A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra.

Journal:	*Journal of Dairy Science*
Manuscript ID	JDS.2020-18870.R1
Article Type:	Research
Date Submitted by the Author:	26-Jul-2020
Complete List of Authors:	Soyeurt, Hélène; University of Liège, Gembloux Agro-Bio Tech, AGROBIOCHEM dept; Statistics, Informatics and Applied modelling Unit Grelet, Clément; Walloon Research Centre, Valorisation of Agricultural Products McParland, Sinéad; Teagasc, Animal and Grassland Research and Innovation Centre, Animal and Bioscience Research Department Coffey, Mike; SRUC, Animal & Veterinary Sciences Group Calmels, Marion; Seenovia Tedde, Anthony; University of Liège, Gembloux Agro-Bio Tech, AGROBIOCHEM dept; Statistics, Informatics and Applied modelling Unit Delhez, Pauline; Université de Liège Gembloux Agro-Bio Tech, Terra Teaching and Research Centre Dehareng, Frédéric; Walloon Agricultural Research Centre Gengler, Nicolas; ULg - Gembloux Agro-Bio Tech (GxABT), Animal Science Unit
Key Words:	milk, lactoferrin, mid-infrared, machine learning

A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra.

H. Soyeurt[1,*], C. Grelet[2], S. McParland[3], M. Calmels[4], M. Coffey[5], A. Tedde[1], P. Delhez[1,6], F. Dehareng[2], N. Gengler[1]

[1] TERRA research and teaching centre, Gembloux Agro-Bio Tech, University of Liège, Gembloux, Belgium

[2] Valorisation of agricultural products, Walloon Research Centre, Gembloux, Belgium

[3] Animal & Grassland Research and Innovation Centre, Teagasc, Moorepark, Fermoy, Co. Cork, Ireland

[4] Research and development, Seenovia, Saint-Berthevin, France

[5] Livestock Breeding, Animal and Veterinary Sciences, Scotland's Rural College, Midlothian, UK

[6] National fund for Scientific Research, Brussels, Belgium

Corresponding author: Prof. Hélène Soyeurt
Gembloux Agro-Bio Tech
Passage des Déportés, 2
5030 Gembloux
Belgium
hsoyeurt@uliege.be
+32/81/62.25.35

ABSTRACT

Lactoferrin (LF) is a glycoprotein naturally present in milk. Its content varies throughout the lactation but also with mastitis, therefore potentially being an additional indicator of udder health beyond somatic cell count. Therefore, there is an interest in quantifying this biomolecule routinely. First prediction equations proposed in the literature to predict the content in milk using milk mid-infrared (MIR) spectrometry were built using Partial Least Square regression (PLSR) due to the limited size of the dataset. Thanks to a large dataset, the current study aimed to test fourth different machine learning algorithms using a large dataset comprising 6,619 records collected across different herds, breeds and countries. The first algorithm was a PLSR as used in past investigations. The second and third algorithms used PLS factors combined with a linear and polynomial Support Vector regression (PLS + SVR). The fourth algorithm also used PLS factors but included in an artificial neural network having one hidden layer (PLS + ANN). The training and validation sets comprised 5,541 and 836 records, respectively. Even if the calibration prediction performances were the best for PLS + polynomial SVR, their validation prediction performances were the worse. The three other algorithms had similar validation performances. Indeed, the validation root mean squared error (RMSEv) ranged between 162.17 and 166.75 mg/L of milk. However, the lower standard deviation of cross-validation RMSE and the better normality of the residual distribution observed for PLS + ANN suggest that this modeling was the more suitable to predict the LF content in milk from milk MIR spectra (R^2v=0.60 and RMSEv=162.17 mg/L of milk). This PLS+ANN model was then applied to almost 6 million spectral records. The predicted

LF showed the expected relationships with milk yield, somatic cell score, somatic cell count and stage of lactation. The model tended to underestimate high LF values (higher than 600 mg/L of milk). However, if the prediction threshold was set to 500 mg/L, 82% of samples from the validation having a content of LF higher than 600 mg/L were detected. Future research should aim to increase the number of those extremely high LF records in the calibration set.

Keywords: milk, lactoferrin, mid infrared, machine learning

INTRODUCTION

Lactoferrin (**LF**), a 80-kDa glycoprotein naturally present in milk, is synthetized by the mammary gland epithelial cells (Molenaar et al., 1996) and has antibacterial, antiviral and antifungal activities potentially interesting to improve the cow's disease resistance. More details about the nutraceutical and pharmaceutical properties of milk bovine LF can be found in the review of Giansanti et al. (2016). These immune effects explain why the synthesis of LF increases in the presence of mastitis infection (Gaunt et al., 1980; Hagiwara et al., 2003) but the responses can be different following the incriminated pathogens (Kawai et al., 1999; Chaneton et al., 2008). Moreover, the content of LF in milk can also vary naturally depending on parity, age, lactation stage (Gaunt et al., 1980; Hagiwara et al., 2003), and breed (Król et al., 2010). Consequently, there is an interest to quantify the LF content in milk at the individual level for different issues related to animal welfare (i.e., early detection of infection) and human health due to the presence of this active bio-molecule in milk. Indeed, humans can take also profit to administer orally LF due to its anti-infective, anti-cancer and anti-inflammatory properties (Wakabayashi et al., 2006).

LF content can be quantified using an immunodiffusion method (Hagiwara et al., 2003) but its quantification is more often based on enzyme-linked immunosorbent assay (**ELISA**) (Chen and Mao, 2004; Chaneton et al., 2013). Unfortunately, those methods are too labour intensive for routine screening of the cow population at the individual scale as desired. Therefore, alternative methods offering a quantification at a low cost must be found. In 2007, a first study was published about the prediction of LF content in milk using milk mid-infrared (**MIR**) spectrometry (Soyeurt et al., 2007), a technology largely implemented in most milk laboratories. This first study measured the content of LF in a limited number of samples (i.e., 69 records). However, this first study and a follow-on one based on a bigger dataset (i.e., 2,499 records) (Soyeurt et al., 2012) validated the potential of MIR to provide a relevant indicator of LF. Indeed, the cross- and external validation coefficients of determination (**R^2**) and root mean square error (**RMSE**) obtained from this second study were of 0.71 and 0.60, and of 50.55 and 58.98 mg/L of milk, respectively. Recently, the European Milk Recording network (EMR, www.milkrecording.eu) has developed also its own equation from more than 2,000 records and offers the prediction service of this biomolecule to its members. By combining all of those datasets, new perspectives are opening to try different machine learning algorithms which could maybe improve the current accuracy of LF prediction. Indeed, all LF prediction equations were built using Partial Least Squares regressions (**PLSR**).

