



Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows

K. Hempstalk,^{*1} S. McParland,[†] and D. P. Berry[†]

^{*}Livestock Improvement Corporation, Private Bag 3016, Hamilton 3240, New Zealand

[†]Animal and Grassland Research and Innovation Center, Teagasc, Moorepark, Fermoy, Co. Cork, Ireland

ABSTRACT

The ability to accurately predict the conception outcome for a future mating would be of considerable benefit for producers in deciding what mating plan (i.e., expensive semen or less expensive semen) to implement for a given cow. The objective of the present study was to use herd- and cow-level factors to predict the likelihood of conception success to a given insemination (i.e., conception outcome not including embryo loss); of particular interest in the present study was the usefulness of milk mid-infrared (MIR) spectral data in augmenting the accuracy of the prediction model. A total of 4,341 insemination records with conception outcome information from 2,874 lactations on 1,789 cows from 7 research herds for the years 2009 to 2014 were available. The data set was separated into a calibration data set and a validation data set using either of 2 approaches: (1) the calibration data set contained records from all 7 farms for the years 2009 to 2011, inclusive, and the validation data set included data from the 7 farms for the years 2012 to 2014, inclusive, or (2) the calibration data set contained records from 5 farms for all 6 yr and the validation data set contained information from the other 2 farms for all 6 yr. The prediction models were developed with 8 different machine learning algorithms in the calibration data set using standard 10-times 10-fold cross-validation and also by evaluating in the validation data set. The area under curve (AUC) for the receiver operating curve varied from 0.487 to 0.675 across the different algorithms and scenarios investigated. Logistic regression was generally the best-performing algorithm. The AUC was generally inferior for the external validation data sets compared with the calibration data sets. The inclusion of milk MIR in the prediction model generally did not improve the accuracy of prediction. Despite the fair AUC for

predicting conception outcome under the different scenarios investigated, the model provided a reasonable prediction of the likelihood of conception success when the high predicted probability instances were considered; a conception rate of 85% was evident in the top 10% of inseminations ranked on predicted probability of conception success in the validation data set.

Key words: prediction, conception, algorithm, dairy cattle

INTRODUCTION

Reproductive management is a key component of profitable dairy production (Cabrerá, 2014; Shalloo et al., 2014), but differences in conception rate exist among individuals and across time. Prior knowledge of the likelihood of conception to a given insemination could be useful in optimizing reproductive herd management. For example, lower cost semen could be used where a reduced likelihood of conception may exist. Multiple semen doses or semen from bulls of known high fertility could be used in such conditions to potentially improve the likelihood of conception. Similarly, the producer may decide not to breed the animal but instead to wait for a subsequent estrus event. In contrast, sexed semen or expensive semen could be used where a greater likelihood of conception is predicted.

Several factors are known to be associated with the likelihood of conception success. In a meta-analysis of 10 studies, Berry et al. (2014) reported that, on average, 2.3% of the phenotypic variation in conception rate (after adjustment for systematic environmental effects) was due to additive genetic effects. Nonadditive genetic effects (i.e., intra- and inter-locus interactions) also contribute to the phenotypic variance (Hoeschele, 1991; Buckley et al., 2014). Known nongenetic effects contributing to differences among animals in conception or pregnancy rate include herd (Jansen and Lagerweij, 1987; Van Doormaal, 1993; Clay et al., 2004; Berry et al., 2011), year (Berry et al., 2011), season of insemination (Everett and Bean, 1986; Murray et al., 1993; Clay et al., 2004; Berry et al., 2011), parity (Everett and Bean, 1986; Jansen and Lagerweij, 1987; Clay et al.,

Received October 16, 2014.

Accepted April 22, 2015.

¹Corresponding author: khempstalk@lic.co.nz

2004; Berry et al., 2011), days postcalving (Jansen and Lagerweij, 1987; Van Doormaal, 1993; Clay et al., 2004; Berry et al., 2011), day of the week (Taylor et al., 1985; Berry et al., 2011), BCS (Roche et al., 2007), as well as the fertility of the mate and the capability of the technician undertaking the insemination (Everett and Bean, 1986; Jansen and Lagerweij, 1987; Berry et al., 2011). Although the coefficients of some of these factors cannot be predicted a priori (e.g., year), quantification of the association between conception rate and other factors can be used to build a biological decision support tool for producers to predict the likelihood of conception given the prevailing conditions and characteristics of the cow. While examples of such models already exist (Shahinfar et al., 2014), of particular interest in the present study was the marginal information content of milk mid-infrared (MIR) data to augment prediction accuracy. McParland et al. (2011, 2012) reported that milk MIR could be used to predict cow energy balance, which is known to be associated with reproductive performance (Beam and Butler, 1999). Our hypothesis was that the milk MIR could provide useful additional information in predicting the likelihood of conception success for an insemination.

Machine learning algorithms have recently been used in the development of conception prediction models (Shahinfar et al., 2014); however, the availability of milk MIR spectral data in the present study poses an additional dimensionality challenge for learning. The spectral data contains 1,060 features, many of which may be unrelated to conception. The algorithm is expected to automatically determine which features best describe the likelihood of conception success. The risk is that the high dimensionality will cause the models to over fit, choosing to learn from features that represent noise; the result could be prediction models that perform poorly during external validation. The machine learning algorithms used in this study exploit feature selection or feature translation to reduce the dimensionality of the calibration data without compromising the predictive power of the model.

The objective of the present study was to develop a biological-based, machine learning model to predict the likelihood of conception success given the herd- and cow-specific attributes; conception success here was defined as conception resulting in an offspring produced or the cow being determined pregnant by ultrasound examination. Particular attention was given to the usefulness of milk MIR spectral data in augmenting the accuracy of prediction. Moreover, prediction accuracy in an external validation data set will provide a truer reflection of the accuracy of predicting the conception success for a given insemination.

MATERIALS AND METHODS

Data

Date of calving ($n = 10,044$), AI ($n = 11,938$), and pregnancy diagnosis ($n = 10,365$) were available on up to 6 parities from 2,933 cows in 7 research farms in Ireland; the data were collected between the years 2009 to 2014, inclusive. Data were also available on parity number, animal breed, and PTA (i.e., measure of genetic merit) for fertility (i.e., calving interval), milk production, live weight, and longevity from the April 2014 national genetic evaluations. The Economic Breeding Index (EBI) value, which is the Irish national dairy cow breeding goal, was also available for each animal from the April 2014 national genetic evaluations. Cows were predominantly Holstein-Friesian (61.1%) but the data set also included purebred Jersey (12%), purebred Norwegian Red (5.8%), and crossbred animals.

