J. Dairy Sci. 97:731–742
http://dx.doi.org/10.3168/jds.2013-6693

Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms

Saleh Shahinfar,*[1] David Page,† Jerry Guenther,* Victor Cabrera,* Paul Fricke,* and Kent Weigel*
*Department of Dairy Science, and
†Department of Biostatistics and Medical Informatics and Department of Computer Science, University of Wisconsin, Madison 53706

ABSTRACT

When making the decision about whether or not to breed a given cow, knowledge about the expected outcome would have an economic impact on profitability of the breeding program and net income of the farm. The outcome of each breeding can be affected by many management and physiological features that vary between farms and interact with each other. Hence, the ability of machine learning algorithms to accommodate complex relationships in the data and missing values for explanatory variables makes these algorithms well suited for investigation of reproduction performance in dairy cattle. The objective of this study was to develop a user-friendly and intuitive on-farm tool to help farmers make reproduction management decisions. Several different machine learning algorithms were applied to predict the insemination outcomes of individual cows based on phenotypic and genotypic data. Data from 26 dairy farms in the Alta Genetics (Watertown, WI) Advantage Progeny Testing Program were used, representing a 10-yr period from 2000 to 2010. Health, reproduction, and production data were extracted from on-farm dairy management software, and estimated breeding values were downloaded from the US Department of Agriculture Agricultural Research Service Animal Improvement Programs Laboratory (Beltsville, MD) database. The edited data set consisted of 129,245 breeding records from primiparous Holstein cows and 195,128 breeding records from multiparous Holstein cows. Each data point in the final data set included 23 and 25 explanatory variables and 1 binary outcome for of 0.756 ± 0.005 and 0.736 ± 0.005 for primiparous and multiparous cows, respectively. The naïve Bayes algorithm, Bayesian network, and decision tree algorithms showed somewhat poorer classification performance. An information-based variable selection procedure identified herd average conception rate, incidence of ketosis, number of previous (failed) inseminations, days in milk at breeding, and mastitis as the most effective explanatory variables in predicting pregnancy outcome.

Key words: machine learning, reproductive management, dairy cattle

INTRODUCTION

Although it is often stated that the decline in reproductive performance of dairy cattle is due to intensive selection for high milk production, it is clear that many environmental features and management practices contribute directly to the insemination outcome. Management features, such as heat detection, nutrition, transition cow management, BCS, semen handling, metabolic disorders, udder health, calving difficulty, reproductive disease, and cow comfort strongly affect reproductive performance (Lucy, 2001; Caraviello et al., 2006; Schefers et al., 2010). Researchers have also reported associations between reproduction traits and genetics (Weigel, 2004; González-Recio and Alenda, 2005; Liu et al., 2008), milk yield (Berry et al., 2003; Windig et al., 2005, 2006; Tiezzi et al., 2011), heat stress (Morton et al., 2007), energy balance (de Vries and Veerkamp, 2000), timing of AI (Cornwell et al., 2006), reproductive health (Sheldon et al., 2002), lameness (Garbarino et al., 2004), quality and quantity of semen (Jaskowski and Szenfeld, 1999), sperm dosage in sex-sorted semen (DeJarnette et al., 2011), rump angle and conformation traits (Wall et al., 2005), and cow health (Chebel et al., 2004). Caraviello et al. (2006) used an alternating decision tree algorithm to identify frequency of hoof trimming, type of bedding in the dry cow pen, type of restraint system, and duration of the voluntary waiting period as key features in predicting first-service conception rate. They also found that bunk space per cow, temperature for thawing semen, percentage of cows with low BCS, number of cows in the maternity pen, strategy for using cleanup bulls, and milk yield at first service were the most informative variables in predicting the insemination outcome at 150 DIM.

Schefers et al., (2010) modeled conception rate and service rate of commercial dairy herds using a model tree algorithm. Their study identified percentage of repeated inseminations between 4 and 17 d post-AI

Received February 13, 2013.
Accepted September 11, 2013.
[1] Corresponding author: shahinfar@wisc.edu

(a measure of breeding protocol compliance), stocking density in the breeding pen, length of the voluntary waiting period, days from insemination to pregnancy check, and SCS as the most important features in predicting herd average conception rate. The most important explanatory variables for predicting herd average service rate were number of cows per breeding technician, resynchronization protocol, use of soakers in the holding area, and bunk space per cow in the breeding pens. The effects of negative energy balance in early lactation have been well studied and seem to be partially responsible for lower conception rates observed in high-producing cows. Oikonomou et al., (2008) showed that BCS, energy content of the diet, cumulative effective energy balance, and blood glucose have favorable genetic relationships with reproduction, whereas BHBA and NEFA are negatively correlated with energy balance and have unfavorable genetic correlations with reproductive traits. In that study, mean daily energy balance, milk protein content, and DMI during the first 28 d postpartum were associated with higher conception rate at first service, whereas cows with high DMI and positive energy balance had a shorter calving-to-conception interval. On the other hand, lower BCS have been associated with a longer calving-to-conception interval (Patton et al., 2007).

Although several studies have attempted to identify specific factors affecting insemination outcome in lactating dairy cattle, few have tried to predict the outcome of individual insemination events based on all health, reproduction, and production data available for each cow at the time of service. Obviously, such a prediction tool could be useful as a decision support system for dairy farmers.

The ability to accommodate large and complex data sets with missing values, as well as the lack of restrictive parametric assumptions, make machine learning methods good candidates for data mining and development of predictive tools in fields such as agriculture. Grzesiak et al. (2010) used artificial neural networks, multivariate adaptive regression splines, logistic regression, classification trees, and classification functions to classify cows with good or poor reproductive performance based on age, calving interval, gestation length, BCS, FCM, and average of fat and protein percentages. They reported classification accuracies of 85 to 86%, with sensitivity and specificity of 85%, for a multilayer perceptron with 2 hidden layers. Among the machine learning methods used in the animal sciences, artificial neural networks are the most frequently used, with applications such as predicting milk yield in dairy cows (Lacroix et al., 1995; Grzesiak et al., 2006; Gianola et al., 2011), classifying mastitis cases (Yang et al., 1999), classifying lameness in horses (Suchorski-Tremblay et al., 2001), predicting the slaughter weight of bull calves (Adamczyk et al., 2005), identifying SNP associated with chicken mortality (Long et al., 2009), and real-time prediction of breeding values in dairy cattle (Shahinfar et al., 2012).

The objective of this study was to compare the performance of different machine learning algorithms for predicting the insemination outcomes of lactating dairy cows using production, reproduction, health, and genetic information. Identification of specific environmental factors or management practices that affect reproductive performance is a by-product of the aforementioned analyses, but in this study our primary goal was to maximize predictive ability for development of a decision support tool.

MATERIALS AND METHODS

Data

The data used in this study were provided by 26 Wisconsin dairy farms that were enrolled in the Alta Genetics (Watertown, WI) Advantage Progeny Testing Program. A general description of these dairy herds can found in Table 1 of Schefers et al. (2010). After editing, the data set contained 129,245 breeding records from primiparous Holstein cows and 195,128 breeding records from multiparous Holstein cows. For each breeding event, there existed corresponding production data, EBV, health events, and reproduction information (Table 1). In terms of reproduction performance, herds in this study are representative of large commercial dairy farms in Wisconsin (Figure 1).