Since the nineties, several pieces of research were conducted in dairy science using artificial neural network (**ANN**), for instance, to analyze breeding dairy patterns (Finn et al., 1996), or to predict the incidence of clinical mastitis (Yang et al., 2000) or milk yield (Grzesiak et al., 2006). At our best knowledge, only 3 articles used mainly or partly the milk MIR spectra as predictors. Those studies concerned the prediction of conception success for a given insemination (Hempstalk et al., 2015), content of blood β-hydroxybutyrate (Pralle et al., 2018) and feed intake (Dórea et al., 2018). Several reasons could explain potentially the large use of PLSR to model milk MIR spectral data. Absorbance values of consecutive spectral data points are highly correlated. Therefore, collinearity problems are present if conventional simple linear regression is employed using all spectral data points as explanatory variables. Fortunately, some solutions exist to solve this problem. The first one consists in selecting a limited number of spectral points by trying to keep the most relevant information in the dataset, while limiting the correlations between them. Then, those low correlated spectral points can be included in a multivariate regression. The second possibility is to reduce the dimensionality of the spectral X matrix by using a principal component analysis (**PCA**). The PCA latent variable (**LV**) can be then used in a multivariate regression, commonly named principal component regression. However, the PCA methodology used to define those LVs takes into account only the spectral variability and not the variability of the trait to be predicted. This could lead to a lack of relevant spectral information to predict the trait of interest. The PLS method solves this problem by defining LVs considering simultaneously the variabilities of X and Y (Despagne et al., 2000). This explains why this methodology was and is still mainly used to develop milk MIR models to predict traits related to milk quality like fatty acids (Soyeurt et al., 2011), cheese making properties (De Marchi et al., 2009), body weight (Soyeurt et al., 2019), fertility (Delhez et al., 2020), traits related to animal welfare (Grelet et al., 2016) and environmental issues (Vanlierde et al., 2016). Unfortunately, PLS is able to consider weakly non-linear relationships by adding LVs, potentially leading to over-fit the developed prediction model (Thissen et al., 2004).

Other machine learning algorithms such as Support Vector Regression (**SVR**) and ANN have the ability to model the non-linear relations (Thissen et al., 2004). Compared to ANN, SVR can deal with a high number of input variables (Thissen et al. 2004). However, after a feature selection, ANN can be also efficient. Moreover, using the priors obtained from a past calibration dataset, the weights used in an ANN network can be updated using a new calibration dataset. This is particularly interesting when a large number of phenotypes useful to predict the trait of interest is available.

The computational methodology differs between SVR and ANN. SVR was created by Vapnik (Thissen et al., 2004) and consists in defining a classification boundary between records in order to minimize the distance between the records and the boundary by taking into account a certain limit of detection (called epsilon). More specifically, SVR is a method that selects a reduced number of samples, the support vectors, defining the best sparse deterministic regression relationship between the MIR data and the reference values. To ensure a global solution, a penalty (called C penalty) is used during the computation. Different kernels can be used to compute the boundary like linear, polynomial or radial kernel; each kernel has its own parameters to be optimized. More details about this method are given by Thissen et al. (2004). ANN, initially introduced by McCulloch and Pitts (1943), is the basis of deep learning which tries to mimic neuronal brain activity (i.e., the computer learns by experience). ANN is composed of different layers including units: one input layer, one output layer and a certain number of hidden layers. The number of hidden layers and its corresponding units must be defined by the user. The higher the number of hidden layers and their corresponding units, the higher the complexity of the model and therefore the higher the potential to over-fit the prediction model. The ANN algorithm aims to estimate the weights of each relation among units by using, for instance, the back propagation methodology. The cost function is related to the minimization of the residual error. The ANN model will provide a response of one or more variables given many explanatory variables (i.e., units) (Beck, 2018). In conclusion, SVR and ANN require the optimization of several parameters. To achieve this objective and to get a global solution, it is important to have a large dataset, explaining potentially why those methods are not often used in milk MIR spectrometry as the size of the dataset is often limited. However, compared to PLSR, SVR and ANN in themselves do not solve the issue of collinearity between spectral data points. So, there is an interest to combine the dimension reduction obtained by PLS with SVR or ANN algorithms. Therefore, the objective of the current study was to compare the accuracy of predictions of milk LF content from milk MIR spectra using 4 different machine learning algorithms: PLSR, linear and polynomial SVR coupled with PLS LVs and ANN coupled with PLS LVs.

MATERIALS AND METHODS

All analyses were performed using R software (version 4.0.1.; https://www.r-project.org/).

Data

The first dataset comprised 3,965 milk samples (50% of morning and 50% of evening milk) preserved with bronopol collected between April 2005 and April 2006 in Belgium, between April 2009 and August 2009 in Ireland, and during August 2009 in Scotland. Part of those samples was also used in a previous study (Soyeurt et al., 2012). The Belgian samples (N=549) were analyzed using one MilkoScan FT6000 spectrometer (Foss, Hillerod, Denmark) located in the milk laboratory "Comité du Lait" (Battice, Belgium). The Irish and Scottish samples (N=3,416) were also analyzed on the same brand of spectrometer at the Animal and Grassland Reasearch and Innovation Centre, Teagasc Moorepark (Fermoy, Co. Cork, Ireland). The spectral data of each sample contained 1,060 wavenumbers. The second dataset comprised 2,654 milk samples (50% of morning milk and 50% of evening milk) collected by the EMR in France (N=1,333), Luxembourg (N=246), England (N=500) and Germany (N=575) between June 2016 and January 2017. The samples in this second dataset were selected based on their LF content predicted using the equation developed by Soyeurt et al. (2012) to increase the variability over what was present in the samples of the first dataset. All samples were analyzed using either FT+ MilkoScan spectrometers (Foss, Hillerod, Denmark) or Bentley spectrometers (Chaska, MN, United States). The spectral data were then standardized based on the procedure explained by Grelet et al. (2017). All aspects related to this standardization was managed by the European Milk recording network and resulted in all samples having 1,060 harmonized wavenumbers and absorbance values available for analysis. A single milk sample per cow was selected from animals of different breeds across several herds and countries right across lactation.

Lactoferrin Quantification

LF concentration was quantified from the milk samples already analyzed by infrared spectroscopy using commercial ELISA kits: Bovine Lactoferrin ELISA Quantification kit from Bethyl Laboratories Inc. (Montgomery, TX, USA) for the first dataset and e_bLF_01 kit from IDBiotech (Issoire, France) for the second dataset. The Belgian samples were analyzed by Gembloux Agro-Bio Tech – University of Liège (Gembloux, Belgium). The ELISA analyses of the Irish and Scottish samples were conducted by Enfer Laboratories (Naas Co. Kildare, Ireland). The ELISA analysis of the second dataset was conducted at Seenovia (Saint-Berthevin, France). The samples were diluted 1:1000; 1:2000; 1:4000; 1:6000; 1:8000 or 1:10000 in sample buffer. The LF concentrations used for the calibration were the average of at least two ELISA measures taken on the same milk sample.

