Phenotypic and Genome-Enabled Prediction of Reproductive Performance in Dairy Cattle Using Machine Learning Algorithms

By

Saleh Shahinfar

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Dairy Science)

at the

UNIVERSITY OF WISCONSIN-MADISON
2014

Date of final oral examination: 6/6/2014

The dissertation is approved by the following members of the Final Oral Committee

Kent Weigel, Professor, Dairy Science
Victor Cabrera, Assistant Professor, Dairy Science
Paul Fricke, Professor, Dairy Science
Daniel Gianola, Professor, Dairy Science
David Page, Professor, Department of Biostatistics and Medical Informatics and Department of Computer Science

Abstract

Fast and cost-effective prediction models are increasingly in demand for commercial use. Prediction of the outcomes of insemination events as successes or failures based on explanatory variables related to genetic predisposition, health history, and lactation performance can have an impact on decision-making on dairy farms. However, interactions between management and physiological features are very complex. Machine learning algorithms can be useful for understanding these complex interactions and developing tools that will help farmers make accurate reproductive management decisions. Results of this study showed that random forests have the best performance in predicting the outcome of an insemination event and that health records of the cow are very important in this prediction. Optimizing classification rate without taking into account the cost of classification errors can be misleading. Nevertheless, the cost of not breeding a cow that would have conceived is much higher than the cost of breeding a cow that would not conceive. The common practice on most commercial dairy farms is to inseminate all cows that are eligible for breeding, which is debatable.

In conjunction with a lift chart analysis, which guides selection of subsets of highly or lowly fertile animals with highest and lowest probabilities of conception, the approach described herein could successfully stratify the pool of eligible cows in order to use different breeding strategies or use semen with different prices in different subsets of eligible cows in order to maximize total economic gain, as well as profit per eligible cow. This approach can enhance profitability of the dairy farm if sufficient data regarding variables that affect insemination outcomes are available.

Fuzzy expert systems are distinguished from other black boxed non-parametric methods, such as random forests and artificial neural networks, because they are easy to understand and interpret. There is lack of research on rule-based methods for genomic selection, because knowledge acquisition in such a complex and highly dimensional space is a limiting factor. In this dissertation, a hybrid fuzzy expert system, which uses genetic algorithms and particle swarm optimization as knowledge acquisition tools from the data was introduced for prediction of daughter pregnancy rate in Holstein bulls.

Table of Content

List of tables

List of Figures

Acknowledgments:

I would like to express my deepest gratitude to my advisor, Professor Kent Weigel, for his excellent advice, patience and support. He generously provided me with a great atmosphere for study and research. His supportive and encouraging attitude, along with his deep and comprehensive knowledge, makes him an outstanding advisor. His guidance is built around flexibility and motivation rather than pressure. In my five years working with him, after each meeting with him I admired his knowledge and manner more than the last time. He spent a lot of time patiently correcting my poor English writings which I appreciate greatly! Without his support I would not be where I am today.

My connection with him goes back to March of 2006 when I was a master's student at the University of Tehran and contacted him to ask for some progeny test data to work with in my Msc thesis. Although he generously provided me format 38 of USDA-ARS data file, I was not able to use it. It wasn't until September of 2008 that I decided to send my application for a PhD position to him. By accepting me as his PhD student he totally changed my path in life and opened the door to a different world for me. I am proud to be his student and will be forever grateful.

Next I would like to thank my committee members. I would like to acknowledge Professor Victor Cabrera's great guidance and suggestions. Seeing him with a happy face, full of energy all the time was enough to give me a better and more energetic feeling, which I will never forget.

I would like to thank Professor David Page for his great advice and suggestions. He was the one who de-mystified machine learning methods and their underlying algorithm for me and gave me the confidence to implement them on my own. This is a life changing skill I learned at UW-Madison and mainly through him.

I would like to thank Professor Daniel Gianola for being a great mentor although I was never smart enough to understand him. I would like to thank Professor Paul Fricke for his deep insights. I learned the concept of 21-day pregnancy rate from him which was a key point in my research.

I would like to thank Jerry Guenther for his unexpected help to teach me how to use DC305, and for providing me data. He motivated me and gave me confidence to develop the cost sensitive and lift chart analysis. His friendship is very important to me.

Next I would like to thank the technical staff: Cathy Rook, Nancy Hilmanowski, Minh Ngo and Steve Swintzer They did their best to facilitate my comfort from the time I got admitted to the program until now. They were very supportive and helpful during my stay in the department.

I would like to thank all of my officemates throughout my PhD, Jon Scheffers, Lydia Hardy, Tom Murphy, Weijing Tang, and Mariana Mamede. Their presence made office #450 the best office in the building which was the envy of everyone else. Our tea time, long challenging conversations, gossips and our songs for each other made it an incredible place to be and work in.

Next I would like to thank my colleagues on the 4th floor those that are here and those that are elsewhere at the moment. For their help here and there, for their presence that made the 4th floor a pleasant environment to work. And most importantly for their valuable friendship. Gota Morota was my "R" reference and more, I learned a lot from him. Paolino Prez for his excellent explanation on Bayesian methods. Afshin Kalantari for his great attitude, and collaboration in research and publication. Heather Adams, she introduced me to scripting in BASH and made my life easier. Bruno Valente for his technical support on the breeder's machine. Yalda Zare for her complains and friendship. Dave Kantor for our conversation about culture, politics, religion and wrestling. Tim Beisinger, Chen Yao, George Shook, Hassan Khatib, Guilherme Rosa, Dave

Thomas, Karen Nielsen, Kati Schopke, Abdalah Emhimad, Marcos Lagrotta, Flavio Cese Arantes, Fabiane Silva, Rostam Abdolahi, Huihui Duan, Yaodong Hu, Fransisco Penagriacano, Llibertat Tusel, Vivian Felipe, and all other graduate students and post docs on our floor for their excellence in friendship.

I would like to thank Alta Genetics for providing data for the first part of my research.

Funding by Agriculture and Food Research Initiative Competitive Grant no. 2010-85122-20612 from the USDA National Institute of Food and Agriculture is highly appreciated.

I would like to thank my boss in Alta genetics Dr. Bob Welper for his support in ordering me to finish my PhD as soon as possible.

I would like to thank my dearest family, for all of their support and encouragement during my life, for all of the sacrifices they made for me to achieve my goals And I dedicate this dissertation to my family, and all of whom I learned from during my life.

Chapter 1

Introduction

A considerable number of articles in the scientific literature on dairy cattle reproduction demonstrate that, along with large increases in the milk production of dairy cattle over the past five decades, the reproductive performance of dairy cattle has declined significantly (Lucy, 2001). Poor reproductive performance in dairy cattle is a complex problem that can be caused by physiology, genetics, management, health, and nutritional factors. Due to low heritability, improvements in reproductive traits by genetic selection tend to occur very slowly. Obviously, the fastest way to enhance reproduction is through alteration of management practices, which requires an understanding of how different environmental and management factors interact with each other to influence reproductive performance on commercial dairy farms. However, genetic selection is the only way to confer permanent improvement in a population that will be transmitted to future generations. Developing an understanding of the interactions between environmental factors, management practices, and genetic predisposition with regard to dairy cow fertility is our primary interest, but these interactions are complex and unknown. Therefore using conventional statistical methods to unveil these underlying relationships is not feasible. On the other hand, machine learning methods can handle complex interactions between variables, accommodate missing values for some variables, and account for multicollinearity between variables when computing predictions.

The objectives of this study were: 1) develop an accurate model for predicting the reproductive performance of dairy cattle using environmental, genetic, and management variables; 2) incorporate economic information regarding the costs and benefits of incorrect and correct predictions for the purpose of creating an on-farm decision-making tool that will help farmers make optimal breeding decisions, and 3) evaluate a novel fuzzy logic approach for the prediction of genetic merit for male fertility and related traits that might accommodate uncertainty about the genetic architecture of these traits.

The structure of this dissertation is as follows: Chapter 2 provides an in-depth literature review about different factors affecting reproductive performance and introduces the concepts associated with the analytical methods and algorithms used in the subsequent chapters. Chapter 3 provides an application of five alternative machine learning algorithms for predicting the outcome of insemination events based on production, reproduction, genetic, and health events. Chapter 4 incorporates cost-sensitive evaluation and lift chart analysis using the best-performing algorithm from the analysis in Chapter 3, namely a random forest algorithm, to show the economic efficiency of different scenarios for inseminating subsets of eligible cows. Chapter 5 describes a novel approach for designing a fuzzy expert system to predict the daughter pregnancy rate and related traits in dairy cattle using information from single nucleotide polymorphism (SNP) genotypes. Chapter 6 is about prediction of retention pay-off using model trees and Chapter 7 provides a summary of entire project and offers general conclusions and some recommendation for future research.

Data for this project were from several different sources. Data used in Chapter 3 were from 26 herds participating in the Alta Genetics (Watertown, WI) Advantage Program. These herds, which are spread throughout Wisconsin, are commercially oriented and tend to collect more detailed information regarding health events than other commercial dairy farms. Data used in Chapter 4 were from three different farms located in Wisconsin, in addition to the 26 herds described in Chapter 3. Data used in Chapter 5 were from 8,010 genotyped Holstein bulls and were provided by the USDA-ARS Animal Improvement Programs Laboratory (Beltsville, MD). Data used in

Chapter 6 were from a dynamic programming simulation, as well as real data from University of Wisconsin-Madison dairy herd.

Several different methods and algorithms were used in this research project, and these fall into four categories of machine learning algorithms. From the probabilistic learning methods, we used naïve Bayes and Bayesian networks. These methods utilize a probabilistic learning algorithm to acquire knowledge from the data during the learning process. Decision trees were used from the information-based category of learning, in which learning is based on information gain theory. From the ensemble group of methods, we used bagging and random forests. Lastly, in Chapter 5, genetic algorithms and particle swarm optimization were used in conducting the rule-based learning and development of membership functions in a fuzzy expert system. One of the most common Bayesian regression methods, known as Bayes B, was used to calculate SNP effects in development of the fuzzy expert system.

WEKA (Version 3.6.4), a data mining software coded in JAVA (Hall et al., 2009), was used to implement the machine learning methods used in Chapter 3. Chapter 4 and 6 utilize a customized implementation of the random forest and model tree algorithm of WEKA, to be able to predict outcomes for unlabeled data, as well as to carry out the cost-sensitive evaluations and lift chart analyses. In Chapter 5, the novel fuzzy expert system was developed by the author and implemented in the Matlab (Matlab, R2013a, Student version) programming environment. The BGLR package (de los Campos and Perez Rodriguez, 2012) version 1.0.2 of the R language was used for genetic evaluation of bulls in Chapter 5, and data preparation and genotype imputation for the genomic data in Chapter 5 were implemented using Plink version 1.07 (Purcell et al., 2007) and Beagle version 3.3.2, (Browning and Yu, 2009) respectively. Data editing and preparation were carried out using several custom software programs written in Java after the raw data were extracted from DairyComp 305 backup files and USDA-ARS Animal Improvement Programs Laboratory databases.

Chapter 2

Literature Review

1. Trends in fertility traits

The decline of reproductive performance in dairy cattle during the past few decades, along with great progress in production traits, is evident. Genetic programs with strong emphasis on milk production have resulted in cows that are genetically vulnerable to greater degrees of negative energy balance (NEB) in early lactation. Basically, NEB occurs because increased milk yield requires nearly double the amount of energy provided by dry matter intake (DMI). Therefore, additional resources are needed to compensate for the peak in milk production, and these are typically provided by enhanced mobilization of adipose tissues and skeletal muscles (Veerkamp, 1998).

Washburn et al., (2002) carried out a comprehensive study of herds in the southeastern US and showed a significant increase in days open, days to first service, and services per conception in Holstein and Jersey cattle from 1976 to 1999. In the same period, milk production and herd size increased significantly for both breeds, and there was an increase from 1.9 to 2.9 in service per conception (Washburn et al., 2002).

Significant declines in reproductive performance, measured as an increase in days to first service from 84 to 104 during the period from 1983 to 2001, a decrease in annual pregnancy rate from 22 to 12% during the period from 1978 to 2001, and an increase in calving interval (CI) from 399 to 429 d during the period from 1970 to 2000 were reported by De Vries and Risco, (2005). In Holstein cows, days to first breeding (DFB) declined from 94 d in 1997 to 86 d in 2006. Jersey cows did not show a similar trend, possibly because synchronized breeding is not as common on farms with Jersey cattle. Additionally, first breeding 70-day non-return rate (NRR70) and first breeding conception rate (CR) declined, and the number of breedings per lactation increased. These declines were more severe in later parity cows than in young cows (De Vries and Risco, 2005). Genetic merit for fertility, as measured by daughter pregnancy rate (DPR), had declined since 1960 , but DPR began to increase in both Holstein and Jersey cattle after the implementation of genetic evaluations for length of productive life (PL) in 1994 and DPR in 2003 (Norman et al., 2009). The use of genetic information for PL in selection programs probably helped to reverse the genetic decline in pregnancy rate because of the strong negative relationship between days open (DO) and PL. Bulls with high genetic merit for DPR, which has a negative relationship with DO, have fewer daughters culled due to reproductive failure and therefore longer PL (Norman et al., 2009).

2. Associations of fertility with other traits and management factors

When the decrease in reproductive performance of dairy cattle was discovered, many researchers attempted to understand the associations between this trend and modern dairy farming practices, with the goal of identifying potential cause-effect relationships between management factors and reproductive performance.

In several studies, the probability of conception was adversely associated with the incidence and severity of NEB in early lactation (Butler and Smith, 1989; Nebel and McGilliard, 1993). Interestingly, an increase in milk yield during the first 120 d of lactation was associated with an increased likelihood of conception, but decreased body condition score (BCS) during the first month of lactation was associated with a decreased likelihood of conception. Together, this shows that the most important factor is not milk production, but rather energy balance, and this is the motivation for using energy corrected milk production in dairy management studies (Domecq et al., 1997).

In another study, no association was found between milk yield and interval to commencement of luteal activity, calving to conception interval, or conception to first service. However, mean daily energy balance, milk protein content, and DMI during the first month of lactation were positively associated with conception to first insemination. The authors also found that cows with poor BCS (≤2.25) at first service had lower conception to first service, whereas cows with higher DMI and more positive energy balance had an increased likelihood of a shorter calving to conception interval (CCI) (Patton et al., 2007). Caraviello et al., (2006) surveyed 153 herds and reported that AI service rate, conception rate, twinning, retained placenta, and mastitis were the most commonly reported factors associated with reproductive management problems in US dairy herds.