Production, reproduction, and health event data were obtained from backup files of the on-farm DairyCOMP 305 herd management software (Valley Agricultural Software, Tulare, CA) of individual farms. Estimated breeding values and calving ease data were extracted from the US Department of Agriculture Agricultural Research Service Animal Improvement Programs Laboratory (Beltsville, MD) database. Each data point in the final data set included 23 or 25 features and 1 binary response variable for primiparous or multiparous cows, respectively. Records were collected over a 10-yr period from 2000 to 2010.

To account for energy balance and reduce the dimensionality of features in the model for analysis, ECM was used as an explanatory variable. The following equation was used to determine the amount of energy needed for producing milk, adjusted to 3.5% fat and 3.2% true protein (Tyrrell and Reid, 1965):

$$\begin{aligned}\text{ECM} = [0.327 \times \text{milk (kg)}] + [12.95 \times \text{fat (kg)}] \\ + [7.65 \times \text{protein (kg)}].\end{aligned}$$

Table 1. Description of features (explanatory variables) used for predicting the outcome of insemination events in cows

			Primiparous cows				Multiparous cows			
No.	Feature	Type	Level	Mean	SD	Missing (%)	Level	Mean	SD	Missing (%)
1	Herd-year-month	Nominal	2,672	—	—	0	2,786	—	—	0
2	Mean conception rate in the herd	Numeric	—	0.36	0.07	0	—	0.31	0.064	0
3	Lactation	Numeric	—	—	—	—	8	2.9	1.176	0
4	Stage of lactation	Numeric	—	4.94	2.4	0	16	4.94	2.34	0
5	Age at calving (mo)	Numeric	—	24.2	2.6	0	115	50.64	16.69	0
6	ECM yield (kg)	Numeric	—	41.69	7.62	0	—	43.74	8.21	0
7	Fat:protein ratio	Numeric	—	117.6	24.2	5	228	116.9	26.4	5
8	SCS	Numeric	—	2.9	1.7	37	87	3.6	1.97	33
9	Previous days open	Numeric	—	—	—	—	371	133.9	73.8	3
10	Previous times bred	Numeric	—	1.7	1.1	14	21	2.74	2.2	4
11	Current times bred	Numeric	—	1.6	1.9	0	10	1.6	1.9	0
12	Mastitis	Binary	2	0.04	0.19	0	2	0.07	0.25	0
13	Lameness	Binary	2	0.01	0.09	0	2	0.02	0.14	0
14	Ketosis	Binary	2	0.001	0.03	0	2	0.00	0.04	0
15	Displaced abomasa	Binary	2	0.001	0.03	0	2	0.00	0.04	0
16	Retained placenta	Binary	2	0.05	0.21	0	2	0.08	0.26	0
17	DIM at breeding	Numeric	—	120.9	54.7	6	—	122.4	54.4	6
18	Breeding protocol	Nominal	13	—	—	9	13	—	—	9
19	Sex of the previous calf	Binary	2	—	—	39	2	—	—	33
20	Birth difficulty in last gestation	Nominal	5	—	—	39	5	—	—	56
21	Multiple birth in the last gestation	Binary	2	—	—	40	2	—	—	56
22	PTA of daughter pregnancy rate for calf	Numeric	—	−0.07	0.89	88	—	0.152	0.88	85
23	PTA of daughter pregnancy rate for sire of the cow	Numeric	—	−0.47	1.52	29	—	−0.38	1.498	41
24	Calf sire conception rate	Numeric	—	0.67	2.018	92	—	0.59	2.04	92
25	Inbreeding coefficient of the cow	Numeric	—	1.5	2.12	88	—	1.5	2.12	85
26	Pregnancy	Binary	2	—	—	0	2	—	—	0

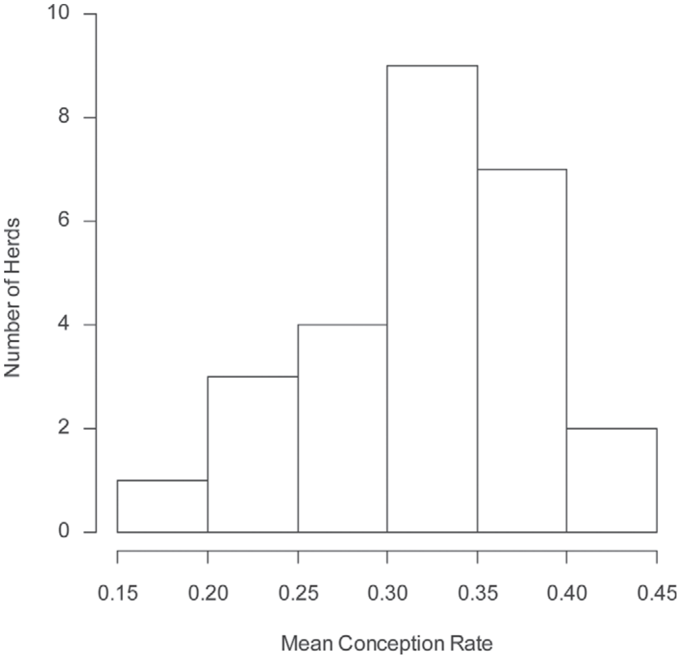

Figure 1. Histogram of mean conception rate in the 26 herds used in this study.

Machine Learning Algorithms

No systematic approach exists that one can use, a priori, to find the most suitable machine learning method for a particular task. Therefore, a common approach in machine learning studies is to test multiple leading algorithms on a new application. In this study, the leading algorithms for learning Bayesian networks and decision trees, including bagging and random forest algorithms that learn ensembles (groups) of trees were tested. The algorithms tested herein are among the most widely used in machine learning today. To classify insemination events into pregnant or nonpregnant outcomes based on the aforementioned explanatory variables, 5 types of machine learning algorithms were used: naïve Bayes, Bayesian networks, decision trees, bagging (ensemble of decision trees), and random forests. A brief explanation of each technique follows.

Naïve Bayes. Naïve Bayes (**NB**) is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. It is a statistical classifier based on Bayes rule (Domingos and Pazzani, 1997), and it is the simplest form of Bayesian network, in which all features are independent, given the value of the outcome (Figure 2a); simplicity, computational feasibility, and robustness make this method suitable for practical use. Naïve Bayes can tolerate dependencies between features very well and can often outperform more-elaborate methods, such as rule learners and decision tree learners (Clark and Niblett 1989; Cestnik, 1990). In addition, NB are quite intuitive and easy to understand, which is a big concern in the machine learning field (Kononenko, 1990). However, linear dependencies between features can reduce the power of NB, and careful selection among highly dependent features can be beneficial. Another concern about NB is that the assumption of normality for numeric features is not always true, and using kernel density estimation has been shown to confer improvement in this case (Witten and Frank, 2005). Suppose **F** is a vector of features ($f1, f2, \ldots, fn$), and c is a class variable with 2 values (O = open; P = pregnant). The probability (p) of the class variable (C) given the feature vector can be calculated as

$$p\left(C = c|f1,f2, \ldots, fn\right) = \frac{p\left(f1,f2, \ldots, fn|C = c\right)p(C = c)}{p(f1,f2, \ldots, fn)}.$$

Because we assume that all features are independent, given the class variable (conditional independence),

$$p\left(f1,f2, \ldots, fn|C = c\right)$$
$$= p\left(f1|C = c\right)p\left(f2|C = c\right) \ldots p(fn \mid C = c);$$

$$p\left(f1,f2, \ldots, fn\right) = p\left(f1\right)p\left(f2\right) \ldots p(fn)$$

can be calculated easily from the training data, as well as the prior probability for a given class, $p(C = c)$.