Spectral Pre-treatment

The spectral data coming from the first dataset were not standardized as this procedure did not exist when the samples were collected. Therefore, in order to correct for a potential baseline drift, the first derivative was applied to the recorded spectra for the dataset 1 and standardized spectra for the dataset 2 using the formula:

$$wavenumber'_i = wavenumber_i - wavenumber_{i+gap}$$

where $wavenumber'_i$ represents the first derivative value of the ith wavenumber, $wavenumber_i$ is the raw value observed for the ith wavenumber and the gap is the windows chosen for the derivation and was equal to 5. Then, the wavenumbers located in the most informative regions were selected. So, a total of 277 wavenumbers were kept for this study and were located from 950 to 1,580 cm^{-1}; from 1,720 to 1,770 cm^{-1}, from 1,780 to 1,850 cm^{-1}, and from 2,800 to 2,970 cm^{-1}.

The presence of potential spectral outliers was assessed by estimating the standardized Mahalanobis distance (also called GH distance) for all recorded spectra. In order to allow the inversion of the matrix needed to calculate the Mahalanobis distance due to the high collinearity existing between some spectral points, a PCA was performed using FactoMineR package (version 1.42; Lê et al., 2008), defining 22 uncorrelated principal components (**PC**) which explained 99.04% of the spectral variability. The formula used to calculate the GH distance was:

$$GH = \sqrt{(x - \mu)^T S^{-1} (x - \mu)} / nPC$$

where, x is the PC scores of a specific spectrum; μ is the mean of PC scores estimated from the entire dataset; S corresponds to the (co-)variance matrix between PC scores estimated from the entire dataset; nPC is the number of PC used in the calculation (i.e., 22 in our case). The PC analysis was performed on a combined dataset containing all the records to improve the certainty of spectral outlier detection. A total of 86 records with a GH higher than 5 were discarded from the dataset. The cleaned dataset contained finally 6,533 records (i.e., 3,931 and 2,602 records of the first and second datasets, respectively).

Data Splitting

To perform a complete external validation, the data coming from 2 different DHI organizations (one in Germany and one in England) were not used to calibrate the model. This external validation dataset represented 836 samples. The remaining samples (N=5,541) were used to calibrate the model and performed two different cross-validations: 10-fold cross-validation where the samples were chosen randomly in the calibration dataset and a leave-one DHI out cross-validation. The leave-one DHI out cross-validation allows to evaluate the models with samples coming from the same context (countries, diets, breeds) as mentioned by Prekopcsak et al. (2010). So, we supposed that the first dataset contained 9 DHI which corresponded in this case to 9 herds. The second dataset contained records coming from 6 different DHI but records coming from 2 DHI were kept for the external validation as explained previously. Therefore, the leave-one DHI out cross-validation procedure considered 13 groups. Two cross-validations were tested and compared in this study as a random N-fold cross-validation could provide over-optimistic prediction performances (Wang and Bovenhuis, 2019).

Machine Learning Algorithms

All machine learning algorithms used in this study were implemented using the CARET package version 6.0-86 (Kuhn, 2008). For all models, the spectral data were scaled and centered before computation.

PLSR was performed on the 277 selected wavenumbers using the method=”pls” as an argument in the train function of CARET package. The maximum number of PLS latent variables was set to 50.

The optimized number of factors was chosen using the selectionFunction="oneSE" and "best" as arguments in the train function of CARET package. The "best" selection function defines the optimal model as the one having the lowest RMSEcv. The "oneSE" selection function allows to select a simpler model having a RMSE lower within one standard error from the lowest obtained RMSEcv. This simpler model is assumed to have a better generalization.

The computation of SVR was based on linear and polynomial kernels and was implemented using the method="svmLinear" or "svmPoly" as arguments in the train function of the CARET package. For both kernels, the expand.grid function was used to test different values to optimize the required parameters. So, for "svmLinear", the tested C values were 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, and 5. For the "svmPoly" kernel, the tested values were 0.25, 0.5, 1 and 2 for C; 1 until 3 for the polynomial degree; and 0,001, 0.01, 0.1 and 1 for the scale. For both kernels, the epsilon parameter was set to 0.1. As for PLSR, the optimal parameters were chosen using the selection function "best" or "oneSE". A total of 26 PLS factors explaining 99% of the spectral variability were combined in SVR to limit the problem of overfitting. The interest in using PLS factors instead of other more conventional selections of features is based on the fact that PLS will extract factors by considering simultaneously the spectral variability and the variability of the trait to be predicted.

ANN seems to be more powerful when the selection of features is made before the modelling (Thissen et al., 2004). So as performed for SVR, ANN included the 26 PLS latent variables instead of the 277 initially selected wavenumbers. ANN based on one-layer perceptron was tested in this study. This ANN architecture is composed of one hidden layer in order to minimize the risk of overfitting. To estimate the weights related to this ANN design, a back propagation was used. This model was performed using method=”nnet” as argument in the train function of the CARET package. Different numbers of units in the unique hidden layer (ranging from 1 to 5) were tested using the expand.grid function. In order to ensure a global solution, a penalty was introduced during the computation of weights (called decay). Decay values of 0, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, and 0.5 were tested using the expand.grid function. As done for other algorithms, the optimized values for the size and decay were chosen using the selection function “best” or “oneSE”. The maximum iteration for the weight estimations was set to 1000 in order to be sure to reach the convergence.

The prediction performances of the different models developed were assessed by estimating the calibration, 10-fold cross-validation, leave-one DHI out cross-validation and external validation coefficients of determination (R^2c, 10-fold R^2cv, DHI R^2cv and R^2v, respectively) as well as their corresponding RMSE (i.e., RMSEc, 10-fold RMSEcv, DHI RMSEcv and RMSEv, respectively). Distributions of residuals were also studied for all models.

Prediction of LF from The Walloon Milk Recording Database

As the calibration and validation sets were composed of a limited number of records and were not representative of the studied dairy population due to the sampling procedure used, there is an interest in observing the behavior of the prediction on a large scale spectral database. Indeed, this allows observing the behavior of the predictions according to known sources of variation like the stage of lactation, parity, breed, and season. So, the machine learning algorithm chosen as the best based on the validation prediction performances was applied on the first derived milk MIR spectra. The spectral database is managed by the Walloon Breeding Association (Awé, Ciney, Belgium). This database is related to the milk recording. A total of 5,651,470 records were collected between January 2007 and March 2020 from 349,396 cows in 1,963 herds. The average values for the predicted LF were estimated according to the stage of lactation, the milk yield and the somatic cell score (SCS; log2(somatic cell count (SCC)/100000)+3). The correlations between the predicted LF and milk yield, fat and protein contents as well as SCC and SCS were also estimated.