Management practices explain part of the observed downward trend in reproductive performance. Some structural changes in dairy operations have been favorable, such as free stalls and automatic milking machines (Huang et al., 2009)(Löf et al., 2007). Cows in low-yielding herds had longer CI due to longer intervals from calving to first AI and from calving to last AI, as well as greater rates of culling due to reproductive problems, as compared with cows in high-producing herds. This suggests that high production is not necessarily associated with poor reproduction. However, high-producing herds had a greater number of AI services per cow, because heat detection and low service rates are among the main causes of poorer reproductive performance low-producing herds (Löf et al., 2007).

Huang et al., (2009) investigated 827,802 AI records from 424,513 first lactation cows born from 1985 to 2000 in 2353 herds in the southeastern US and found an overall increase in milk production and decrease in CR over time. Least-square means analysis of month by milk production interactions showed that in cool months (November to April) the deterioration of CR over time was small for cows with low or intermediate milk production, and there was virtually no deterioration among high-producing cows. On the other hand, there was significant decline for CR in May and June for cows of all production levels. This research suggests that there has been a decline over time in the ability of cows to handle the onset of heat stress, or that cows have not adapted well to intensive management and high production under conditions of heat stress.

Heat stress is one of the most important environmental problems that can affect dairy cow performance, especially reproduction. Heat stress can cause a significant reduction in CR from 42 d before to 40 d after insemination, and it has been suggested that fresh transfer of embryos collected from donor heifers housed in sheds with a cooling system can enhance conception rate in hot months of the year (Jordan, 2003). Heat stress is the most important causal factor in the seasonality of DO. Therefore selection for smaller monthly fluctuations in DO within sire families may increase heat tolerance of dairy cows (Oseni et al., 2003, 2004a, 2004b).

Infectious and metabolic diseases are often found to be associated with poor reproductive performance. A dramatic decline in CR in cows that had clinical mastitis between 14 and 35 d after AI was reported by Hertl et al., (2010) . The greatest effect, which was an 80% reduction in CR, was associated with an infection from gram-negative bacteria during the week after insemination. Clinical mastitis caused by gram-positive bacteria had a smaller effect on CR, but the reduction was still significant. Hudson et al., (2012) also showed that there was a clear unfavorable association between clinical and subclinical mastitis and reproductive performance. Giuliodori et al., (2013) showed that cows with clinical or puerperal mastitis produced less milk and had lower risk for pregnancy by 100 and 150 d postpartum than healthy cows. Furthermore, prepartum NEB and calving problems were associated with an increased risk of mastitis (Giuliodori et al., 2013). Clinical ketosis, dystocia, and retained placenta are associated with longer days to first service (6 to 12 additional days on average) and lower conception rate in subsequente services. (Fourichon et al., 2000) (Fourichon et al., 2000).

Dairy cows pass through a period of negative energy balance during the transition period as the demand for production cannot be met by feed intake (Herdt, 2000; McArt et al., 2012). Cows adapt to this high energy demand by mobilizing lipid reserves, which circulate in the blood as non-esterified fatty acids (NEFA) a cause the liver to produce ketone bodies or triglycerides. Excess NEFA, overwhelms the liver, therefore, ketone bodies will be produced and the cow will become hyperketonemic (Herdt, 2000). Cows with hyperketonemia will show weight loss, decreased appetites, and therefore decreases in milk production and BCS. Subclinical ketosis (SCK) is defined as an increase in ketone bodies in blood without clinical signs of hyperketonemia (Andersson, 1988). Cows with SCK are pre-disposed to higher risk of other postpartum disease such as displaced abomasum (DA) and metritis (Duffield et al., 2009; Ospina et al., 2010) which increase their risk of culling in early lactation. In another study McArt et al., (2012) showed that intensive detection followed by treatment of SCK decreased the risk of development of DA and culling risk in the first 30 DIM and an increase in conception rate to first service in the herd, as well as a decrease in days to first conception.

Retained placenta was shown to decrease the probability of conception by 25% (Hertl et al., 2010), and cows with metritis in early lactation suffered from significant delays in time to first estrus and significant declines in the probability of success at first insemination (Coleman et al., 1985; Fourichon et al., 2000; Elkjær et al., 2013). Földi et al., (2006) found that clinical metritis affects reproduction in two ways, first by postponing the regeneration of the endometrium, therefore disturbing uterine function, and second, by unexpected endocrine signals, therefore interfering with normal ovarian function.

An investigation of factors influencing the probability of pregnancy by 30 d after the conclusion of the voluntary waiting period (VWP) in Sweden revealed that high heat detection efficiency, longer VWP (> 51 d), and the presence of free stalls were associated with higher chance of pregnancy. On the other hand, the probability of pregnancy was lower for cows with severe foot problems. Cows with a history of reproductive-related diseases and other diseases, as well as cows with dystocia, twin calving, high milk production, high fat/protein ratio, high somatic cell count (SCC), or summer breeding also had decreased chance of pregnancy (Löf et al., 2014). In another Swedish study using logistic and Poisson regression, mammary and foot problems were associated with a lower probability of pregnancy and also a greater number of AI services per eligible cow (Lomander et al., 2013). High prevalence of lameness has been associated with poor reproductive performance, as these herds tend to have longer calving to conception interval and longer overall calving interval (Chapinal et al., 2013).

3. Genetic selection for reproductive traits

In terms of improving fertility through genetic selection, there are several challenges. First, the complexity of the trait leads to confusion in determining how to record a cow's phenotype to optimally reflect her fertility. Second, different measures of fertility have been collected in different countries, and in some cases these traits have low or negative correlations with one another. Inconsistencies between recorded fertility traits among countries is one of the biggest challenges in international genetic evaluation of fertility (Philipsson et al., 2007). For example, DO and CI can be calculated from milk recording data, and these are probably the most readily available fertility traits in dairy cattle. However, interval between first and last insemination (IFL), number of inseminations per conception (INS), days to first service (DFS), and time from first insemination to conception are also recorded in many countries with well-established data recording programs. A third challenge is the low heritability of fertility traits, which makes their improvement through selection slow and inefficient. A final challenge is the need for specialized statistical modeling of fertility data (González-Recio and Alenda, 2005). Linear models are not appropriate for binary traits, and data for continuous measures of fertility are often badly skewed and frequently censored (Weigel, 2004).

Fertility of service sire also affects conception rate in the herd (Kuhn et al., 2006). However, zero (Kuhn and Hutchison, 2008) or very low (0.01 to 0.04) heritability of male reproductive traits and an unfavorable correlation with production traits have limited their usage in selection programs (Liu et al., 2008).

Most selection pressure has focused on improvement of fertility in lactating cows, and heifer CR has a positive correlation of 0.39 with cow CR (Kuhn et al., 2006). This moderate correlation between the same measurement in heifers and cows suggests that one should consider traits related to fertility in heifers and cows separately (Philipsson et al., 2007).

Calving interval is often reported to have an unfavorable genetic correlation with milk yield, ranging from 0.19 to 0.35,(Kuhn et al., 2006; Wall et al., 2005). However, a high positive correlation between CI and longevity (0.59) may explain the decline in fertility has slowed recently, noting that PL has been included in the lifetime net merit index used for selection of US dairy cattle since 1994 (VanRaden et al., 2004b).

Wall et al., (2005) reported an unfavorable genetic correlation between CI and rump angle, udder support, and mammary system, as well as a favorable correlation between feet and legs score and non-return rate. These findings suggest that more selection emphasis on feet and legs and less emphasis on mammary systems in a selection index might help to improve reproductive traits in Holstein dairy cattle indirectly.

González-Recio and Alenda (2005) and González-Recio et al. (2006) estimated low heritability parameters (0.02 to 0.06) for a range of fertility traits. Strong genetic correlations were found among most fertility traits (0.89 to 0.99), and a fertility index composed of DFS and pregnancy within 56 d postpartum was shown to achieve the fastest genetic gain, in terms of reducing fertility costs, reducing days to first service, and reducing the number of inseminations per lactation by $8.60, -1.31 d, -0.03 AI, respectively.

It is commonly accepted that 21-day pregnancy is the best measurement for assessing the current reproductive performance of a dairy herd. Unlike other measurements that are based on outdated information, such as CI, or measures that consider only inseminated cows, such as services per conception, 21-day pregnancy rate provides a timely and comprehensive evaluation of how quickly open cows become pregnant (LeBlanc, 2005).

With regard to specific genes, polymorphisms in two genes, FGF2 and STAT5A, have been associated with estimated relative conception rate, a measure of male fertility (Khatib et al., 2010). These genotypes have been reported to be associated with milk composition traits in previous studies, which provides molecular evidence for an genetic relationships between milk production and fertility in dairy cattle (Wang et al., 2008; Khatib et al., 2008).

Using a stringent 5% genome-wide significance level, 22 and 5 SNPs were found to be associated with fertilization rate and blastocyst rate of *in vitro* embryos, respectively (Huang et al., 2010). Male fertility, namely sire conception rate (SCR), was reported to be associated with 8 SNPs. Some of these significant SNPs were located very close to, or in the middle of, genes with known roles in male fertility, such as sperm acrosome reaction, chromatin remodeling during spermatogenesis, or meiotic processes during male germ cell maturation (Peñagaricano et al., 2012).

By investigating common haplotypes of ≤75 markers that were not observed in homozygous form in 58,453 Holstein, 5,288 Jersey, and 1,991 Brown Swiss cows in North America, VanRaden et al. (2011) found that 5 of 11 candidate haplotypes were detrimental to conception rate. The corresponding reductions in conception rate ranged from 3.0 to 3.7 percentage points and were confirmed using 14,911,387 Holstein, 930,391 Jersey, and 68,443 Brown Swiss conception rate records. The estimated effects of carrier status on stillbirth rate were small. The authors concluded that these lethal effects may include conception, gestation and stillbirth losses. Three defective haplotypes were found in Holstein and named as HH1, HH2, HH3; these had carrier frequencies of 2.7 to 6.5% in the current population. In Brown Swiss and Jersey cattle, one defective haplotype was found for each breed and named as BH1 and JH1; these had carrier frequencies of 16 and 23% in the current population, respectively. The effects of these haplotypes on non-return rate at 60 DIM were investigated, and reductions ranged from 1.1 to 3.1% for Holsteins and 2.5 –to 3.7% for Brown Swiss and Jerseys, respectively (VanRaden et al., 2011).

4. Economics of reproductive traits

Although reproductive performance has a large impact on dairy farm profitability, there are relatively few studies investigating its impact on profit when compared with the number of studies on the economics of nutrition or mastitis, as examples. Improvement in reproduction cause higher milk production and thereby, higher income over feed cost, lower culling rate and breeding cost, and more replacement candidates (Cabrera, 2014). The objective of reproductive efficiency is to maximize the profit per stall per year, by maximizing the proportion of a cow's lifetime that is spent in the most profitable part of the lactation curve (Ferguson and Galligan, 2011; Cabrera, 2014) which minimizes the costs associated with replacement and mortality and also minimizes the relative costs associated with reproduction (Giordano et al., 2012).

The economics of reproductive performance are strongly affected by the shape of the lactation curve of a dairy cow, and income over feed costs is greatest around the time of peak milk production. Insemination rate is often the starting point as an intervention to improve reproductive efficiency, because cows cannot get pregnant unless they are inseminated at the proper time, but service rate is also extremely important. Most of commercial farms in the United States take advantage of a type of synchronization protocol along with oestrous detection in their reproduction management (Caraviello et al., 2006; Giordano et al., 2012; Cabrera, 2014). The most common measurement for assessing reproductive performance in commercial dairy farms is the 21-d pregnancy rate. Although it does not give any insight into economic efficiency of reproduction programs (Cabrera, 2014). Because of the complex nature of reproduction management and its relationship with replacement rate, production, and profitability of the farm, understanding its economic impact, is difficult in reality. A few simulation studies have tried to address this need by providing some decision support tools to help farmers compare economic returns of different scenarios of reproduction programs (Giordano et al., 2011, 2012; Galvao et al., 2013). Cabrera (2012) introduced the concept of economic value of a cow, which is very similar to the concept of retention pay off introduced by De Veries (2006) that can be used to optimize the reproduction performance of the farm at the cow level, which is the focus of this dissertation.

A major determinant of the economic value of a pregnancy is expected future milk production of the cow relative to its herd mates. This, in turn, is a function of cow's current production, age, probability of culling, probability of pregnancy in next cycle (if non-pregnancy), and breeding policy of the herd.

In general, the economic value of reproductive performance can be considered as a function of two opportunity costs:

1) The forgone profit associated with a cow that remains open beyond the optimal DO (approximately 120 DIM in a typical herd) and therefore spends a greater proportion of time in the late, low-production part of lactation.
2) Failure to realize all of the lifetime profit a cow is genetically predisposed to provide, because her inability to become pregnant leads to premature culling from the herd.

The opportunity costs of prolonged days open and involuntary culling are typically greater than the direct costs of increased labor and additional AI semen. The opportunity cost of an additional day open, as well as the value of pregnancy at a certain interval after calving, depend substantially on milk yield, lactation persistency, milk price, the availability and cost of replacement heifers, and herd's pregnancy rate (LeBlanc, 2007). Parity, age, stage of lactation, carcass value, herd breeding policy, and discount rate also influence the economics of reproduction (De Vries, 2006), and the cost of an additional day open has been estimated from $0.44 at 130 DIM to $4.70 at 150 DIM (LeBlanc, 2007).

González-Recio et al., (2004) showed that the number of inseminations per cow has a major impact on dairy farm profitability. When the average number of inseminations increases beyond 3.0, herd profit decreases by more than $205/cow/year. Cows that require more inseminations per pregnancy tend to have higher milk production but also higher culling risk, which leads to shorter productive life, less lifetime production, and lower profit. Economic values for calving interval and number of times bred were estimated as $-4.90 and $-67.32/cow/year per unit of change, respectively.

The cost per AI event on farms using the Ovsynch protocol was estimated to be $20.50/cow, and this number can increase to $24.20 on small farms, in addition to semen and heat detection costs, whereas the cost per AI event on farms using the Cosynch protocol was estimated to be $21.00 on small farms (Olynk and Wolf, 2009).In another study, Lima et al. (2010) estimated a per timed AI event of $22.07 when the semen cost was assumed to be $6.00 per unit.