Bayesian Network. A Bayesian network (**BN**) represents the joint probability distribution of a set of variables$\{X1, X2, \ldots, Xn\}$ as a discrete acyclic graph and a set of conditional probability distributions that correspond to specific features (Figure 2b). The joint probability distribution can be calculated as $P(X_i) = \prod_i P(X_i \mid X_{Pa(i)})$, where $X_{Pa(i)}$ denotes a set of parent variables for node i (Kjærulff and Madsen, 2007). A small set of parent variables is preferred, because the network requires a parameter space that is exponential in the number of parents of each node (Lowed and Domingos, 2005).

When the structure of the network is known, learning reduces to estimating conditional probability distribution parameters; otherwise, the structure can be found by using a greedy hill-climbing search, starting from an empty network. Missing values can be estimated using an expectation maximization or maximum a posteriori probability algorithm (Jensen, 2001). Unlike NB, BN have no strong independence assumption between features. Bayesian networks are especially useful, because by using the Bayes theorem, it is easy to compute the probability distribution of children given the values of their parents, as well as the probability distribution of

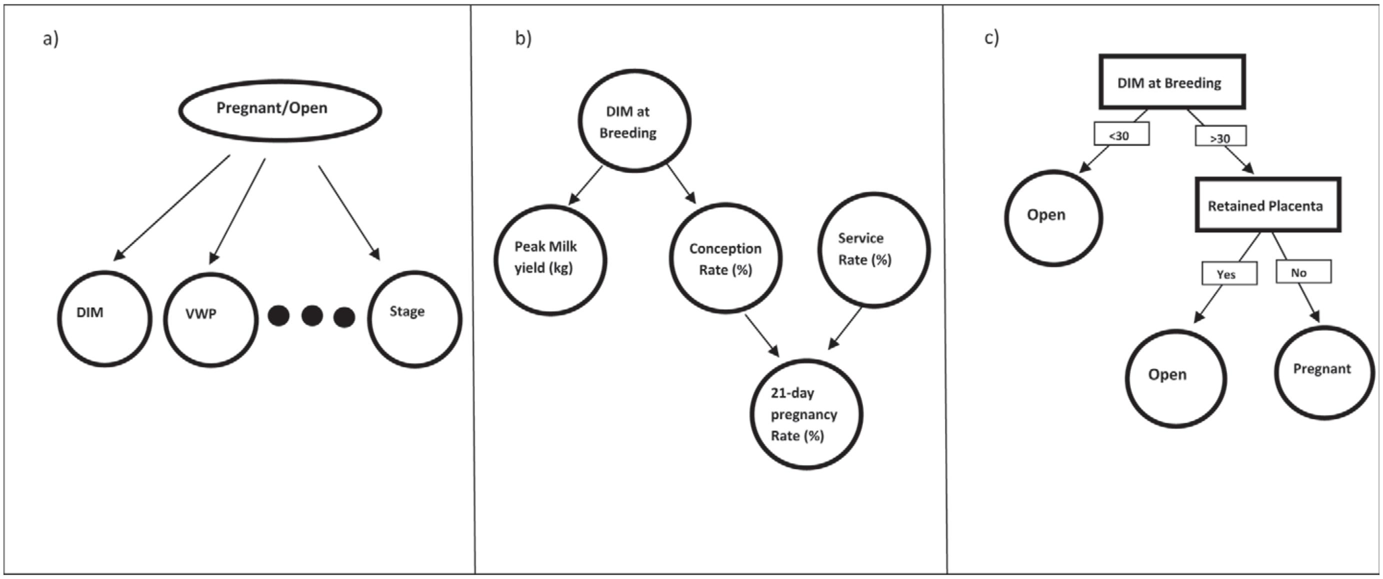

Figure 2. Schematic illustration of (a) Naïve Bayes classifiers, (b) Bayesian networks, and (c) decision trees. VWP = voluntary waiting period; stage = stage of lactation.

parents given the values of their children. That is, BN can proceed not only from causes to consequence, but also from consequence to causes (Uusitalo, 2007). Bayesian networks are well suited for small and incomplete data and can achieve good prediction accuracy in such situations (Kontkanen et al., 1997). Structural learning is possible with BN, and combining different sources of knowledge is a powerful property of BN, because prior knowledge can come from many different sources with different distributions. An additional advantage of BN is that, once the structure and parameters are known, any query can be done very quickly. Nevertheless, a few challenges exist with using BN in practice, including discretizing of continuous variables, collecting and structuring expert knowledge, and implementing recursive feedback loops (Uusitalo, 2007).

Decision Tree. Decision trees (**DT**) are among the simplest, most intuitive, easily interpretable, and widely used machine learning algorithms. They are tree-shaped models with tests carried out on feature values in the internal nodes and class labels in the leaves (Figure 2c). New instances are classified by passing through the nodes and corresponding tests within the tree, and the label of the final leaf reached by each instance is the predicted class for that example. The C4.5 algorithm (Quinlan, 1993) is a well-known algorithm for constructing decision trees. This algorithm builds the tree by a divide-and-conquer approach, choosing the most informative feature in each iteration using a heuristic function called the gain ratio (**GR**; described in a subsequent section of this paper). Features that correspond to the largest information gain (**IG**) ratios will be chosen by the algorithm in earlier iterations (at higher levels of the tree) to divide the instances. This process repeats recursively to construct the sub-trees (lower levels of the tree). The C4.5 algorithm will create a leaf and stop building the sub-tree if all of the remaining instances in the training set belong to the same class, or if the number of instances remaining is fewer than the minimum defined in the algorithm. Eventually, after building a tree to the maximum size, the C4.5 algorithm will prune the tree backward to reduce classification error due to overfitting. Decision trees are divide-and-conquer methods, and they are very useful for approximating discrete-valued functions. They are robust to noisy data and are capable of learning the disjunctives of conjunctions. Despite their simplicity and ease of understanding, some challenges exist in learning a DT, including determining the optimum depth of the tree, choosing an appropriate attribute selection method, and handling data with missing features (Mitchell, 1997) .