RESULTS AND DISCUSSION

Descriptive Statistics and Data Cleaning

The LF content measured in the samples included in the first dataset (N=3,931) ranged from 3 to 2,038 mg/L of milk, with an average of 202 ± 170 mg/L of milk. LF content in the second dataset (N=2,602) varied from 6 to 1,299 mg/L of milk, with an average of 325 ± 257 mg/L of milk. The average content observed in the first dataset is within the expected range compared to other published articles. For instance, Gaunt et al. (1980) found an average content of LF of 266 ± 136 mg/L of milk from a first set of 4 herds and 228 mg ± 112 mg/L of milk from a second set of 4 herds. Cheng et al. (2008) found a slightly lower content of LF in bovine milk (177 ± 120 mg/L) from samples collected on cows without mastitis infection. In a previous article using a part of the first dataset (N=2,499), Soyeurt et al. (2012) found an average content of LF equal to 163 ± 103 mg/L of milk. The content observed in the second dataset seemed to be high compared to the literature. Moreover, the standard deviation was also higher compared to the first dataset. This can be related to the sample selection. Indeed, the samples included in the second dataset did not come from entire herds as they were selected based on a past LF MIR predictive model in order to cover as much as possible the LF content and spectral variation. However, the ranges of variation observed in both datasets were very high, especially extreme high LF measurements were obtained. This could be related to the fact that some samples could be collected from cows having subclinical mastitis. Indeed, some authors found a positive relationship between the content of LF and the presence of mastitis (Kawai et al., 1999) even if the response differs following the incriminated pathogens (Chaneton et al., 2008). For instance, Gaunt et al. (1980) measured average LF content of 222 ± 168 mg/L of milk for healthy cows to 640 ± 250 mg/L of milk for cows presenting mastitis. Cheng et al. (2008) obtained similar values (i.e., 742 ± 374 mg/L of milk for cows suspected of having mastitis on the basis of the SCC of the milk). The distribution of LF (Figure 1) and spectra (data not shown) from the 2 datasets were complementary. This was expected since samples of the dataset 2 were selected to complement the dataset 1.

Lactoferrin Predictions Using Milk MIR Spectrometry

Two kinds of cross-validation procedures were tested in this study to fix the model parameters (i.e., number of LVs for PLS regressions; C value for linear SVR; C, scale and degree for polynomial SVR; and size and decay for ANN). The leave-one DHI out cross-validation leads to the model being under-fit resulting a higher prediction error. Indeed, RMSEv values were higher than 175 mg/L of milk and were always greater for models developed using the leave-one DHI out cross-validation when compared to models built using the 10-fold cross-validation (Table 1). Moreover, the high RMSEcv SD for models building using the leave-one DHI out cross-validation (i.e., values higher than 90 mg/L of milk) confirmed the low robustness of the developed models. This suggests that too many informative samples were taken out from the calibration set. For instance, during a 10-fold cross-validation, samples coming from the same herd can be in the training and validation set involving relevant information to provide a better prediction. Consequently, the use of a 10-fold cross-validation to parametrize a model is still relevant to limit the under-fitting but as mentioned by Wang and Bovenhuis (2019), this procedure leads to be over-optimistic concerning the prediction performances. Indeed, the observed RMSEcv were always lower than the one observed for the validation set (Table 1). R^2cv values obtained from models developed using 10-fold cross-validation were similar between models used and ranged from 0.51 to 0.56. Similarly, the observed RMSEcv were also globally the same and ranged from 138.40 to 144.60 mg/L (Table 1). This suggests similar prediction performances. However, RMSEcv SD was higher for PLS + polynomial SVR compared to other tested algorithms. Based on the external validation, PLSR, PLS + polynomial SVR and PLS + ANN showed similar validation prediction performances with respective RMSE values of 163.76, 166.75 and 162.17 mg/L of milk. However, the correlation values between predictions on the validation set (N=836) suggested some differences. Indeed, higher correlations were observed between the predictions given by the PLSR and PLS + linear SVR models (0.99) compared to PLS + polynomial SVR or PLS + ANN (0.95 for both algorithms). The correlation between PLS + ANN and PLS + polynomial SVR was 0.94. From Figure 2, it is clear that the relationships between the predictions made from PLSR and PLS + linear SVR models is strong. However, the relationship of those models with other tested ones was not linear. There appears to be a saturation for low and high prediction values (i.e., S shape).

Therefore, even though the predictions made by the 4 models were highly correlated (i.e., higher than 0.94), the low and high values behaved differently. Moreover, the range of predictions is really different, with PLS + linear SVM and PLS + ANN having a reduced range compared to PLSR and PLS + polynomial SVR. Moreover, except PLS + ANN, all other tested algorithms had the tendency to predict negative values (Table 2). The correlation between residuals and predicted content of LF ranged from 0.64 for PLS + ANN to 0.77 for PLS + linear SVM based on the training set. From the validation set, these correlation values were comprised between 0.60 for PLS + ANN to 0.83 for PLS + linear SVM. These correlation values were lower with the squared residuals (from 0.49 for PLS + ANN to 0.61 for PLS + linear SVR and from 0.30 for PLS + ANN to 0.68 for PLS + linear SVR based on the training and validation sets, respectively). This suggests that higher errors were made for samples having a high content of LF. The validation prediction performances, the robustness (low RMSEcv SD) (Table 1), the prediction of positive values (Table 2) and the lowest correlation between squared residuals and LF content observed for PLS + ANN suggest that this modeling is the most relevant to predict daily LF content in milk from milk MIR spectrometry. For another application dedicated to dairy science, Dórea et al. (2018) obtained also better prediction performances using ANN including one hidden layer after a selection of input variables compared to PLSR to predict feed intake. Pralle et al. (2018) obtained similar performances for PLSR and ANN including also one hidden layer to predict blood β-hydroxybutyrate.

Compared to the previous studies published by our team on the same topic, the prediction error observed in the current study is higher than the one observed in the past (RMSEv = 77.26 mg/L of milk in Soyeurt et al. (2012) vs. 163.76 mg/L in the current study based on PLSR or 162.17 mg/L of milk for PLS+ANN). However, this is difficult to compare as the validation dataset was not the same and all spectra were not standardized. If the equation published in 2012 (Soyeurt et al., 2012) is applied on the current validation set, the validation prediction error is of 462 mg/dL with a R^2v equal to 0.02. Even if this old equation was built from a part of the dataset 1 (2499 samples), the variability in the current datasets is higher and the past prediction equation is not suitable to predict those records. This could also be related to the fact that 2 different ELISA kits were used. However, the residual distributions obtained after using the PLS + ANN model were similar for datasets 1 and 2 (data not shown). Moreover, an analysis of variance also confirmed that the differences between residuals observed from datasets 1 and 2 were not significant. It is interesting to note that 50% of records had a residual error between –69.06 and 51.66 mg/L of milk and between -85.78 and 91.37 mg/L of milk for the training and validation set, respectively.

