI) is an important performance indicator in livestock production and could be used to determine feed efficiency and to optimise feed utilisation. Knowledge of DMI is also crucial to avoid overfeeding, which could lead to undesirable consequences such as metabolic disease in dairy cows, feed inefficiency and associated effects on production economics, as well as problems with calving in overweight animals. However, it is often not possible to measure feed intake at individual level, since dairy cows are often group-fed and since the required equipment is usually not available on farms. In conventional dairy production, milk yield (MY), body weight (BW) and days in milk (DIM) are sometimes used to estimate DMI. However, many modern dairy farms are now participating in milk recording schemes that use mid-infrared spectroscopy (MIRS) analysis of milk samples to measure milk composition (fat, protein and lactose content). This is based on the fact that MIRS on milk samples provides information about the chemical bonds present in the milk, and thus indicates the types of molecules present in the samples.

As milk is individually collected, and composition in addition to yield is related to DMI, there may be possibilities to gain information about feed intake from individual cows through milk collection. The usage of MIRS in predicting DMI could be used as a strategy in selecting cows that have high efficiency in the utilisation of feed nutrients in relation to milk production, for breeding purpose. Therefore, it is important to find a robust strategy incorporating parameters and methods that can be used in predicting DMI. Equations using milk MIRS to predict feed intake-related parameters have been developed by several researchers using different methods (Shetty *et al.*, 2017; Wallén *et al.*, 2018; Lahart *et al.*, 2019; Grelet *et al.*, 2020). Partial least square (PLS) regression is commonly used to develop prediction models, as this method is suitable for multivariate data (Soyeurt *et al.*, 2006; Eskildsen *et al.*, 2014; Parrini *et al.*, 2019). Other studies have attempted to use various types of machine learning (ML) algorithms (Ghasemi and Tavakoli, 2013; Contla Hernández *et al.*, 2021; Meza Ramirez *et al.*, 2021). Among these, support vector machine (SVM) and random forest (RF) are widely used supervised learning models that were initially developed to learn the algorithms of non-linear relationship data in correspondence to certain discrete and continuous output. While the PLS method is most widely used in development of prediction models, it is of interest to compare the performance of different approaches and tools. Therefore, the present study compared

three different approaches (PLS, SVM and RF regression) in terms of their ability to predict DMI in Swedish dairy cattle using milk MIRS data.

Materials and methods

Data collection and pre-processing

Data on milk MIRS and on DMI (forage DMI + concentrate DMI) in the years 2017–2021 were collected from cows in the research herd, containing approximately 240 places for lactating cows, at the Swedish Livestock Research Centre, Lövsta, Sweden. All cows were either Swedish Red or Swedish Holstein breed. All cows were attached to an ID system (DelPro^{TM}) that logged feed intake and was also linked to the automatic milking system (DeLaval International AB, Tumba, Sweden), which recorded other milking-related parameters (date, time, time since the last milking, yield). The details of the feed intake and milk data records can be found in online Supplementary Text S1.

The DMI data were pre-processed before use in developing and validating models for predicting DMI. Total DMI exceeding 40 kg/d was filtered out from the dataset. The common management practice in the research barn is for cows in mid- to late lactation to be moved to a different part of the barn where forage DMI is not recorded. Therefore, DMI data were available only between 0 and 180 DIM. All cows were also kept partly on pasture during the summer (May–August) and total DMI for the animals was not measured within this period. Daily DMI was averaged over the 3 d immediately preceding the date of MIRS data.

Information about DIM was also included in the dataset, so the cows were categorised according to stage of lactation. Predictive model development was performed with all data (3–180 DIM), and with early (3–30 DIM) and mid- (30–180 DIM) lactation data separately.

Data analyses

Models were developed with three different tools, PLS regression, SVM regression and RF regression, as explained in online Supplementary Text S1. All analyses were performed using R software version 4.2.0 (R Core Team, 2022).

The prediction models were developed with data from 2017 to 2020, which comprised 1323 datalines for the full data. All models were validated using data from the most recent year (2021, 471 datalines). Coefficient of determination (R^2), RMSE of prediction (RMSEP) and mean absolute error (MAE) were used to evaluate and compare the performance of the prediction models.

Results

Descriptive statistics on the data used in the analysis are presented in online Supplementary Text S2, Table S1 and Figure S1.

