J. Dairy Sci. 105:3615–3632
https://doi.org/10.3168/jds.2021-21516
© 2022, The Authors. Published by Elsevier Inc. and Fass Inc. on behalf of the American Dairy Science Association®.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Pregnancy status predicted using milk mid-infrared spectra from dairy cattle

K. M. Tiplady,[1,2*] M.-H. Trinh,[1] S. R. Davis,[1] R. G. Sherlock,[1] R. J. Spelman,[1] D. J. Garrick,[2] and B. L. Harris[1]
[1]Research and Development, Livestock Improvement Corporation, Private Bag 3016, Hamilton 3240, New Zealand
[2]School of Agriculture, Massey University, Ruakura, Hamilton 3240, New Zealand

ABSTRACT

Accurate and timely pregnancy diagnosis is an important component of effective herd management in dairy cattle. Predicting pregnancy from Fourier-transform mid-infrared (FT-MIR) spectroscopy data is of particular interest because the data are often already available from routine milk testing. The purpose of this study was to evaluate how well pregnancy status could be predicted in a large data set of 1,161,436 FT-MIR milk spectra records from 863,982 mixed-breed pasture-based New Zealand dairy cattle managed within seasonal calving systems. Three strategies were assessed for defining the nonpregnant cows when partitioning the records according to pregnancy status in the training population. Two of these used records for cows with a subsequent calving only, whereas the third also included records for cows without a subsequent calving. For each partitioning strategy, partial least squares discriminant analysis models were developed, whereby spectra from all the cows in 80% of herds were used to train the models, and predictions on cows in the remaining herds were used for validation. A separate data set was also used as a secondary validation, whereby pregnancy diagnosis had been assigned according to the presence of pregnancy-associated glycoproteins (PAG) in the milk samples. We examined different ways of accounting for stage of lactation in the prediction models, either by including it as an effect in the prediction model, or by pre-adjusting spectra before fitting the model. For a subset of strategies, we also assessed prediction accuracies from deep learning approaches, utilizing either the raw spectra or images of spectra. Across all strategies, prediction accuracies were highest for models using the unadjusted spectra as model predictors. Strategies for cows with a subsequent calving performed well in herd-independent validation with sensitivities above 0.79, specificities above 0.91 and area under the receiver operating characteristic curve (AUC) values over 0.91. However, for these strategies, the specificity to predict nonpregnant cows in the external PAG data set was poor (0.002–0.04). The best performing models were those that included records for cows without a subsequent calving, and used unadjusted spectra and days in milk as predictors, with consistent results observed across the training, herd-independent validation and PAG data sets. For the partial least squares discriminant analysis model, sensitivity was 0.71, specificity was 0.54 and AUC values were 0.68 in the PAG data set; and for an image-based deep learning model, the sensitivity was 0.74, specificity was 0.52 and the AUC value was 0.69. Our results demonstrate that in pasture-based seasonal calving herds, confounding between pregnancy status and spectral changes associated with stage of lactation can inflate prediction accuracies. When the effect of this confounding was reduced, prediction accuracies were not sufficiently high enough to use as a sole indicator of pregnancy status.

Key words: Fourier-transform mid-infrared spectra, pregnancy prediction, milk composition, dairy cattle, machine learning

INTRODUCTION

Knowledge of pregnancy status for dairy cattle is an important component of an efficient and productive herd management system. In an ideal seasonal calving system, estrus is reliably detected during the mating period, so that animals are inseminated and conceive in a timely manner, resulting in a herd average 365-d calving interval. Knowing a cow is pregnant during the mating period avoids wasted re-inseminations, and early identification of nonpregnant cows could provide an opportunity to shorten interbreeding intervals and result in an increase in herd profitability (Ferguson and Galligan, 2011; Giordano et al., 2013). Moreover, knowledge of nonpregnant status beyond the mating period plays a role in herd management and culling decisions. Pregnancy status during the mating period is crudely determined by nonreturn to estrus, and later in lactation is ascertained by indirect methods such as

Received November 2, 2021.
Accepted December 27, 2021.
*Corresponding author: Kathryn.Sanders@lic.co.nz

those measuring milk progesterone levels, or pregnancy-associated glycoproteins (**PAG**) in blood or milk, or direct methods such as transrectal palpation and ultrasonography. Direct pregnancy testing is costly and may also require additional animal-handling, and detection of pregnancy status based only on nonreturn to estrus is unreliable unless estrus detection monitoring and recording is of a high standard. Further, in instances of embryonic loss after initial pregnancy establishment, nonpregnant cows may not all return to estrus due to the extended presence of a corpus luteum (Ricci et al., 2017). For these reasons, a methodology for determining pregnancy status using Fourier-transform mid-infrared (**FT-MIR**) spectroscopy is of interest, because the data are often already available from routine milk testing at 30- or 60-d intervals.

Pregnancy results in changes to metabolism and energy requirements and leads to a repartitioning of resources to different physiological functions, compared with a nonpregnant lactating animal, and has a consequent influence on milk composition in dairy cattle, particularly in mid to late lactation (Olori et al., 1997; Loker et al., 2009; Penasa et al., 2016). Previous studies have examined the effect of pregnancy stage on detailed milk composition as determined by FT-MIR spectra (Lainé et al., 2017), and have reported the ability to predict conception outcomes (Hempstalk et al., 2015) or pregnancy (Toledo-Alvarado et al., 2018; Delhez et al., 2020; Brand et al., 2021) from FT-MIR spectra. Improvements in accuracy from incorporating FT-MIR data into pregnancy prediction models vary between studies. Toledo-Alvarado et al. (2018) assessed and compared the ability to predict pregnancy from milk components (fat, protein, lactose and casein) or from a single wavenumber or a full FT-MIR spectra, using a Bayesian variable selection model. The best predictions of pregnancy in that study were obtained when full FT-MIR spectra were incorporated into prediction models, with the area under the receiver operating characteristic curve (**AUC**) values of around 0.6. Delhez et al. (2020) investigated the potential of FT-MIR to predict pregnancy status of dairy cows with partial least squares discriminant analysis (**PLS-DA**) using residual FT-MIR spectra, evaluated from the difference between the spectra before and after insemination at a specific stage of lactation; and predicting pregnancy within lactation stage classes, to account for the effect of lactation stage on milk composition. They found that prediction accuracies for models developed using FT-MIR spectra across different stages of lactation were limited, with AUC values of around 0.6, but that models using data after 150 d of pregnancy had promising prediction accuracies with AUC values of around 0.78. The use of deep learning models to establish pregnancy status were examined by Brand et al. (2021). They compared prediction accuracies between models developed using genetic algorithms for feature selection and network design, and transfer learning models that used a pretrained dense convolutional network (DenseNet) model. The former of these approaches resulted in high validation accuracies of 0.89, but loss values of 0.18, which were considered too high for useful application in the industry. However, models using transfer learning whereby FT-MIR spectra was converted to gray-scaled images, resulted in accuracy and loss values of 0.97 and 0.08, respectively, indicating that transfer learning can provide pregnancy prediction models with high enough accuracies for industry application.

In previous studies where FT-MIR spectra had been used to predict pregnancy, there were key differences in the manner in which records were selected for inclusion in the analysis, and how records were classified as pregnant or nonpregnant. The purpose of this study was to investigate pregnancy prediction accuracy from FT-MIR spectra in a data set of NZ seasonal calving herds, when differing strategies for classifying pregnant or nonpregnant records, broadly similar to those from previous studies, were used across the same data set. We assess 3 strategies for partitioning records, 2 of which use records for cows with a subsequent calving only, whereas the third includes records for cows without a subsequent calving. We examine the impact of different ways of accounting for the effect of stage of lactation in these models, either by including days in milk as a model predictor, or by preadjusting the spectra for DIM; and for a subset of models, we compare prediction accuracies from PLS-DA models to those from alternative models developed using deep learning approaches. Finally, we investigate the relationship between FT-MIR spectra and lactation stage by assessing how well DIM can be predicted from spectral data.

MATERIALS AND METHODS

Ethics Statement

All data were collected as part of routine on-farm activities and thus did not require formal ethics approval.

Data

Fourier-Transform Mid-Infrared Spectra. Fourier-transform mid-infrared spectra were from a wider set of 2,044,094 routine milk test samples for 1,877,456 animals, collected from Bentley FTS (Chaska, MN) instruments by Livestock Improvement Corporation, as previously described in Tiplady et al. (2019). Briefly, FT-MIR spectra from milk samples analyzed between

September 2017 and May 2018 were preprocessed to remove outliers and standardized to account for differences between instruments. Outliers were removed according to the squared Mahalanobis distance between each spectrum and the average within-instrument spectrum from each analyzer, and standardization was performed using piecewise direct standardization (Grelet et al., 2015). Spectral data consisted of light absorbance values for 899 spectral wavenumbers across the range from 649.03 to 3,998.59 cm^{-1}. These were restricted to exclude wavenumbers within noise regions as defined by Tiplady et al. (2019) (649–970, 1,608–1,682, and ≥3,021 cm^{-1}). This resulted in 528 wavenumbers for use in the development of prediction equations. Exclusions were applied to remove records with high SCC (≥400,000 cells/mL) or records where there had been less than 30 samples processed for the herd on a day. Additionally, records were restricted to those for spring-calving animals that calved between June and November, and where the sample took place between 5 and 300 DIM. This resulted in a data set of 1,853,771 spectral records for 1,375,227 cows, across 5,529 herds.

Glycoprotein-Based Pregnancy Diagnosis. Pregnancy-associated glycoproteins are macromolecules produced by placental tissue and can provide a good indication of pregnancy (Green et al., 2005; Sousa et al., 2006; Commun et al., 2016), with accurate indication of pregnancy status achievable from PAG in milk samples as early as 25 d after successful AI (Ricci et al., 2015; Commun et al., 2016). Assessment of PAG in milk samples was undertaken at Livestock Improvement Corporation's Animal Health laboratory, with cows assigned as pregnant, not pregnant, or unconfirmed. In total, there were 25,493 records among available spectral records for which there was a PAG result, of which 22,235 were assigned as pregnant, 2,032 were assigned as not pregnant, and the remaining 1,226 with an unconfirmed result. Records were restricted to those that were definitively assigned as pregnant or not pregnant, and where the test date of the PAG result was ≥28 d after the last AI for the cow. This resulted in a data set of 24,063 records, representing 24,004 cows in 202 herds. At the time of the PAG assessment, the average DIM was 186, ranging from 42 to 299 d; and the average number of days since the last mating was 103 d, ranging from 28 to 222 d.