In a simulation study, the average value of a new pregnancy was estimated to be $278, and the average cost of pregnancy loss due to abortion was estimated to be $555 for Holstein cows under typical US dairy farming conditions (De Vries, 2006).

5. Overview of Machine Learning methods used in this research

There is no systematic approach 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 the field of machine learning today. In order to classify insemination events into pregnant or non-pregnant outcomes based on a set of explanatory variables, five 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:

5.1 Naïve Bayes (NB): Naïve Bayes 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; simplicity, computational feasibility, and robustness make this method suitable for practical use. NB 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 …fn), and c is a class variable with two values (O = Open, P = Pregnant). The probability of the class variable given the feature vector can be calculated as:

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

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

$$p(f1, f2, \dots ., fn|C = c) = p(f1|C = c)p(f2|C = c) \dots \dots p(fn|C = c)$$

$$p(f1, f2, \dots fn) = \ p(f1)p(f2) \dots p(fn)$$

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

5.2 Bayesian Network (BN): A Bayesian network represents the joint probability distribution of a set of variables{X1, X2,…. Xn} as a discrete acyclic graph and a set of conditional probability distributions that correspond to specific features. 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* (Kjaerulff 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. BN are especially useful, because by using 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 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). BN 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, there are a fewer challenges with using BN in practice, including discretizing of continuous variables, collecting and structuring expert knowledge, and implementing recursive feedback loops (Uusitalo, 2007).

5.3 Decision tree (DT): Decision trees 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. 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 (described in a subsequent section of this chapter). Features that correspond to the largest information gain 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 backwards in order to reduce classification error due to over-fitting. DT 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, there are some challenges 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) .

5.4 Bootstrap Aggregation (BG): 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, bagging 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, bagging of stable models, like NB, will not improve predictive performance, but bagging of unstable methods, such as tree learners, will improve the predictability of the ensemble model (Brieman, 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).

5.5 Random forest (RF): Random forest 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 (Breiman, 2001; Ho, 1995). Thus, a 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.

5.6 Model tree (MT): They are similar to decision trees because they use a divide and conquer approach to partition the multi-dimensional state space of the problem (Quinlan, 1992). The model tree (MT) predicts the values for test instances by the linear model (LM) stored in each leaf based on the attributes of instances that reach that specific leaf. The MT and alternating decision trees have been used in assessment of reproductive performance of dairy herds (Caraviello et al., 2006; Schefers et al., 2009).

5.7 Artificial life: This set of methods uses informational concepts and computer modeling to study life in general (Keil and Wilson, 1999). A-life has been used to describe man-made systems that mimic some essential aspects of living organisms. It has two main streams: how computational techniques can help us understand life and how biological phenomena can help us overcome computational problems (http://www.alife.org). In this dissertation we use three techniques: genetic algorithm, particle swarm optimization, and fuzzy logic from the second main stream.

5.8 Genetic algorithm (GA): Fitness landscape, introduced by (Wright, 1932), was an important starting point in the development of the genetic algorithm (GA). In the 1950's statisticians such as Box (1957) and Friedman (1959) used evolutionary algorithms for optimization purposes. This didn't gain much attention until evolution strategies were used to solve optimization problems for airfoils by Rechenberg (1965). In the 1970's, Holland (1975, 1995) championed the GA both as an optimization tool and as a method to expand our understanding of biological adaptation. The Alife movement (Langton, 1995) developed this idea to the next level by considering the product of a genetic algorithm as an organism, not only a solution to an optimization problem.

Genetic algorithm is a stochastic hill-climbing search in which a large population of states are maintained. New states (population) are generated by selection, crossing over, and mutation.

It is a variant of stochastic beam search in which, instead of modifying a single state, a successor state is generated by combining two parent states through genetic operations. GA is analogoues to natural selection as stochastic beam search, except that it deals with sexual rather than asexual reproduction. Similar to beam searches, GA initiates the solution space with a set of k randomly generated solutions, called the population. Each state is called an individual and is represented as a string of 0's and 1's to represent a solution.

The next generation of states will be produced by selection after a fitness value has been determined by the fitness function. The probability in which each current individual will be selected to produce the next generation of solutions is proportional to the fitness value of each individual. After the parents of the next generation are selected, in order to generate the next generation of states, each pair of individuals will be mated, a crossing over point will be chosen on the string heuristically, and individuals of the next generation will be generated. Conceptually, when the two parent states are quite different, the crossover operation can produce a state (child) that is far away from each parent. It is often the case when the population is very heterogeneous in early iterations, so crossovers frequently take large steps in state space and explore the entire space for the optimum solutions. However, as the algorithm continues and diversity decreases in the population, crossovers will take smaller steps and exploit the solution space for the optimality. Finally, in each iteration after crossing-over, each location is exposed to random mutation with a small probability to prevent rapid convergence and local optima.

Similar to stochastic beam search, a GA utilizes hill climbing with random exploration and exchange of information between parallel intermediate solutions. The primary advantage of GA comes from the crossover operation, which combines the suboptimal solution together to create higher individual fitness in the next generation of state space (Russell, and Norvig, 2002).

5.9 Particle swarm optimization (PSO): This is a social-biological system inspired by the social behavior of a swarm of birds or fish when looking for food or escaping from predators. It is a technique used to explore the search space of the problem in order to find the optimum solution. This technique was originated based on social habits of bird swarm and evolutionary computing (Kennedy and Eberhart, 1995).

Suppose a group of bids are searching randomly for food in an area where there is only one piece of food. Birds do not know where the food is, but in each iteration they will know how far they are from the food and also which one of them is the closest to the food. So, the best strategy here would be to follow the bird that is closest to the food. Here, each bird is a single solution in the state space and is called a particle. Each particle has a fitness value associated with it in each iteration. Also, each particle has the memory of its own fitness so far, as well as the knowledge of the best fitted bird in the swarm. The first one is the best position that the particle has achieved so far and is called particle best. The second one is the best position obtained by any particle so far in all iterations and is called global best (http:\\swarm intelligence.org).

In particle swarm optimization (PSO), initial solutions generate randomly throughout the algorithm. In each iteration particles pass through three stages as follow:

1- Evaluate fitness

2- Update particle best and global best fitness and positions

3- Velocity and position updating of each particle

However, the last step is not trivial. Particle’s velocity is updated as follows:

$$v_{t+1}^{i} = wv_{t}^{i} + c_1 r_1 [\hat{x}_{t}^{i} - x_{t}^{i}] + c_2 r_2 [g_t - x_{t}^{i}]$$

where v_{t+1}^{i} is the velocity of particle i in the iteration $t+1$, v_{t}^{i} is the velocity if particle i in iteration t. $w, c_1,$ and c_2 are user specified coefficients between 0 and 2 which are also subject to fine tuning of the algorithm. r_1 and r_2 are random numbers between 0 and 1 that are regenerated in each iteration to add stochasticity to the process. The $\hat{x}_{t}^{i}$ is the particle best solution up to iteration *t* and g_t is the global (swarm) best solution so far.

Each of the three terms in the equation above has a different role in the convergence of the algorithm. The first term, wv_{t}^{i} , is called the inertia component and keeps the particle movement in the initial direction. If the value of the inertial coefficient *w* is less than 1 it can suppress the particle's inertia, or it can increase the speed of particle in the original direction if it is between 1 and 2 (Shi and Eberhart, 1998).

Basically, smaller values of the inertial coefficient increase the convergence speed of the swarm, and bigger inertial coefficients encourage exploration of the entire search space. The second term $c_1 r_1 [\hat{x}_{t}^{i} - x_{t}^{i}]$, called the cognitive component, is essentially the particle's memory, forcing the particle to return to the regions of the search space in which it has experienced high individual fitness. In other words, it forces the particle to search locally and explore the neighborhood. The cognitive coefficient, *c*, is usually close to 2 and the size of the step of the particle toward its individual best candidate solution $\hat{x}^{i}$ is affected by it.

The third term, $c_2 r_2 [g_t - x_{t}^{i}]$, called the social component, leads the particle to the best region that the swarm has found so far. The social coefficient c_2 is usually close to 2 and indicates the size of the step the particle takes toward the global best, g_{t},. In other words, this term is responsible for the exploitation of the whole search space.

In order to bound the particles to move too far beyond the search space a technique called velocity clamping can be used to limit the maximum velocity of each particle. For a search space in the range $[x_{min}\ ,\ x_{max}]$, velocity clamping keeps the velocity in the range $[v_{min}\ ,\ v_{max}]$, where $vmax\ =\ k * x_{max}$, and k is a user specified coefficient between 0 and 1.

After the velocity for each particle is calculated and checked to be within the acceptable boundaries, each particle's position will be calculated and updated by the equation below:

$$x_{t+1}^{i} = x_{t}^{i} + \ v_{t+1}^{i}$$

This process is repeated as long as a stopping condition is not satisfied or the maximum iteration number is reached (Blondin, 2009).

5.10 Fuzzy Logic (FL): Fuzzy logic is a type of multi-valued logic, dealing with reasoning that is uncertain rather than fixed and precise. Against traditional binary sets, which variables take either true or false values, in fuzzy logic variables are associated with a truth value that ranges between 0 and 1. Fuzzy logic has been developed to handle the concept of partial truth, in which the truth value may fall between completely true and completely false (Novák et al., 1999).

It is an approach to reasoning based on "degrees of truth" rather than absolute "true or false" (1 or 0) in Boolean logic on which the modern computers work. Fuzzy logic was first introduced by Dr. Lotfi Zadeh of the University of California at Berkeley (Zadeh, 1965). The idea emerged when he was working on the problem of computer understanding of natural language. Like most other activities in life, natural language cannot be easily translated into the crisp terms of 0 or 1. In fuzzy logic 0 and 1 would be considered as extreme cases of truth, and it also includes the various states of truth in between. For instance, the result of a comparison between the height of two persons would be 0.38 of tallness rather than crisp "tall" or "short". Fuzzy logic seems closer to the state of nature and the way our brains inference. Data will be aggregated, and a number of partial truths will be formed, which could be aggregated further into higher truths. It may be easier to visualize fuzzy logic as the way reasoning really works and the fact that binary or Boolean logic is simply a special case of it (Rouse, 2006)

5.10.1 Fuzziness vs probability: Fuzziness describes the ambiguity of an event. What it measure is the degree to which an event happens, not whether it happens or not. On the other hand, randomness describes the uncertainty of an event's occurrence. Randomness describes how frequently an event occurs or not and how you can bet on it. In another words, "whether an event occurs is "random" and the degree it occurs is "fuzzy""(Kosko, 1990).

5.10.2 Fuzzy expert systems: A fuzzy expert system is simply an expert system which uses a set of if-then rules and respective membership functions to reason about data. The general form of a rule in a fuzzy expert system is as follow:

$$R_i: if\ x_{q1}\ is\ A_{i1}\ and\ \ldots..and\ x_{qn}\ is\ A_{in}\ then\ class\ is\ C_i$$

where $A_{i1} \ldots \ldots A_{in}$ are antecedent fuzzy sets of the inputs $x_{q1} \ldots .. x_{qn}$ and C_i is the output class label. A set of such rules forms the rule base for the fuzzy expert system in which qualitative reasoning to infer the results is performed. Fuzzy if-then rules along with membership functions form the core part of a fuzzy inference system.

Fuzzy inference is the process of the mapping from a set of given input to an output using fuzzy logic. This mapping then provides a basis which can be used in decisions making or pattern recognition. Different types of fuzzy reasoning have been developed. Depending on the types of fuzzy reasoning and knowledge base (fuzzy if-then rules) used, most fuzzy inference systems (FIS) would be classified into two types:

Mamdani: This was among the first control systems built using fuzzy set theory. It was formed as an attempt to control a steam engine and boiler combination by synthesizing a set of linguistic control rules obtained from experienced human operators (Mamdani, 1977). The overall fuzzy output is obtained by applying "max" operation to the fuzzy outputs (each of which is derived by applying "min" operation to the firing strength and the output membership function of each rule). Different methods have been proposed to calculate the final crisp output based on the overall fuzzy output. These could be centroid of area, bisector of area, mean of maxima, and maximum criterion, to name a few.

Takagi-Sugeno: In Takagi and Sugeno's fuzzy if-then rules (Takagi and Sugeno, 1983) a linear combination of input variables plus a constant term. The final output of the system would be the weighted average of all rule's output. Most of the differences come from the specification of the consequent part (monotonically non-decreasing or bell-shaped membership functions, or crisp function) and also the de-fuzzification (weighted average, centroid of area, etc.) schemes (Jang, 1993).

FISs are also known as fuzzy-rule-based systems as well. Generally an FIS is contain of five functional parts:

1) A knowledge base.

- A rule base that contains a set of fuzzy if-then rules;
- A database which defines the membership functions of the fuzzy sets used in the fuzzy rules.

2) A fuzzification interface scheme which transforms the crisp inputs into degrees of belongingness to the linguistic variables.

3) A decision-making engine which performs the inference operations on the rules.

- Combines the membership values on the antecedent part of the rule to get the firing strength (weight) of each rule.
- Generates the qualified consequent which can be either fuzzy or crisp for each rule based on the firing strength.

4) A de-fuzzification interface which transforms the fuzzy results of the inference into a crisp output by aggregating the qualified consequents.

Mamdani and Sugeno fuzzy inference systems (Jang, 1993; Pal, 1999) were used in this research and their performances were compared against each other's as well as genomic best linear unbiased prediction (GBLUP), RF, ordinary least squares (OLS), and support vector machines (SVM).

In the model developed herein, a set of input SNPs were compared against the antecedent part of each if-then rule, and in the case of a match the response (output) was obtained by fuzzy implication operation. The extent to which each rule is fired indicates the weight of that rule. Next, the response of fired rules for a particular input are combined to obtain the final class prediction of that input set.

In the development of a fuzzy inference system knowledge acquisition is the most critical part. It can be done practically by using an expert's knowledge through a questionnaire or interview and/or by the means of data mining tools applied to existing databases. In the current study, because of the large number of SNPs, there would be an infinite number of possible rules and therefore no expert exists to define a complete rule set as well as an optimum membership function for each SNP that is included in antecedent of the rules.

One or more machine learning method can be used to tackle this problem. The hybrid proposed fuzzy system uses genetic algorithms to come up with an optimum set of if-then rules and particle swarm optimization algorithm to tune membership function points.