Bootstrap Aggregation. Bootstrap aggregation, which is also known as bagging (**BG**), is an ensemble method in which multiple versions of a predictor will be generated to drive an aggregated predictor. For predicting numeric values, aggregating would be a simple average over all models, whereas for classification purposes, it would be the majority vote of classes (Breiman, 1996). Essentially, BG improves prediction performance by building several models and letting them vote. Bagging can be used with any type of classifier, and it is easy to implement and computationally feasible. However, the black box nature of a bagged model makes it difficult to understand and interpret explicitly. Bagging uses the instability of a model to improve predictability. Therefore, BG of stable models, such as NB, will not improve predictive performance, but BG of unstable methods, such as tree learners, will improve the predictability of the ensemble model (Breiman, 1994). In BG, using a training set with n features and M instances, the algorithm will create m new training sets by taking bootstrap samples from the original data set. Then, using any machine learning technique, m models will be trained using the m bootstrap samples, and the final value will be generated by averaging the model predictions or voting for the majority class. In this paper, a tree learner category of classifiers that relies on a reduced-error pruning approach, commonly known as RepTree, was used (Witten and Frank, 2005).

Random Forest. Random forest (**RF**) is another ensemble method, in which training many classifiers using m bootstrap samples from the training set is combined with random selection of a subset of features for generating each of those classifiers (Ho, 1995; Breiman, 2001). Thus, an RF algorithm is very similar to BG, except that in each iteration of building a tree, RF picks a random subset of features and divides the instances based on the most-informative feature. Because the task is limited to a small subset of features and instances, RF is a computationally efficient technique that can be used with highly dimensional data sets. One of the biggest advantages of RF, and the main reason it was chosen in this study, is that RF is very efficient for estimating missing values and can maintain high accuracy when a large proportion of the data are missing; this is a common situation when analyzing producer-reported health data of dairy cattle.

Model Assessment

Receiver Operating Characteristic Curves. Traditionally, model performance assessment has relied on metrics derived from the confusion matrix. However, a scalar metric often provides a poor summary of the performance of a model, especially for nonparametric models. In addition, some performance metrics are sensitive to data discrepancies, such as skewness in class frequencies. In this case, the receiver operating characteristic (**ROC**) curve offers the same information as the confusion matrix, but in a much more intuitive and robust fashion (Hamel, 2008). In our study, we compared the performance of different models based on the area under the ROC curve, which allows comparison of different classifiers in terms of their misclassification costs (Provost and Fawcett, 1997). The ROC curve maps false-positive (**FP**) rate versus true-positive (**TP**) rate. The FP rate is the proportion of negative examples (nonpregnant cows) that are predicted incorrectly as positive examples (pregnant cows). The TP rate is the proportion of actual positive examples that are predicted correctly as positive examples (pregnant cows). Each point on the ROC curve corresponds to a threshold that can be used to classify examples into 2 classes. The upper left corner of the curve (FP = 0 and TP = 1) is the ideal point with respect to performance of a classifier. In this paper, we report the area under the curve (**AUC**), which is defined as the area between the ROC curve and the horizontal axis (FP rate). The closer that AUC is to 1, the better the performance of the classifier. In this study, the ROC curve corresponding to each model was obtained by 5-fold cross-validation, and we compared the AUC of different classifiers statistically using t-tests.

Feature Selection. To evaluate the relevance of features in our analysis, we used IG as the criterion and ranked features based on their IG. The IG statistic evaluates features by measuring their entropy or uncertainty with respect to the instance label or outcome.

Imagine a binary classification problem with positive and negative instances (N^+ and N^-, respectively) in the training set, such that the sum of N^+ and N^- equals N. The total entropy (H) contained within a data set described by a binary class variable can be calculated from the proportion of negative and positive examples in the training set as follows:

$$H\left(\frac{N^+}{N},\frac{N^-}{N}\right)=-\frac{N^+}{N}\log_2\frac{N^+}{N}-\frac{N^-}{N}\log_2\frac{N^-}{N}.$$

Let the entropy contained within the subset of data that correspond to level i of feature f be defined as

$$H\left(\frac{N_i^+}{N_i},\frac{N_i^-}{N_i}\right)=-\frac{N_i^+}{N_i}\log_2\frac{N_i^+}{N_i}-\frac{N_i^-}{N_i}\log_2\frac{N_i^-}{N_i},$$

where N_i^+ is the number of positive instances (pregnant cows) with level i of feature f and positive class variable, N_i^- is the number of negative instances (open cows) with level i of feature f and negative class variable, and N_i is the number of instances with level i of feature f with both positive and negative class variables. The change in entropy attributed to feature f with c levels is called IG (Russell and Norvig, 2002) and is calculated as

$$\mathrm{IG}(f)=H\left(\frac{N^+}{N},\frac{N^-}{N}\right)-\sum_{i=1}^{c}\frac{N_i^+ + N_i^-}{N}H\left(\frac{N_i^+}{N_i},\frac{N_i^-}{N_i}\right).$$

The larger the IG, the greater the relevance of that feature to the class variable. Note that H is a measure of impurity in the data and IG is a measure of the reduction in impurity obtained by dividing the data based on that feature. Dividing IG by the H of each feature will give us the proportion of IG that is explained by increasing each unit of H for that specific feature. This ratio is defined as the GR:

$$\mathrm{GR}\left(f\right)=\frac{H\left(\frac{N^{+}}{N},\frac{N^{-}}{N}\right)-\sum_{i=1}^{c}\frac{N_i^{+}+N_i^{-}}{N}H\left(\frac{N_i^{+}}{N_i},\frac{N_i^{-}}{N_i}\right)}{\sum_{i=1}^{c}\frac{N_i^{+}+N_i^{-}}{N}H\left(\frac{N_i^{+}}{N_i},\frac{N_i^{-}}{N_i}\right)}.$$

To assess the statistical significance of differences between algorithms in the AUC and the proportion of correctly classified instances (**CCI**), we used a 2-tailed paired t-test. Paired sample values were computed for each of the 5 algorithms using 5-fold cross-validation.

Lesion Approach. The complexity of some machine learning algorithms sometimes leads to difficulty in interpreting the results intuitively. Therefore, to gain insight about the interrelationships among variables, a lesion approach was used. In a lesion approach, potential explanatory variables are removed from the model one at a time to determine their contributions to performance of the model.

RESULTS AND DISCUSSION

As shown in Table 2, results of the IG analysis indicate that mean within-herd conception rate in the past 3 mo, herd-year-month (**HYM**) of breeding, DIM at breeding, number of inseminations in the current lactation, and stage of lactation appeared among the top 10 features for predicting insemination outcome in primiparous and multiparous cows. However, results of the GR analysis suggested greater impact of health traits, with ketosis, mastitis, retained placenta, and lameness among the top 10 features in primiparous cows, and with mastitis, displaced abomasa, and retained placenta among the 10 most important features in multiparous cows.

Classification Accuracy

Table 3 shows the AUC and the proportion of CCI for all 5 folds of the cross-validation, as well as the average of AUC and CCI across folds, for the 5 machine learning methods used in this paper. Likewise, Figure 3 compares the 5 machine learning techniques for classification accuracy between pregnant and nonpregnant primiparous cows. The RF algorithm outperformed all other methods, with a cross-validation AUC of 0.756, which was significantly ($P < 0.001$) better than BG, DT, BN, and NB. Very similar results were obtained for multiparous cows, with a cross-validation AUC of 0.736 for RF, which exceeded ($P < 0.001$) other methods (Figure 4), including BG, DT, BN, and NB. A similar trend was observed for CCI (Table 3). Although RF showed the best performance among methods considered in this study, a considerable number of misclassified instances remained. This lack of accuracy can be explained, in part, by the nature of reproductive data. The insemination outcome phenotype is typically characterized by low heritability, and it is highly affected by many environmental factors, such as breeding practices, health events, nutrition, and production level. Therefore, a very comprehensive data set is needed to predict the outcome of an insemination event accurately, and many farms do not record all variables consistently or completely.