Consolidation of Spectral Records with AI and Calving Data. Records of AI events and those of subsequent calvings were obtained for all cows with spectral records. Validated AI events were assigned where calving took place between 271 and 293 d after an AI event. To reduce the risk of assigning a record incorrectly as pregnant, if there was more than one potential AI date within the window of 271 to 293 d before successful calving, all records for that animal were excluded. Similarly, if there was a subsequent calving but no validated AI event between 271 and 293 d before calving, all records for the animal were removed. The resulting data set was filtered to exclude all herds with animals that had a pregnancy diagnosis based on PAG, to enable the latter data set to be used as an external herd-independent validation data set. This resulted in a final data set of 1,161,436 records for 863,982 animals, across 5,170 herds for generating and evaluating pregnancy prediction models. The median calving date across these records was August 11, 2017, and the median parity of cows was 3 with a range of 1 to 9. The breed composition comprised 277,658 cows with ≥14/16 Holstein or Friesian composition; 87,111 cows with ≥14/16 Jersey composition, 446,136 cows with ≥3/16 Holstein-Friesian and ≥3/16 Jersey composition; and 53,077 cows from other breeds or crosses.

Strategies for Classifying Pregnancy Status

Three different strategies that broadly reflected those from previous studies were used to select and classify records into pregnant and nonpregnant groups. For all of these strategies, records were only defined as pregnant if there was a validated AI event and a subsequent calving (n = 700,332 records). The strategies varied in the manner in which nonpregnant records were assigned. Specifically, the strategies were defined as follows: (1) records before the first mating were assigned as nonpregnant (n = 164,537); (2) records after the first mating but before the validated AI event were assigned as nonpregnant (n = 14,778); and (3) in addition to nonpregnant records used in (2), records for cows without a subsequent calving were assigned as nonpregnant (n = 197,624). Strategy 1 was similar to that defined in the study by Brand et al. (2021), whereas strategy 2 was similar to that defined by Toledo-Alvarado et al. (2018), except that in their study they only retained records within 90 d after each insemination, and classified records without a subsequent insemination within 90 d as pregnant, and records with a subsequent insemination within 90 d as nonpregnant. Strategy 3 was similar to that defined in the study by Delhez et al. (2020) in that records were not restricted to those for cows with a subsequent calving, but differed in that our data set was not restricted to using only a single-spectral record after each insemination.

Partial Least Squares Discriminant Analysis Model Development and Validation

Animals with confirmed pregnant or nonpregnant status based on PAG formed the basis of an external

validation data set (**VAL-PAG**). For each pregnancy classification strategy, spectra from the remaining records were partitioned into training and validation data sets. Each training data set consisted of records for cows from a random sample of 80% of herds, with the remaining spectra assigned to validation (**VAL-Test**). Random sampling with replacement was conducted to augment the minority class (nonpregnant) to be the same size as the majority class (pregnant) in the training data set. The PLS-DA models were developed from training data with 10 repeats of 10-fold cross-validation using the caret package in R (Kuhn, 2008). For each pregnancy classification strategy, 3 types of models were evaluated: (1) models using unadjusted spectra wavenumbers as predictors; (2) models using unadjusted spectra wavenumbers but including DIM as a predictor, where DIM was fitted as a class variable representing 30-d windows from the start of lactation; and (3) models using adjusted spectra wavenumbers as predictors, where the spectra had been pre-adjusted for DIM (30-d window classes) using repeated measures models in ASReml-R (Butler et al., 2009).

To assess the effect of augmenting the data using upsampling of the minority nonpregnant class, a secondary set of models were developed using a downsampled training data set whereby random sampling was conducted to reduce the majority class (pregnant) to be the same size as the minority class (nonpregnant). Additionally, for models using unadjusted spectral wavenumbers as predictors, we assessed the effect of excluding records classified as pregnant where the test date was within a short time period after a validated successful AI event. Specifically, we evaluated alternative models whereby spectral records classified as pregnant were removed if the test date associated with the record was within 7, 14, or 21 d of a cow's successful AI.

Model performance for each pregnancy classification strategy and model type was assessed according to the sensitivity and specificity of pregnancy prediction, and from the AUC values when the trained model was applied to each of the 2 validation data sets (VAL-Test and VAL-PAG). Sensitivity was defined as the proportion of pregnant records that were correctly assigned as pregnant by the model, and specificity was defined as the proportion of nonpregnant records that were correctly assigned as nonpregnant by the model. Receiver operating characteristic (**ROC**) curves represent the relationship between a model's true positive rate (records correctly assigned as pregnant) and the false positive rate (records incorrectly assigned as pregnant), for different classification thresholds. The AUC measures the area under the ROC curve when values of the true positive rate are plotted against values of the false positive rate on a continuous scale, providing a consolidated measure of model performance across all possible classification thresholds. Values of AUC range from 0 to 1, with an AUC value of 0.5 indicating that the model is only able to classify records as well as random allocation of pregnant and nonpregnant status.

Deep Learning Models

Two different deep learning approaches were developed for a subset of models using training and test data sets as defined by strategy 3. The first approach used a multilayer perceptron (**MLP**) feed-forward artificial neural network to classify pregnancy status based on raw spectra, whereas the second used a convolutional neural network (**CNN**). A diagram showing the pipeline for the 2 approaches is in Figure 1. Both deep learning approaches were implemented with PyTorch (v1.7.1; Paszke et al., 2019), and one-hot encoding was used to transform the categorical predictor DIM, which was classified by 30-d windows from the start of lactation. For the MLP network, we used 2 fully connected layers with leaky rectified linear activation (LeakyReLU) and batch normalization (BatchNorm) to accelerate convergence speed. When trained with DIM, the obtained one-hot encoded tensor was concatenated with the spectral wavenumbers tensor before input to the fully connected layers. The CNN network shared the same fully connected layers design as MLP, but the spectral wavenumbers tensor went through a dense convolutional network architecture, extracting 1,024 features that were then concatenated with one-hot encoded DIM. PyTorch Image Models (Wightman, 2019) were used to generate the DenseNet121 (Huang et al., 2018) feature extractor layers. The original DenseNet121 architecture was designed for images with 3 channels, 224 rows, and 224 columns as input, whereas the input of our 528 spectral wavenumbers were an image with one channel, one row, and 528 columns. To accommodate the difference in the image size of the spectra data, the number of input channels was changed from 3 to 1 in the first convolution layer; and the kernel size and stride of all average pooling layers was changed from (2,2) to (2,1). We applied adaptive average pooling (AdaptiveAvgPool) to the output of our 1D DenseNet121 network to reduce the overall number of parameters, and applied LeakyReLU and BatchNorm to accelerate the convergence of the stochastic gradient descent. Across both deep learning approaches, networks were trained for 50 epochs on computers equipped with NVIDIA Titan XP or RTX 5000 graphics cards, using a batch size of 1,024 and randomized initial weights. Stochastic gradient descent was used to minimize binary cross entropy loss, with the learning rate starting at 1e-03 and reduced

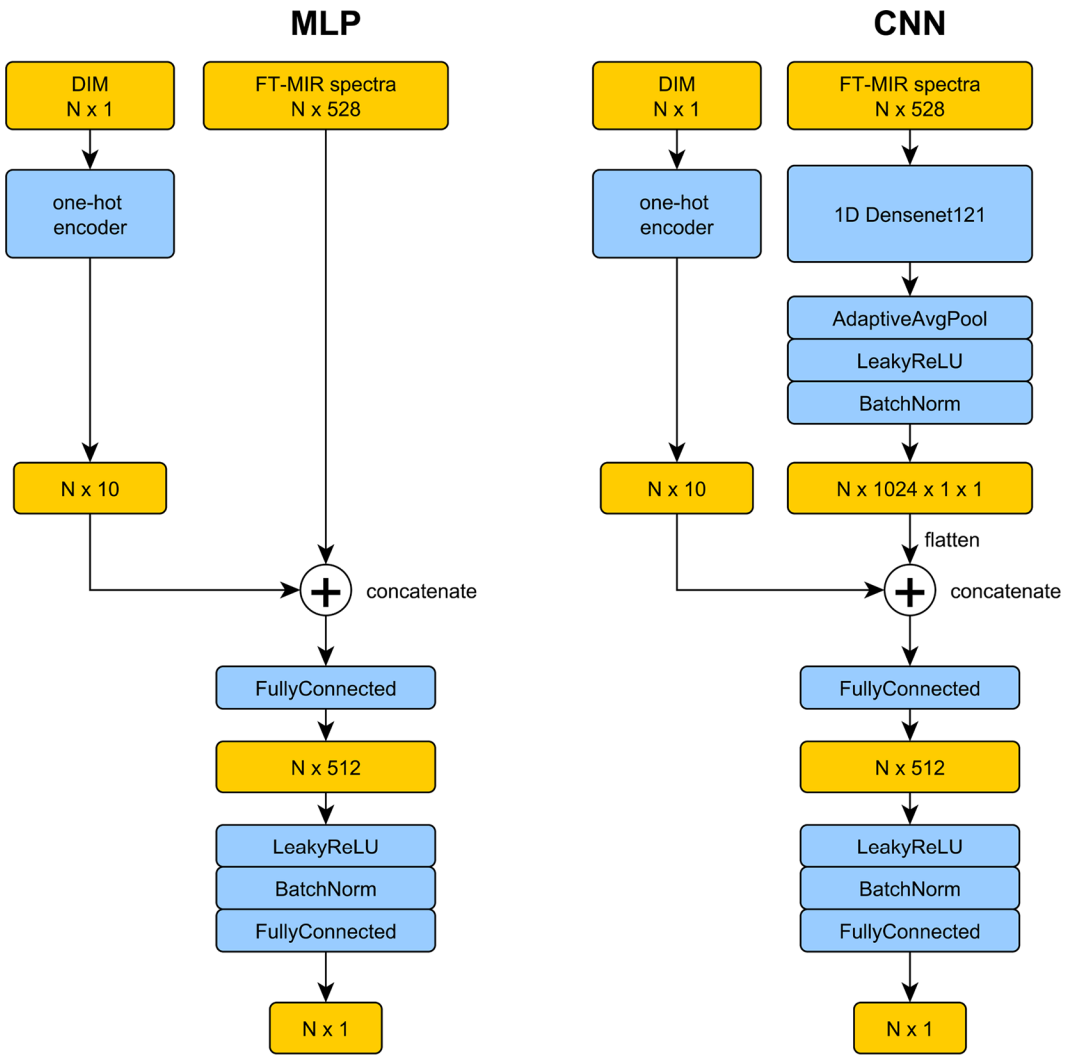

Figure 1. Architecture of the multilayer perceptron (MLP) feed-forward artificial neural network and the convolutional neural network (CNN) used to classify pregnancy status. N = batch size, representing the number of samples processed before the model is updated. FT-MIR = Fourier-transform mid-infrared. One-hot encoding used to transform the categorical predictor DIM; for the CNN, spectral wavenumbers tensor passed through a dense convolutional network architecture (1D Densenet 121) and adaptive average pooling (AdaptiveAvgPool) applied, followed by Leaky Rectified Linear Unit (LeakyReLU) and batch normalization (BatchNorm); for both networks, one-hot encoded DIM tensor concatenated with spectral wavenumbers tensor to form fully connected layers which were passed through a further round of LeakyReLU and BatchNorm.

by a factor of 10 at epoch 15, 25, and 35. Validation of each deep learning approach was conducted in the same way as for the PLS-DA models, using test data (VAL-Test) and the separate cow-independent data set whereby pregnancy diagnosis had been assigned according to PAG in the milk sample (VAL-PAG).