Besides the interest to predict the quantity of lactoferrin in milk, it could be useful to know if the models were able to detect extreme values. Indeed, even if a high prediction error exists for the records having a high content of lactoferrin, the prediction increased as expected but with a lower intensity. PLS + ANN model gave predictions allowing detection of 65% of the records with a content of LF higher than 600 mg/L of milk in the training set if we fixed the prediction limit to 500 mg/L of milk. This proportion reached to 82% for the samples in the validation set. The threshold of 600 mg/L was used as the RMSE started to be related to the content of LF from this content and because authors like Gaunt et al. (1980) and Kawai et al. (1999) mentioned that a such high content is potentially related to cows having mastitis.

Due to the distribution of LF observed in Figure 1, we have also tested the log transformation but the results were not better (data not shown).

PLS + ANN Model Applied to a Large Spectral Database

PLS + ANN model was applied to 5,651,470 records from cows within the first 365 days in milk. The obtained average prediction was 307.80 mg/L of milk with a SD of 209.17 mg/L of milk.

The minimum and maximum values were 19.74 and 1121.00 mg/L, respectively. So as observed on the training and validation datasets, no negative predictions were observed.

The LF content predicted using MIR varied according to the stage of lactation (Figure 3). This variation was already observed by Gaunt et al. (1980). The differences per lactation stages of the log transformed LF contents obtained by predictions and found by Cheng et al. (2008) and Hagiwara et al. (2003) were very similar (Table 3) even if the contents observed in this study were closer to the ones obtained by Hagiwara et al. (2003).

We also observed a negative correlation with milk yield (-0.24) but positive with fat and protein contents (0.11 and 0.28, respectively). Cheng et al. (2008) also found a strongly positive relationship with protein (r=0.48) but they mentioned that the correlation with fat content was not significantly different from 0. However, the negative relationship between LF and milk yield was stronger for Cheng et al. (2008) (r=-0.47). The difference of LF after log transformation and observed by the level of milk yield was similar even if the contents found in this study were higher (Table 4). As expected, positive correlations with predicted LF were observed for SCC and SCS (0.21 and 0.30, respectively; N= 5,477,197). Cheng et al. mentioned that the correlation between LF and SCC was not significantly different than 0 but the correlation found by these authors for SCS (r=0.37) was similar to the one estimated in the current study. The evolution of log-transformed predicted LF was also in agreement with the results found by Hagiwara et al. (2003) and Cheng et al. (2008) (Table 5).

CONCLUSIONS

This study tried 4 examined machine learning algorithms to predict the daily content of lactoferrin in cow's milk from milk MIR spectral data. It found, based on the validation prediction performances, that PLS, PLS + polynomial SVR and PLS + ANN provided similar results, but that the model using PLS factors combined with an ANN was the best. This model was then applied to the Walloon milk recording spectral database to observe the relationships between predicted LF content and the main milk components as well as SCC and SCS which were in line with the literature. However, the model still had some difficulties in predicting extremely high values. Indeed, the observed RMSE increased strongly once LF content exceeded 600 mg/L of milk; 12 percent of records coming from the Walloon dairy cow population reached this level of LF production. Including extreme values of milk LF content to the calibration set could help to improve the prediction models. We now have the possibility to directly predict the content of LF in the milk lab allowing identification of those specific samples. Moreover, as the quantity of milk required for the ELISA analysis is quite low, the same sample as the one analyzed for the routine milk recording could be used. This could be an appropriate task to improve in the future the ability of the ANN network to discriminate low and high LF samples. Until now, no implementation of this LF prediction is done by DHI. However, there is an interest for them to use this molecule information to improve the detection of subclinical mastitis. The inclusion of the LF trait into breeding program to improve the cow robustness or the milk nutritional quality has not been investigated yet.

ACKNOWLEDGMENTS

This research received a financial support from the European Commission, Directorate-General for Agriculture and Rural Development, under Grant Agreement 211708 and from the Commission of the European Communities, FP7, KBBE-2007-1. This paper does not necessarily reflect the view of these institutions and in no way anticipates the Commission's future policy in this area. The Ministry of Agriculture of the Walloon Region of Belgium [Service public de Wallonie, Direction générale de l'Agriculture, des Ressources naturelles et de l'Environnement, Direction de la Recherche] is acknowledged for their financial support through the project NovaUdderHealth (D31-1207). Scottish Agricultural College receives funding from the Scottish Government and they are acknowledged for funding the long-term selection experiment producing milk samples used in this analysis. The work of Sinead McParland was funded by a Science Foundation Ireland Starting Investigator Research Grant (18/SIRG/5562).

References

Beck, M.W. 2018. NeuralNetTools: Visualization and analysis tools for neural networks. J. Stat. Softw. 85:1–20. doi:10.18637/jss.v085.i11.

Chaneton, L., M. Bontá, M. Pol, L. Tirante, and L.E. Bussmann. 2013. Milk lactoferrin in heifers: Influence of health status and stage of lactation. J. Dairy Sci. 96:4977–4982. doi:10.3168/jds.2012-6028.

Chaneton, L., L. Tirante, J. Maito, J. Chaves, and L.E. Bussmann. 2008. Relationship between milk lactoferrin and etiological agent in the mastitic bovine mammary gland. J. Dairy Sci. 91:1865–1873. doi:10.3168/jds.2007-0732.

Chen, P.W., and F.C. Mao. 2004. Detection of lactoferrin in bovine and goat milk by enzyme-linked immunosorbent assay. J. Food Drug Anal. 12:133–139.

Cheng, J.B., J.Q. Wang, D.P. Bu, G.L. Liu, C.G. Zhang, H.Y. Wei, L.Y. Zhou, and J.Z. Wang. 2008. Factors Affecting the Lactoferrin Concentration in Bovine Milk. J. Dairy Sci. 91:970–976. doi:10.3168/JDS.2007-0689.

Delhez, P., P.N. Ho, N. Gengler, H. Soyeurt, and J.E. Pryce. 2020. Diagnosing the pregnancy status of dairy cows: How useful is milk mid-infrared spectroscopy?. J. Dairy Sci. 103:3264–3274. doi:10.3168/jds.2019-17473.

Despagne, F., D. Luc Massart, and P. Chabot. 2000. Development of a robust calibration model for nonlinear in-line process data. Anal. Chem. 72:1657–1665. doi:10.1021/ac991076k.

Dórea, J.R.R., G.J.M. Rosa, K.A. Weld, and L.E. Armentano. 2018. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. J. Dairy Sci. 101:5878-5889.

Finn, G.D., R. Lister, T. Szabo, D. Simonetta, H. Mulder, and R. Young. 1996. Neural Networks Applied to a Large Biological Database to Analyse Dairy Breeding Patterns. Neural Comput. & Applic. 4:237-253.