Only insemination records from parities with either a pregnancy diagnosis or a subsequent calving were retained; 10,633 insemination records remained. Conception status to a given insemination was determined based on whether or not a subsequent insemination existed for that lactation, the pregnancy diagnosis outcome, or subsequent calving date, where available. No records of embryonic loss were available so conception success considered here does not consider embryo loss.

Animals diagnosed as pregnant were assumed to have conceived to the most recent insemination; this was (in)validated by subsequent calving dates where available. Earlier inseminations were assumed to not result in conception. Insemination dates within 14 d of the estimated date of conception (estimated as 282 d before the subsequent calving) were assumed to have resulted in the conception, unless an additional insemination also existed within the interval. If more than one insemination existed within the 14-d interval, all inseminations within the interval were discarded. Final insemination dates more than 14 d from the estimated date of conception were assumed not to have resulted in pregnancy and were also discarded. Also, only insemination records between 10 and 150 d postcalving were retained ($n = 6,689$).

Cow live weight and BCS records were also available for each cow. Live weight and BCS were recorded approximately every 3 wk on most farms, although 3 farms only recorded live weight and BCS at calving. The live weight and BCS record closest to the time of insemination, but within 30 d before insemination, was provided to the prediction model. A total of 991 inseminations had no live weight record in close proximity and 1,873 insemination records had no BCS in close proximity. Changes in BCS and live weight were

calculated as the difference between the closest record to insemination (within 30 d of insemination) and the previous record if measured 90 d before insemination.

Spectral Data

All cows were milked twice a day. Milk yield was measured daily and milk samples were taken weekly for separate composition analyses on consecutive evening and morning milkings. All samples were analyzed by a MIR spectrometer (MilkoScan FT6000, Foss, Hillerød, Denmark) and the predicted fat, protein, and lactose concentrations were stored. The raw MIR spectrum from the Foss spectrometer was also stored. For a single sample, the spectrometer recorded 1,060 data points, each point representing the absorption of infrared light through the milk sample at a particular wavelength in the 900 to 5,000 cm^{-1} region. The MIR spectra of the most recent morning (a.m.) and afternoon (p.m.) milk samples from the 7 d before each insemination record (including the day of insemination itself) were retained.

Data Editing

Several additional variables for consideration in the prediction model were calculated: DIM at insemination, the number of previous inseminations in the current lactation, the number of inseminations in the previous lactation (or missing, if there were none), DIM when the animal conceived in the previous lactation (or missing if this was the animal's first lactation), day of the week, month of insemination, and the previous year's conception rate of the herd (from all inseminations). The probability of conception based on DIM was calculated using the R statistical package (R Core Team, 2013) by fitting a quadratic polynomial function to the data, which resulted in the following equation:

$$\text{DIM probability} = 0.0473 + 0.0105 \cdot \text{DIM} - 0.00005878 \cdot \text{DIM}^2.$$

This function was applied to each record to obtain a new feature, DIM probability. This variable was calculated because the DIM is known to have a nonlinear association with conception outcome (Berry et al., 2011), and not all of the machine learning algorithms tested are able to effectively learn nonlinear variables; transforming the variable into a probability and providing it as an additional feature resolves this issue. The same technique was also performed on the DIM when the animal conceived in the immediately previous parity, adding the feature "previous parity's probability," using the derived equation:

$$\begin{aligned} \text{Previous parity's probability} &= 0.5394 \\ &- 0.0012 \cdot \text{Previous parity DIM}. \end{aligned}$$

Energy balance (**EB**) of each animal was predicted using the milk yield and MIR spectral data as outlined in detail by McParland et al. (2012). The correlation between actual and predicted EB was 0.69. Predicted EB within 7 d before insemination was available for 3,405 inseminations.

Machine Learning Algorithms

Machine learning is a way of automatically creating statistical models that can generate predictions based on what they have learned. The algorithms are provided with data records (known as "instances") containing many variables (known as "attributes") and they use this information to learn a pattern that describes a target variable (known as the "class variable"). This process of learning is known as "training." The models are evaluated (or "tested") using data records that were not included in the training data set.

In the present study, models to predict the likelihood of conception were generated using machine learning algorithms from the WEKA Machine Learning Workbench (Witten et al., 2011). The algorithms evaluated were C4.5 decision trees, naïve Bayes, Bayesian network, logistic regression, support vector machine, partial least squares regression, random forest, and rotation forest. The models used here are classifiers; they learn a pattern for the binary class variable conception status based on the examples provided during the training phase. These 8 particular models were selected because they are commonly used in classification tasks, and represent different machine learning approaches to classification.

C4.5 Decision Tree. The C4.5 decision tree (Quinlan, 1993) builds a tree by evaluating the information gain of each feature (i.e., independent variable) and then creates a split (or decision) by choosing the most informative feature and dividing the records into left and right nodes of the tree. This process repeats until all of the records at a node belong to a single class (i.e., conceived or not) or the number of records reaches the threshold defined in the algorithm (i.e., a minimum of 2 instances per leaf). A prediction is made by traversing the tree using the values from the current instance and returning the majority class at the leaf node reached by the traversal. The tree prevents over-fitting by performing pruning to remove nodes that may cause error in the final model. The specific version of C4.5 used in the present study was J48, provided by WEKA (Witten et al., 2011). Default parameters were used, which were

a minimum of 2 instances per leaf ($M = 2$), and a confidence factor ($C = 0.25$) of 0.25 for pruning.

Naïve Bayes. The naïve Bayes algorithm “naïvely” assumes each feature is independent and builds a model based on Bayes’ rule. It multiplies the probabilities of each feature belonging to each class (i.e., conceived or not) to generate a prediction. All the features in the present study available for learning were numeric (with the exception of the class variable), and for numeric features, the implementation used in this study assumed each feature was normally distributed. The probability for each feature is calculated by supplying the mean and standard deviation to a Gaussian probability density function, which are then multiplied together using Bayes’ rule. The implementation of naïve Bayes used in this study was Naïve Bayes in WEKA (Witten et al., 2011).

Bayesian Network. A Bayesian network classifier represents each feature as a node on a directed acyclic graph, each node containing the conditional probability distribution that can be used for class prediction. A Bayesian network assumes that each node is conditionally independent of its nondescendants, given its immediate parents. During calibration, the network structure is built by searching through the space of all possible edges and computing the log-likelihood of each resulting network as a measure of quality. In the present study, a hill-climbing algorithm (K2 in WEKA) was used to search for the best network structure, and the probability tables within each node were calculated using “SimpleEstimator”; the default parameters for the algorithm “Bayes Network” in WEKA (Witten et al., 2011) were used.