Table 1 shows the performance of the prediction models in predicting DMI with different types of predictors for the PLS, SVM and RF regression, respectively, when using data from 2021 as the external validation dataset (in total corresponding to 26% of all available data).

In all of the three prediction approaches, it can be seen from Table 1, the best predictions were achieved using DIM 3–30 (early lactation) data. The best coefficient of determination (R^2) were observed in PLS regression approach (0.65) followed by RF regression (0.62) and SVM regression (0.55).

Generally, it was found that using the full milk MIRS data alone in the model predicting DMI provided low-to-moderately good prediction accuracy (R^2 = 0.07–0.40, MAE = 2.65–3.22). When including more variables together with the milk MIRS data, e.g. MY and concentrate DMI, the R^2 of the model improved and the prediction error (RMSE and MAE) were reduced.

Discussion

In the present study, we used the classical PLS method to predict DMI from milk MIRS and compared the prediction accuracy performance with that of two other non-linear ML methods (SVM and RF) that can also be used to predict regression data. The prediction accuracy (R^2) is an important measurement in evaluating as well as applying a prediction equation of a trait or parameter. For example, methods and equations to estimate BW in cattle based on prediction accuracy from body size measurements are well established and widely used by farmers and researchers (Heinrichs *et al.*, 1992; Bozkurt, 2006). Bozkurt (2006) did show R^2-values of 0.69 for prediction of BW from heart girth measurements, that predictive ability is close to the highest ones in the present study (0.65 when using the PLS method and with MY and concentrate intake as well as MIRS as predictors).

In the present study, among the three approaches, PLS regression provided the best prediction accuracy. The R^2 were also quite good for the RF and SVM regression approaches. This agrees with findings by Ghasemi and Tavakoli (2013), who concluded that the RF regression tool has potential and gives good prediction accuracy for non-linear multivariate data. However, their plot of RF predicted *vs.* measured values showed a similar pattern to that seen in the present study, where the range of predicted values tended to be quite narrow compared with those obtained using the PLS regression method. RF regression is much easier to perform than SVM regression, because the SVM approach requires more fine-tuning of the hyperparameters in the model to choose the best values for cost and gamma functions and get good prediction accuracy. Both these non-linear ML approaches have been used successfully for classification of types of data and analysis, e.g. predicting pregnancy status (Brand *et al.*, 2021) and metabolic status (Grelet *et al.*, 2019). Non-linear ML algorithms can be used for both classification and regression of predictive models, depending on the nature of the study. However, according to Meza Ramirez *et al.* (2021), SVM and RF are more commonly used for classification predictive models than for regression.

Although the SVM and RF regression models provided good prediction accuracy when the milk MIRS data were used together with additional variables, the more conventional PLS regression method still provided the best outcome. However, there are several other options or packages available for ML approaches that can be tested to better explore the possibility of using such approaches on multivariate data to predict DMI. SVM and RF regression were selected for comparison in this study, since they are both user-friendly tools that can easily be employed by users with different backgrounds.

With any approach, validation is important to ensure a reliable output. In this study, the data from the last year (2021) were used as the external test data to validate the models. This choice of test set may have resulted in lower prediction accuracy compared with a test set randomly selected from the full dataset, as there is a risk that time will cause a bias in the data. However, using the latest collected data reflected the situation where a predictive model is used on data generated after the model was built.

https://doi.org/10.1017/S0022029923000171 Published online by Cambridge University Press

Table 1. Prediction accuracy of PLS, SVM and RF regression analysis. Coefficient of determination (R^2) for validation/test dataset, RMSEP (kg/d) and MAE (kg/d) between predicted and actual observations of DMI (kg/d)