Prediction Models for Stage of Lactation

To investigate the relationship between FT-MIR spectra and lactation stage, a partial least squares (**PLS**) model was developed to predict actual DIM, using the strategy 3 data set. Spectra from a random sample of 80% of herds were assigned as a training data set (n = 724,864) to develop the model with 10 repeats of 10-fold cross-validation, using the caret package in R (Kuhn, 2008). The remaining spectra (n = 187,870) were used for validation, comprising 145,014 pregnant and 42,856 nonpregnant records. A secondary set of models were also developed and validated, whereby the model was trained on only records classified as pregnant (n = 555,318). For each model, performance was assessed according to the relative root mean square error (**RMSE**) between actual and predicted DIM,

and according to the correlation between actual and predicted DIM in the validation data sets.

RESULTS AND DISCUSSION

Data Description

Table 1 shows the number of records and cows by pregnancy status for each classification strategy, and the mean DIM values for records in each class. A large difference was observed between the average DIM of nonpregnant and pregnant records for strategy 1 and strategy 2, with values of 55 to 89 for nonpregnant records, compared with 170 to 171 for pregnant records. This difference was smaller for strategy 3, where the average DIM for nonpregnant records was 150, compared with 170 to 171 for pregnant records; and in the external PAG validation data set the average DIM for nonpregnant and pregnant records were 176 and 187, respectively. These differences in the distribution of DIM for records in each pregnancy status group are further demonstrated in Figure 2. Figures 2a–c show the DIM distribution of records for the training data set of each strategy. The distribution of DIM for the VAL-PAG data set are shown in Figure 2d. The distributions for strategy 1 and strategy 2 were similar in the early stages of lactation, in that there was a good representation of nonpregnant records, but beyond ~120 d, there were few nonpregnant records (Figure 2a and b). Strategy 3 differed in that there was representation of both pregnancy classifications across lactation (Figure 2c). Similarly, in the VAL-PAG data set, both pregnancy classifications were well represented across lactation (Figure 2d).

Diagnosis of Pregnancy Status Using PLS-DA Models

In this study, we compared 3 strategies for selecting FT-MIR spectral records for analysis and partitioning records into nonpregnant and pregnant groups. Table 2 shows prediction accuracies for PLS-DA models within the training, VAL-Test and VAL-PAG data sets for each strategy and model type. For each strategy, models that used unadjusted FT-MIR spectra as predictors outperformed the prediction accuracy of models that used spectra that had been pre-adjusted for DIM. Boxplots representing prediction probabilities for nonpregnant and pregnant records using unadjusted FT-MIR spectra for each strategy are provided in Figure 3.

Strategy 1 was comparable to the approach used by Brand et al. (2021), where only cows with a subsequent calving were included, with records before the first mating assigned as nonpregnant, and records after a validated AI event assigned as pregnant. Brand et al. (2021) reported promising predictive ability to classify pregnancy status in a large data set of FT-MIR spectra from UK herds using PLS-DA models, with accuracy, sensitivity, and specificity values of 0.77, 0.73, and 0.82, respectively. In our study, the model using unadjusted FT-MIR spectra with strategy 1 data had sensitivity of 0.94, specificity of 0.96, and AUC values of 0.99 for the VAL-Test data set (Table 2), higher than those reported by Brand et al. (2021). However, for this model, the specificity to correctly classify nonpregnant cows in the VAL-PAG data set was poor (0.002). This lack of consistency in prediction accuracy across the training and validation data sets for the strategy 1 data set is clearly demonstrated in Figure 3a–c, where we

Table 1. Total number of records and cows, and descriptive statistics (mean ± SD) for DIM across pregnancy classification strategies for training and validation data sets

	Pregnant			Nonpregnant		
Data set[1]	No. of records	No. of cows	DIM	No. of records	No. of cows	DIM
Strategy 1						
Training data	552,263	440,083	170 (56.0)	131,056	130,167	55 (20.8)
Test validation	148,069	116,048	171 (56.1)	33,481	33,170	55 (20.5)
Strategy 2						
Training data	560,215	444,687	170 (56.0)	11,781	11,746	89 (22.0)
Test validation	140,117	111,407	170 (56.0)	2,997	2,987	89 (22.1)
Strategy 3						
Training data	557,440	443,574	170 (55.9)	167,945	141,407	150 (56.5)
Test validation	142,892	112,552	171 (56.6)	44,457	36,912	150 (57.6)
PAG validation (VAL-PAG)	22,117	22,068	187 (36.7)	1,946	1,936	176 (39.3)

[1]For all strategies, records defined as pregnant if there was a validated AI event and a subsequent calving (n = 700,332 records). Nonpregnant records defined for each strategy as follows: (1) Records before the first mating assigned as nonpregnant (n = 164,537); (2) records after the first mating but before the validated AI event assigned as nonpregnant (n = 14,778); and (3) in addition to nonpregnant records used in strategy 2, records for cows without a subsequent calving assigned as nonpregnant (n = 197,624). PAG validation (VAL-PAG) = pregnancy-associated glycoproteins validation data set.

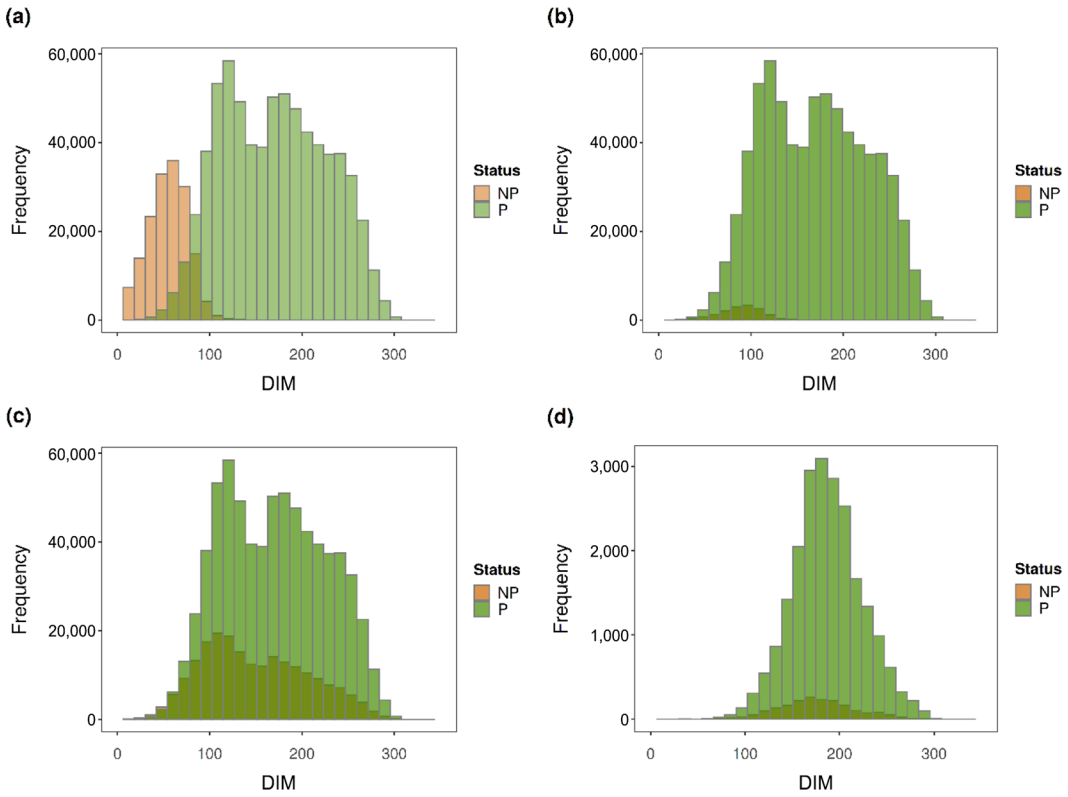

Figure 2. Frequency of pregnant (P) and nonpregnant (NP) records across DIM for training and validation records for (a) strategy 1; (b) strategy 2; (c) strategy 3; and (d) pregnancy-associated glycoproteins (PAG) records. For all strategies, records defined as P if there was a validated AI event and a subsequent calving (n = 700,332 records). Records were defined as NP for each strategy as follows: (1) records before the first mating (n = 164,537); (2) records after the first mating but before the validated AI event (n = 14,778); and (3) in addition to nonpregnant records used in strategy 2, records for cows without a subsequent calving were assigned as nonpregnant (n = 197,624).

observed good partitioning between the distribution of prediction probabilities in the training and VAL-Test data sets (Figure 3a and b), but a tendency to predict nonpregnant records as pregnant in the VAL-PAG data set (Figure 3c). When DIM was included as a predictor for strategy 1 models, accuracies in the VAL-Test and VAL-PAG data sets were relatively unchanged, compared with fitting FT-MIR spectra alone (Table 2). However, pre-adjusting spectra for DIM resulted in prediction accuracies that were more consistent across training and validation data sets, with sensitivity of 0.66, specificity of 0.61 and AUC values of 0.68 in the VAL-Test data set; and 0.63, 0.42, and 0.53 in the VAL-PAG data set, respectively. Although the training and VAL-Test accuracies were lower in the model that used spectra pre-adjusted for DIM, the improved consistency in results across training and both validation data sets indicated that using pre-adjusted spectra was at least partially effective at removing some of the confounding effect between stage of lactation and pregnancy status.