Caraviello et al. (2006) used more than 300 potential explanatory variables and an alternating decision tree algorithm in a herd-based study and showed that hoof trimming, bedding in the dry cow pen, cow restraint system, and length of the voluntarily waiting period were the key variables affecting first-service conception rate. They also found that bunk space per cow, temperature for thawing AI semen, percentage of low-BCS cows within the herd, number of cows per maternity pen, strategy for using cleanup bulls, and milk yield at first AI were the most important explanatory variables for predicting pregnancy status at 150 DIM. They reported 75.6 and 71.4% accuracies of classification for predicting first-service conception rate and pregnancy status at 150 DIM, respectively, which are close to the accuracy levels achieved in the present study.

In a recent herd-based study, Schefers et al. (2010) reported coefficient of determination values of 35 and 40% for explaining the observed variation in conception rate and service rate, respectively. They reported that the percentage of repeated inseminations between 4 and 17 d post-AI, stocking density in the breeding pen, length of the voluntary waiting period, days from breeding to pregnancy check, and SCS were the most important features affecting conception rate. Number of lactating cows per breeding technician, use of a resynchronization program, utilization of soakers in the holding area during the summer, and bunk space per cow in the breeding pens were the most important features affecting service rate. The focus of the present study was slightly different, because the objective was to predict the outcome of each insemination event rather than identify significant environmental factors, but our results suggest a very similar conclusion in that plays a major role in determining reproductive outcomes.

Table 2. Features used to predict insemination outcomes in primiparous and multiparous Holstein cows, as ranked by information gain and gain ratio

	Primiparous		Multiparous	
Rank	Information gain[1]	Gain ratio[2]	Information gain	Gain ratio
1	Herd-year-month	Mean conception rate in the herd	Herd-year-month	Mean conception rate in the herd
2	Mean conception rate in the herd	Ketosis	Mean conception rate in the herd	Stage of lactation
3	Stage of lactation	Current times bred	Stage of lactation	Herd-year-month
4	DIM at breeding	Stage of lactation	DIM at breeding	DIM at breeding
5	Current times bred	Herd-year-month	Current times bred	Mastitis
6	Breeding protocol	DIM at breeding	Previous times bred	Current times bred
7	PTA of daughter pregnancy rate for sire of the cow	Mastitis	Previous days open	Previous times bred
8	Age at calving (mo)	Retained placenta	Age at calving (mo)	Previous days open
9	ECM yield (kg)	Lameness	Breeding protocol	Displaced abomasum
10	SCS	Breeding protocol	Lactation number	Retained placenta
11	Birth difficulty in last gestation	PTA of daughter pregnancy rate for sire of the cow	SCS	Age at calving (mo)
12	Previous times bred	Age at calving (mo)	Mastitis	Breeding protocol
13	Mastitis	ECM yield (kg)	PTA of daughter pregnancy rate for sire of the cow	ECM yield (kg)
14	Retained placenta	Birth difficulty in last gestation	Retained placenta	Lactation number
15	PTA of daughter pregnancy rate for calf	Previous times bred	Fat:protein ratio	SCS
16	Sex of the previous calf	SCS	PTA of daughter pregnancy rate for calf	PTA of daughter pregnancy rate for sire of the cow
17	Fat:protein ratio	Sex of the previous calf	ECM yield (kg)	Fat:protein ratio
18	Lameness	PTA of daughter pregnancy rate for calf	Birth difficulty in last gestation	Lameness
19	Ketosis	Fat:protein ratio	Multiple birth in the last gestation	Birth difficulty in last gestation
20	Inbreeding coefficient of the cow	Calf's sire, sire conception rate	Sex of the previous calf	PTA of daughter pregnancy rate for calf
21	Calf's sire, sire conception rate	Inbreeding coefficient of the cow	Lameness	Multiple birth in the last gestation
22	Displaced abomasum	Displaced abomasum	Displaced abomasum	Sex of the previous calf
23	Multiple birth in the last gestation	Multiple birth in the last gestation	Inbreeding coefficient of the cow	Inbreeding coefficient of the cow
24			Calf's sire, sire conception rate	Calf's sire, sire conception rate
25			Ketosis	Ketosis

[1] Information gain (IG) is a measure of entropy or uncertainty with respect to the instance outcome.

[2] Gain ratio is the result of dividing IG by the entropy of the feature, which is the proportion of IG that is explained by increasing each unit of the entropy of that specific feature.

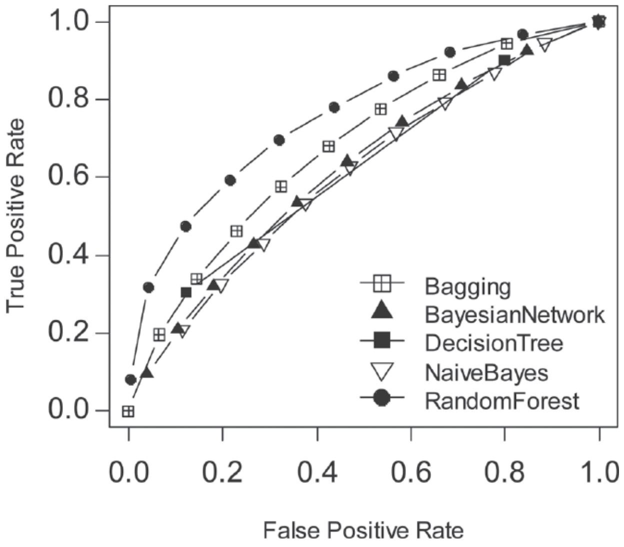

Figure 3. Receiver operating characteristic curves for 5 types of machine learning algorithms used to predict insemination outcomes in primiparous cows.

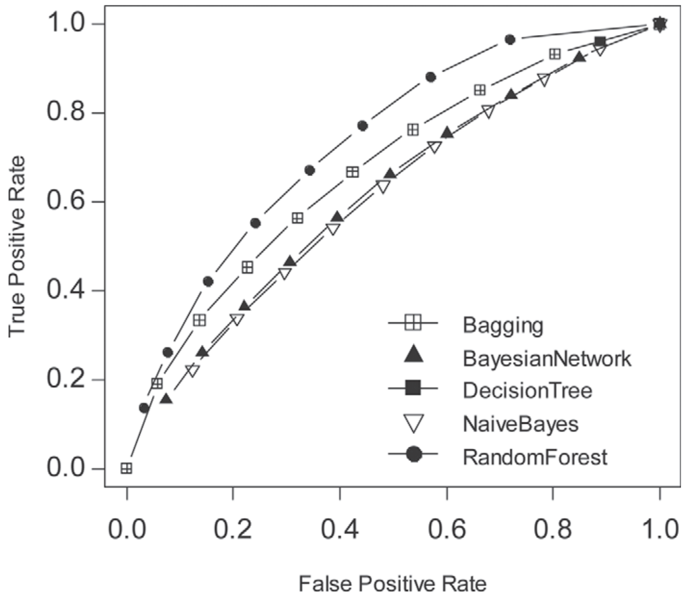

Figure 4. Receiver operating characteristic curves for 5 types of machine learning algorithm used to predict pregnancy outcomes in multiparous cows.