Chapter3

Prediction of Insemination Outcomes in Holstein Dairy Cattle using Alternative Machine Learning Algorithms

1. 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 ten-year 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 USDA ARS 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(25) explanatory variables and one binary outcome for primiparous (multiparous) cows, respectively. The best performance was exhibited by a random forest algorithm with 0.723±0.21 (0.736±0.28) classification accuracy and area under the curve of 0.756±0.005 (0.736±0.005) for primiparous (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.

2. INRODUCTION

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, body condition score, 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.). 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, J. M., and J. 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 body condition scores, 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 days in milk (**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 days post-AI (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 somatic cell score 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 body condition score, energy content of the diet, cumulative effective energy balance, and blood glucose have favorable genetic relationships with reproduction, whereas B-hydroxybutyrate and non-esterified fatty acids 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 dry matter intake during the first 28 days postpartum were associated with higher conception rate at first service, whereas cows with high dry matter intake and positive energy balance had a shorter calving to conception interval. On the other hand, lower body condition scores 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, body condition score, fat corrected milk, and average of fat and protein percentages. They reported classification accuracies of 85 to 86%, with sensitivity and specifity 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, R. et al., 1995; Grzesiak et al., 2006; Gianola et al., 2011), classifying mastitis cases (Yang, X. Z. et al., 1999), classifying lameness in horses (Suchorski-Tremblay, A. M. et al., 2001), predicting slaughter weight of bull calves (Adamczyk et al., 2005), identifying SNPs 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.

3. MATERIALS AND METHODS

3.1 Data

The data utilized 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, estimated breeding values, health events, and reproduction information (Table 3.1). In term of reproduction performance herds in this study are representative of large commercial dairy farms in Wisconsin (Figure 3.1).

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

In order to account for energy balance and reduce the dimensionality of features in the model for analysis, energy corrected milk (**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 (Tyrell and Reid, 1965):

$$\text{ECM} = (0.327 * \text{Milk kg}) + (12.95 * \text{Fat kg}) + (7.65 * \text{Protein kg}).$$

3.2 Machine Learning Algorithms

There is no systematic approach 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. In order to classify insemination events into pregnant or non-pregnant outcomes based on the aforementioned explanatory variables, five 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:

3.2.1 Naïve Bayes (NB). Naïve Bayes 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 3.2a); simplicity, computational feasibility, and robustness make this method suitable for practical use. NB 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 …fn), and c is a class variable with two values (O = Open, P = Pregnant). The probability of the class variable given the feature vector can be calculated as:

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

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

$$p(f1, f2, \dots., fn|C = c) = p(f1|C = c)p(f2|C = c) \dots \dots p(fn|C = c)$$

$$p(f1, f2, \dots fn) = \ p(f1)p(f2) \dots p(fn)$$

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

3.2.2 Bayesian Network (BN). A Bayesian network represents the joint probability distribution of a set of variables {X1, X2,…. Xn} as a discrete acyclic graph and a set of conditional probability distributions that correspond to specific features (Figure 3.2b). 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* (Kjaerulff 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. BN are especially useful, because by using 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 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). BN are well suited for small and incomplete data and can achieve good prediction accuracy in such situations (Kontkanen, 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, there are a few challenges with using BN in practice, including discretizing of continuous variables, collecting and structuring expert knowledge, and implementing recursive feedback loops (Uusitalo, 2007).

3.2.3 Decision tree (DT). Decision trees 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 3.2c). 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 (described in a subsequent section of this paper). Features that correspond to the largest information gain 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 backwards in order to reduce classification error due to over-fitting. DT 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, there are some challenges 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) .

3.2.4 Bootstrap Aggregation (BG). 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, bagging 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, bagging of stable models, like NB, will not improve predictive performance, but bagging 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).

3.2.5 Random forest (RF). Random forest 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 (Breiman, 2001; Ho, 1995). Thus, a 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.

3.3 Model Assessment

3.3.1 Receiver Operating Characteristic (ROC) 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 ROC curve offers the same information as the confusion matrix, but in a much more intuitive and robust fashion (Hamel, L., 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 (non-pregnant 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 two classes. The upper left corner of the curve (FP=0, 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.

3.3.2 Feature selection. In order to evaluate the relevance of features in our analysis, we used information gain (**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 information gain (Russell and Norvig, 2002) and is calculated as:

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

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 Gain Ratio (**GR**).

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

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

3.3.3 Lesion approach. The complexity of some machine learning algorithms sometimes leads to difficulty in interpreting the results intuitively. Therefore, in order to gain insight about the interrelationships among variables, a lesion approach was employed. 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.

4. RESULTS AND DISCUSSION

As shown in Table 3.2, results of the IG analysis indicate that mean within-herd conception rate in the past three months, herd-year-month (**HYM**) of breeding, DIM at breeding, number of inseminations in the current lactation, and stage of lactation appeared among the top ten 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 ten features in primiparous cows, and with mastitis, displaced abomasums, and retained placenta among the ten most important features in multiparous cows.

4.1 Classification Accuracy

Table 3.3 shows the AUC and the proportion of CCI for all five-folds of the cross validation, as well as the average of AUC and CCI across folds, for the five machine learning methods used in this paper. Likewise, Figure 3.3 compares the five machine learning techniques for classification accuracy between pregnant and non-pregnant primiparous cows. The RF algorithm outperformed all other methods, with 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 cross-validation AUC of 0.736 for RF, which exceeded ($P < 0.001$) other methods (Figure 3.4); including BG, DT, BN, and NB. A similar trend was observed for CCI (Table 3.3). 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 dataset 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 clean-up 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 R^2 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 days 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.

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 two 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 over-fit small data sets (Mitchell, 1997), it is possible that their model was over-trained on that specific dataset and lacks generality for widespread use. A common practice in machine learning analyses is to evaluate several different learning algorithms on each specific task of interest, because the performance of these algorithms may change due to differences in the size, structure, and other characteristics of the data set.

4.2 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 3.5 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 3.5), 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 3.6), eliminating HYM caused AUC to change from 0.74 to 0.76 and CCI to change from 0.73 to 0.75. As shown in Figure 3.6, 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 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.

5. CONCLUSION

The machine learning algorithms considered in this study were effective in predicting pregnant versus non-pregnant 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 three months, 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 gain ratio analysis suggested greater importance of health traits in explaining insemination outcomes relative to the information gain analysis. Based on the gain ratio, the incidence of ketosis, mastitis, retained placenta, and lameness (for primiparous cows) and the incidence of mastitis, displaced abomasums, 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 utilize gender-enhanced semen or expensive semen from high merit sires.

6. AKNOWLEDGMENT

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. Data for this study was provided by Alta Genetics (Watertown, WI) Advantage Progeny Testing Program and USDA-ARS Animal Improvement Programs Laboratory (Beltsville, MD). Authors acknowledge the technical comments and support from John Schefers, Daniel Gianola, and Gilherme Rosa. Saleh Shahinfar acknowledges Paulino Pérez Rodríguez, Marion Dutreuil, and Lydia Hardie for their assistance in making graphs. Kent Weigel acknowledges partial financial support from the National Association of Animal Breeders (Columbia, MO).

Table 3.1. Description of features (explanatory variables) used for predicting the outcome of insemination events cows.

				Primiparous cows				Multiparous cows		
No	Feature	Type	Levels	Mean	SD	Missing %	Levels	Mean	SD	Missing %
1	Herd-Year-Month	Nominal	2672	-	-	0	2786	-	-	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 (month)	Numeric	-	24.2	2.6	0	115	50.64	16.69	0
6	Energy corrected milk 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 Breed	Numeric	-	1.7	1.1	14	21	2.74	2.2	4
11	Current Times Breed	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 Abomasums	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

Table 3.2. Features used to predict insemination outcomes in primiparous and multiparous Holstein cows, as ranked by information gain and gain ratio.

	Primiparous		Multiparous	
	Information Gain[1]	Gain Ratio[2]	Information Gain	Gain Ratio
Rank				
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 Breed	Stage of Lactation	Herd-Year-Month
4	DIM at Breeding	Stage of Lactation	DIM at Breeding	DIM at Breeding
5	Current Times Breed	Herd-Year-Month	Current Times Breed	Mastitis
6	Breeding Protocol	DIM at Breeding	Previous Times Breed	Current Times Breed
7	PTA of Daughter Pregnancy Rate for Sire of the Cow	Mastitis	Previous Days Open	Previous Times Breed
8	Age at calving (month)	Retained Placenta	Age at calving (month)	Previous Days Open
9	Energy corrected milk 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 (month)
12	Previous Times Breed	Age at calving (month)	Mastitis	Breeding Protocol
13	Mastitis	Energy corrected milk yield (kg)	PTA of Daughter Pregnancy Rate for Sire of the Cow	Energy corrected milk yield (kg)
14	Retained Placenta	Birth Difficulty in Last Gestation	Retained Placenta	Lactation Number
15	PTA of Daughter Pregnancy Rate for Calf	Previous Times Breed	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	Energy corrected milk 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**

Table 3.3. Area under the ROC curve (AUC) and proportion of correctly classified instances (CCI) for each of the five classification algorithms1 by 5-fold cross validation in primiparous and multiparous cows.

	Area Under Curve						Correctly classified instances					
Algorithm						Folds						
	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]Classification algorithms: Naïve Bayes (NB), Bayesian networks(BN), Decision Trees(DT), Bagging(BG), and Random Forests(RF).

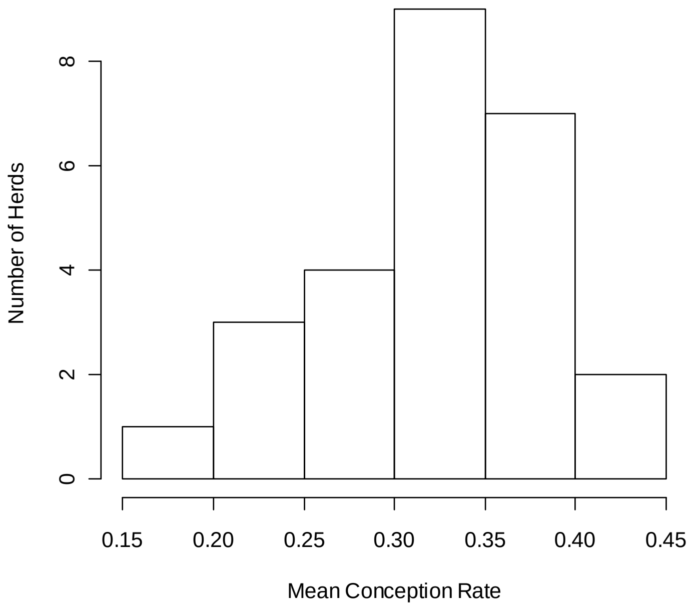

Figure 3.1. Histogram of mean conception rate in the 26 herds used in this study.

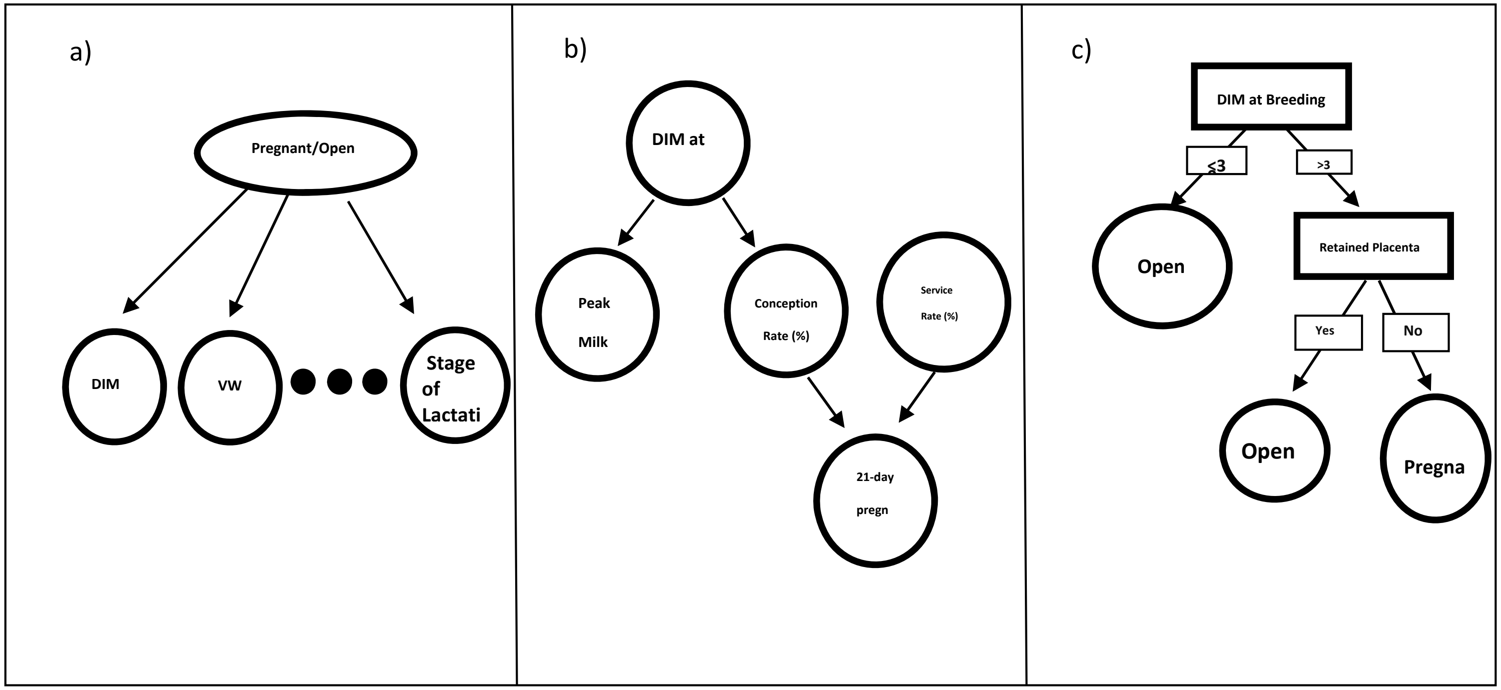

Figure 2. Schematic illustration of a) Naïve Bayes classifiers, b) Bayesian networks, c) Decision Trees. Days in milk (DIM), voluntiary waiting period (VWP),stage of lactation (Stage).