Strategy 2 for classifying pregnancy status was comparable to the approach used by Toledo-Alvarado et al. (2018), whereby only cows with a subsequent calving were included, and records between the first mating and a validated AI event were assigned as nonpregnant, and records after a validated AI event were assigned as pregnant. Toledo-Alvarado et al. (2018) classified pregnancy status using FT-MIR spectra from cattle raised in heterogeneous farming systems in northeastern Italy using a Bayesian model, and reported cross-validation AUC values of ~0.6 to 0.66. In our study, strategy 2 models using unadjusted spectra had comparatively higher AUC values for the VAL-Test data set (0.91 to 0.93). However, in the VAL-PAG data set these dropped to between 0.57 and 0.59, and the specificity to correctly classify nonpregnant cows in the VAL-PAG data set was poor (0.02–0.04). Similar to the observations for strategy 1, we observed good partitioning between the distribution of prediction probabilities in the training and VAL-Test data sets (Figure 3d and e), but a tendency to predict nonpregnant records as pregnant in the VAL-PAG data set (Figure 3f). A small improvement was observed in the specificity to correctly classify nonpregnant records in the VAL-PAG data set when spectra were pre-adjusted for stage of lactation, however the AUC value for this model was still low

Table 2. Model performance for partial least squares discriminant analysis models with upsampling[1]: accuracy (Acc), sensitivity (Sens), specificity (Spec), and area under the receiver operating characteristic curve (AUC) values within the training, herd-independent validation (VAL-Test) and pregnancy-associated glycoproteins validation (VAL-PAG) data sets

	Training				Test validation (VAL-Test)				Glycoprotein-based validation (VAL-PAG)			
Classification strategy[2] and model[3]	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
Strategy 1												
FT-MIR spectra	0.938	0.932	0.966	0.987	0.941	0.936	0.961	0.987	0.918	0.998	0.002	0.559
FT-MIR spectra + DIM	0.940	0.934	0.966	0.991	0.946	0.940	0.972	0.993	0.917	0.997	0.014	0.544
FT-MIR spectra (pre-adjusted for DIM)	0.672	0.675	0.660	0.723	0.654	0.664	0.606	0.676	0.613	0.630	0.421	0.532
Strategy 2												
FT-MIR spectra	0.807	0.805	0.931	0.922	0.801	0.799	0.906	0.914	0.912	0.990	0.018	0.572
FT-MIR spectra + DIM	0.800	0.797	0.943	0.932	0.794	0.791	0.923	0.926	0.907	0.984	0.037	0.585
FT-MIR spectra (pre-adjusted for DIM)	0.716	0.716	0.709	0.770	0.705	0.708	0.604	0.701	0.734	0.776	0.262	0.523
Strategy 3												
FT-MIR spectra	0.596	0.594	0.603	0.637	0.599	0.600	0.596	0.636	0.665	0.673	0.571	0.668
FT-MIR spectra + DIM	0.617	0.626	0.588	0.649	0.618	0.628	0.586	0.649	0.697	0.711	0.536	0.677
FT-MIR spectra (pre-adjusted for DIM)	0.568	0.567	0.571	0.596	0.568	0.572	0.554	0.589	0.573	0.569	0.622	0.639

[1]Upsampling undertaken using random sampling with replacement to augment the minority class (nonpregnant) to be the same size as the majority class (pregnant).

[2]For all strategies, records defined as pregnant if there was a validated AI event and a subsequent calving (n = 700,332 records). Nonpregnant records defined for each strategy as follows: (1) Records before the first mating assigned as nonpregnant (n = 164,537); (2) records after the first mating but before the validated AI event assigned as nonpregnant (n = 14,778); and (3) in addition to nonpregnant records used in strategy 2, records for cows without a subsequent calving assigned as nonpregnant (n = 197,624).

[3]Fourier-transform mid-infrared (FT-MIR) spectra models use spectral wavenumbers as predictors only; FT-MIR spectra + DIM models use spectral wavenumbers and DIM (30-d window class) as predictors; FT-MIR spectra (pre-adjusted for DIM) models use spectral wavenumbers pre-adjusted for DIM (30-d window class).

Table 3. Model performance for multilayer perceptron (MLP) and convolutional neural network (CNN) approaches based on strategy 3 data[1]: accuracy (Acc), sensitivity (Sens), specificity (Spec), and area under the receiver operating characteristic curve (AUC) values within the training, herd-independent validation (VAL-Test) and pregnancy-associated glycoproteins validation (VAL-PAG) data sets

	Training				Test validation (VAL-Test)				Glycoprotein-based validation (VAL-PAG)			
Deep learning approach[2] and model[3]	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
MLP approach												
FT-MIR spectra	0.592	0.574	0.611	0.628	0.586	0.580	0.607	0.632	0.664	0.672	0.569	0.669
FT-MIR spectra + DIM	0.594	0.621	0.566	0.631	0.614	0.629	0.564	0.635	0.692	0.709	0.499	0.647
FT-MIR spectra (pre-adjusted for DIM)	0.559	0.554	0.564	0.583	0.562	0.567	0.547	0.581	0.554	0.547	0.636	0.636
CNN approach												
FT-MIR spectra	0.625	0.625	0.625	0.675	0.611	0.620	0.582	0.641	0.684	0.696	0.554	0.676
FT-MIR spectra + DIM	0.645	0.670	0.620	0.700	0.636	0.659	0.563	0.654	0.723	0.741	0.519	0.685
FT-MIR spectra (pre-adjusted for DIM)	0.982	0.975	0.988	0.998	0.668	0.790	0.273	0.551	0.759	0.805	0.266	0.564

[1]Strategy 3: Records defined as pregnant if there was a validated AI event and a subsequent calving (n = 700,332 records). Records after the first mating but before a validated AI event assigned as nonpregnant (n = 14,778); records for cows without a subsequent calving assigned as nonpregnant (n = 197,624).

[2]The multilayer perceptron (MLP) feed-forward artificial neural network classified pregnancy status based on raw spectra; the convolution neural network (CNN) used an image-based approach to classify pregnancy status.

[3]Fourier-transform mid-infrared (FT-MIR) spectra models use spectral wavenumbers as predictors only; FT-MIR spectra + DIM models use spectral wavenumbers and days in milk (30-d window class) as predictors; FT-MIR spectra (pre-adjusted for DIM) models use spectral wavenumbers pre-adjusted for DIM (30-d window class).

(0.52). The lack of consistency in prediction accuracies across training and validation data sets for strategy 2 were similar to those for strategy 1, indicating a lack of robustness in the models, likely due to confounding between pregnancy status and stage of lactation in the training data set, and a lack of representation of nonpregnant and pregnant records across lactation.

Prediction accuracies for strategy 3 models were relatively consistent across the training and validation data sets for unadjusted and adjusted spectra. For models including unadjusted spectra only, sensitivity was 0.67, specificity was 0.57, and AUC values were 0.67 for the VAL-PAG data set. Unlike the other 2 strategies, we did not observe clear partitioning between the distribution of prediction probabilities in the training and VAL-Test data sets (Figure 3g and h), however, the observed trend was consistent in the VAL-PAG data set (Figure 3i). For models using unadjusted spectra that also included DIM as a predictor, AUC values in the VAL-PAG data set increased from 0.67 to 0.68. Notably, a decline in prediction accuracy was observed for models that used spectra pre-adjusted for DIM, with overall accuracy and sensitivity dropping to 0.57, and AUC values dropping to 0.64. Strategy 3 models were comparable to the approach used by Delhez et al. (2020), whereby records for cows were included (and assigned as nonpregnant) if they did not have a subsequent calving. Delhez et al. (2020) classified pregnancy status in a data set of Australian Holstein cattle using PLS-DA models with FT-MIR spectra as independent predictors, and observed validation AUC values of 0.63 to 0.65. They also examined the effect of using residual spectra, evaluated as the difference between a nonpregnant record and pregnant record for the same animal, but did not see an improvement in results. However, they did observe an improvement in prediction accuracy for models developed from spectra in different stages of lactation, particularly for spectra recorded after 150 d of lactation, with validation AUC values of 0.76 to 0.78. We also undertook a similar approach for the strategy 3 data set whereby we fitted separate models for different stages of lactation (Appendix Table A1) and observed a consistent increase in AUC values after 210 d of lactation in the VAL-PAG data set (0.68–0.76), but the overall prediction accuracy after 210 d of lactation remained low (0.55–0.64).

Prediction accuracies for PLS-DA models where the majority class (pregnant) was reduced to be the same size as the minority class (nonpregnant) are shown in Appendix Table A2. Prediction accuracy metrics across all strategies and model types were only marginally different to those presented in Table 1. This indicated that augmenting the minority class to be the same size as the majority class did not introduce bias to the results. Moreover, it indicated that the reduced data set where the majority class was downsampled to be the same size as the minority class, sufficiently captured the extent of the relationships between spectral wavenumbers and pregnancy classification.

Prediction accuracies for PLS-DA models based on unadjusted spectra wavenumbers as predictors are shown in Appendix Table A3, whereby records classified as pregnant were removed if the test date was within 7, 14, or 21 d of successful AI. The premise for this analysis was that changes in an animal's physiological status might not be detectable in milk composition for some time after pregnancy is established, and that including those records could lower prediction accuracy. Of the total 700,332 pregnant records, 22,338 had a test date within 7 d after a validated AI event; 48,573 had a test date within 14 d after a validated AI event; and 81,581 head a test date within 21 d after a validated AI event. For all strategies, prediction accuracy metrics were relatively unchanged by removing these records (Appendix Table A3). This indicated that although there may be changes in the physiological status of an animal that are not detectable in milk composition shortly after successful AI, including and classifying these records as pregnant did not affect pregnancy prediction accuracies, compared with completely ignoring those records.