	R^2 (test)			RMSEP			MAE		
Predictor	PLS	SVM	RF	PLS	SVM	RF	PLS	SVM	RF
DIM 3–180									
MIRS	0.19	0.16	0.18	3.67	3.80	3.75	2.96	3.05	3.05
MY	0.31	0.25	0.23	5.03	4.90	5.18	4.19	4.02	4.16
Conc	0.46	0.38	0.38	3.73	3.71	4.06	3.00	2.90	3.19
MIRS + MY	0.43	0.34	0.33	3.19	3.66	3.82	2.48	2.91	3.00
MIRS + MY + Lact stage	0.44	0.33	0.36	3.07	3.64	3.75	2.39	2.89	2.93
MIRS + MY + Lact stage + Par	0.44	0.36	0.41	3.09	3.76	3.54	2.42	2.97	2.73
MIRS + Conc	0.44	0.42	0.53	3.86	3.18	2.81	2.99	2.50	2.16
MIRS + MY + Conc	0.62	0.52	0.62	2.71	2.88	2.78	2.13	2.26	2.17
DIM 3–30									
MIRS	0.40	0.30	0.28	3.19	3.45	3.68	2.65	2.67	2.92
MY	0.44	0.44	0.41	3.44	3.48	3.66	2.63	2.65	2.89
Conc	0.58	0.52	0.36	2.92	2.99	3.66	2.29	2.27	2.77
MIRS + MY	0.55	0.42	0.46	2.96	3.20	3.34	2.30	2.52	2.63
MIRS + Conc	0.58	0.46	0.55	2.88	3.07	2.93	2.21	2.41	2.30
MIRS + MY + Conc	0.65	0.55	0.62	2.65	2.85	2.69	2.14	2.25	2.09
DIM 30–180									
MIRS	0.20	0.08	0.07	3.49	3.87	3.93	2.87	3.13	3.22
MY	0.24	0.19	0.19	5.18	5.07	5.27	4.31	4.13	4.26
Conc	0.49	0.49	0.46	3.90	3.19	5.39	3.21	2.51	4.34
MIRS + MY	0.40	0.26	0.28	3.10	3.63	3.96	2.42	2.90	3.12
MIRS + Conc	0.43	0.35	0.56	3.99	3.22	2.67	3.13	2.55	2.11
MIRS + MY + Conc	0.60	0.44	0.54	2.62	3.00	3.16	2.05	2.36	2.46

PLS, Partial Least-Squares regression; SVM, Support Vector Machine regression; RF, Random Forest regression; MIRS, full milk mid-infrared spectra (935 wavenumbers); MY, average daily milk yield; DIM, days in milk; Conc, concentrate DMI; Lact stage, lactation stage; Par, parity; RMSEP, root mean square error of prediction; MAE, mean absolute error.

Many modern farms have good data/recordkeeping, to measure the performance of the farm and to ensure optimum profit in parallel with sustainable production (Soyeurt *et al.*, 2019). To our knowledge, the highest prediction accuracy to date has been obtained by Shetty *et al.* (2017) ($R^2 = 0.81$), who included MY and BW in a model containing the full MIRS data as a predictor for DMI. However, information on BW is not easily available on every farm. Therefore, different types of models with different parameters included would provide choices and enable users to predict DMI with good accuracy. Basic records on dairy cows, e.g. date of birth, parity number and milking records, are usually available. In development of prediction models for DMI, parameters such as these, which are also easy to retrieve, could be included to improve the prediction accuracy and reduce the prediction error. We included MY, lactation stage, parity and concentrate DMI in the models and found that the predictive ability of PLS, SVM and RF models was improved when more variables were included together with the milk MIRS data. Most Swedish dairy farms also have the possibility to adjust all or most of the concentrate allowance for their animals based on MY and stage of lactation, while the amount of concentrates consumed or delivered to each individual is often available for use as input data. Concentrate intake makes up a part of DMI and thus a relationship with total DMI can be expected. Therefore, it could be useful to include this information in models for predicting DMI.

There was an obvious pattern in the lactation curve, separating early and mid-lactation, so the data for these lactation stages were categorised and analysed separately. Using data for 3–30 DIM gave the highest prediction accuracy in all approaches, possibly due to the shorter range of days and strong linear relationship within this timeline (Rachah *et al.*, 2020). Overall, the first month of lactation is very crucial as it increases the probability of negative energy balance, since the animal would probably have low DMI compared with the amount of milk produced and body reserves will be used to compensate for this. Later in lactation, there are probably also other mechanisms behind the relationship between milk MIRS and DMI.

In general, MIRS in combination with other easily available data provides good prediction accuracy in predicting DMI. However, with current advances in precision livestock farming, it would be interesting to combine current developed sensor technology, for example 3-dimensional camera and triaxial accelerometer data to estimate feed intake in cows with the MIRS data to enhance the precision in predicting the feed intake. As MIRS

https://doi.org/10.1017/S0022029923000171 Published online by Cambridge University Press

from milk is a completely different measure to these measures, there is a good possibility that they may complement each other.