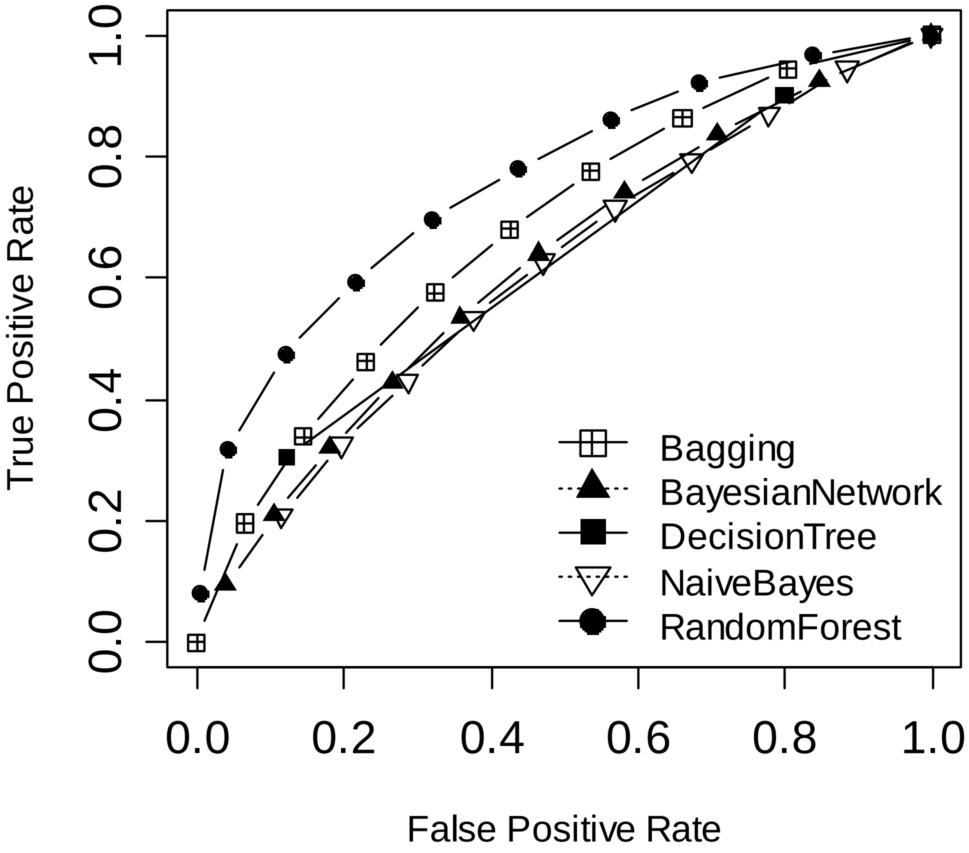

Figure 3.3. Receiver operating characteristic curves for five types of machine learning algorithms used to predict insemination outcome in primiparous cows.

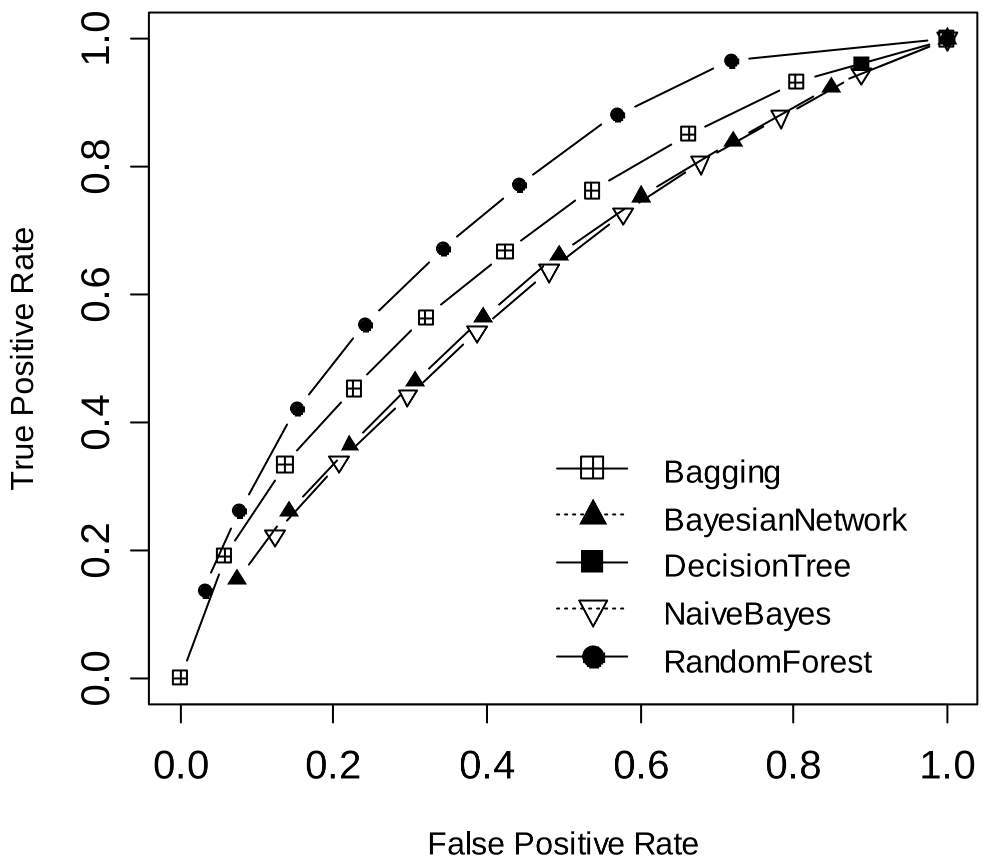

Figure 3.4. Receiver operating characteristic curves for five types of machine learning algorithm used to predict pregnancy outcome in multiparous cows.

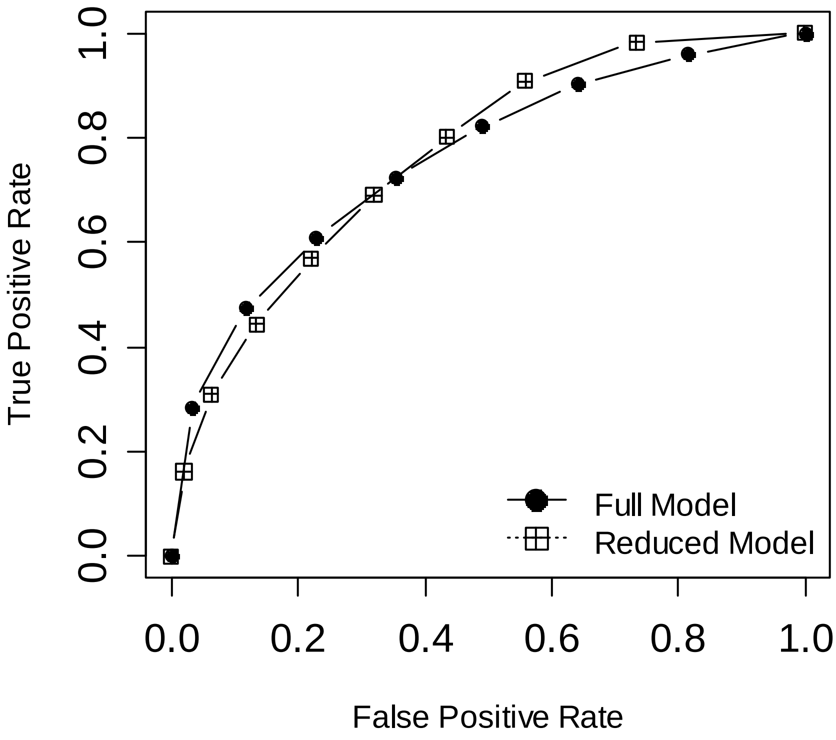

Figure 3.5. Receiver operating characteristic curves for full and reduced (without HYM of insemination) models of RF in primiparous cows in a lesion analysis.

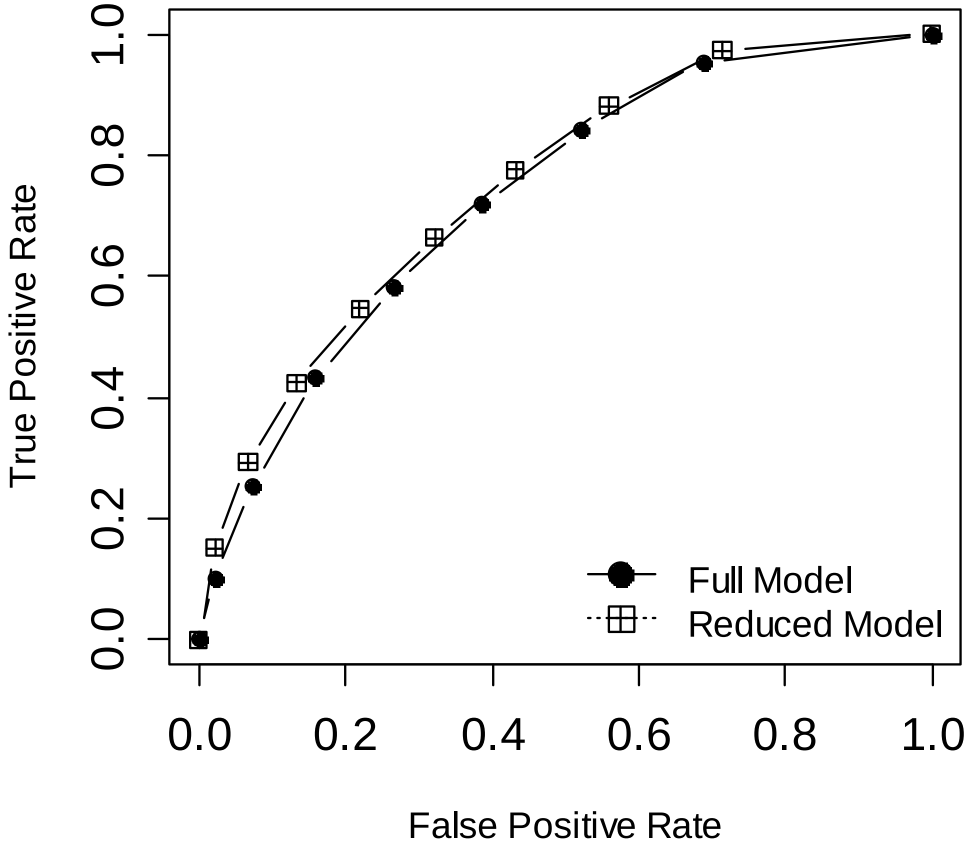

Figure 3.6. Receiver operating characteristic curves for full and reduced (without HYM of insemination) models of RF in multiparous cows in a lesion analysis.

Chapter4

Optimization of Reproductive Management Programs using Lift Chart Analysis and Cost-Sensitive Evaluation of Classification Errors

1. ABSTRACT

The common practice on most commercial dairy farms is to inseminate all cows that are eligible for breeding, while ignoring (or absorbing) the costs associated with semen and labor directed toward lowly fertile cows that are unlikely to conceive. Modern analytical methods, such as machine learning algorithms, can be applied to cow-specific explanatory variables for the purpose of computing the probabilities of success or failure associated with upcoming insemination events. Lift chart analysis can identify subsets of high fertility cows that are likely to conceive, and are therefore appropriate targets for insemination (e.g., with conventional AI semen or expensive gender-enhanced semen), as well as subsets of low fertility cows that are unlikely to conceive and should therefore be passed over at that point in time. While such a strategy might be economically viable, the management, environmental, and financial conditions on one farm might differ widely from conditions on the next, and hence the reproductive management recommendations derived from such a tool may be suboptimal for specific farms. When coupled with cost-sensitive evaluation of misclassified and correctly classified insemination events can be potentially powerful tool for optimizing the reproductive management of individual farms. In the present study, lift chart analysis and cost-sensitive evaluation were applied to a data set consisting of 54,806 insemination events of primiparous Holstein cows on 26 Wisconsin farms, as well as a data set with 17,197 insemination events of primiparous Holstein cows on 3 Wisconsin farms, where the latter had more detailed information regarding health events of individual cows. In the first data set, the gains in profit associated with limiting inseminations to subsets of 79 to 97% of the most fertile eligible cows ranged from \$0.44 to 2.18 per eligible cow, depending on days in milk at breeding and milk yield relative to contemporaries. In the second data set, the strategy of inseminating only a subset consisting of 59% of the most fertile cows conferred a gain in profit of

$5.21 per eligible cow. These results suggest that, when used with a powerful classification algorithm, lift chart analysis and cost-sensitive evaluation of correctly classified and misclassified insemination events can enhance the performance and profitability of reproductive management programs on commercial dairy farms.

Key words: machine learning, reproductive management, lift chart analysis, cost-sensitive evaluation, dairy cattle.

2. INTRODUCTION

The underlying physiological factors that determine the success or failure of an insemination event in dairy cows are complex and often unknown. Conception rate has a major impact on farm profitability, because it affects the number of replacement heifers needed to maintain herd size, the capacity to generate and sell extra heifers, the proportion of pregnant cows in the herd, average milk production of these cows, insemination and veterinary costs, involuntary culling rate, and other factors (Britt, 1985; Meadows et al., 2005; Giordano et al., 2011).

Previous studies have attempted to predict the outcome of an insemination event in lactating dairy cows based on results of ovarian palpation (Ludwich and Rader, 1967), interactions between stage of lactation and pregnancy status (Sharma et al., 1990) , or levels of nonesterified fatty acids and blood glucose (Garverick et al., 2013). Other studies have considered reproductive management programs from an economic point of view, for example, by estimating the economic value of a pregnancy (De Vries, 2006) or carrying out an economic comparison of natural service and timed AI programs (Lima et al., 2010; Giordano et al., 2012).

On most commercial dairy farms it is common practice to inseminate all cows that are eligible for breeding, with the hope that modest conception rates coupled with high service rates will lead to efficient overall reproductive performance. This strategy may be economically sensible when insemination costs, specifically semen and technician service, are low. However, evidence that this strategy maximizes farm profitability is lacking, and insemination costs can be relatively high if gender-enhanced semen is utilized. Several studies have indicated that targeted reproductive management of specific (groups of) cows could be more profitable (e.g., Giordano et al., 2012), but no study to date has attempted to predict the outcomes of insemination events for individual cows prior to the breeding while subsequently evaluating the benefits and costs of correct and incorrect breeding decisions, respectively.