Diagnosis of Pregnancy Status Using Deep Learning Models

Deep learning is a subclass of machine learning that uses neural networks with multiple layers to extract features from data. These neural networks consist of densely interconnected processing nodes arranged into layers, with each node receiving information from nodes in the layer beneath it and sending data to the nodes in the layer above it. The complexity of these networks enable training models to be developed on data sets with multiple connections, making them a good choice for managing high-dimensional data sets such as those presented from FT-MIR spectra. Previous studies have established that it is possible to use artificial neural networks to identify features in spectra relating to pregnancy status (Brand et al., 2018), and that this could be extended to predict bovine tuberculosis status of individual cows (Denholm et al., 2020). More recently, Brand et al. (2021) assessed the accuracy of predicting pregnancy status using a deep learning image-based approach with a pretrained dense convolutional network (DenseNet), compared with a PLS-DA approach. In that study, when a deep learning image-based approach was employed, they observed an increase in sensitivity

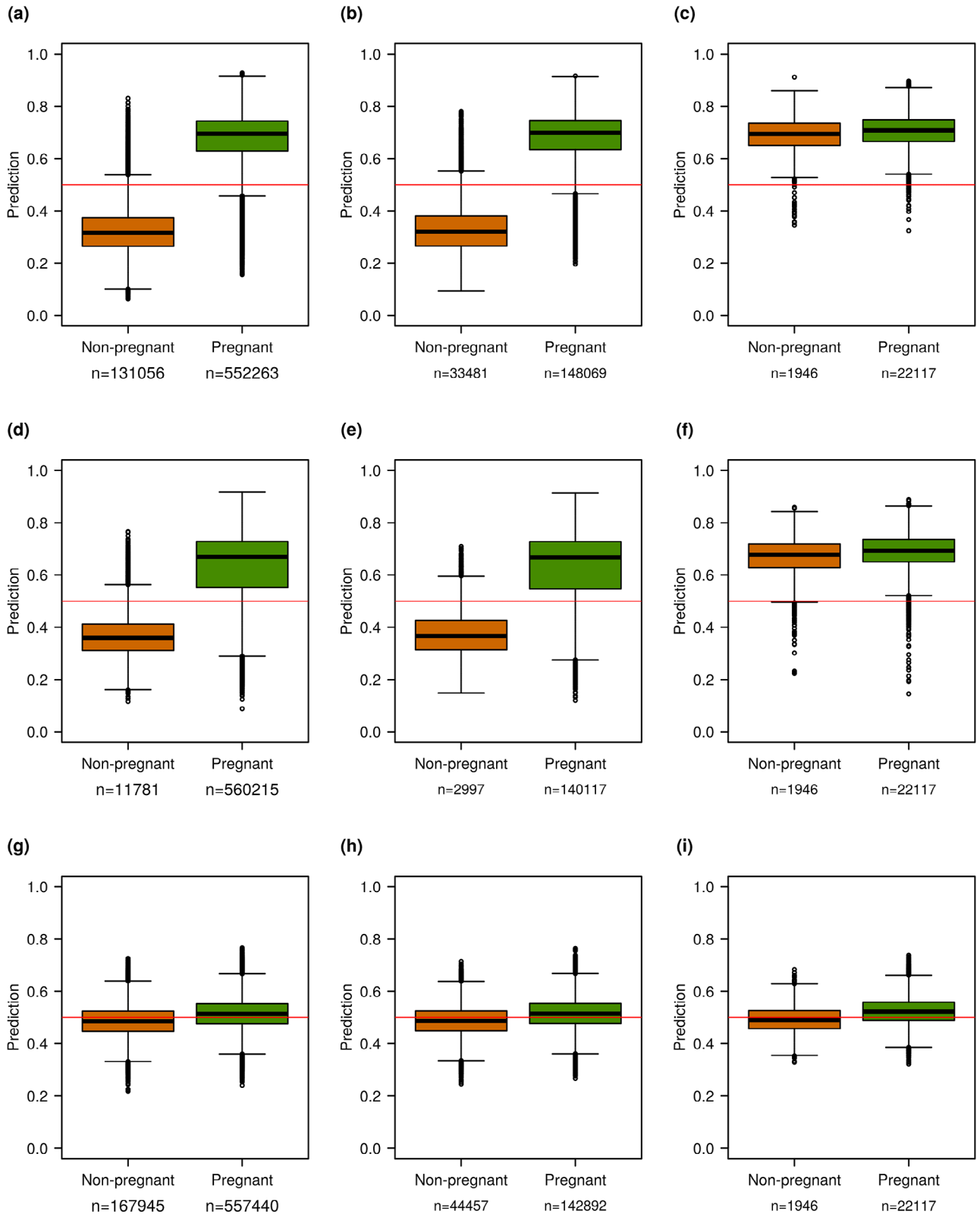

Figure 3. Summary of prediction probabilities for nonpregnant and pregnant records for training and validation data sets based on differing strategies for record selection and pregnancy status classification. Panels show (a) strategy 1: training; (b) strategy 1: herd-independent validation (VAL-Test); (c) strategy 1: external validation data set based on pregnancy-associated glycoproteins (VAL-PAG); (d) strategy 2: training; (e) strategy 2: VAL-Test; (f) strategy 2: VAL-PAG; (g) strategy 3: training; (h) strategy 3: VAL-Test; and (i) strategy 3: VAL-PAG.

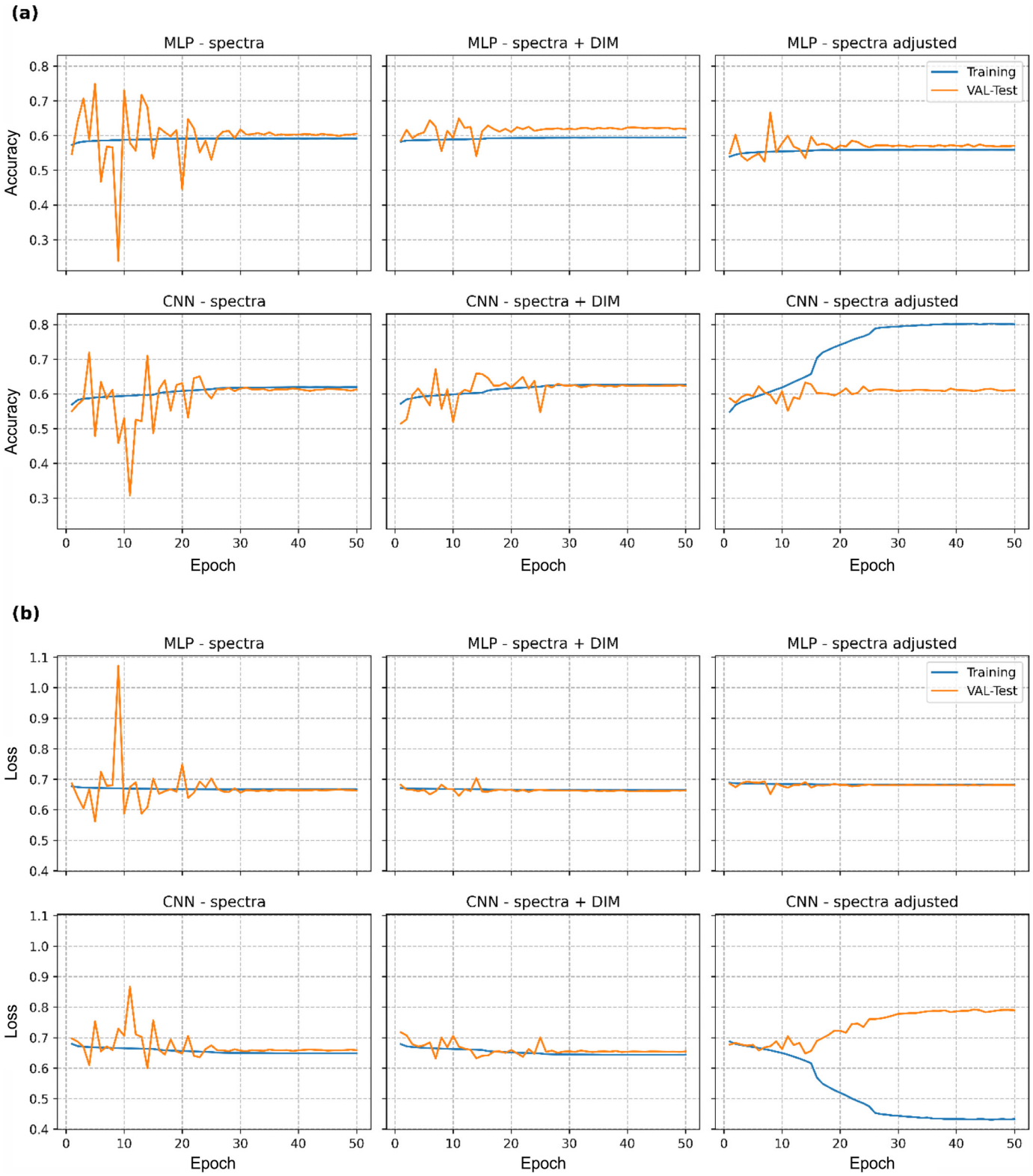

Figure 4. Accuracy (a) and loss (b) values for deep learning approaches, assessed using training and herd-independent validation (VAL-Test) data sets. MLP = multilayer perceptron; CNN = convolutional neural network.

from 0.77 to 0.88, an increase in specificity from 0.73 to 0.89, and an increase in the AUC value from 0.82 to 0.89.

In this study, we assessed pregnancy status prediction accuracies for 2 deep learning approaches, and compared these to the accuracies achieved from PLS-DA models. The first approach used a simple MLP with one hidden layer, using 4,600 parameters, whereas the second imaged-based CNN approach was significantly more complex with up to 7.4 million parameters. Convolutional neural networks are widely used in the computer vision domain and achieve the best performance when applied on image inputs. This type of architecture can efficiently extract local features, patterns and textures which are very common in natural images generated from optical cameras. In this study, adjacent spectral wavenumbers also present high correlations and thus may contain local patterns that a CNN model can learn and extract. In contrast to the Brand et al. (2021) study, we did not apply transfer learning as we deemed this unnecessary because most pretrained DenseNet networks are trained on images generated from optical cameras, (e.g., ImageNet), whereas our spectral data were acquired from a different physical process, namely FT-MIR spectroscopy, and we use a large training data set with more than 700,000 samples.