Logistic Regression. Linear regression is a common statistical technique used to express a class variable as a linear combination of the features. However, it is designed to predict a real numeric value and cannot handle a categorical or binary class (i.e., conceived or not). To overcome this, a model can be built for each class value that ideally predicts 1 for that class value, and 0 otherwise, and at prediction time assigns the class value whose model predicts the greatest probability. Unfortunately, regression functions are not guaranteed to produce a probability between 0 and 1, and so the target class must first be transformed into a new space before it is learned. This is achieved using a log-transform, and this regression method is known as logistic regression (Witten et al., 2011). In logistic regression, the weights are chosen to maximize the log-likelihood (instead of reducing the squared error), by iteratively solving a sequence of weighted least-squares regression problems until the log-likelihood converges on the maximum. One algorithm in WEKA (and the algorithm that was used in this research) that performs

this type of logistic regression is SimpleLogisticRegression, which by default uses boosting ($M = 500$) to find the maximum log-likelihood, and cross-validation with greedy stopping ($H = 50$) to ensure the algorithm stops boosting if no gains have been made in the last H iterations.

Support Vector Machine. Support vector machines (SVM) can produce nonlinear boundaries (between classes) by constructing a linear boundary in a large, transformed version of the feature space (Hastie et al., 2009). In practice, a soft margin boundary (Cortes and Vapnik, 1995) is used to prevent over-fitting; however, a hard margin is easier to visualize when describing SVM. In the hard margin case, the algorithm assumes that classes in the transformed space are linearly separable, and it is possible to generate a hyperplane that completely separates them. By employing a technique known as the kernel trick (Aizerman et al., 1964), SVM are able to generate nonlinear decision boundaries. This is possible because the kernel trick reduces the computational effort by estimating similarities of the transformed instances as a function of their similarities in the original space. The SVM used in this research was SMO, sequential minimal optimization (Platt, 1998), from WEKA (Witten et al., 2011), which uses the sequential minimal optimization algorithm to increase the speed of finding the maximum-margin hyperplane. The parameters were set to the default complexity ($C = 1.0$) and a polynomial kernel of degree 1 was used (`weka.classifiers.supportVector.PolyKernel`). Although other kernels [including the radial basis function (RBF) kernel] and complexities (0.1, 1.0, 10, 100) were tested, the greatest accuracy was obtained using the default settings.

Partial Least Squares Regression. Partial least squares regression (PLS; Geladi and Kowalski, 1986) can be performed as a preprocessing step before training a machine learning algorithm; it works like principal component analysis (PCA) in that it transforms the data set into a new projection that represents the entire data set, and then chooses the C most informative axes (or “components”) in the new projection as features in the transformed data set. Where the PCA and PLS algorithms differ is that PLS takes into consideration the dependent variable when constructing its projection, but PCA does not. One advantage of using the dependent variable during learning is that the algorithm is able to perform regression using the projections it has calculated. A binary prediction (i.e., conceived or not) can be made by creating a regression model that predicts the probability (of conception) and returning true if the probability reaches a set threshold, or false otherwise. In this research, the WEKA classifier `ClassificationViaRegression` was used to provide the bi-

nary prediction from the PLSClassifier. The number of components used was 20 ($C = 20$), except when using only non-MIR data for prediction, in which case $C = 10$ so that the number of components was not larger than the number of features. Partial least squares regression cannot handle categorical (nominal) features, so these features had to be removed from all data sets when using this algorithm for prediction. These changes may result in a reduction in accuracy compared with algorithms with the full feature set available.

Random Forest. Random forest (Breiman, 2001) is an ensemble learner that creates a “forest” of decision trees, and predicts the most popular class estimated by the set of trees. Each tree is provided with a random set of training instances sampled with replacement from the entire training set. The intention of this step is to create a diverse set of trees. The algorithm differs from bagged decision trees (which also provide randomly selected subsets to each tree) because during training the algorithm randomly selects a subset of features available for selection at each split in the tree. The implementation used in this study is RandomForest in WEKA, which by default has an unlimited tree depth ($maxDepth = 0$) and the number of features randomly selected into each subset = $\log_2(\text{total number of features}) + 1$. By default, this algorithm creates a forest of 10 trees ($numTrees = 10$); however, this was increased to 1,000 ($numTrees = 1000$) because the algorithm had poor accuracy when considering only 10 trees. The effect of increasing this parameter is that accuracy is improved, but also that the algorithm takes much longer to run than any others evaluated here.

Rotation Forest. Rotation forest (Rodríguez et al., 2006) is an ensemble learner similar to random forest except that PCA is applied to select the features for each tree (instead of random selection), and the components are all kept when the base classifier is trained. The classifier sees a “rotated” set of features in each tree in its forest. The intention is to create individual accuracy in the tree and diversity in the ensemble, compared with random forest, which aims only to create diversity in the ensemble. Results for a rotation forest learner have been shown to be as good as those of other ensemble learning schemes such as bagging, boosting, and random forests (Rodríguez et al., 2006). The implementation of rotation forest used in this study was RotationForest in WEKA, with default parameters: *projection = PCA*, *base classifier = J48*.

Analysis

The final data set consisted of 4,341 insemination records from 2,874 lactations on 1,789 cows. This data

set was separated into a calibration data set and a validation data set using either of 2 approaches: (1) the calibration data set consisted of records from all 7 farms for the year 2009 to 2011, inclusive, and the validation data set included data from the 7 farms for the years 2012 to 2014 inclusive (hereafter referred to as the “year-out analysis”) or (2) the calibration data set contained information from all 6 yr on 5 farms, whereas the validation data set contained the information from all 6 years on 2 farms (hereafter referred to as the “farm-out analysis”). The prediction models were evaluated twice, once using the calibration data set and standard 10-times 10-fold cross-validation, and then also using the full calibration set for training, and the validation set for evaluation. The 10-times 10-fold cross-validation was initialized with a different random seed for each 10-fold evaluation, and the standard deviation between the different random initializations was recorded as a measure of consistency.

Variables considered for inclusion in the models were the non-MIR spectral values described in Table 1 and all 1,060 MIR wavelength values. The Foss calculated values for fat, protein, and lactose were included as non-MIR spectral features; although the values are derived from the spectral information, they do not represent all the spectra information or any particular one wavelength. Our hypothesis was that the spectral information itself might add to the prediction accuracy, rather than values routinely calculated from them.

The prediction models were developed using 3 sets of features: (1) MIR spectral variables only, (2) non-MIR variables only, and (3) all variables. The predictive ability of the different algorithms on these feature sets was determined based on area under the receiver operating characteristics (ROC) curve (AUC) to predict conception success. The significance of the difference of the cross-validation AUC from 0.5 was based on the mean and standard deviation of the AUC for the different folds.