In conclusion, all tools tested were able to predict DMI with moderate performance. Overall, PLS regression analysis gave better results than the other machine learning tools, although the differences between the tools were small. The RF regression approach gave similar accuracy as the more complicated SVM regression. Early lactation DMI gave better prediction results compared with mid-lactation DMI. Inclusion of additional variables, especially MY and concentrate DMI, improved the predictions for both lactation stages (early, mid-) examined in the present study.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0022029923000171.

References

Bozkurt Y (2006) Prediction of body weight from body size measurements in Brown Swiss feedlot cattle fed under small-scale farming conditions. *Journal of Applied Animal Research* **29**, 29–32.

Brand W, Wells A, Smith S, Denholm S, Wall E and Coffey M (2021) Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. *Journal of Dairy Science* **104**, 4980–4990.

Contla Hernández B, Lopez-Villalobos N and Vignes M (2021) Identifying health status in grazing dairy cows from milk mid-infrared spectroscopy by using machine learning methods. *Animals* **11**, 2154.

Eskildsen CE, Rasmussen M, Engelsen S, Larsen L, Poulsen N and Skov T (2014) Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: understanding predictions of highly collinear reference variables. *Journal of Dairy Science* **97**, 7940–7951.

Ghasemi JB and Tavakoli H (2013) Application of random forest regression to spectral multivariate calibration. *Analytical Methods* **5**, 1863–1871.

Grelet C, Vanlierde A, Hostens M, Foldager L, Salavati M, Ingvartsen KL, Crowe M, Sorensen M, Froidmont E and Ferris C (2019) Potential of milk mid-IR spectra to predict metabolic status of cows through blood components and an innovative clustering approach. *Animal* **13**, 649–658.

Grelet C, Froidmont E, Foldager L, Salavati M, Hostens M, Ferris CP, Ingvartsen KL, Crowe MA, Sorensen MT and Pierna JF (2020) Potential of milk mid-infrared spectra to predict nitrogen use efficiency of individual dairy cows in early lactation. *Journal of Dairy Science* **103**, 4435–4445.

Heinrichs AJ, Rogers GW and Cooper JB (1992) Predicting body weight and wither height in Holstein heifers using body measurements. *Journal of Dairy Science* **75**, 3576–3581.

Lahart B, McParland S, Kennedy E, Boland T, Condon T, Williams M, Galvin N, McCarthy B and Buckley F (2019) Predicting the dry matter intake of grazing dairy cows using infrared reflectance spectroscopy analysis. *Journal of Dairy Science* **102**, 8907–8918.

Meza Ramirez CA, Greenop M, Ashton L and Rehman IU (2021) Applications of machine learning in spectroscopy. *Applied Spectroscopy Reviews* **56**, 733–763.

Parrini S, Acciaioli A, Franci O, Pugliese C and Bozzi R (2019) Near infrared spectroscopy technology for prediction of chemical composition of natural fresh pastures. *Journal of Applied Animal Research* **47**, 514–520.

Rachah A, Reksen O, Afseth NK, Tafintseva V, Ferneborg S, Martin AD, Kohler A and Prestløkken E (2020) Fourier transform infrared spectroscopy of milk samples as a tool to estimate energy balance, energy- and dry matter intake in lactating dairy cows. *Journal of Dairy Research* **87**, 436–443.

R Core Team (2022) R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/.

Shetty N, Løvendahl P, Lund M and Buitenhuis A (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**, 253–264.

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

Soyeurt H, Froidmont E, Dufrasne I, Hailemariam D, Wang Z, Bertozzi C, Colinet F, Dehareng F and Gengler N (2019) Contribution of milk mid-infrared spectrum to improve the accuracy of test-day body weight predicted from stage, lactation number, month of test and milk yield. *Livestock Science* **227**, 82–89.

Wallén S, Prestløkken E, Meuwissen T, McParland S and Berry D (2018) Milk mid-infrared spectral data as a tool to predict feed intake in lactating Norwegian Red dairy cows. *Journal of Dairy Science* **101**, 6232–6243.

https://doi.org/10.1017/S0022029923000171 Published online by Cambridge University Press