(Shahinfar et al., 2014) evaluated the classification accuracy of several machine learning algorithms when predicting the outcome of an insemination event for an individual cow based on the production, reproduction, health, and genetic information available prior to the insemination. The proportion of insemination events that were classified correctly as successes or failures was 72.3% for primiparous cows and 73.6% for multiparous cows in 5-fold cross-validation using a random forest algorithm. Classification accuracies of this magnitude bring into question the practice of inseminating every eligible cow, while ignoring the vast amount of information available about each cow's genetic potential, health history, and lactation performance. Furthermore, we can carry out a lift chart analysis (Witten and Frank, 2005) to identify a subset of cows with much greater probability of conception than the entire pool of eligible cows. While such an analysis would, for example, facilitate the use of expensive gender-enhanced or high genetic merit semen for mating the subset of cows most likely to conceive, it would be limited by the fact that evaluation of algorithms based only on classification accuracy implies that the benefits and costs associated with all correctly classified or misclassified events are equal. In reality, the costs and benefits may differ considerably. For example, if insemination costs are low, the cost of failing to inseminate a cow that would have conceived (i.e., a false negative (**FN**)) is likely to be greater than the cost of inseminating a cow that will not conceive (i.e., a false positive (**FP**)). And, for example, the benefit derived from inseminating a cow that will conceive (i.e., a true positive (**TP**)) may be greater than the benefit associated with failing to inseminate a cow that would not have conceived (i.e., a true negative (**TN**)).

The objective of this study was to build upon the work of Shahinfar et al. (2014a), which showed that relatively high accuracy of classifying insemination events as successes or failures could be achieved prior to insemination, by using lift chart analysis to identify subsets of cows with high probability of conception and cost-sensitive evaluation to assess the economic benefits one might achieve by implementing such an approach for reproductive management of lactating dairy cows on commercial farms.

3. MATERIALS AND METHODS

3.1 Data

Two data sets were used in the analyses presented herein. The first, which will be denoted as Data_26, consisted of 54,806 insemination events from 22,210 primiparous Holstein cows that were inseminated between 2000 and 2010 in 26 Alta Genetics (Watertown, WI) Advantage Program herds located throughout Wisconsin. These breeding records represented a subset of the data used by Shahinfar et al. (2014), in which insemination events corresponded to three intervals of DIM, 60-90, 90-120, or 120-150, and three ranges of relative milk yield (**RMY**), between 18 and 6% below the within-herd mean, between 6% below and 6% above the within-herd mean, or between 6 and 18% above the within-herd mean. Mean conception rate in Data_26 was 32%, and individual herds ranged from 18 to 44%. Additional details about genetic predisposition, health history, and lactation performance of herds in Data_26 are provided in Shahinfar et al. (2014). The second, which will be denoted as Data_3, consisted of 17,197 insemination events from 5,356 primiparous Holstein cows that were inseminated between 2002 and 2013 in three additional herds that were also located in Wisconsin. These herds were chosen, because they had more complete data regarding common early postpartum health disorders (i.e., mastitis, ketosis, lameness, displaced abomasum, metritis, and retained placenta) than the herds in Data_26. In contrast to Data_26, the insemination events in Data_3 were not stratified by DIM or RMY, and all insemination events were analyzed jointly. The range in DIM at time of insemination was 12 to 479 in Data_3. Mean across-herd conception rate in Data_3 was 42%, and means for individual herds ranged from 38 to 45%.

As in Shahinfar et al. (2014), a total of 25 potential explanatory variables related to genetic predisposition, health history, and lactation performance that were known for each cow prior to insemination were considered when classifying insemination outcomes in Data_26 and Data_3. These included: mean conception rate within the herd during the previous 3 months, age at calving, stage of lactation, DIM at first breeding, number of times bred during the current lactation, number of times bred as a yearling heifer, breeding protocol utilized, sire's PTA for daughter pregnancy rate, service sire's PTA for daughter pregnancy rate, service sire's PTA for sire conception rate, inbreeding coefficient of the cow, sex of previous calf, fat:protein ratio at last test, ECM at last test, SCS at last test, and incidence of dystocia, mastitis, ketosis, retained placenta, lameness, displaced abomasum, and twinning in the current lactation. In the present study, herd, year, and month of insemination were considered as separate explanatory variables, whereas the study of Shahinfar et al. (2014) utilized herd-year-month as an explanatory variable.

3.2 Probability of Conception

Because of its superior classification accuracy relative to other machine learning algorithms considered in Shahinfar et al. (2014), a random forest algorithm was used in the present study. Since this is an ensemble algorithm in which numerous classifiers are trained by repeatedly sampling from the original data, numerous decision trees are generated. Therefore, the probability of conception for a particular insemination event corresponding to a given cow represents the aggregated probability across all trees and forests generated by the algorithm. In the present study, 10-fold cross-validation was carried out within each of the nine subclasses in Data_26 that corresponded to a given range in DIM and a given level of RMY, and all 54,806 insemination events were included in the analysis. Within each subclass, the proportion of correctly classified insemination events was calculated within each fold, and these were subsequently averaged across folds to determine classification accuracy. In addition, the area under the receiver operating characteristic (**ROC**) curve, which has FP rate on the x-axis and TP rate on the y-axis, was computed for each of the nine subclasses.

In contrast, Data_3 was randomly subdivided into two partitions. The first contained 14,000 insemination events, and this partition of the data was used to train the model and learn its parameters using 10-fold cross-validation, as described previously for Data_26. The second contained 3,197 randomly chosen insemination events, and the purpose of creating this partition was to mimic a real-life situation in which a decision-making tool was initially developed and tuned using an established database (e.g., at a breeding company) and subsequently applied to new data on a commercial farm (e.g., by a field technician). Therefore, the outcomes of the 3,197 insemination events in the second partition were masked throughout the training process; at the time of insemination data regarding the corresponding explanatory variables were fed into the model developed from the 14,000 events in the first partition in order to compute the probability of success or failure for each insemination event.

All analyses were carried out on a monthly basis, due to the fact that variables such as fat:protein ratio and ECM are typically measured monthly in DHI milk recording programs. Therefore, for instance, it was assumed that a cow inseminated in one month could be declared as pregnant or re-inseminated in the following month and, for instance, a cow that was not inseminated in the current month could be inseminated or passed over a second time in the following month. All explanatory variables regarding the genetic potential, health history, and lactation performance of a given cow were updated monthly, such that a cow with low probability of conception in a given month would tend to have low probability of conception in the subsequent month, but not strictly so. Likewise, a cow with high probably of conception in one month would tend to have a high probability of conception in the following month, but this probability could decrease due to lameness, high SCS, or other factors. In practice, one could consider a longer or shorter interval (e.g., 21 days) for compiling data regarding explanatory variables and generating predictions of insemination outcomes, and in herds with daily milk recording and automated measurement of physical activity and rumination patterns it would be possible to carry out such an analysis on a daily basis.

3.3 Lift Chart Analysis

In commercial applications of data mining algorithms that involve the marketing of consumer products, a lift chart analysis is often used to predict the increase in response rate that would be achieved by targeting a specific demographic group, as opposed to blanket marketing of a given product to the entire population of consumers. In the context of reproductive management of dairy herds, a lift chart analysis can be used to predict the gain in conception rate that would be achieved by inseminating a selected subset of cows with high probability of conception, as opposed to inseminating all eligible cows in the herd. For example, consider a herd with 1,000 cows that are eligible for insemination, and assume that the mean conception rate for the entire herd is 0.30. If we inseminate all of the eligible cows, 70% will not conceive, and our 1,000 inseminations will yield 300 pregnancies. However, if we have a classification algorithm that is capable of identifying a subset of 400 cows for which the expected conception rate will be 0.60, we can limit our inseminations to this highly fertile subset, and by doing so we can generate 240 pregnancies from only 400 inseminations.

In practical applications of such an analysis, we would have a classification algorithm that would predict the probability of success or failure for each potential insemination event, and we could rank these events according to their respective probabilities of success. Therefore, we could consider a wide range of scenarios and, for example, we may be faced with the decision of whether to inseminate a subset of 400 cows with an expected conception rate of 0.60 or a subset of 700 cows with an expected conception rate of 45%. This concept is illustrated in the lift chart shown in Figure 4.1, in which the proportion of cows to be inseminated is shown on the x-axis, and the number of anticipated pregnancies is shown on the y-axis. The gain in conception rate within a given subset, relative to the conception rate of the entire population of eligible cows, is known as the lift factor, which for the examples shown above would be (0.60 / 0.30) = 2.0 and (0.45 / 0.30) = 1.5, respectively.

3.4 Cost-Sensitive Evaluation

Optimizing classification accuracy in a lift chart analysis, as described above, can be misleading if one does not consider the benefits and costs associated with the various types of correctly classified or misclassified outcomes, specifically TP, TN, FP, and FN. It is naïve to assume that the costs of FP and FN errors are equal, and in practice the benefits of TP and TN outcomes may also differ. In a two-class problem, one can construct a cost-benefit matrix that has diagonal elements representing the benefits of correct classification decisions (TP and TN) and off-diagonal elements corresponding to the costs of incorrect classification decisions (FP and FN), as shown in Table 4.1. In the case of insemination events for lactating dairy cows, TP corresponds to the correct decision to inseminate a cow that will conceive, TN corresponds to the correct decision to refrain from inseminating a cow that would not have conceived, FP corresponds to the incorrect decision to inseminate a cow that will not conceive, and FN corresponds to the incorrect decision to refrain from inseminating a cow that would have conceived.

The precise assignment of costs and benefits associated with TP, TN, FP, and FN decisions can be subject of considerable debate, but it is clear that these values will be influenced by cow-level factors such as DIM and RMY, herd-level factors such as 21-day pregnancy rate and culling rate, and population-level parameters such as heifer price and salvage value. In the present study, we sought to demonstrate the concepts of lift chart analysis and cost-sensitive evaluation of classifiers using realistic economic values for commercial herds in Wisconsin, and we recognize that determination of the exact costs and benefits for a specific herd could easily justify another research study.

In the present study, the cost-benefit matrix was constructed using the estimates of pregnancy values and days open from a dynamic programming model developed by Kalantari and Cabrera, (2012). As noted earlier, medium RMY corresponds to cows with milk yield between 6% below and 6% above the within-herd mean, whereas low RMY corresponds to cows with milk yield between 18 and 6% below the within-herd mean, and RMY refer to cows with milk yield between 6 and 18% above the within-herd mean. Figure 4.2 illustrates the conceptual trajectories during the lactation for the value of a non-pregnant cow (solid line) and the value of a pregnant cow (dashed line); actual values for these curves were calculated by Kalantari et al. (2012) and were used to derive the cost-benefit matrix in the following manner. For a TP classification, the benefit corresponds to the difference between the value of a pregnant cow at point C on the dashed curve (one month after the insemination, when diagnosed pregnant) and the value of a non-pregnant cow at point B on the solid curve (one month after the insemination, when diagnosed as non-pregnant). In the case of a TN classification, the benefit corresponds to the difference between the value of a non-pregnant cow at point B on the solid line (one month after the insemination, when diagnosed as non-pregnant) and the value of a non-pregnant cow at point A on the solid line (at the time the cow was passed over for insemination). Moving from point A in a given to point B in the subsequent month is an inevitable outcome that is associated with a reduction in the value of a non-pregnant cow, but no misclassification occurred due to a poor prediction or an errant management intervention, so a benefit of $0 was assumed.

Table 4.2 shows the pregnancy values derived from the dynamic programming model of Kalantari et al. (2012) for each of the nine DMI by RMY subclasses of Data_26, as well as pregnancy values for 150 to 180 DIM at each level of RMY, which contribute to costs assigned to FN classifications. A pregnancy value and loss of pregnancy cost of $88 and $24.30 were assumed for the cost-sensitive evaluation in Data_3, based on the average of values given in Table 4.2, respectively. As noted earlier, and insemination cost of $25 was used for both analyses.

With regard to misclassified outcomes, the cost of a FP decision is depicted as the difference between point B on the solid curve for a non-pregnant cow (one month from now, when diagnosed as non-pregnant), and the point denoted as "-IC", which represents the insemination cost. The insemination cost, which was incurred unnecessarily due to an incorrect model prediction, was assumed to be roughly $25 (Olynk and Wolf, 2009). Conceptually, the most challenging case is the FN classification. In this case, the cow would have become pregnant if she had been inseminated, but she was not inseminated due to an incorrect model prediction. The opportunity cost for of incorrect decision can be calculated as the difference between points C and D on the dashed curve for a pregnant cow, where point D represents the value of a pregnancy two months after the foregone insemination. The assumption behind this is, if we forego the opportunity to inseminate a cow that would have become pregnant, we will have another chance to inseminate that cow in the following month. The insemination in the following month may or may not be successful, and the probability of conception at the subsequent insemination may differ if the cow's health status or other explanatory variables change from one month to the next.

3.5 Integration of Costs and Benefits into Classification Decisions

Given a set of explanatory variables, a typical classification algorithm will assign a positive or negative outcome to a given insemination event based on an underlying probability threshold. In other words, if the probability of conception exceeds a certain value the event will be classified as a success, and if it fall short of that value it will be classified as a failure. A decision threshold of 0.5 is typically used to classify instances as success or failures. Alternatively this threshold, which is known as the score threshold in the machine learning literature, can be optimized for a given set of cost values for incorrect decisions and benefit values for correct decisions. If a cost-benefit matrix with information about the values of TP, TN, FP, and FN classifications is available, this information can be coupled with the probability threshold to compute the costs associated with a particular classifier and probability threshold. More importantly, the optimum probability threshold at which benefits are maximized (or costs are minimized) can be determined. For example, if the decision in question is whether or not to inseminate a given cow, a cost-sensitive evaluation of the classifier can tell us the optimum number (and identities) of high fertility cows that should be inseminated, as well as the corresponding number (and identities) of low fertility cows that should not be inseminated, in order to maximize profitability of the reproductive management program. Or, more generally, such an evaluation could help us identify a subset of cows that should be inseminated with gender-enhanced semen, a subset that should be inseminated with conventional semen, and a subset that should not be inseminated. In the present study, benefits and costs associated with TP, TN, FP, and FN classifications were taken into account when evaluating new insemination decisions, but not when training the classifier; this is what differentiates cost-sensitive evaluation from cost-sensitive learning (Witten and Frank, 2005). This study focused on cost-sensitive evaluation, because previous studies (Elkan, 2001) have shown that the vast majority of the gain associated with incorporating a cost-benefit matrix comes in the evaluation phase, rather than the learning phase (by cost sensitive learning procedures, such as changing the balance of positive and negative training examples). Elkan (2001) recommends a two-step process of training a classifier without considering cost and benefit information, followed by computing optimal decisions using probability estimates given by the classifier based on the cost-benefit matrix.