RESULTS

The mean conception rate in the entire data set was 47.3%. Average days postcalving at insemination was 88. The mean number of inseminations per lactation was 1.8, and the mode was 1. The percentage of animals in parities 1, 2, 3, 4, 5, and 6+ was 26.9, 24.0, 18.8, 14.2, 8.7, and 7.4%, respectively.

Year-Out Analysis

The AUC for the different algorithms and different sets of features (i.e., MIR spectral variables only, non-

Table 1. The set of non-mid-infrared (i.e., not spectral) features presented to the machine learning algorithms and their statistical properties

| Variable | Type | Unique values | Missing | Minimum | Maximum | Mean | SD |
|---|----------------|---------------|---------|---------|---------|-------|-------|
| Parity | Nominal | 6 | | 1 | 6+ | | |
| Days since sample (d) | Nominal | 7 | | | | | |
| DIM (d) | Numeric | 140 | | | | | |
| DIM probability | Numeric | | | 10 | 149 | | |
| No. of inseminations last year | Nominal | 4 | 1,369 | 1 | 4+ | | |
| No. of inseminations this year | Nominal | 5 | | 0 | 4+ | | |
| Last year DIM (d) | Numeric | 140 | 1,697 | 10 | 149 | | |
| Last year DIM probability | Numeric | | | | | | |
| Day of week | Nominal | 7 | | | | | |
| Month of year | Nominal | 10 | | | | | |
| No. of days since last service | Numeric | | 2,556 | 1 | 96 | 24.9 | 14 |
| Last year conception rate | Numeric | | 279 | 30 | 75.4 | 38.3 | 8.9 |
| Dairy EBI (€) | Numeric | | 56 | -124 | 308 | 152.4 | 51.2 |
| Fertility EBV (d) | Numeric | | 56 | -9.9 | 6.2 | -4 | 1.9 |
| Survival EBV (proportion) | Numeric | | 56 | -2.8 | 5.6 | 2 | 1.1 |
| Milk yield EBV (kg) | Numeric | | 56 | -884 | 626.5 | 70.2 | 201.2 |
| Live weight EBV (kg) | Numeric | | 56 | -85 | 37 | -10.4 | 19.2 |
| Breed proportion | Nominal | 11 | | | | | |
| Second breed proportion | Nominal | 6 | 3,829 | | | | |
| BCC (scale 1 to 5) | Numeric | | 1,873 | 2 | 4.4 | 2.9 | 0.2 |
| Change in BCS (scale 1 to 5) | Numeric | | 2,192 | -0.75 | 1 | 0.1 | 0.2 |
| Live weight (kg) | Numeric | | 991 | 263 | 780 | 501.6 | 77.9 |
| Change in weight (kg) | Numeric | | 1,413 | -357 | 199 | -9.3 | 32.5 |
| Milk yield (kg) | Numeric | | 230 | 0.1 | 30.9 | 15.3 | 4.1 |
| Fat yield (kg) | Numeric | | 34 | 0.6 | 13.9 | 3.6 | 0.9 |
| Protein yield (kg) | Numeric | | 34 | 0.9 | 4.9 | 3.4 | 0.3 |
| Fat to protein ratio | Numeric | | 34 | 0.2 | 4.7 | 1 | 0.3 |
| ECM yield (kg) | Numeric | | 230 | 0.12 | 35 | 16.1 | 4.3 |
| Predicted energy balance (unite fourragere lait/d) | Numeric | | 936 | -47 | 106.5 | 32.4 | 21.2 |
| Pregnant this insemination (1/0) | Binary (class) | 2 | | | | | |

MIR variables only, all variables) considered is in Table 2. Across all sets of features and algorithms investigated, the AUC achieved varied from 0.492 to 0.665; the standard deviation of the AUC in cross-validation for the different models varied from 0.023 to 0.035.

Although not always consistent across either the calibration or validation data set, or which set of features were considered, we detected a general tendency for

the logistic regression approach to achieve the greatest AUC whereas the C4.5 decision trees achieved the least. The AUC achieved for logistic regression in cross-validation was always greater ($P < 0.05$) than 0.5. Of the 6 different scenarios investigated (i.e., calibration or validation data set by 3 different sets of potential features), the difference in AUC between the best and worst algorithm varied from 0.036 (only MIR spectral

Table 2. The area under the curve (AUC) for 10-times, 10-fold cross-validation (CV; calibration data set), standard deviation (SD) of AUC between folds for cross-validation, and the AUC for external validation (Ext) with different algorithms using milk mid-infrared information only as predictor variables (MIR only), using predictor variables other than milk MIR (no MIR), or both combined (both) based on a calibration data set consisting of all herds for the years 2009–2011 and a validation data set of all herds for the years 2012–2014

| Algorithm | MIR only | | | No MIR | | | Both | | |
|------------------------|----------|-------|--------------------|----------|-------|--------------------|----------|-------|--------------------|
| | CV | SD | Ext | CV | SD | Ext | CV | SD | Ext |
| C4.5 decision trees | 0.527 | 0.030 | 0.492 | 0.589** | 0.033 | 0.573 | 0.564* | 0.030 | 0.556 |
| Naïve Bayes | 0.548 | 0.034 | 0.492 | 0.649*** | 0.034 | 0.607 | 0.555 | 0.030 | 0.501 |
| Bayes network | 0.564* | 0.031 | 0.504 ¹ | 0.660*** | 0.035 | 0.626 | 0.579* | 0.032 | 0.519 |
| Logistic regression | 0.571** | 0.024 | 0.499 ¹ | 0.665*** | 0.028 | 0.614 ¹ | 0.661*** | 0.025 | 0.622 |
| Support vector machine | 0.521 | 0.028 | 0.500 | 0.610*** | 0.025 | 0.586 | 0.605*** | 0.023 | 0.588 |
| Partial least squares | 0.563* | 0.030 | 0.504 ¹ | 0.623*** | 0.033 | 0.572 | 0.625*** | 0.028 | 0.540 ¹ |
| Random forest | 0.569* | 0.029 | 0.526 | 0.657*** | 0.027 | 0.640 | 0.594*** | 0.029 | 0.552 |
| Rotation forest | 0.557 | 0.033 | 0.490 ¹ | 0.651*** | 0.029 | 0.608 | 0.617*** | 0.032 | 0.557 |

¹External validation AUC different ($P < 0.05$) from cross-validation AUC.

* $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$: significance of difference from 0.5.

data considered as features in the validation data set) to 0.121 (all variables considered as predictors in the validation data set). The difference between algorithms was least when just the MIR spectral variables were considered as features and was greatest when all potential predictor variables were considered.