In brief, the superior performance of BG and RF can be explained by the power of ensemble methods to generate high-performance classifiers by training a collection of individual classifiers. Among the two, RF had slightly better performance in this study due to its property of random selection of feature subsets. Note that this discussion does not imply that RF is always the best algorithm for predicting insemination outcomes in dairy cattle. Grzesiak et al. (2010) attempted to classify cows into 2 categories based on reproductive performance using several environmental and phenotypic variables: "good" cows, with ≤2 AI services per conception and "poor" cows, with >2 AI services per conception. They found that a multilayer perceptron neural network with 2 hidden layers was the best predictor, with 85.7% accuracy. Because they had a very small data set (768 training instances and 150 testing instances), and because artificial neural networks tend to overfit small data sets (Mitchell, 1997), it is possible that their model was overtrained on that specific data set and lacks generality for widespread use. A common practice in machine learning analyses is to evaluate several different learning algorithms on each specific task

Table 3. Area under the receiver operating characteristic (ROC) curve (AUC) and proportion of correctly classified instances (CCI) for each of the 5 classification algorithms by 5-fold cross-validation in primiparous and multiparous cows

	Fold cross-validation											
	AUC						CCI					
Algorithm[1]	1	2	3	4	5	Mean ± SE	1	2	3	4	5	Mean ± SE
Primiparous												
NB	0.61	0.61	0.60	0.61	0.61	60.8 ± 0.004^{a}	60.5	61.0	60.6	60.7	60.7	60.7 ± 0.19^{a}
BN	0.62	0.62	0.62	0.62	0.62	62.0 ± 0.00^{a}	62.8	63.9	63.8	63.8	63.4	63.5 ± 0.46^{a}
DT	0.64	0.65	0.64	0.66	0.64	64.6 ± 0.009^{a}	66.7	66.7	66.7	66.3	66.1	66.5 ± 0.29^{a}
BG	0.68	0.68	0.67	0.68	0.67	67.6 ± 0.005^{a}	67.0	67.7	67.2	67.6	66.9	67.3 ± 0.36^{a}
RF	0.75	0.76	0.75	0.76	0.76	75.6 ± 0.005^{b}	72.1	72.6	72.3	72.1	72.2	72.3 ± 0.21^{b}
Multiparous												
NB	0.61	0.61	0.61	0.60	0.61	60.8 ± 0.004^{a}	63.3	63.7	63.3	63.7	63.5	63.5 ± 0.2^{a}
BN	0.62	0.61	0.61	0.62	0.62	61.6 ± 0.005^{a}	68.0	68.2	67.8	68.3	68.0	68.1 ± 0.19^{a}
DT	0.61	0.60	0.62	0.60	0.61	60.8 ± 0.008^{a}	68.6	69.2	68.9	68.9	68.8	68.9 ± 0.22^{a}
BG	0.67	0.67	0.67	0.67	0.67	67.0 ± 0.00^{a}	70.2	70.5	70.3	70.4	70.5	70.4 ± 0.13^{a}
RF	0.74	0.74	0.74	0.73	0.73	73.6 ± 0.005^{b}	73.4	73.9	73.9	73.6	73.3	73.6 ± 0.28^{b}

[a,b] Means within a column with different superscripts differ ($P < 0.001$).

[1] NB = naïve Bayes; BN = Bayesian network; DT = decision tree; BG = bagging; RF = random forest.

of interest, because the performance of these algorithms may change due to differences in the size, structure, and other characteristics of the data set.

Lesion Results

In this study, a lesion approach was carried out using the best-performing algorithm (RF) to further investigate the relative contributions of specific variables to predict the phenotype. The results of this analysis are shown in Figures 5 and 6 for primiparous and multiparous cows, respectively. Considering the inclusion of all variables in the model as the baseline, as in the case of primiparous cows (Figure 5), eliminating features other than HYM caused a change in predictive ability of <1% relative to the baseline model. In the case of HYM, removing it from the list of input features increased the AUC from 0.76 to 0.77 and increased the CCI from 0.72 to 0.73. In looking at the curves carefully, none dominates the other over the entire decision space. In the high-specificity region (left half of the plot), the baseline model (full model) is dominant. Conversely, in the high-sensitivity region (right half of the plot), the model without HYM performs better. In this case, the decision will be based on the region in which one wants to operate, as well as the cost of the misclassified instances (FP and FN). In our study, the cost of FN instances (failing to inseminate a cow that would have become pregnant) is much greater than the cost of FP instances (inseminating a cow that will not become pregnant). Therefore, FN should be avoided more precisely than FP; in other words operating in the high-sensitivity region is of greater interest and, therefore, excluding HYM will be beneficial. In the case of multiparous cows (Figure 6), eliminating HYM caused the AUC to change from 0.74 to 0.76 and CCI to change from 0.73 to 0.75. As shown in Figure 6, the reduced model dominates the full model in the entire plot; hence, eliminating HYM from the model is helpful in both the high-specificity and high-sensitivity regions.

The decision of whether or not to include HYM in the model represents a balance between explaining variation in the current data set and generalizing our results to other data sets. When included, HYM accounts for significant variation in fertility, because it describes the management practices and environment to which the cow is exposed at that point in time. However, many other features in the model can explain known management and environmental factors, and the marginal value of including HYM is accounting for unknown or unreported factors that are not repeatable across herds.

After removing HYM from the feature list, a second round of the lesion analysis was carried out recursively. Results of the second round of the lesion analysis were not clear in terms of model performance and did not provide insight toward an interpretive conclusion. It is worth noting that many of the explanatory variables are interrelated. Therefore, excluding one variable from the model will not have a major impact on model performance, because other correlated features in the model will compensate the effect of the excluded

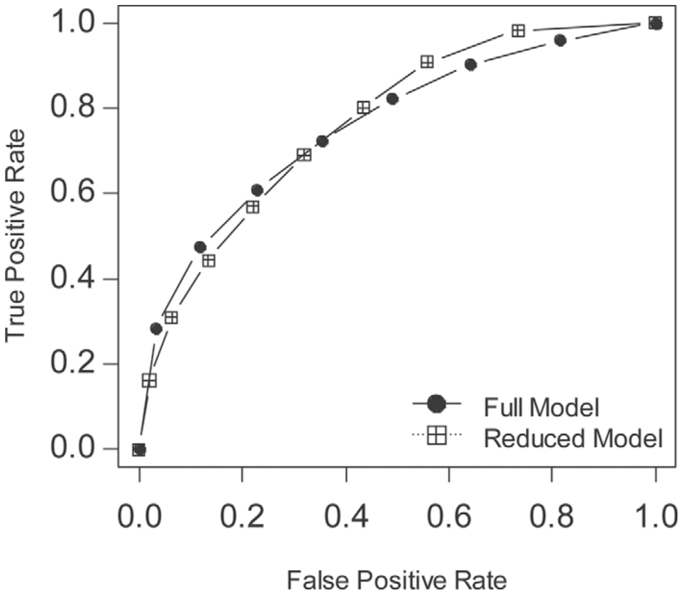

Figure 5. Receiver operating characteristic curves for full and reduced [without herd-year-month (HYM) of insemination] random forests (RF) models in primiparous cows in a lesion analysis.