Training and validation accuracy and loss values for deep learning approaches are shown in Figure 4. Accuracies for the VAL-Test data set stabilized gradually after epochs 15, 25, and 35, corresponding to the reduction of the learning rate. In the case of our CNN trained on adjusted spectra, the model started overfitting from around epoch 10, with an increasing accuracy of the training data set, whereas accuracy for the VAL-Test data set stayed the same. This was despite the usage of batch normalization for regularization. Prediction accuracies for the MLP and CNN approaches within the training, VAL-Test, and VAL-PAG data sets are shown in Table 3. For each approach, the best performing models used unadjusted spectra and DIM as predictors. Across all models, prediction accuracies for the MLP approach were similar to those from PLS-DA models, however we observed a marginal increase in prediction accuracy for the models using an image-based CNN approach. For the image-based model that used unadjusted spectra and DIM as predictors, the overall prediction accuracy for the external PAG validation data set increased from 0.70 to 0.72, sensitivity increased from 0.71 to 0.74, and the AUC value increased from 0.68 to 0.69, but the specificity decreased from 0.54 to 0.52. Notably, gains in prediction accuracy from adopting a deep learning approach were lower than those previously reported by Brand et al. (2021).

Prediction Models for Stage of Lactation

To understand the magnitude of the effect that stage of lactation may be having on pregnancy prediction, we used the strategy 3 data set to develop and validate a PLS model for predicting DIM from FT-MIR spectra. Records from a random sample of 80% of herds were used as a training data set to develop the prediction model, with the remaining strategy 3 records used as herd-independent validation data sets for pregnant and nonpregnant records, respectively. The relationships between predicted and actual DIM values for the training data set and 2 validation data sets are shown in Figure 5. Consistently high correlations between actual and predicted DIM were observed across the training and validation data sets when the DIM prediction model was developed across all (pregnant and nonpregnant) records (0.89–0.90; Figure 5a–c). When only records assigned as pregnant were used to develop the DIM prediction model, correlations between actual and predicted DIM remained high (0.89–0.90; Figures 5d–f). Relative RMSE values for the validation data set of pregnant records dropped marginally from 14.5 to 14.4% when only records assigned as pregnant were used to develop the model (Figure 5b and e), whereas relative RMSE values for the validation data set of nonpregnant records increased marginally from 17.3 to 17.6% when only records assigned as pregnant were used to develop the model (Figure 5c and f). These marginal shifts in validation prediction accuracy when models were developed on all records versus only records assigned as pregnant, highlighted that in pasture-based seasonal calving systems the underlying relationship between FT-MIR spectra and DIM was upheld regardless of pregnancy status.

Confounding Between Pregnancy Status and Stage of Lactation Effects

Changes in dairy cattle milk composition across lactation are more noticeable in seasonal pasture-based farming systems (compared with nonseasonal systems), where compact calving periods are used so that peak lactation volumes are matched with peak grass growth (Timlin et al., 2021). Although NZ dairy systems are mainly pasture-based, intensification has resulted in widespread use of supplement feed to offset the effect of high-stocking rates and ensure that the nutritional requirements of cows are met. In particular, there has been an increased use of palm kernel extract and maize silage as supplements in NZ dairy systems over the last 2 decades (Ministry for Primary Industries NZ, 2017). Palm kernel extract is associated with an increase in

milk fat content (DairyNZ, 2017; Van Wyngaard and Meeske, 2017) and changes to milk fatty acid composition (Dias, 2010; Oliveira et al., 2015) and has resulted in the introduction of a Fat Evaluation Index by Fonterra in 2018 to assess the suitability of milk for processing (DairyNZ, 2017). More generally, fatty acids have been shown to change with different dietary systems (Elgersma, 2015) and levels of pasture in the diet (White et al., 2001; Couvreur et al., 2006; Butler et al., 2011; O'Callaghan et al., 2016). Nevertheless, across different diets, as lactation progresses, consistently lower milk volumes (McAuliffe et al., 2016) and higher concentrations of fat and CP have been reported (O'Callaghan et al., 2016). This can be problematic for the prediction of indirect traits such as pregnancy status from FT-MIR spectra, particularly when the spectra are from seasonal calving herds, because as lactation progresses, changes in milk composition coincide with the advent of a cow becoming pregnant, and most cows do become pregnant. In seasonal calving pasture-based systems, there are also other changes that are confounded with lactation stage such as climatic changes and the use of dietary supplements to ensure that the energy requirements of cows are met at times when pasture growth is low. It is thus important to ensure that pregnancy prediction models based on FT-MIR spectra include a consistent representation of pregnant and nonpregnant records across all stages of lactation. In this study, the most robust prediction accuracies were achieved when the prediction model was developed on a data set with a good representation of pregnant and nonpregnant records across lactation. Contrary to this, when we devel-

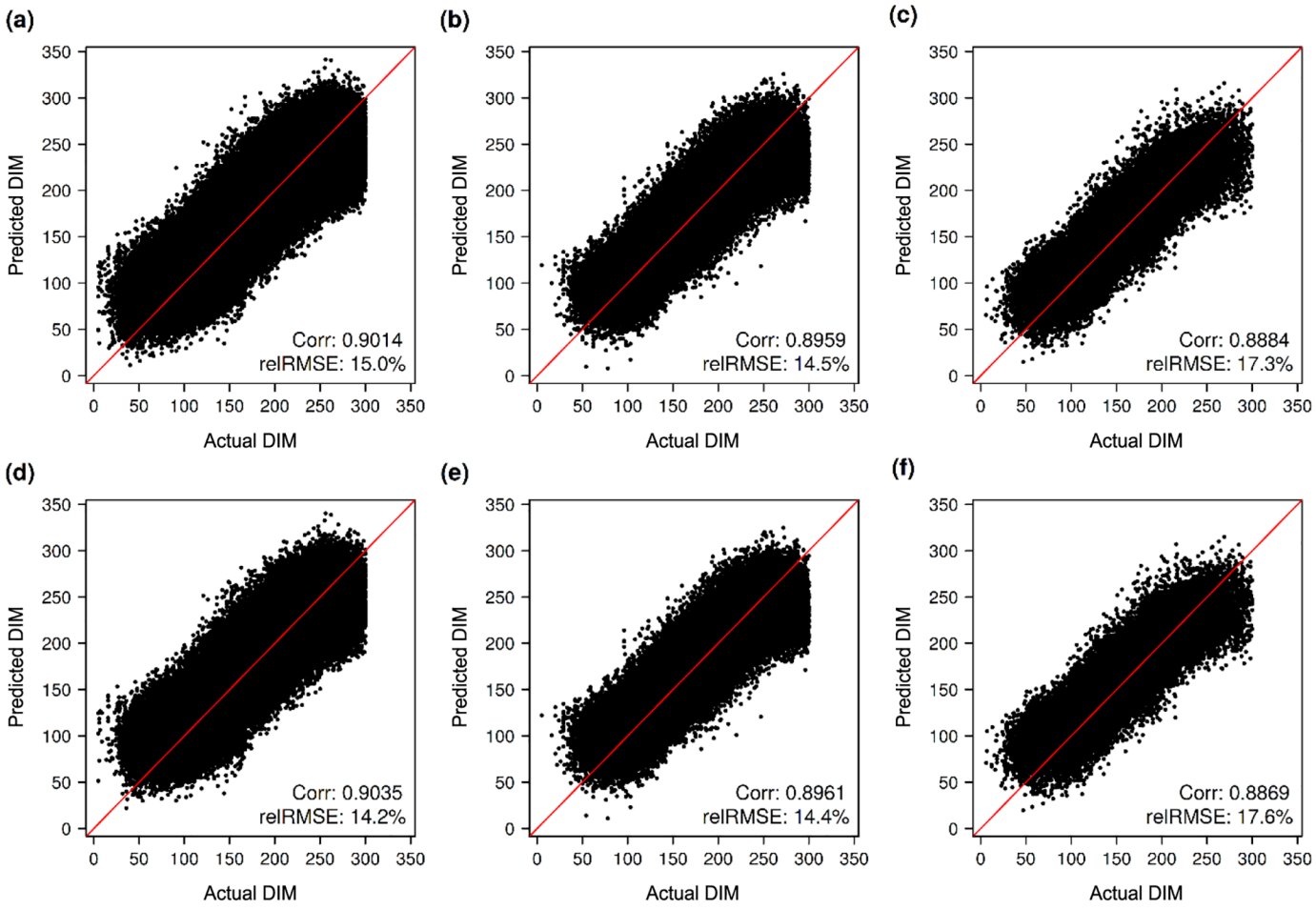

Figure 5. Summary of predicted versus actual DIM for training and validation data sets from partial least squares prediction models based on pregnant and nonpregnant records for (a) training data (n = 724,864), (b) validation: pregnant (N = 145,014), and (c) validation: nonpregnant (n = 42,856); and prediction models based on pregnant records only for (d) training data (n = 555,318), (e) validation: pregnant (n = 145,014), and (f) validation: nonpregnant (n = 42,856). Continuous red lines represent y = x. Corr. is the correlation between actual and predicted DIM; relRMSE is the relative root mean square error between actual and predicted DIM.

oped prediction models on data sets that did not have a good representation of pregnant and nonpregnant records across lactation, prediction accuracies appeared promising in initial validation, but did not perform well in the external PAG validation data set. In future research, to improve prediction accuracies for pregnancy status from FT-MIR spectra within a seasonal calving pasture-based context, it is important that careful consideration is given to how models can account for the confounding effects of factors such as lactation stage, feed management and seasonality. Although some of this could be addressed by including multiple seasons of data, including other information such as knowledge of feed management and supplementation may also play an important role.

Prediction Model Validation Strategies

This is not the first study to highlight differences in prediction accuracy for FT-MIR predicted traits, depending on the validation data set and strategy used. Wang and Bovenhuis (2019) observed that using a random cross-validation approach to predict methane (CH_4) emissions from FT-MIR spectra resulted in overoptimistic results. Other studies show that prediction accuracies can be inflated by the split-data strategy used for validation, with cow-independent validation having lower accuracies compared with record-independent cross-validation (Shetty et al., 2017; Smith et al., 2019), and trial- or herd-independent validation having lower accuracies compared with record- or cow-independent validation (Dórea et al., 2018; Lahart et al., 2019; Luke et al., 2019). Recently, Bresolin and Dórea (2020) have reviewed the effect of validation strategies on predictive quality for a range of FT-MIR predicted milk composition and animal health traits, and highlight the value of an external validation data set whereby the external validation uses data from a different herd, trial, or season. In 67 of the 113 studies they reviewed, internal validation (holdout, leave-one-out, k-fold) was performed. Of the 32 papers they reviewed that used an external validation, only 17 conducted validation using an independent data set based on herd, trial, or season, whereas the other 15 used cow-independent validation. In our study, we demonstrate that in some instances, even a herd-independent validation approach can overestimate prediction accuracies, if there is systemic confounding between the trait of interest and other underlying factors in the FT-MIR spectra. Specifically, where there were divergent DIM characteristics between pregnancy status groups, we were not only predicting changes in milk composition due to pregnancy, but also changes due to other factors.