The AUC in the validation data set for the different scenarios investigated was always numerically worse (i.e., 0.008 to 0.085 units lower) than that of the respective scenario in the calibration data set. However, the difference between calibration and validation was only significant at the 5% level in 3 of the 24 cases analyzed [MIR logistic regression ($P = 0.003$), MIR PLS ($P = 0.049$), both PLS ($P = 0.002$)]. The average difference in AUC between the calibration and validation data sets was least (i.e., 0.008) for the C4.5 decision trees and greatest (i.e., 0.085) for the PLS algorithm. The ROC from predictions developed using simple logistic regression for the year-out analysis is in Figure 1; the AUC was 0.661 and 0.622 for the calibration and validation data sets, respectively.

The AUC achieved when only the MIR spectral data were considered in the prediction model was always worse (i.e., 0.060 to 0.122 units worse) than that achieved using all other predictor variables except the MIR spectral data. In general, the AUC from prediction models with all potential features considered, including

the MIR spectral data, was not better than the AUC based on all features other than the MIR spectral data; the exception was predictions in the external validation using either logistic regression or SVM.

Farm-Out Analysis

Table 3 summarizes the AUC for the different algorithms and prediction model scenarios investigated when the 2 farms excluded from the calibration data set were used in the external validation. The AUC varied from 0.485 to 0.675; the standard deviation of the AUC in cross-validation for the different models varied from 0.023 to 0.034. Although not always consistent across the 6 scenarios investigated (i.e., calibration or validation data set by 3 sets of features considered), we observed a tendency for predictions using logistic regression to out-perform the other algorithms, whereas prediction models developed using C4.5 decision trees or naïve Bayes algorithms performed the worst. The AUC achieved for logistic regression in cross-validation was always greater ($P < 0.05$) than 0.5. The difference in AUC between the algorithms for each of the 6 scenarios investigated varied from 0.059 (calibration data set using just MIR spectral data) to 0.125 (validation data set using predictor variables other than MIR spectral data).

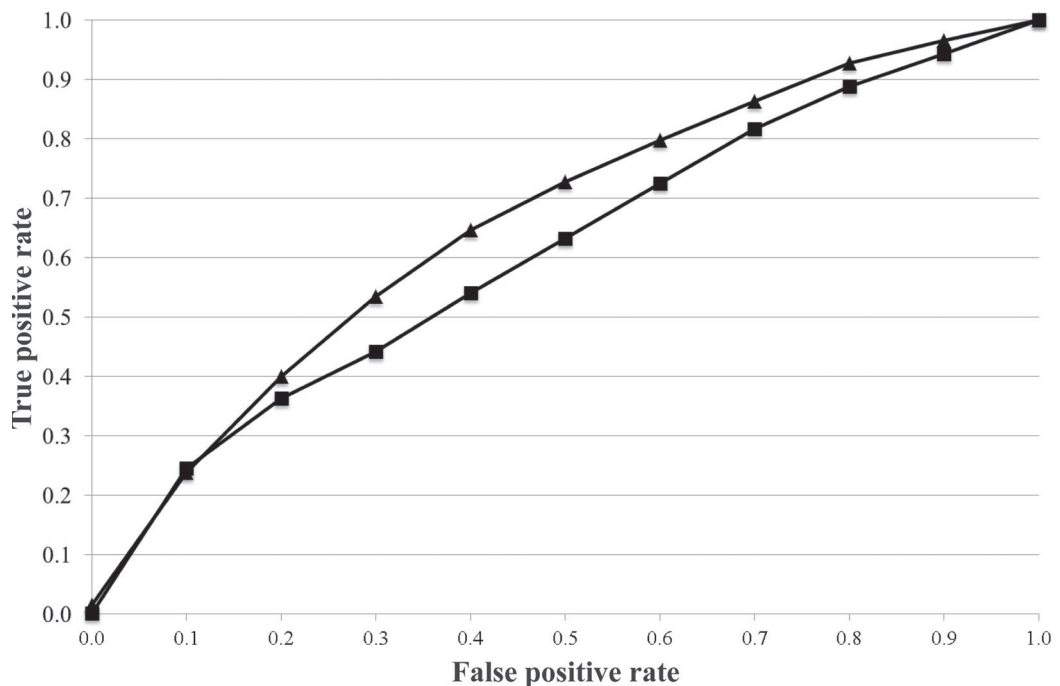


Figure 1. Receiver operator characteristic (ROC) curve for the cross-validation (triangles) and external validation (squares) for the year-out combined (mid-infrared and non-mid-infrared) analysis developed using simple logistic regression.

Table 3. The area under the curve (AUC) for 10-times, 10-fold cross-validation (CV; calibration data set), standard deviation (SD) of AUC between folds for cross-validation, and the AUC for external validation (Ext) with different algorithms using milk mid-infrared information only as predictor variables (MIR only), using predictor variables other than milk MIR (no MIR), or both combined (both) based on a calibration data set consisting of 5 herds for all years and a validation data set of the 2 other herds for all years

| Algorithm | MIR only | | | No MIR | | | Both | | |
|------------------------|----------|-------|--------------------|----------|-------|--------------------|----------|-------|--------------------|
| | CV | SD | Ext | CV | SD | Ext | CV | SD | Ext |
| C4.5 decision trees | 0.531 | 0.028 | 0.514 | 0.613*** | 0.028 | 0.493 ¹ | 0.580** | 0.030 | 0.487 ¹ |
| Naïve Bayes | 0.563 | 0.034 | 0.501 | 0.664*** | 0.027 | 0.541 ¹ | 0.573* | 0.034 | 0.502 ¹ |
| Bayes network | 0.562* | 0.031 | 0.507 | 0.661*** | 0.029 | 0.581 ¹ | 0.572* | 0.033 | 0.509 |
| Logistic regression | 0.570* | 0.030 | 0.537 | 0.668*** | 0.028 | 0.618 | 0.667*** | 0.026 | 0.598 ¹ |
| Support vector machine | 0.530 | 0.029 | 0.485 | 0.614*** | 0.023 | 0.558 ¹ | 0.610*** | 0.026 | 0.549 ¹ |
| Partial least squares | 0.561* | 0.031 | 0.528 | 0.633*** | 0.026 | 0.540 ¹ | 0.626*** | 0.028 | 0.536 ¹ |
| Random forest | 0.589** | 0.032 | 0.487 ¹ | 0.675*** | 0.024 | 0.571 ¹ | 0.615*** | 0.031 | 0.497 ¹ |
| Rotation forest | 0.578** | 0.027 | 0.508 ¹ | 0.658*** | 0.028 | 0.574 ¹ | 0.628*** | 0.026 | 0.578 |

¹External validation AUC different ($P < 0.05$) from cross-validation AUC.

* $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$: significance of difference from 0.5.