Gaunt, S.N., N. Raffio, E.T. Kingsbury, R.A. Damon, W.H. Johnson, and B.A. Mitchell. 1980. Variation of Lactoferrin and Mastitis and Their Heritabilities. J. Dairy Sci. 63:1874–1880. doi:10.3168/jds.S0022-0302(80)83154-7.

Giansanti, F., G. Panella, L. Leboffe, and G. Antonini. 2016. Lactoferrin from milk: Nutraceutical and pharmacological properties. Pharmaceuticals 9:1–15. doi:10.3390/ph9040061.

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

Grelet, C., J.A.F. Pierna, P. Dardenne, H. Soyeurt, A. Vanlierde, F. Colinet, C. Bastin, N. Gengler, V. Baeten, and F. Dehareng. 2017. Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models. J. Dairy Sci. 100:7910–7921. doi:10.3168/jds.2017-12720.

Grzesiak, W., P. Blaszczyk, and R. Lacroix. 2006. Methods of predicting milk yield in dairy cows—Predictive capabilities ofWood's lactation curve and artificial neural networks (ANNs). Computers and Electronics in Agriculture 54:69–83.

Hagiwara, S.I., K. Kawai, A. Anri, and H. Nagahata. 2003. Lactoferrin concentrations in milk from normal and subclinical mastitic cows. J. Vet. Med. Sci. 65:319–323. doi:10.1292/jvms.65.319.

Hempstalk, K., S. McParland, and D.P. Berry. 2015. Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. J. Dairy Sci. 98:5262-5273.

Kawai, K., S. Hagiwara, A. Anri, and H. Nagahata. 1999. Lactoferrin Concentration in Milk of Bovine Clinical Mastitis. Vet. Res. Commun. 23:391–398. doi:10.1023/A:1006347423426.

Król, J., Z. Litwińczuk, A. Brodziak, and J. Barłowska. 2010. Lactoferrin, lysozyme and immunoglobulin G content in milk of four breeds of cows managed under intensive production system. Pol. J. Vet. Sci. 13:357–361.

Kuhn, M. 2008. caret Package. J. Stat. Softw. 28:1–26.

Lê, S., J. Josse, and F. Husson. 2008. FactoMineR: An R package for multivariate analysis. J. Stat. Softw. 25:1–18. doi:10.18637/jss.v025.i01.

De Marchi, M., C.C. Fagan, C.P. O'Donnell, A. Cecchinato, R. Dal Zotto, M. Cassandro, M. Penasa, and G. Bittante. 2009. Prediction of coagulation properties, titratable acidity, and pH of bovine milk using mid-infrared spectroscopy. J. Dairy Sci. 92:423–432. doi:10.3168/jds.2008-1163.

McCulloch, W.S., and W. Pitts. 1943. A logical calculus of the ideas imminent in nervous activity. Bulletin of Mathematical Biophysics 5:115-133. doi:10.1007/bf02478259.

Molenaar, A.J., Y.M. Kuys, S.R. Davis, R.J. Wilkins, P.E. Mead, and J.W. Tweedie. 1996. Elevation of Lactoferrin Gene Expression in Developing, Ductal, Resting Regressing Parenchymal Epithelium of the Ruminant Mammary Gland. J. Dairy Sci. 79:1198–1208. doi:10.3168/jds.S0022-0302(96)76473-1.

Pralle, R.S., K.W. Weigel, and H.M. White. 2018. Predicting blood β-hydroxybutyrate using milk Fourier transform infrared spectrum, milk composition, and producer-reported variables with multiple linear regression, partial least squares regression, and artificial neural network. J. Dairy Sci. 101:4378-4387.

Prekopcsak, Z., T. Henk, and C. Gaspar-Papanek. 2010. Cross-validation : the illusion of reliable performance estimation. RCOMM RapidMiner Community Meet. Converence 1–6.

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

Soyeurt, H., F.G. Colinet, V.M.R. Arnould, P. Dardenne, C. Bertozzi, R. Renaville, D. Portetelle, and N. Gengler. 2007. Genetic variability of lactoferrin content estimated by mid-infrared spectrometry in bovine milk. J. Dairy Sci. 90:4443–4450. doi:10.3168/jds.2006-827.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D.P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 94:1657–1667. doi:10.3168/jds.2010-3408.

Soyeurt, H., E. Froidmont, I. Dufrasne, D. Hailemariam, Z. Wang, C. Bertozzi, F.G. Colinet, F. Dehareng, and N. Gengler. 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. Livest. Sci. 227:82–89. doi:10.1016/j.livsci.2019.07.007.

Thissen, U., M. Pepers, B. Üstün, W.J. Melssen, and L.M.C. Buydens. 2004. Comparing support vector machines to PLS for spectral regression applications. Chemom. Intell. Lab. Syst. 73:169–179. doi:10.1016/j.chemolab.2004.01.002.

Vanlierde, A., M.L. Vanrobays, N. Gengler, P. Dardenne, E. Froidmont, H. Soyeurt, S. McParland, E. Lewis, M.H. Deighton, M. Mathot, and F. Dehareng. 2016. Milk mid-infrared spectra enable prediction of lactation-stage-dependent methane emissions of dairy cattle within routine population-scale milk recording schemes. Anim. Prod. Sci. 56:258–264. doi:10.1071/AN15590.

Wang, Q., and H. Bovenhuis. 2019. Validation strategy can result in an overoptimistic view of the ability of milk infrared spectra to predict methane emission of dairy cattle. J. Dairy Sci. 102:6288–6295. doi:10.3168/jds.2018-15684.

Wakabayashi, H., K. Yamauchi, and M. Takase. 2006. Lactoferrin research, technology and applications. Int. Dairy J. 16:1241-1251.

Yang, X.Z., R. Lacroix, and K.M. Wade. 2000. Investigation into the production and conformation traits associated with clinical mastitis using artificial neural networks. Can. J. Anim. Sci. 80: 415–426.

Table 1. Ten-fold cross-validation and external validation performances for predicting lactoferrin content in milk using four different machine learning algorithms.

		PLSR	PLS + Linear SVR	PLS + Polynomial SVR	PLS + ANN
Selection function		oneSE	oneSE	best	best = oneSE*
Calibration (N=5541)	parameters	nLV=23	C=5	degree=3; scale=0.01; C=1	size=4; decay=0.5
	R^2c	0.53	0.53	0.64	0.60
	RMSEc	140.94	144.32	125.89	130.59
Cross-validation	R^2cv	0.51	0.53	0.56	0.55
	R^2cv SD	0.03	0.03	0.03	0.03
	RMSEcv	144.31	144.60	138.40	139.01
	RMSEcv SD	5.77	5.61	8.08	5.05
	RPD	1.43	1.42	1.49	1.48
External validation (N=836)	R^2v	0.61	0.63	0.62	0.60
	RMSEv	163.76	174.92	166.75	162.17

PLSR= Partial least squares regression; PLS + Linear SVR = Linear Support Vector Regression based on 26 PLS latent variables; PLS + Polynomial SVR = Linear SVR based on 26 PLS latent variables; PLS + ANN = modeling based on artificial neural network including 26 PLS latent variables in the input layer and one hidden layer; nLV= number of PLS latent variables; C= cost penalty used for SVR; * = the selection function 'best' and 'oneSE' provided the same results.