CONCLUSIONS

We have assessed and compared pregnancy prediction accuracy from FT-MIR spectra using different strategies for classifying pregnancy status and accounting for the effect of stage of lactation. We have also compared prediction accuracies from PLS-DA models to alternative models developed using deep learning approaches. We have shown that the ability to predict pregnancy status from FT-MIR spectra is influenced by which records are used, and how these records are partitioned into pregnant and nonpregnant groups. Prediction models developed on data sets without adequate representation of nonpregnant and pregnant status across lactation, led to misleading results, whereby prediction accuracies were high in the training and herd-independent validation data set, but were not upheld for an external validation data set where pregnancy status was assigned according to PAG in milk samples. This demonstrated that even with herd-independent validation, prediction accuracies can be misleading where there is systematic confounding between pregnancy status and other factors such as stage of lactation. For models where the effect of this confounding was reduced, prediction accuracies were not sufficiently high to be used as a sole indicator of pregnancy status within a seasonal calving herd management context.

ACKNOWLEDGMENTS

The authors acknowledge Livestock Improvement Corporation (Hamilton, New Zealand) herd-testing staff for the processing and analysis of milk samples. Kathryn also acknowledges and thanks the wider Livestock Improvement Corporation research and development team and fellow students for underlying technical support and thoughtful discussion, and Tod Schilling (Bentley Instruments Inc., Chaska, MN) and Pierre Broutin (Bentley Instruments Inc., Lille, France) for their help with obtaining FT-MIR spectra from Bentley instruments. With gratitude, we also recognize the use of New Zealand eScience Infrastructure (NeSI) high-performance computing for this research. The funding for this research was provided by Livestock Improvement Corporation (Hamilton, New Zealand) and the New Zealand Ministry for Primary Industries (Wellington, New Zealand), within the Resilient Dairy Programme through Sustainable Food and Fibre Futures (Funding No: PGP06-17006).

REFERENCES

Brand, W., A. T. Wells, and M. Coffey. 2018. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. J. Dairy Sci. 101(Suppl. 2):347. (Abstr

Brand, W., A. T. Wells, S. L. Smith, S. J. Denholm, E. Wall, and M. P. Coffey. 2021. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. J. Dairy Sci. 104:4980–4990. https://doi.org/10.3168/jds.2020-18367.

Bresolin, T., and J. R. R. Dórea. 2020. Infrared spectrometry as a high-throughput phenotyping technology to predict complex traits in livestock systems. Front. Genet. 11:923. https://doi.org/10.3389/fgene.2020.00923.

Butler, D. G., B. R. Cullis, A. R. Gilmour, B. J. Gogel, and R. Thompson. 2009. ASReml-R reference manual (version 3). The State of Queensland, Department of Primary Industries and Fisheries: Brisbane, Queensland.

Butler, G., J. H. Nielsen, M. K. Larsen, B. Rehberger, S. Stergiadis, A. Canever, and C. Leifert. 2011. The effects of dairy management and processing on quality characteristics of milk and dairy products. NJAS Wagening. J. Life Sci. 58:97–102. https://doi.org/10.1016/j.njas.2011.04.002.

Commun, L., K. Velek, J.-B. Barbry, S. Pun, A. Rice, A. Mestek, C. Egli, and S. Leterme. 2016. Detection of pregnancy-associated glycoproteins in milk and blood as a test for early pregnancy in dairy cows. J. Vet. Diagn. Invest. 28:207–213. https://doi.org/10.1177/1040638716632815.

Couvreur, S., C. Hurtaud, C. Lopez, L. Delaby, and J. L. Peyraud. 2006. The linear relationship between the proportion of fresh grass in the cow diet, milk fatty acid composition, and butter properties. J. Dairy Sci. 89:1956–1969. https://doi.org/10.3168/jds.S0022-0302(06)72263-9.

DairyNZ. 2017. Palm Kernel Extract (PKE). Accessed July 19, 2021. https://www.dairynz.co.nz/feed/supplements/palm-kernel-extract-pke/.

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. https://doi.org/10.3168/jds.2019-17473.

Denholm, S. J., W. Brand, A. P. Mitchell, A. T. Wells, T. Krzyzelewski, S. L. Smith, E. Wall, and M. P. Coffey. 2020. Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning. J. Dairy Sci. 103:9355–9367. https://doi.org/10.3168/jds.2020-18328.

Dias, F. N. 2010. Supplementation of palm kernel expeller to grazing dairy farms in New Zealand. PhD thesis. Massey University, Palmerston North, New Zealand.

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. https://doi.org/10.3168/jds.2017-13997.

Elgersma, A. 2015. Grazing increases the unsaturated fatty acid concentration of milk from grass-fed cows: A review of the contributing factors, challenges and future perspectives. Eur. J. Lipid Sci. Technol. 117:1345–1369. https://doi.org/10.1002/ejlt.201400469.

Ferguson, J. D., and D. T. Galligan. 2011. The value of pregnancy diagnosis – A revisit to an old art. Clin. Theriogenology 3:559–578.

Giordano, J. O., P. M. Fricke, and V. E. Cabrera. 2013. Economics of resynchronization strategies including chemical tests to identify nonpregnant cows. J. Dairy Sci. 96:949–961. https://doi.org/10.3168/jds.2012-5704.

Green, J. A., T. E. Parks, M. P. Avalle, B. P. Telugu, A. L. McLain, A. J. Peterson, W. McMillan, N. Mathialagan, R. R. Hook, S. Xie, and R. M. Roberts. 2005. The establishment of an ELISA for the detection of pregnancy-associated glycoproteins (PAGs) in the serum of pregnant cows and heifers. Theriogenology 63:1481–1503. https://doi.org/10.1016/j.theriogenology.2004.07.011.

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

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. https://doi.org/10.3168/jds.2014-8984.

Huang, G., Z. Liu, L. van der Maaten, and K. Q. Weinberger. 2018. Densely connected convolutional networks. Pages 4700–4708 in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Kuhn, M. 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28:1–26. https://doi.org/10.18637/jss.v028.i05.

Lahart, B., S. McParland, E. Kennedy, T. M. Boland, T. Condon, M. Williams, N. Galvin, B. McCarthy, and F. Buckley. 2019. Predicting the dry matter intake of grazing dairy cows using infrared reflectance spectroscopy analysis. J. Dairy Sci. 102:8907–8918. https://doi.org/10.3168/jds.2019-16363.

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

Loker, S., F. Miglior, J. Bohmanova, J. Jamrozik, and L. R. Schaeffer. 2009. Phenotypic analysis of pregnancy effect on milk, fat, and protein yields of Canadian Ayrshire, Jersey, Brown Swiss, and Guernsey breeds. J. Dairy Sci. 92:1300–1312. https://doi.org/10.3168/jds.2008-1425.

Luke, T. D. W., S. Rochfort, W. J. Wales, V. Bonfatti, L. Marett, and J. E. Pryce. 2019. Metabolic profiling of early-lactation dairy cows using milk mid-infrared spectra. J. Dairy Sci. 102:1747–1760. https://doi.org/10.3168/jds.2018-15103.

McAuliffe, S., T. J. Gilliland, and D. Hennessy. 2016. Comparison of pasture-based feeding systems and a total mixed ration feeding system on dairy cow milk production. Page 289 in Sustainable Meat and Milk Production from Grassland. Proceedings of the 27th General Meeting of the European Grassland Federation. Teagasc Animal and Grassland Research and Innovation Centre.

Ministry for Primary Industries NZ. 2017. Feed Use in the NZ Dairy Industry. DairyNZ.

O'Callaghan, T. F., D. Hennessy, S. McAuliffe, K. N. Kilcawley, M. O'Donovan, P. Dillon, R. P. Ross, and C. Stanton. 2016. Effect of pasture versus indoor feeding systems on raw milk composition and quality over an entire lactation. J. Dairy Sci. 99:9424–9440. https://doi.org/10.3168/jds.2016-10985.

Oliveira, R., M. Faria, R. Silva, L. Bezerra, G. Carvalho, A. Pinheiro, J. Simionato, and A. Leão. 2015. Fatty acid profile of milk and cheese from dairy cows supplemented a diet with palm kernel cake. Molecules 20:15434–15448. https://doi.org/10.3390/molecules200815434.

Olori, V. E., S. Brotherstone, W. G. Hill, and B. J. McGuirk. 1997. Effect of gestation stage on milk yield and composition in Holstein Friesian dairy cattle. Livest. Prod. Sci. 52:167–176. https://doi.org/10.1016/S0301-6226(97)00126-7.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, ed. Curran Associates Inc.

Penasa, M., M. De Marchi, and M. Cassandro. 2016. Short communication: Effects of pregnancy on milk yield, composition traits, and coagulation properties of Holstein cows. J. Dairy Sci. 99:4864–4869. https://doi.org/10.3168/jds.2015-10168.

Ricci, A., P. D. Carvalho, M. C. Amundson, R. H. Fourdraine, L. Vincenti, and P. M. Fricke. 2015. Factors associated with pregnancy-associated glycoprotein (PAG) levels in plasma and milk of Holstein cows during early pregnancy and their effect on the accuracy of pregnancy diagnosis. J. Dairy Sci. 98:2502–2514. https://doi.org/10.3168/jds.2014-8974.

Ricci, A., P. D. Carvalho, M. C. Amundson, and P. M. Fricke. 2017. Characterization of luteal dynamics in lactating Holstein cows for 32 days after synchronization of ovulation and timed artificial insemination. J. Dairy Sci. 100:9851–9860. https://doi.org/10.3168/jds.2017-13293.

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

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

Sousa, N. M., A. Ayad, J. F. Beckers, and Z. Gajewski. 2006. Pregnancy-associated glycoproteins (PAG) as pregnancy markers in the ruminants. J. Physiol. Pharmacol. 57(Suppl 8):153–171.

Timlin, M., J. T. Tobin, A. Brodkorb, E. G. Murphy, P. Dillon, D. Hennessy, M. O'Donovan, K. M. Pierce, and T. F. O'Callaghan. 2021. The impact of seasonality in pasture-based production systems on milk composition and functionality. Foods 10:607. https://doi.org/10.3390/foods10030607.