The AUC from cross-validation in the calibration data set was 0.017 to 0.123 units greater than the AUC for the same algorithm in the validation data set, and the difference was significant ($P < 0.05$) in 15 of 24 cases tested. The AUC from predictions based on just the MIR spectral data were up to 0.101 units worse than predictions using variables other than the MIR spectral data; the exception was predictions using C4.5 decision trees, which was slightly better using just the MIR spectral data as predictors compared with using all possible predictor variables other than the MIR spectral data. With the exception of logistic regression and SVM in the calibration data set, the AUC using all possible features including the MIR spectral data was always inferior to the AUC achieved when all features except the MIR spectral data were considered.

Model Solutions

The regression coefficients on the terms in the logistic regression model varied slightly between the year-out and farm-out analyses, but the general model was the same. The probability of conception increased with increasing DIM and greater BCS, but decreased with increasing parity, increasing number of inseminations in the previous parity, and increasing number of inseminations in the current parity. The likelihood of conception was least in the latter months of the mating season. The likelihood of conception was lesser on Mondays and Thursdays, and greatest on Saturday. The strongest association in both analyses was DIM (coefficient = 1.21) followed by BCS (coefficient = 0.35); the largest negative association was with the number of inseminations so far this lactation totaling 4 (or more) (coefficient = -0.45). Some breeds had a greater probability of conception than others, but some breeds (i.e., Holstein

cross, Montbéliarde cross, Montbéliarde) represented <50 records in the calibration data set.

Ranked Predictions

The accuracy of predicting conception outcome for percentiles of records ranked on predicted probability of conception by the model in the year-out validation data set is in Table 4. Only logistic regression and random forest were considered in the ranked prediction analysis. The number of features considered in the prediction model development was reduced by using a correlation-based feature subset selection algorithm (CfsSubsetEval in WEKA; Witten et al., 2011) to find a minimal feature set that achieved similar predictive power to the full feature set. Breed and previous season's herd conception rate were excluded from the feature set. The accuracy of prediction with the reduced feature set was lower than with the full (non-MIR) feature set due primarily to a single farm having a high conception rate, and a single breed with few records in the data set having a high conception rate, and the full model therefore assigning high probabilities to these records. The AUC for the reduced set was 0.656 using cross-validation on the calibration set, and 0.574 in the validation set. In contrast, the AUC for the full feature set was 0.665 for cross-validation and 0.614 for validation using the non-MIR data in the year-out data set.

Using simple logistic regression in the reduced data set, 72% of the inseminations in the top 5% of predicted probability of conception success resulted in a successful conception; mean conception rate in the validation data set was 54.6%. The lowest ranked predictions across all models had an accuracy that was no better than random guessing; the model could not accurately determine when an animal was not likely to conceive.

Table 4. Cumulative accuracy of conception predictions in the year-out validation data set for different feature sets using the top 25 percent of predictions ranked by probability of conception from simple logistic regression and rotation forest models

| Percentile | Simple logistic regression | | | Rotation forest | |
|------------|----------------------------|-------------|------------------|-----------------|------------------|
| | Full set | Reduced set | Reduced with MIR | Reduced set | Reduced with MIR |
| 5 | 84.0 | 72.0 | 56.0 | 64.0 | 66.0 |
| 10 | 85.0 | 65.0 | 58.0 | 63.0 | 63.0 |
| 15 | 76.7 | 62.7 | 58.7 | 62.7 | 57.3 |
| 20 | 73.5 | 64.0 | 57.5 | 62.5 | 56.5 |
| 25 | 71.2 | 64.8 | 58.8 | 62.0 | 57.2 |

When MIR data were included in the logistic regression prediction model, the accuracy of prediction was reduced, and even the records in the top 5% of predicted probability were no better than those in the top 25% on predicted probability of conception success. Because such a reduction in prediction accuracy could be caused by a dimensionality problem with logistic regression, rotation forest (which selects a subset of features using PCA) was also evaluated. Although the ranking results were superior with MIR when using rotation forest for prediction, the predictions were still inferior to using the reduced sets of variables (i.e., excluding MIR data); the addition of MIR data did not add any value to the ranked predictions.

DISCUSSION

The ability to accurately predict conception outcome to a given insemination is a useful tool for producers who are considering using premium (e.g., more expensive) semen products such as sexed semen. Previous research has suggested some promise in using animal- and herd-level data to predict conception outcome in a year-round mating production system (Shahinfar et al., 2014). This research, however, is not directly applicable to seasonal mating production systems, such as exist in Ireland and New Zealand, because the most predictive attributes suggested by Shahinfar et al. (2014) pertain to prevailing seasonal data from the year-round US system. Moreover, the study of Shahinfar et al. (2014) documented an accuracy of conception outcome from cross-validation that may be an overestimation of actual accuracy once applied to an independent data set. As well as developing a prediction model using seasonal breeding production systems and evaluating its accuracy in a validation data set, of particular interest in the present study was the potential predictive ability of conception outcome from information contained within the milk MIR spectrum; this is because of the known predictive ability of both energy balance (McParland et al., 2011) and BCS change (McParland et al., 2014)

from milk MIR. Although milk MIR is currently only undertaken on routinely recorded milk samples, several weeks apart, the potential could exist to implement in-line near-infrared spectrometers. Near-infrared spectrometry measures overtones of the primary vibrational absorptions detected by MIR (as well as higher energy primary vibrations).

Algorithms

When based on the AUC of the prediction model, in general, the prediction models developed using logistic regression were superior to the alternative algorithms investigated. Logistic regression has been used successfully elsewhere in the prediction of conception outcome (Buckley et al., 2003; Roche et al., 2007). Although naïve Bayes performed well when the MIR data were not considered in the prediction model, it performed poorly when the MIR features were considered. This is not unexpected because naïve Bayes assumes each feature is independent and contributes equally to the prediction result; many of the wavelength features are, however, not informative. The Bayesian network and random forest algorithm also suffer from this concern. The other algorithms—C4.5 decision trees, SVM, PLS, and rotation forest—all use some form of feature selection or feature weighting to try and discard uninformative features from the data set. All of these algorithms performed similarly when all features, including MIR, were considered versus only the non-MIR features. There was one exception: C4.5 decision trees using the MIR data performed slightly worse than using the non-MIR data alone.