By applying the elements of the cost-benefit matrix to different probability thresholds along the ROC curve, the costs and benefits of each potential decision were evaluated. In Data_26, 10-

fold cross-validation was used in conjunction with the cost-benefit matrix to determine the optimum probability threshold for each of the nine DIM by RMY subclasses, assuming a strategy of maximizing total profit of the reproductive program. Estimates of the lift factor, as well as estimates of financial gains or losses, were subsequently evaluated in each of the nine subclasses. In Data_3, the optimal probability threshold was also determined by 10-fold cross-validation, and model results were saved and applied to the aforementioned partition of 3,197 independent insemination events for estimation of financial gains and losses, in order to mimic a real application of this decision-making tool on a commercial farm.

4. RESULTS AND DISCUSSION

Results of the Data_26 analysis, including classification accuracy, lift chart analysis, and cost-sensitive evaluation, are shown in Table 4.3 for each of the nine DIM by RMY subclasses. As noted earlier, results are from 10-fold cross-validation within the full population of eligible cows in each of the nine subclasses. When the benefits and costs associated with various types of correctly classified and misclassified events were not considered, the accuracy of classifying upcoming insemination events as successes or failures (i.e., the proportion of correctly classified events) ranged from 0.58 to 0.80, whereas area under the ROC curve ranged from 0.61 to 0.80. No clear trends were observed between classification accuracy and DIM or RMY.

In the subsequent lift chart analysis of Data_26, which is also shown in Table 4.3, conception rates in selected subpopulations of highly fertile cows within each DIM by RMY subclass ranged from 0.35 to 0.51. These values were consistently greater than conception rates in the respective populations of all eligible cows, which ranged from 0.32 to 0.44. The corresponding lift factors ranged from 1.02 to 1.15, which suggests that financial gains might be achieved by focusing semen, labor, and supplies (e.g., for timed AI) on a selected subset of highly fertile cows rather than the entire population of eligible cows.

Table 4.3 also shows results of the cost-sensitive evaluation of correctly classified and misclassified insemination events for each of the nine DIM by RMY subclasses, which was implemented using the cost-benefit matrix data shown in Table 4.2. When costs and benefits associated with classification decisions were considered, the accuracies of prediction achieved by 10-fold cross-validation ranged from 0.42 to 0.57. These accuracies tended to be lower than accuracies that were achieved when costs and benefits were ignored, because in this case the algorithm chooses a score threshold that maximizes the profit derived from correct and incorrect decisions, rather than the score threshold that simply maximizes prediction accuracy. The optimum probability threshold (also known as the score threshold) reflects the point at which total profit of the reproductive management program is maximized, given a particular cost-benefit matrix and considering the algorithm's ability to identify various highly fertility subsets of cows in the lift chart analysis. The selected subsets of highly fertile cows represented from 79.0 to 97.2% of the eligible cows in the population, depending on the DIM by RMY subclass. This means that breeding the vast majority of cows was typically justified, presumably because insemination costs were low relative to the value of a pregnancy, noting that low insemination costs reflect the fact that semen of AI bulls used on commercial dairy farms in the US tends to be inexpensive. Higher insemination costs, for example due to the use of gender-enhanced semen, would likely result in the selection of a smaller percentage of cows for insemination. The target population refers to the group of cows that would conceive if they were inseminated, and this would be comprised of TP and FN classifications. In our cost-sensitive evaluation, 90.3 to 99.0% of cows in the target population were identified as candidates to be inseminated by the classification algorithm, so very few cows were missed as FN classifications. The apparent contradiction between the identification of a very high percentage of the target population, while achieving relatively modest accuracy of prediction, reflects the large number of FP classifications. Such classifications occur frequently due to low insemination costs coupled with high pregnancy values, as noted earlier, which mean that FP classifications tend to be inexpensive mistakes. Profit was calculated for the optimal reproductive management strategy, as determined by the lift analysis and cost-sensitive evaluation process, as well as for the base strategy in which all eligible cows were inseminated. Gains in profit were achieved in all nine DIM by RMY subclasses by focusing inseminations on the subset of highly fertility cows, as identified by the algorithm described herein. These gains ranged from $2783 for the population of 2291 eligible high RMY cows at 120 to 150 DIM, to $15,123 for the population of 13,745 eligible low RMY cows at 60 to 90 DIM. When expressed as the gain in profit per eligible cow, the gains ranged from a low of $0.44 per eligible medium RMY cow at 60 to 90 DIM, to a high of $2.18 per eligible high RMY cow at 90 to 120 DIM. As noted earlier, the actual gain that could be achieved on a particular farm will depend on the cost-benefit matrix associated with that farm, which will be a function of pregnancy rate, culling rate, replacement heifer inventory, and related factors. The gain will also depend on the algorithm's ability to accurately differentiate successes and failures prior to the insemination, given a set of explanatory variables, and this will depend critically on the accuracy and completeness of data recording for these explanatory variables.

Because the accuracy and completeness of data recording differs widely between farms, and because routine implementation of the decision-support tool described herein will likely involve centralized model development and parameter tuning with subsequent application by field consultants, we carried out a second analysis using Data_3. That analysis, for which the results are given in Table 4.4, involved development and tuning of the prediction model in a randomly selected partition of 14,000 insemination events on three farms with high quality data regarding infectious diseases, such as mastitis, and early postpartum metabolic disorders, such as ketosis. Very high accuracy of prediction was observed in these herds, with proportion of correctly classified insemination events of 0.80 and area under the ROC curve of 0.89. Note that accuracies and areas under the ROC curve ranged from 0.58 to 0.80 and from 0.61 to 0.80, respectively, in Data_26.

Greater accuracy of prediction should lead to a more informative lift chart, and Table 4.4 shows the lift chart analysis and cost-sensitive evaluation results corresponding to implementation of the aforementioned model in the 3197 randomly chosen independent insemination events that comprised the validation population. One can see that conception rate in the selected subpopulation of highly fertile cows was 0.66, as compared with 0.43 in the entire population of eligible cows, which results in a remarkably high lift factor of 1.55. When the costs and benefits of misclassified and correctly classified insemination events were taken into account, using a pregnancy value averaged across DIM and RMY subclasses in Table 4.2, the accuracy of prediction was 0.76 for an optimum probability (or score) threshold of 0.33. In this application, only 59% of the eligible population of cows were inseminated, yet these represented 92% of the target population of cows that would have conceived if inseminated. In terms of profit, the total gain achieved by using the optimized reproductive management strategy derived herein, relative to the alternative of breeding all 3197 eligible cows, was $16,660. On a per cow basis, the gain in profit was $5.21 per eligible cow, which was several-fold greater than the gains achieved with lower quality data, and hence poorer accuracy of prediction, in Data_26. This difference demonstrates the importance of accuracy and completeness of recording of potential explanatory variables, with respect to devising more profitable reproductive management programs.

While it is true that data on many farms are incomplete or inaccurate, especially as regards health events, there exist a substantial number of modern dairy farms with strictly enforced standard operating procedures, meticulous recording of data, and new technologies for automated capture of health, management, and performance information. Thus, although the results for Data_3 reflected the benefit of more accurate and complete health history data for individual cows as compared with herds in Data_26, there is still significant room for improvement in herds such as those represented in Data_3. For example, while the present study considered the genetic predisposition of individual cows using PTA for daughter pregnancy rate, further gains could be achieved by using genomic predictions if most or all eligible cows were genotyped. Similarly, while the present study considered each cow's health history as recorded in the on-farm herd management software program, additional gains could be achieved by using health-related indicators captured by physical activity and rumination monitoring systems. And finally, although the present study considered variables such as milk yield and fat:protein ratio at the most recent DHI test, future applications of this methodology might utilize milk weights, electrical conductivity, or mid-infrared spectral data recorded on a daily basis.

Lastly, as noted earlier, strategies for maximizing the profit of reproductive management programs may differ widely between farms, depending on current performance, financial position, or future plans. For example, a herd that is expanding may face a shortage of replacement heifers, and this could change the structure of the cost-benefit matrix for that herd. Conversely, if replacement heifers are plentiful and inexpensive, this could lead to a much different cost-benefit matrix and, in turn, a different optimization strategy. In theory, a 3x3 (or larger) cost-benefit matrix could be implemented, for example, to classify eligible cows or heifers into subclasses corresponding to gender-enhanced semen, conventional semen, or do-not-breed (or use beef semen) reproductive management actions at a given time. However, this would increase the complexity, particularly as regards determination of appropriate cost and benefit parameters. Fruitful topics for future research projects could include an assessment of the extent of variation in cost and benefit parameters between herds, as well as variation in these parameters within a herd over time, in addition to an evaluation of the sensitivity of recommended reproductive management actions to errors in the cost-benefit matrix. Previous authors have recommended a delay in inseminating high-producing primiparous cows, because the pregnancy value associated with these highly persistent and fertile cows tends to be negative in early lactation (DeVries, 2006). Furthermore, McCullock et al., (2013) showed that profitability is enhanced by using gender-enhanced semen on yearling heifers and a fraction of genetically superior cows, and Seidel, (2003) noted that gender-enhanced semen could be used to improve the genetic merit of herd replacements when its usage was targeted toward genetically superior cows or heifers. Although the present study did not consider genetic merit of the cows when constructing the cost-benefit matrix, optimizing costs and benefits of insemination decisions involving cows of high or low genetic merit in the presence of gender-enhanced semen could be the subject of a future study.

5. CONCLUSIONS

Although prediction of the outcome of an insemination event for a lactating dairy cow from dozens of potential explanatory variables is difficult, information regarding genetic predisposition, health history, lactation performance, and other factors can be used with modern analytical tools, such as machine learning algorithms, to identify cow with substantially higher or lower probability of conception, as compared with the average of the population of eligible cows. When semen is inexpensive and technician costs are modest, the common practice is to simply inseminate all eligible cows and ignore the costs associated with breeding lowly fertile cows with a history of health and reproductive problems. Lift chart analysis can be used to stratify the pool of eligible cows into those that are likely to conceive and those that should not be inseminated at a given point in time. When coupled with cost-sensitive evaluation of misclassified and correctly classified insemination events, the approach provides a potentially powerful tool optimizing the reproductive management strategy of a given farm under that farm's economic conditions. Implementation of such a decision-making tool would likely require the assistance of a professional reproductive management consultant, but perhaps key features of this approach could be incorporated into existing dairy herd management software programs in the future. The rapid development and adoption of systems and devices for real-time monitoring of animal health, activity, and performance will enhance our ability to refine, and derive greater benefit from, the types of decision-support tools described herein.

6. 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. Data for this study were provided by Alta Genetics (Watertown, WI) and the USDA-ARS Animal Improvement Programs Laboratory (Beltsville, MD). Authors acknowledge the technical comments and support from Jon Schefers and David Vanness. Kent Weigel acknowledges partial financial support from the National Association of Animal Breeders (Columbia, MO).

Table 4.1. Conceptual illustration of cost-benefit matrix for classifying upcoming insemination events as successes or failures based on a set of explanatory variables, including benefits assigned to true positive (TP) classification of a cow that was inseminated and became pregnant, a true negative (TN) classification of a cow that was inseminated and would not have become pregnant if inseminated, a false positive (FP) classification of a cow that was inseminated but did not become pregnant, and a false negative (FN) classification of a cow that was not inseminated but would have become pregnant if inseminated. The benefits and costs assigned to each type of classification decision in the present study are shown in parentheses.

		Predicted Classification	
		Pregnant	Non-Pregnant
Actual Outcome	Pregnant	TP (Pregnancy Value)	FN (Difference between Two Subsequent Pregnancy Values)
	Non-pregnant	FP (Insemination cost)	TN (0)

Table 4.2. Pregnancy values ($) assigned to primiparous Holstein cows in the Data_26 analysis, according to days in milk (DIM) and relative milk yield (RMY), as derived from the dynamic programming model of Kalantari et al. (2012). Low RMY cows were between 6 and 18% below the within-herd mean, medium RMY cows were between 6% below and 6% above the within-herd mean, and high RMY cows were between 6 and 18% above the within-herd mean.

		RMY		
		Low	Medium	High
DIM	60-90	66	65	63
	90-120	89	87	86
	120-150	124	123	122
	150-180	149	148	147

Table 4.3. Classification accuracy and results of the lift chart analysis and cost-sensitive evaluation in Data_26, according to days in milk (DIM) at and relative milk yield (RMY), when maximizing total profit of the reproductive management program.

	60 to 90 DIM			**90 to 120 DIM**			**120 to 150 DIM**		
	Low RMY	Medium RMY	High RMY	Low RMY	Medium RMY	High RMY	Low RMY	Medium RMY	High RMY
					Classification Accuracy				
Eligible Population (n)	13,745	11,015	5239	6109	5271	2372	4576	4188	2291
Accuracy of Prediction	0.68	0.58	0.72	0.74	0.72	0.74	0.75	0.76	0.80
Area under the ROC[1] Curve	0.75	0.61	0.80	0.78	0.77	0.77	0.76	0.78	0.80
					Lift Chart Analysis				
CR[2] in the Eligible Population	0.44	0.44	0.44	0.37	0.36	0.34	0.33	0.32	0.34
CR in the Selected Subpopulation	0.49	0.45	0.51	0.43	0.41	0.41	0.36	0.36	0.35
Lift Factor	1.11	1.02	1.15	1.15	1.13	1.20	1.09	1.13	1.04
					Cost-Sensitive Evaluation				
Accuracy of Prediction	0.56	0.46	0.57	0.50	0.49	0.52	0.42	0.45	0.46
Optimal Probability Threshold	0.11	0.10	0.14	0.10	0.09	0.10	0.05	0.06	0.05
% of Eligible Population	85.5	97.2	84.5	83.6	85.3	79.0	89.9	85.6	87.0
% of Target Population	95.7	99.0	97.2	96.1	96.7	95.0	97.8	96.9	90.3
Profit with Optimal Strategy ($)	260,334	196,495	96,841	130,413	105,504	43,210	130,963	114,116	67,573
Profit with Inseminating All Cows ($)	245,211	191,661	86,548	124,196	97,619	38,047	125,932	107,883	64,790
Total Gain in Profit ($)	15,123	4834	10,293	6217	7885	5163	5031	6233	2783
Gain in Profit per Eligible Cow ($)	1.10	0.44	1.96	1.19	1.50	2.18	1.10	1.49	1.21

[1] Receiver operating characteristic

[2] Conception rate

Table 4.4. Results of cost-sensitive evaluation and lift chart analysis for 3,197 independent insemination events in Data_3, with the objective of maximizing total profit of the reproductive management program.