Tiplady, K. M., R. G. Sherlock, M. D. Littlejohn, J. E. Pryce, S. R. Davis, D. J. Garrick, R. J. Spelman, and B. L. Harris. 2019. Strategies for noise reduction and standardization of milk mid-infrared spectra from dairy cattle. J. Dairy Sci. 102:6357–6372. https://doi.org/10.3168/jds.2018-16144.

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

Van Wyngaard, J. D. V., and R. Meeske. 2017. Palm kernel expeller increases milk fat content when fed to grazing dairy cows. S. Afr. J. Anim. Sci. 47:219–230. https://doi.org/10.4314/sajas.v47i2.14.

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. https://doi.org/10.3168/jds.2018-15684.

White, S. L., J. A. Bertrand, M. R. Wade, S. P. Washburn, J. T. Green Jr., and T. C. Jenkins. 2001. Comparison of fatty acid content of milk from Jersey and Holstein cows consuming pasture or a total mixed ration. J. Dairy Sci. 84:2295–2301. https://doi.org/10.3168/jds.S0022-0302(01)74676-0.

Wightman, R. 2019. PyTorch Image Models. GitHub Repos. https://doi.org/10.5281/zenodo.4414861.

ORCIDS

K. M. Tiplady https://orcid.org/0000-0002-3307-9208
S. R. Davis https://orcid.org/0000-0002-4942-1055
D. J. Garrick https://orcid.org/0000-0001-8640-5372

APPENDIX

Table A1. Record numbers and model performance for partial least squares discriminant analysis models fitted within stage of lactation classes for accuracy (Acc), sensitivity (Sens), specificity (Spec), and area under the receiver operating characteristic curve (AUC) values within the training, herd-independent validation (VAL-Test) and pregnancy-associated glycoproteins validation (VAL-PAG) data sets

Stage of lactation class	Training						Test validation (VAL-Test)						Glycoprotein-based validation (VAL-PAG)					
	Pregnant	Nonpregnant	Acc	Sens	Spec	AUC	Pregnant	Nonpregnant	Acc	Sens	Spec	AUC	Pregnant	Nonpregnant	Acc	Sens	Spec	AUC
5–30 d	127	261	0.956	0.937	0.966	0.989	10	52	0.677	0.700	0.673	0.688	0	0	—	—	—	—
31–60 d	4,315	4,783	0.581	0.582	0.581	0.616	1,189	1,351	0.543	0.573	0.517	0.554	1	0	—	—	—	—
61–90 d	32,034	20,045	0.576	0.581	0.569	0.608	8,420	5,583	0.549	0.553	0.541	0.564	77	28	0.590	0.597	0.571	0.666
91–120 d	96,794	36,670	0.582	0.589	0.563	0.607	24,176	9,913	0.575	0.595	0.525	0.583	661	124	0.508	0.474	0.694	0.648
121–150 d	93,889	29,094	0.602	0.608	0.583	0.632	23,618	7,399	0.601	0.618	0.549	0.617	2,616	346	0.478	0.436	0.792	0.679
151–180 d	88,318	25,074	0.606	0.610	0.593	0.640	21,742	6,380	0.578	0.581	0.566	0.606	6,243	574	0.614	0.615	0.605	0.662
181–210 d	92,083	23,008	0.609	0.612	0.596	0.644	24,433	5,966	0.598	0.601	0.584	0.626	7,104	522	0.656	0.665	0.536	0.640
211–240 d	76,476	16,610	0.618	0.620	0.611	0.660	19,683	4,233	0.600	0.606	0.572	0.623	3,593	220	0.588	0.579	0.736	0.681
240–270 d	59,902	10,199	0.624	0.623	0.633	0.678	15,487	2,890	0.608	0.610	0.596	0.640	1,476	124	0.554	0.539	0.726	0.691
271–300 d	13,502	2,201	0.674	0.670	0.701	0.747	4,134	690	0.631	0.638	0.587	0.652	346	8	0.644	0.645	0.625	0.759
Overall	557,440	167,945	0.604	0.609	0.588	0.639	142,892	44,457	0.589	0.599	0.555	0.609	22,117	1,946	0.600	0.595	0.647	0.669

Table A2. Model performance for partial least squares discriminant analysis models with downsampling[1] for accuracy (Acc), sensitivity (Sens), specificity (Spec), and area under the receiver operating characteristic curve (AUC) values within the training, herd-independent validation (VAL-Test), and pregnancy-associated glycoproteins validation (VAL-PAG) data sets

Classification strategy[2] and model[3]	Training				Test validation (VAL-Test)				Glycoprotein-based validation (VAL-PAG)			
	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
Strategy 1												
FT-MIR spectra	0.937	0.930	0.966	0.987	0.940	0.935	0.961	0.987	0.918	0.998	0.002	0.560
FT-MIR spectra + DIM	0.940	0.934	0.966	0.991	0.946	0.940	0.972	0.992	0.917	0.997	0.014	0.544
FT-MIR spectra (pre-adjusted for DIM)	0.671	0.673	0.661	0.723	0.651	0.661	0.608	0.676	0.608	0.624	0.432	0.531
Strategy 2												
FT-MIR spectra	0.803	0.800	0.930	0.920	0.798	0.795	0.909	0.912	0.911	0.990	0.020	0.572
FT-MIR spectra + DIM	0.796	0.793	0.945	0.929	0.792	0.789	0.925	0.923	0.907	0.982	0.046	0.584
FT-MIR spectra (pre-adjusted for DIM)	0.701	0.701	0.712	0.763	0.692	0.694	0.616	0.696	0.719	0.757	0.277	0.524
Strategy 3												
FT-MIR spectra	0.594	0.592	0.604	0.637	0.598	0.599	0.595	0.636	0.663	0.671	0.578	0.669
FT-MIR spectra + DIM	0.616	0.624	0.591	0.649	0.618	0.627	0.588	0.649	0.694	0.707	0.544	0.679
FT-MIR spectra (pre-adjusted for DIM)	0.566	0.564	0.573	0.596	0.567	0.571	0.554	0.588	0.571	0.566	0.639	0.643

[1]Downsampling undertaken whereby random sampling was conducted to reduce the majority class (pregnant) to be the same size as the minority class (nonpregnant).

[2]For all strategies, records defined as pregnant if there was a validated AI event and a subsequent calving (n = 700,332 records). Nonpregnant records defined for each strategy as follows: (1) Records before the first mating assigned as nonpregnant (n = 164,537); (2) records after the first mating but before the validated AI event assigned as nonpregnant (n = 14,778); and (3) in addition to nonpregnant records used in strategy 2, records for cows without a subsequent calving assigned as nonpregnant (n = 197,624).

[3]FT-MIR = Fourier-transform mid-infrared. FT-MIR spectra models use spectral wavenumbers as predictors only; FT-MIR spectra + DIM models use spectral wavenumbers and DIM (30-d window class) as predictors; FT-MIR spectra (pre-adjusted for DIM) models use spectral wavenumbers pre-adjusted for DIM (30-d window class).

Table A3. Model performance for partial least squares discriminant analysis models with FT-MIR spectral wavenumbers as predictors, excluding records classified as pregnant if the test date was within 7, 14, or 21 d after a validated AI event for accuracy (Acc), sensitivity (Sens), specificity (Spec), and area under the receiver operating characteristic curve (AUC) values within the training, herd-independent validation (VAL-Test), and pregnancy-associated glycoproteins validation (VAL-PAG) data sets

	Training				Test validation (VAL-Test)				Glycoprotein-based validation (VAL-PAG)			
Classification strategy[1] and data set[2]	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
Strategy 1												
FT-MIR spectra (all records)	0.938	0.932	0.966	0.987	0.941	0.936	0.961	0.987	0.918	0.998	0.002	0.559
FT-MIR spectra (excluding records within 7 d post-AI)	0.932	0.923	0.972	0.986	0.938	0.930	0.972	0.987	0.918	0.998	0.001	0.556
FT-MIR spectra (excluding records within 14 d post-AI)	0.928	0.916	0.974	0.986	0.933	0.923	0.973	0.987	0.917	0.998	0.001	0.556
FT-MIR spectra (excluding records within 21 d post-AI)	0.913	0.897	0.982	0.985	0.920	0.907	0.981	0.986	0.917	0.998	0.002	0.556
Strategy 2												
FT-MIR spectra	0.807	0.805	0.931	0.922	0.801	0.799	0.906	0.914	0.912	0.990	0.018	0.572
FT-MIR spectra (excluding records within 7 d post-AI)	0.803	0.801	0.933	0.920	0.797	0.794	0.929	0.919	0.910	0.988	0.022	0.566
FT-MIR spectra (excluding records within 14 d post-AI)	0.797	0.794	0.937	0.920	0.790	0.787	0.935	0.920	0.909	0.987	0.023	0.569
FT-MIR spectra (excluding records within 21 d post-AI)	0.794	0.791	0.941	0.918	0.787	0.784	0.942	0.918	0.910	0.988	0.023	0.565
Strategy 3												
FT-MIR spectra	0.596	0.594	0.603	0.637	0.599	0.600	0.596	0.636	0.665	0.673	0.571	0.668
FT-MIR spectra (excluding records within 7 d post-AI)	0.594	0.592	0.602	0.635	0.601	0.600	0.603	0.642	0.676	0.687	0.548	0.668
FT-MIR spectra (excluding records within 14 d post-AI)	0.591	0.588	0.602	0.633	0.598	0.597	0.603	0.639	0.687	0.701	0.532	0.662
FT-MIR spectra (excluding records within 21 d post-AI)	0.588	0.585	0.600	0.628	0.596	0.595	0.599	0.636	0.710	0.729	0.496	0.662

[1]For all strategies, records defined as pregnant if there was a validated AI event and a subsequent calving (n = 700,332 records). Nonpregnant records defined for each strategy as follows: (1) records before the first mating assigned as nonpregnant (n = 164,537); (2) records after the first mating but before the validated AI event assigned as nonpregnant (n = 14,778); and (3) in addition to nonpregnant records used in strategy 2, records for cows without a subsequent calving assigned as nonpregnant (n = 197,624).

[2]FT-MIR = Fourier-transform mid-infrared. FT-MIR spectra models (all records): All records included; FT-MIR spectra (excluding records within [7,14,21] days post-AI): Records removed if the test date was within 7, 14, or 21 d after a validated AI event. Number of records excluded: test date within 7 d of validated AI (n = 22,338); test date within 14 d of validated AI (n = 48,573); and test date within 21 d of validated AI (n = 81,581).