Ensemble learning schemes such as bagging are often used to increase the accuracy of base learning schemes, such as decision trees, by creating a variety of models and combining the results from all the models for a prediction. In bagging, this is achieved by supplying a subset of the calibration data set to each model, and returning the modal binary prediction of all the models as its output. The rotation forest and random forest al-

gorithms used in the present study are ensemble learning schemes that employ a combination of bagging and feature selection. In random forest, the feature selection is random, and in the rotation forest algorithm used in the present study, the feature selection is performed by PCA. Random forest was used by Shahinfar et al. (2014), who also attempted to predict conception outcome in dairy cows. Using the year-out and farm-out data sets, the random forest algorithm (after parameter tuning of the number of trees) in the present study performed better than rotation forest using the MIR data only or the non-MIR data in both the calibration and validation sets, but performed worse in the combined data sets. The random forest can suffer from dimensionality issues on the MIR data; each feature is selected at random and it is possible to generate every tree entirely from noise. This can be mitigated against by increasing the number of trees used, but even with a 1,000-tree forest, the algorithm was not statistically ($P > 0.05$) better than a single C4.5 decision tree on the combined data sets.

Shahinfar et al. (2014) documented that random forests prediction models achieved the greatest AUC (0.76) for prediction of conception outcome in US dairy cows, whereas naïve Bayes achieved the least (0.61); all AUC statistics were based on cross-validation and were superior to those observed in the present study. The other prediction algorithms evaluated by Shahinfar et al. (2014) included naïve Bayes, Bayesian networks, and decision trees. Bagged decision trees achieved an AUC of 0.67 (Shahinfar et al., 2014). Shahinfar et al. (2014) did not evaluate logistic regression or rotation forest algorithms, which were generally the best performing algorithms in the current study.

Several notable differences exist, however, between the study of Shahinfar et al. (2014) and the present study, potentially contributing to the lesser accuracy of prediction in the present study. First, the number of insemination records included in the present study ($n = 4,341$) was considerably less than used ($n = 129,245$) by Shahinfar et al. (2014); larger data sets are likely to result in superior prediction models because more examples are available for learning. The “herd-year-month” (HYM) feature used by Shahinfar et al. (2014) provided the best information gain for both primiparous and multiparous animals in their study; however, this feature was not considered in the present study as would not be available in the field because of the seasonal breeding system in Ireland (Berry et al., 2013). Although month was considered as a feature in the present study, herd and year were removed to ensure the model generalized across herds and years so that when a novel farm or new mating season is presented,

the model should be able to make useful predictions for these data. Finally, Shahinfar et al. (2014) presented only prediction accuracies from cross-validation while, in reality, accuracy of prediction using an external validation data set would be more appropriate; nonetheless, the accuracy of prediction in the US study of Shahinfar et al. (2014) was superior to the accuracy of prediction in cross-validation in the present study.

Although the objective of the present study was to quantify the ability of statistical models to predict the outcome of conception rather than the actual factors associated with conception outcome, the importance of the features DIM at breeding, number of inseminations in the current lactation, and genetic merit for reproductive performance corroborate previously reported associations (Scheffers et al., 2010; Berry et al., 2011; Shahinfar et al., 2014).

Accuracy Statistics

An ROC curve (Figure 1) is a graphical representation of the performance of a binary classifier test (e.g., statistical model) as its discrimination threshold is altered. Performance is measured as both sensitivity (i.e., true positives) and $1 - \text{sensitivity}$ (i.e., false positives). The AUC of the ROC is the probability that a classifier (e.g., prediction probability from a statistical model) will rank a randomly chosen positive outcome higher than a randomly chosen negative outcome; the AUC is similar to the Mann-Whitney U statistical test, which evaluates whether the model ranking of positive outcomes is higher than the ranking of negative outcomes. An AUC of 0.50 indicates no discriminative ability of the classifier (e.g., statistical model). An AUC of 0.50 to 0.75 is assumed to be fair, 0.75 to 0.92 is assumed to be good, 0.92 to 0.97 is assumed to be very good, and 0.97 to 1.00 is assumed to be excellent (Brubaker, 2008). The AUC, however, only summarizes model performance across all ranges of sensitivity and specificity values. What may be more of interest to producers is the accuracy of a model at predicting animals that have a high likelihood of conceiving versus those with a low likelihood of conceiving. One would expect that extreme high and low of probabilities produced from the prediction model would be more likely to be correct than predicted probabilities close to 0.5; such a phenomenon was evident in the present study (Table 4).

Predictive Ability of Conception

The AUC achieved in the present study for either cross-validation or external validation was never great-

er than 0.675, suggesting that, overall, the prediction model may be classified as just fair. The prediction of success to conception using field data is unlikely to be accurate for a multitude of reasons, especially in an external validation data set. First, heat detection is not completely accurate. Royal et al. (2002) using information from milk progesterone profiles suggested that 5% of inseminations in dairy cows from 8 commercial herds in the UK were undertaken at the incorrect time of the estrus cycle, thereby considerably reducing the likelihood of conception; this will affect both the development of the prediction model and the accuracy of validation. Moreover, both insemination sire and the technician undertaking the AI process are well known to affect pregnancy success (Berry et al., 2011). Herd-level factors such as contemporary group of herd-year-month of insemination are likely to contribute significantly to the variability in success to conception outcome. In practice, estimates of such effects are unlikely to exist especially in the early period of the breeding season in seasonal breeding herds. Finally, although conception may have occurred to a given insemination, the variable being predicted in the present study was pregnancy outcome. Thus, factors affecting pregnancy establishment and embryo mortality are also likely to contribute to variation not detected by the prediction models; for example, the mating of animals carrying known lethal recessive genomic polymorphisms (VanRaden et al., 2011) will influence pregnancy outcome yet are not directly related to the metabolic and physiological status of the cow herself.

Our hypothesis that the inclusion of MIR spectral information in the prediction models would improve the accuracy predictions is rejected on the basis of these results. This hypothesis was based on the documented ability of milk MIR to predict EB and BCS change (McParland et al., 2011, 2014). Research (Beam and Butler, 1999) has shown that measures such as EB, which can be predicted from milk MIR (McParland et al., 2014), affect the postpartum interval to return to cyclicity but not subsequent conception rates once the cow is cycling. The present study assumed that the animal was already cycling and that heat detection was accurate; therefore, including information such as EB, fat, protein, and lactose may be unhelpful. This was substantiated by the non-MIR models. For example, logistic regression chose not to include EB or milk composition when the model was built using the entire training set (for both the farm-out and year-out analyses). Our hypothesis, therefore, that information in the individual MIR spectra could increase the accuracy of prediction was not substantiated by the results of the present study.

CONCLUSIONS

The predictive ability of conception outcome to a given insemination, based on herd- and cow-level factors was, on average, just fair. This is not unexpected given the many other factors that contribute to the success of pregnancy establishment from a given insemination, which are not generally known a priori; these factors include herd-year-season of insemination, insemination technician capability, and mate fertility. The prediction model, however, could identify with reasonable accuracy the small proportion of cows that had a high probability of establishing a successful pregnancy to a given insemination. Such information could be useful in decision support tools in aiding the selection of cows (taking cognizance of other cow features such as genetic merit) for using more expensive semen. The algorithms with the best performance of predicting conception outcome (measured by AUC) were logistic regression, partial least squares, and rotation forest. Milk MIR spectral data did not contribute to the accuracy of predicting conception outcome over and above other routinely available herd- and animal-level factors.