Table 2. Data distribution of the reference lactoferrin contents as well as the predictions obtained from the developed models after a 10-fold cross-validation.

		Data distribution (mg/L of milk)								
		0%	5%	10%	25%	50%	75%	90%	95%	100%
Training set	Lactoferrin	3.01	34.00	57.22	99.84	179.72	302.69	532.39	669.97	2038.27
	PLSR	-277.95	35.10	77.14	142.16	218.59	308.55	430.42	521.86	1244.38
	PLS + linSVR	-195.05	37.97	74.54	128.68	194.55	271.75	385.47	466.19	1060.68
	PLS + polSVR	-47.40	51.77	74.18	119.48	187.15	271.52	413.47	532.49	1423.58
	PLS + ANN	31.27	77.53	95.71	135.09	188.21	289.78	472.00	605.24	1053.65
Validation set	Lactoferrin	6.00	32.32	47.46	122.30	280.58	476.69	707.86	860.52	1286.86
	PLSR	-162.60	13.98	70.29	165.56	320.69	420.68	518.84	591.00	912.53
	PLS + linSVR	-129.09	22.84	68.17	150.40	290.59	367.41	455.30	520.24	795.47
	PLS + polSVR	-72.59	27.30	55.13	123.53	263.11	381.74	533.62	626.71	1200.00
	PLS + ANN	36.67	67.71	85.11	127.48	313.83	483.87	633.11	682.23	851.38

PLSR= Partial least squares regression; PLS + linSVR = linear Support Vector Regression based on 26 PLS latent variables; PLS + polSVR = polynomial SVR based on 26 PLS latent variables; PLS + ANN = modeling based on an artificial neural network including 26 PLS latent variables in the input layer and one hidden layer.

Table 3. Evolution of lactoferrin content predicted by MIR following the stage of lactation and comparison with the literature.

		MIR lactoferrin mg/L of milk		log(lactoferrin)		log(lactoferrin) [Cheng et al.,2008]			log(lactoferrin) [Hagiwara et al., 2003]*		
	N	Mean	SD	Mean	SD	N	Mean	SD	N	Mean	SD
DIM <= 20	317,322	229.75	181.22	2.23	0.33						
20 > DIM <= 100	1,486,501	245.45	176.28	2.29	0.30	49	1.90	0.12	8	2.06	0.43
100 > DIM <= 200	1,736,757	297.04	198.64	2.37	0.30	45	2.03	0.28	59	2.23	0.38
200 > DIM <= 365	2,110,890	372.31	223.49	2.48	0.29	28	2.20	0.20	32	2.30	0.45

DIM = days in milk; * The range of DIM was slightly different

Table 4. Evolution of lactoferrin content predicted by MIR following the milk yield (kg/day) and comparison with the literature.

		MIR lactoferrin mg/L of milk		log(lactoferrin)		log(lactoferrin) [Cheng et al.,2008]		
	N	Mean	SD	Mean	SD	N	Mean	SD
milk yield < 20 kg	1,611,617	374.43	228.40	2.48	0.30	36	2.18	0.25
20 kg <= milk yield < 25 kg	1,333,645	308.94	204.31	2.39	0.30	34	1.99	0.23
25 kg <= milk yield < 30 kg	1,201,798	283.77	194.82	2.35	0.30	33	1.93	0.16
30 kg >= milk yield	1,504,410	254.63	181.96	2.30	0.30	19	1.89	0.18

Table 5. Evolution of lactoferrin content predicted by MIR following the somatic cell score (SCS) and comparison with the literature.

		MIR lactoferrin		log(lactoferrin)		log(lactoferrin) [Cheng et al.,2008]			log(lactoferrin) [Hagiwara et al., 2003]		
SCS	N	Mean	SD	Mean	SD	N	Mean	SD	N	Mean	SD
0	163,716	236.23	166.91	2.28	0.29	12	1.91	0.14	36	2.18	0.19
1	996,710	233.82	169.95	2.27	0.29	20	2.02	0.17	28	2.16	0.42
2	1,365,672	267.61	184.44	2.33	0.30	50	1.98	0.19	39	2.27	0.51
3	1,164,258	305.80	199.50	2.39	0.30	40	2.06	0.26			
4	860,534	347.21	214.41	2.45	0.30	34	2.10	0.25			
5	510,941	382.85	227.91	2.49	0.30	20	2.26	0.32			
6	292,339	403.68	234.51	2.52	0.30	22	2.28	0.30			
7	160,828	421.79	234.70	2.54	0.30						
8	83,575	461.05	235.88	2.59	0.28						
9	42,340	569.97	244.31	2.70	0.25						

* 10,557 records were deleted because the SCS had a negative value.