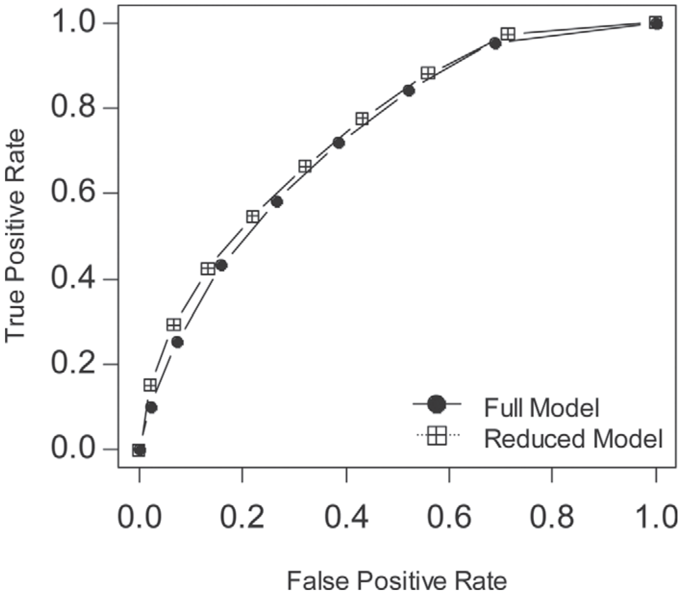

Figure 6. Receiver operating characteristic curves for full and reduced [without herd-year-month (HYM) of insemination] random forests (RF) models in multiparous cows in a lesion analysis.

variable. However, inclusion of more variables in the model may enhance the robustness and generality of the model in the case of missing values (which are common in field data), because other variables can explain the effect of the missing value through their underlying relationships. Finally, this area should be explored further to gain additional insight into the interrelationships between features and the corresponding impact on predictability of the models.

CONCLUSIONS

The machine learning algorithms considered in this study were effective in predicting pregnant versus nonpregnant cows at the time of insemination. Among the algorithms considered in this paper, RF was significantly better in terms of classification accuracy (72.3 and 73.6% for primiparous and multiparous cows, respectively) and area under the ROC curve (75.6 and 73.6%, respectively). Evaluation of features by information gain indicated that the mean within-herd conception rate in the past 3 mo, HYM of breeding, DIM at breeding, number of inseminations in the current lactation, and stage of lactation when the breeding occurred were the most informative features for predicting insemination outcome. In addition, the GR analysis suggested greater importance of health traits in explaining insemination outcomes relative to the information gain analysis. Based on the GR, the incidence of ketosis, mastitis, retained placenta, and lameness (for primiparous cows), and the incidence of mastitis, displaced abomasa, and retained placenta (multiparous cows) were the most important explanatory variables. The results of the lesion analysis suggested that excluding the HYM of insemination from the feature set may improve the predictability of these models in independent data sets. Overall, results of this paper suggest that, although prediction of the insemination outcome for individual lactating dairy cows is extremely difficult, information regarding health, reproductive history, production level, and other environmental features can be used to identify highly fertile subsets of cows. Decision support tools developed using this methodology may allow dairy farmers to optimize their breeding programs by targeting animals that are most likely to become pregnant. Such tools could be especially valuable in herds that use sex-enhanced semen or expensive semen from high-merit sires.

ACKNOWLEDGMENTS

This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2010-85122-20612 from the USDA National Institute of Food and Agriculture (Washington, DC). Data for this study was provided by Alta Genetics (Watertown, WI) Advantage Progeny Testing Program and USDA-ARS Animal Improvement Programs Laboratory (Beltsville, MD). The authors acknowledge the technical comments and support from John Schefers, Daniel Gianola, and Gilherme Rosa (all from Department of Dairy Science, University of Wisconsin-Madison). Saleh Shahinfar acknowledges Paulino Pérez Rodríguez, Marion Dutreuil, and Lydia Hardie (all from Department of Dairy Science, University of Wisconsin-Madison) for their assistance in making graphs. Kent Weigel acknowledges partial financial support from the National Association of Animal Breeders (Columbia, MO).

REFERENCES

Adamczyk, K., K. Molenda, J. Szarek, and G. Skrzyński. 2005. Prediction of bulls' slaughter value from growth data using artificial neural network. J. Cent. Euro. Agric. 6:133–142.

Berry, D. P., F. Buckley, P. Dillon, R. D. Evans, M. Rath, and R. F. Veerkamp. 2003. Genetic parameters for body condition score, body weight, milk yield, and fertility estimated using random regression models. J. Dairy Sci. 86:3704–3717.

Breiman, L. 1994. Bagging predictors. Technical report. Department of Statistics, University of California Berkeley, CA.

Breiman, L. 1996. Bagging predictors. Mach. Learn. 24:123–140.

Breiman, L. 2001. Random forests. Mach. Learn. 45:5–32.

Caraviello, D. Z., K. A. Weigel, M. Craven, D. Gianola, N. B. Cook, K. V. Nordlund, P. M. Fricke, and M. C. Wiltbank. 2006. Analysis of reproductive performance of lactating cows on large dairy farms using machine learning algorithms. J. Dairy Sci. 89:4703–4722.

Cestnik, B. 1990. Estimating probabilities: A crucial task in machine learning. Pages 147–149 in Proc. 9th European Conference on Artificial Intelligence, Stockholm, Sweden. Pitman Publishing Ltd., London, UK.

Chebel, R. C., J. E. P. Santos, J. P. Reynolds, R. L. Cerri, S. O. Juchem, and M. Overton. 2004. Features affecting conception rate after artificial insemination and pregnancy loss in lactating dairy cows. Anim. Reprod. Sci. 84:239–255.

Clark, P., and T. Niblett. 1989. The CN2 induction algorithm. Mach. Learn. 3:261–283.

Cornwell, J. M., M. L. McGilliard, R. Kasimanickam, and R. L. Nebel. 2006. Effect of sire fertility and timing of artificial insemination in a Presynch + Ovsynch protocol on first-service pregnancy rates. J. Dairy Sci. 89:2473–2478.

de Vries, M. J., and R. F. Veerkamp. 2000. Energy balance of dairy cattle in relation to milk production variables and fertility. J. Dairy Sci. 83:62–69.

DeJarnette, J. M., M. A. Leach, R. L. Nebel, C. E. Marshall, C. R. McCleary, and J. F. Moreno. 2011. Effects of sex-sorting and sperm dosage on conception rates of Holstein heifers: Is comparable fertility of sex-sorted and conventional semen plausible? J. Dairy Sci. 94:3477–3483.