ACKNOWLEDGMENTS

Funding from the INTERREG (Brussels, Belgium) project, OptiMIR (<http://www.optiMIR.eu>), is gratefully acknowledged.

REFERENCES

- Aizerman, M. A., E. M. Braverman, and L. I. Rozoner. 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Autom. Remote Control* 25:821–837.
- Beam, S. W., and W. R. Butler. 1999. Effects of energy balance on follicular development and first ovulation in postpartum dairy cows. *J. Reprod. Fertil. Suppl.* 54:411–424.
- Berry, D. P., M. P. Coffey, J. E. Pryce, Y. de Haas, P. Lovendahl, N. Krattenmacher, J. J. Crowley, and Z. Wang. 2014. International genetic evaluations for feed intake in dairy cattle through the collation of data from multiple sources. *J. Dairy Sci.* 97:3894–3905.
- Berry, D. P., R. D. Evans, and S. McParland. 2011. Evaluation of bull fertility in dairy and beef cattle using cow field data. *Theriogenology* 75:172–181.
- Berry, D. P., J. F. Kearney, K. Twomey, and R. D. Evans. 2013. Genetics of reproductive performance in seasonal calving dairy cattle production systems. *Ir. J. Agric. Food Res.* 52:1–16.
- Breiman, L. 2001. Random forests. *Mach. Learn.* 45:5–32.
- Brubaker, P. H. 2008. Do not be statistically xenophobic: Time to ROC and roll. *J. Cardiopulm. Rehabil. Prev.* 28:420–421.
- Buckley, F., N. Lopez-Villalobos, and B. J. Heins. 2014. Cross-breeding: Implications for dairy cow fertility and survival. *Animal* 8(Suppl. 1):122–133.
- Buckley, F., K. O'Sullivan, J. F. Mee, R. D. Evans, and P. Dillon. 2003. Relationships among milk yield, body condition, cow weight, and reproduction in spring-calving Holstein-Friesians. *J. Dairy Sci.* 86:2308–2319.
- Cabrera, V. E. 2014. Economics of fertility in high-yielding dairy cows on confined TMR systems. *Animal* 8(Suppl. 1):211–221.

- Clay, J. S., B. T. McDaniel, and C. H. Brown. 2004. Variances of and correlations among progeny tests for reproductive traits of cows sired by AI bulls. *J. Dairy Sci.* 87:2307–2313.
- Cortes, C., and V. Vapnik. 1995. Support vector networks. *Mach. Learn.* 20:273–297.
- Everett, R. W., and B. Bean. 1986. Semen fertility—An evaluation system for artificial insemination sires, technicians, herds and systematic fixed effects. *J. Dairy Sci.* 69:1630–1641.
- Geladi, P., and B. R. Kowalski. 1986. Partial least squares regression: A tutorial. *Anal. Chim. Acta* 185:1–17.
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY.
- Hoeschele, I. 1991. Additive and non-additive genetic variance in female fertility of Holsteins. *J. Dairy Sci.* 74:1743–1752.
- Jansen, J., and G. W. Lagerweij. 1987. Adjustment of non-return rates for AI technicians and dairy bulls. *Livest. Prod. Sci.* 16:363–372.
- McParland, S., G. Banos, B. McCarthy, E. Lewis, M. P. Coffey, B. F. O'Neill, M. O'Donovan, E. Wall, and D. P. Berry. 2012. Validation of mid-infrared spectrometry in milk for predicting body energy status in Holstein-Friesian cows. *J. Dairy Sci.* 95:7225–7235.
- McParland, S., G. Banos, E. Wall, P. Coffey, H. Soyeurt, R. F. Veerkamp, and D. P. Berry. 2011. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. *J. Dairy Sci.* 94:3651–3661.
- McParland, S., E. Lewis, E. Kennedy, S. G. Moore, B. McCarthy, M. O'Donovan, S. T. Butler, and J. E. Pryce. 2014. Mid-infrared spectrometry of milk as a predictor of energy intake and efficiency in lactating dairy cows. *J. Dairy Sci.* 97:5863–5871.
- Murray, B. B., L. R. Schaeffer, and E. B. Burnside. 1993. Heritability of nonreturn rate of Canadian Holstein-Friesian bulls. *Can. J. Anim. Sci.* 63:39–48.
- Platt, J. 1998. Fast training of support vector machines using sequential minimal optimization. Pages 185–208 in *Advances in Kernel Methods: Support Vector Learning*. B. Schölkopf, C. J. Burges, and A. J. Smola, ed. MIT Press, Cambridge, MA.
- Quinlan, R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.
- R Core Team. 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Roche, J. R., K. A. MacDonald, C. R. Burke, J. M. Lee, and D. P. Berry. 2007. Associations among body condition score, body weight and reproductive performance in seasonal-calving dairy cattle. *J. Dairy Sci.* 90:376–391.
- Rodríguez, J. J., L. I. Kuncheva, and C. J. Alonso. 2006. Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.* 28:1619–1630.
- Royal, M. D., A. P. F. Flint, and J. A. Woolliams. 2002. Genetic and phenotypic relationships among endocrine and traditional fertility traits and production traits in Holstein-Friesian dairy cows. *J. Dairy Sci.* 85:958–967.
- Scheffers, J. M., K. A. Weigel, C. L. Rawson, N. R. Zwald, and N. B. Cook. 2010. Management practices associated with conception rate and service rate of lactating Holstein cows in large, commercial dairy herds. *J. Dairy Sci.* 93:1459–1467.
- Shahinfar, S., D. Page, J. Guenther, V. Cabrera, P. Fricke, and K. A. Weigel. 2014. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. *J. Dairy Sci.* 97:731–742.
- Shalloo, L., A. Cromie, and N. McHugh. 2014. Effect of fertility on the economics of pasture-based dairy systems. *Animal* 8(Suppl. 1):222–231.
- Taylor, J. F., R. W. Everett, and B. Beam. 1985. Systematic environmental, direct, and service sire effects on conception rate in artificially inseminated Holstein cows. *J. Dairy Sci.* 68:3004–3022.
- Van Doormaal, B. J. 1993. Linear model evaluations of non-return rates for dairy and beef bulls in Canadian AI. *Can. J. Anim. Sci.* 73:795–804.
- VanRaden, P. M., K. M. Olson, D. J. Null, and J. L. Hutchison. 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. *J. Dairy Sci.* 94:6153–6161.
- Witten, I. H., E. Frank, and M. A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, USA.