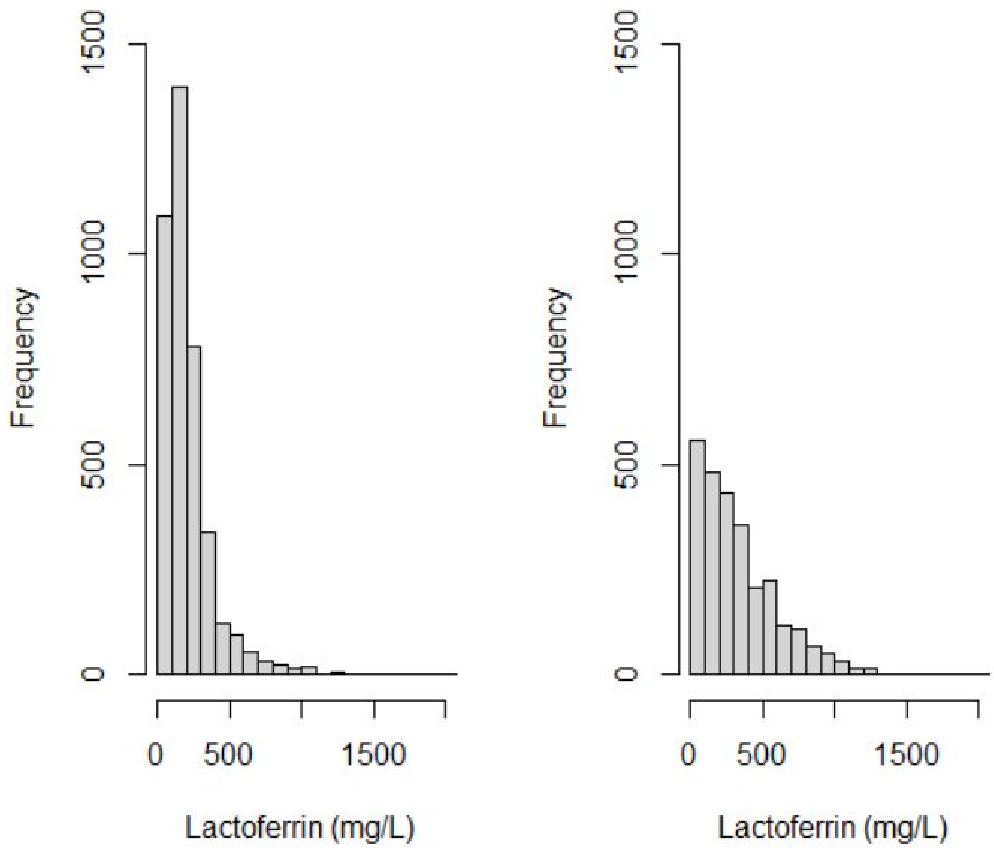


Figure 1. Distribution of ELISA Lactoferrin quantifications in the first dataset on the left and the second dataset on the right.

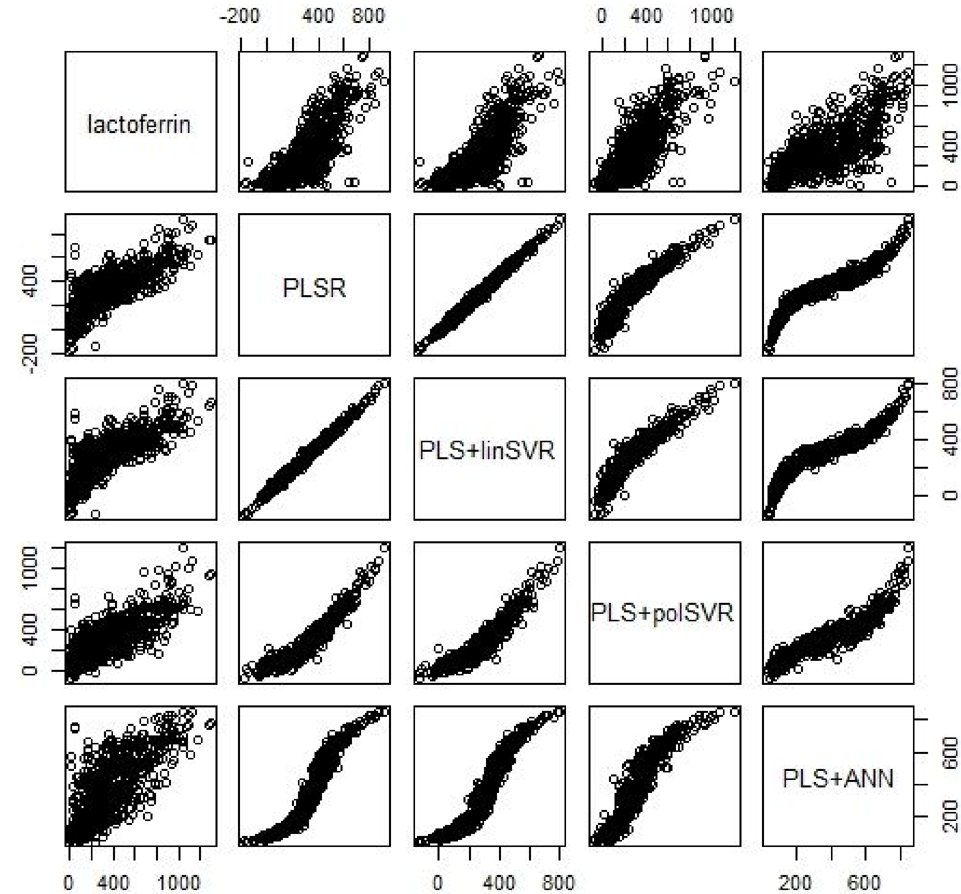
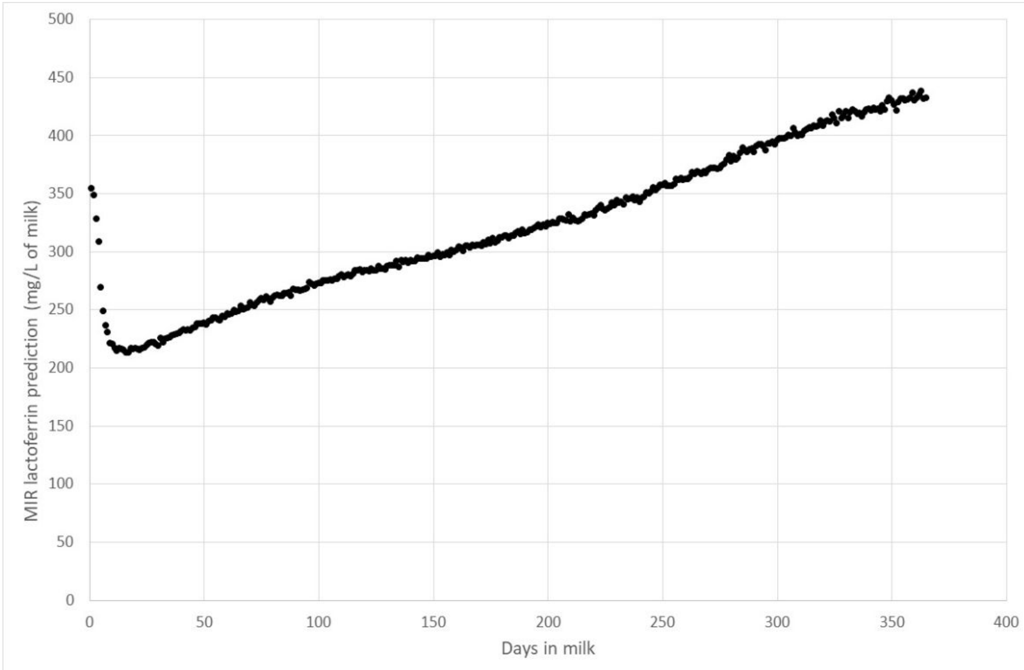


Figure 2. Relationships between reference lactoferrin content (mg/L of milk) and the predictions obtained using fourth different machine learning approaches applied on the validation set (PLSR = Partial Least Squares Regression; PLS+linSVR = 26 PLS factors included in a linear Support Vector Regression; PLS+polSVR= 26 PLS factors included in a polynomial SVR; PLS+ANN = 26 PLS factors included in an artificial neural network having one hidden layer).



560

561 Figure 3. Evolution of lactoferrin content predicted by mid-infrared spectrometry following the stage of
562 lactation.