Domingos, P., and M. J. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. 29:103–130.

Garbarino, E. J., J. A. Hernandez, J. K. Shearer, C. A. Risco, and W. W. Thatcher. 2004. Effect of lameness on ovarian activity in postpartum Holstein cows. J. Dairy Sci. 87:4123–4131.

Gianola, D., H. Okut, K. A. Weigel, and G. J. M. Rosa. 2011. Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows and wheat. BMC Genet. 12:87–101.

González-Recio, O., and R. Alenda. 2005. Genetic parameters for female fertility traits and fertility index in Spanish dairy cattle. J. Dairy Sci. 88:3282–3289.

Grzesiak, W., P. Błaszczyk, and R. Lacroix. 2006. Methods of predicting milk yield in dairy cows—Predictive capabilities of Wood's lactation curve and artificial neural networks (ANNs). Comput. Electron. Agric. 64:69–83.

Grzesiak, W., D. Zaborski, P. Sablik, A. Żukiewicz, A. Dybus, and I. Szatkowska. 2010. Detection of cows with insemination problems using selected classification models. Comput. Electron. Agric. 74:265–273.

Hamel, L. 2008. Model Assessment with ROC Curves. Pages 1316–1323 in The Encyclopedia of Data Warehousing and Mining. 2nd ed. Information Science Reference, Hershey, PA..

Ho, T. K. 1995. Random decision forest. Pages 278–282 in Proc. 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada.

Jaskowski, J. M., and J. Szenfeld. 1999. The influence of the quantity and quality of semen and insemination techniques on results of pregnancies in cows. Med. Weter. 55:160–162.

Jensen, F. V. 2001. Bayesian Networks and Decision Graphs. Springer-Verlag, New York, NY.

Kjærulff, U. B., and A. L. Madsen. 2007. Bayesian Networks and Influence Diagrams. Springer, New York, NY.

Kononenko, I. 1990. Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. Pages 190–197 in Current Trends in Knowledge Acquisition. B. Wielinga, J. Boose, B. Gaines, G. Schreiber, and M. van Someren, ed. IOS Press, Amsterdam, the Netherlands.

Kontkanen, P., P. Myllymäki, T. Silander, H. Tirri, and P. Grunwald. 1997. Comparing predictive inference methods for discrete domains. Pages 311–318 in Proc. 6th International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, FL.

Lacroix, R., K. M. Wade, R. Kok, and J. F. Hayes. 1995. Prediction of cow performance with a connectionist model. Trans. ASAE 38:1573–1579.

Liu, Z., J. Jaitner, F. Reinhardt, E. Pasman, S. Rensing, and R. Reents. 2008. Genetic evaluation of fertility traits of dairy cattle using a multiple-trait animal model. J. Dairy Sci. 91:4333–4343.

Long, N., D. Gianola, G. J. M. Rosa, K. A. Weigel, and S. Avendaño. 2009. Comparison of classification methods for detecting associations between SNPs and chick mortality. Genet. Sel. Evol. 41:18–32.

Lowed, D., and P. Domingos. 2005. Naïve Bayes models for probability estimation. Pages 529–536 in Proc. 22nd International Conference on Machine Learning, Bonn, Germany. ACM, New York, NY.

Lucy, M. C. 2001. Reproductive loss in high-producing dairy cattle: Where will it end? J. Dairy Sci. 84:1277–1293.

Mitchell, T. M. 1997. Artificial neural networks. Pages 111–112 in Machine Learning. McGraw-Hill International Edition, New York, NY.

Morton, J. M., W. P. Tranter, D. G. Mayer, and N. N. Jonsson. 2007. Effects of environmental heat on conception rates in lactating dairy cows: Critical periods of exposure. J. Dairy Sci. 90:2271–2278.

Oikonomou, G., G. Arsenos, G. E. Valergakis, A. Tsiaras, D. Zygoyiannis, and G. Banos. 2008. Genetic relationship of body energy and blood metabolites with reproduction in Holstein cows. J. Dairy Sci. 91:4323–4332.

Patton, J., D. A. Kenny, S. McNamara, J. F. Mee, F. P. O'Mara, M. G. Diskin, and J. J. Murphy. 2007. Relationships among milk production, energy balance, plasma analytes, and reproduction in Holstein-Friesian cows. J. Dairy Sci. 90:649–658.

Provost, F. J., and T. Fawcett. 1997. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. Pages 14–17 in Proc. 3rd International Conference on Knowledge Discovery and Data Mining, Newport Beach, CA. AAAI Press, Menlo Park, CA.

Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Series in Machine Learning. John Wiley & Sons, New York, NY.

Russell, S., and P. Norvig. 2002. Artificial Intelligence: A Modern Approach. 3rd ed. Pearson, Upper Saddle River, NJ.

Schefers, 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., H. Mehrabani-Yeganeh, C. Lucas, A. Kalhor, M. Kazemian, and K. A. Weigel. 2012. Prediction of breeding values for dairy cattle using artificial neural networks and neuro-fuzzy systems. Comput. Math. Meth. Med. 2012:127130.

Sheldon, I. M., D. E. Noakes, A. N. Rycroft, D. U. Pfeiffer, and H. Dobson. 2002. Influence of uterine bacterial contamination after parturition on ovarian dominant follicle selection and follicle growth and function in cattle. Reproduction 123:837–845.

Suchorski-Tremblay, A. M., R. Kok, and J. J. Thompson. 2001. Modelling horse hoof cracking with artificial neural networks. Can. Biosyst. Eng. 43:715–722.

Tiezzi, F., C. Maltecca, M. Penasa, A. Cecchinato, Y. M. Chang, and G. Bittante. 2011. Genetic analysis of fertility in the Italian Brown Swiss population using different models and trait definitions. J. Dairy Sci. 94:6162–6172.

Tyrrell, H. F., and J. T. Reid. 1965. Prediction of the energy value of cow's milk. J. Dairy Sci. 48:1215–1223.

Uusitalo, L. 2007. Advantages and challenges of Bayesian networks in environmental modeling. Ecol. Modell. 203:312–318.

Wall, E., I. M. S. White, M. P. Coffey, and S. Brotherstone. 2005. The relationship between fertility, rump angle, and selected type information in Holstein-Friesian cows. J. Dairy Sci. 88:1521–1528.

Weigel, K. A. 2004. Improving the reproductive efficiency of dairy cattle through genetic selection. J. Dairy Sci. 87(E. Suppl.):E86–E92.

Windig, J. J., M. P. L. Calus, B. Beerda, and R. F. Veerkamp. 2006. Genetic correlation between milk production and health and fertility depending on herd environment. J. Dairy Sci. 89:1765–1775.

Windig, J. J., M. P. L. Calus, and R. F. Veerkamp. 2005. Influence of herd environment on health and fertility and their relationship with milk production. J. Dairy Sci. 88:335–347.

Witten, I. H., and E. Frank. 2005. The explorer. Pages 407–408 in Data Mining. 2nd ed. Elsevier, San Francisco, CA.

Yang, X. Z., R. Lacroix, and K. M. Wade. 1999. Neural detection of mastitis from Dairy Herd Improvement records. Trans. ASAE 42:1063–1072.