Classification Accuracy	
Eligible Population for Training (n)	14,000
Accuracy of Prediction	0.80
Area under the ROC[1] Curve	0.89
Lift Chart Analysis	
CR[2] in the Eligible Population	0.43
CR in the Selected Subpopulation	0.66
Lift Factor	1.55
Cost-Sensitive Evaluation	
Eligible Population for Validation (n)	3197
Accuracy of Prediction	0.76
Optimal Probability Threshold	0.33
% of Eligible Population	0.59
% of Target Population	0.92
Profit with Optimal Strategy ($)	91,354
Profit with Inseminating All Cows ($)	74,694
Total Gain in Profit ($)	16,660
Gain in Profit Per Eligible Cow ($)	5.21

[1] Receiver operating characteristic

[2] Conception rate

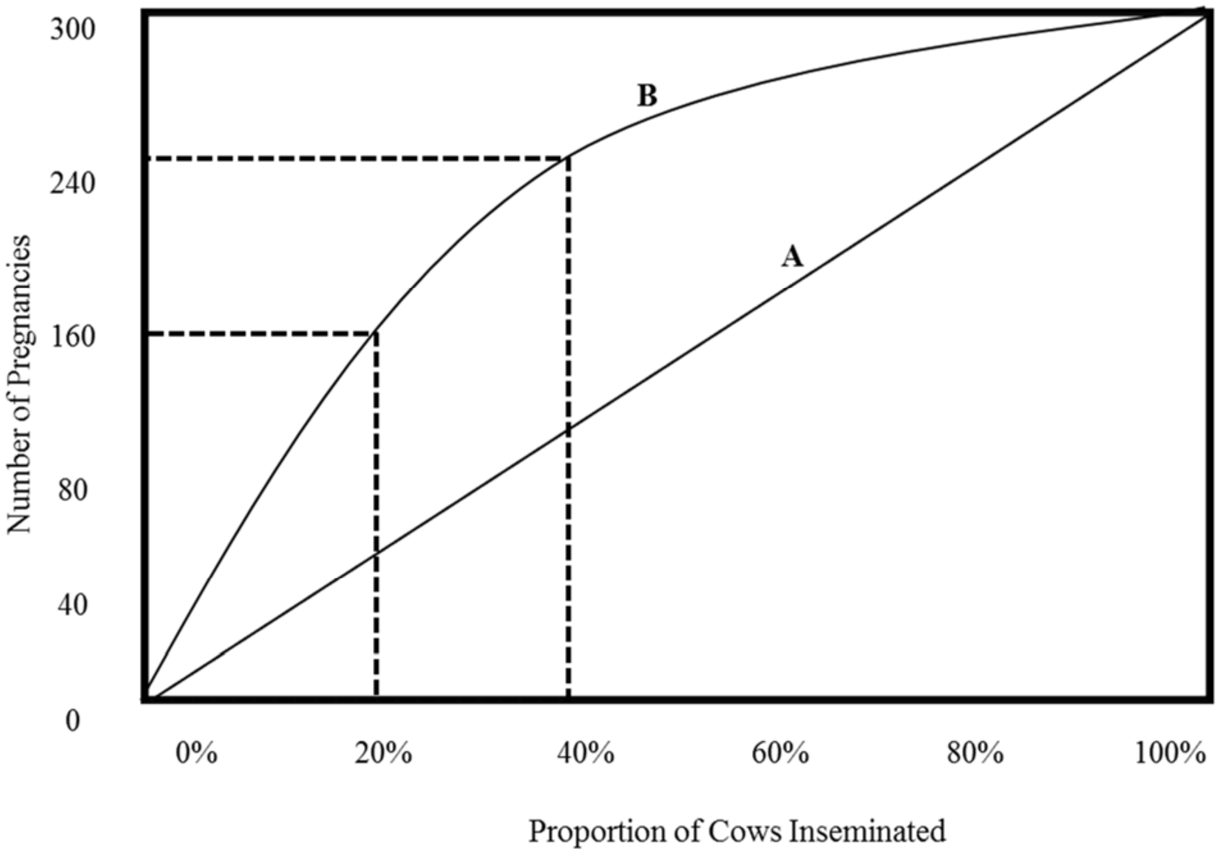

Figure 4.1. Conceptual illustration of a lift chart analysis for the classification problem of predicting the success or failure of an insemination event, as carried out in the present study. Line A shows the number of pregnancies that would be generated by inseminating a random subset of cows of a given size, whereas curve B shows the number of pregnancies that would be generated by inseminating subsets of cows of varying size (e.g., 20 or 40%) after ranking the insemination events from highest to lowest based on probability of conception.

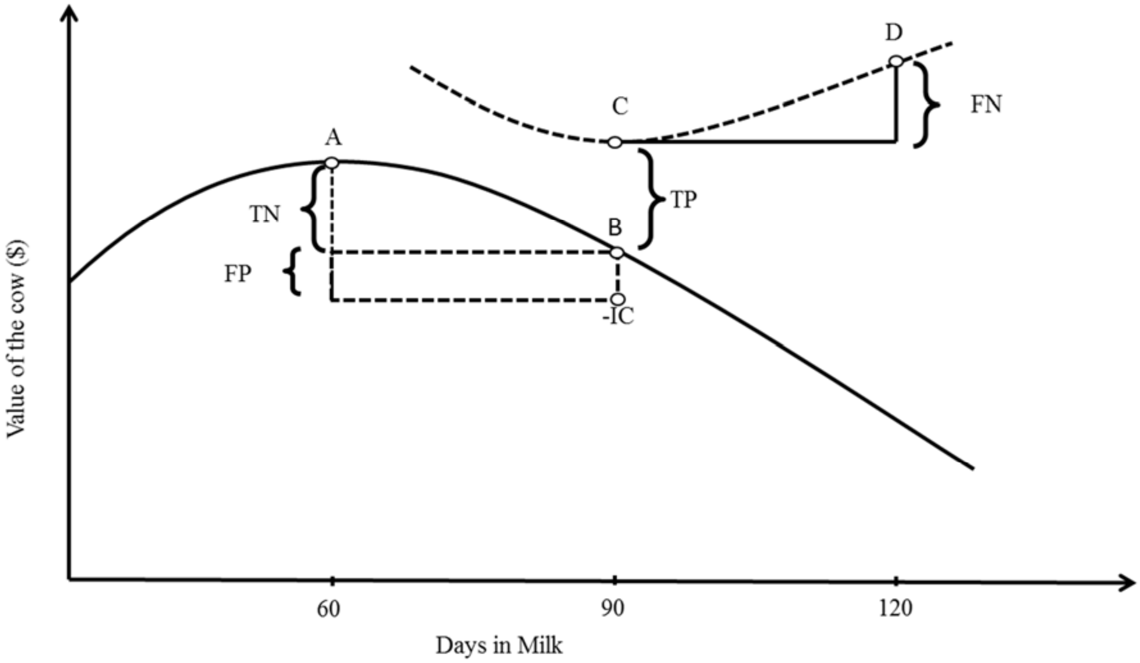

Figure 4.2. Conceptual curves that depict the value of a pregnant cow (dashed curve) and value of a non-pregnant cow (solid line) and value of a pregnant cow (dashed line), as used to assign the benefits of true positive (TP) and true negative (TN) classifications and costs of false positive (FP) and false negative (FN) classifications in the present study. Point A is the value of the cow at the time of insemination. Points B and C are the values of non-pregnant and pregnant cows one month after the insemination, respectively. Point D is the value of a pregnant cow two months after the insemination, and IC is the insemination cost.

Chapter 5:

Introducing a Hybrid Genetic-Swarm Fuzzy Expert System for Genome-Enabled Prediction

1. Abstract:

Fast and cost-effective prediction models are increasingly in demand for commercial use. The fuzzy expert system is distinguished from other black boxed non-parametric methods, such as random forests and artificial neural networks, because it is easy to understand and interpret. Information gained through a genetic algorithm and particle swarm optimization techniques were used to develop a fuzzy expert system to classify bulls based on a selected subset of 100 single nucleotide polymorphisms (SNPs) for three phenotypic traits. Advanced genetic algorithm operators were used for the construction of a rule set, and particle swarm operators were used to find optimal membership functions. The performance of the proposed model was suboptimal, presumably because of the structure of the data used in this analysis.

2. Introduction

Since the groundbreaking work of Meuwissen et al., (2001) there have been extensive efforts in animal and plant breeding to develop practical methods for selecting parents of the next generation based on predictions of their genetic merit derived from genotypes rather than traditional pedigree based estimated breeding values (EBVs).

In animal and plant breeding the general belief is that most economically important traits have a complex heritage and are controlled by an infinite number of genes, each having a small effect. Therefore, to predict these complex traits a large number of variants must be considered simultaneously. The continuing advances in high-throughput genotyping and sequencing technology have led to the discovery of millions of SNP markers in the genomes of organisms. This allows exploitation of multi-locus linkage disequilibrium (LD) between quantitative trait loci

(QTL) and SNPs to predict genetic merit (de los Campos et al., 2013). Other than a few researchers that have used machine learning (Long et al., 2009; Yao et al., 2013) and semi-parametric procedures (Gianola et al., 2006; de los Campos et al., 2010; Gianola et al., 2011), the majority of researchers have focused on statistical methods such as Bayesian regression (Meuwissen et al., 2001; Habier et al., 2011) and best linear unbiased prediction (BLUP) (Fernando and Grossman, 1989; VanRaden, 2008). In addition to this, a few researchers have published review papers that compared different methods introduced for the purpose of genomic selection (GS) and investigated the pros and cons of each approach (Calus, 2009; Gianola et al., 2009; Campos et al., 2013). One reason for the lack of research on rule-based methods for GS is that knowledge acquisition in such a complex and highly dimensional space is the most limiting factor, and methods cannot be implemented by traditional procedures such as interviewing domain experts to extract the optimum set of rules and membership functions. Nevertheless, this obstacle can be overcome by the means of global search optimization methods, such as evolutionary algorithms.

Fuzzy expert systems have been successfully developed and used for discovering gene regulatory networks (Woolf and Wang, 2000), classification (Chen and Tsai, 2008; Schaefer et al., 2009; Ho et al., 2006), and clustering (Maraziotis, 2012) of gene expression data in conjunction with other machine learning methods (Schaefer and Nakashima, 2010; Ganesh Kumar et al., 2012). In all cases, fuzzy based methods outperformed classical classification or statistical methods. However, there has not been any research to investigate the ability of a fuzzy expert system in GS in animal or plant breeding.

Fuzzy set theory provides an approximate, yet effective, means for describing the characteristics of a system that is too complex or ill-defined to admit precise mathematical analysis (Zadeh, 1975, 1978).The fuzzy approach is based on the fact that the key elements in human thinking are a class of objects, not just numbers, in which the transition from membership to non-membership is smooth rather than crisp. Most of the logic behind daily based human reasoning is not the traditional binary or even multivalued logic, but rather logic with fuzzy truth and fuzzy rules of inference (Pal, 1999).

The objective of this study was to develop a fuzzy expert system empowered by two stochastic machine learning methods, in order to construct the knowledge base for genome enabled selection of dairy cattle for daughter pregnancy rate (DPR). Furthermore, this method's performance was compared with the performance of one commonly used method in the field of animal breeding, known as genomic BLUP (GBLUP) and methods used commonly in the field of machine learning, such as random forest (RF), ordinary least square (OLS), and support vector machine (SVM) algorithms.

3. Material and Methods

3.1 Data

Genotypes of 10,618 Holstein bulls using the Illumina BovineSNP 50k Bead Chip were provided by USDA-ARS Animal Genomics and Improvement Laboratory (Beltsville, MD). Daughter yield deviations (DYDs) for DPR were considered as the phenotype. After removing bulls without daughter information, 8,010 bulls born from 1952 to 2005 were kept in the dataset for later analysis. The dataset was divided into two portinos based on birthdate; 7,000 bulls born before July 1, 2005 were kept as the training set, and the remaining 1,010 bulls born after this date were considered as the testing set.

A brief description of the data distribution is presented in Table 5.1. SNP markers with a call rate lower than 90% or minor allele frequency below 5% were removed using Plink (Purcell et al., 2007). After SNP editing, 40,812 SNPs were available for the analysis. In order to decrease the dimension of the dataset and exclude redundant SNPs that were in linkage disequilibrium (LD) with each other, a LD pruning procedure was performed using Plink. A subset of 17,310 LD-pruned SNPs were obtained using $r^2 \leq 0.4$, and only this subset of SNPs was included in subsequent analyses and model development.

3.2 Definition of the phenotypes used in this study:

Yield deviation represents the yield of an animal adjusted for all effects other than genetic merit and error. It contributes indirectly to the breeding value of its sire, after it has been combined with information from its parents and offspring. Thus, the progeny contribution is a regressed measure, and it is not independent of information from parents. Therefore, an independent and unregressed measure of progeny performance is the DYD (VanRaden and Wiggans, 1991) which is basically a weighted average of the yield deviations of all progeny after correction for known fixed effects and the breeding values of mates (Mrode, 2005).

3.3 Daughter pregnancy rate:

Genetic evaluations for DPR of US dairy bulls have been available since 2003. Because of the low heritability of fertility traits, approximately 0.04, the predicted transmitting abilities (PTAs) for daughter fertility will have a poor reliability for bulls that don't have a large number of progeny. For bulls with only first-crop daughters (from the initial progeny testing), reliability is about 60%, and parent average still provides a major contribution. As computed presently, pregnancy rate (PR) and days open are, in fact, the same trait; a 1% increase in PR is roughly equal to 4 additional days open. PR is commonly used for herd management purposes, as a measure of how quickly cows in a herd become pregnant after the voluntary waiting period, and it is computed as the percentage of eligible non-pregnant cows that become pregnant during each 21-day time period. The linear approximation was introduced by VanRaden et al., (2004) and is as follows:

DPR = 0.25 (233 – DO)

The PTA for DPR is a measure of a sire's effect on the female fertility of his daughters, opposed to sire conception rate (SCR), which measures male fertility and is computed as the difference in conception rate of the bull's semen in comparison with other bulls in the population.

3.4 Estimates of SNP effects:

In order to estimate each of the 17,310 SNP effects on the trait of interest, a whole genome Bayesian regression method was used. The BGLR R-package (de los Campos and Perez Rodriguez, 2012) was used to carry out the analysis.

The general structure of standard Bayesian Linear Model used in GS is:

$$P(\mu, \beta, \sigma^2 | y, \omega)$$

$$\propto p(y|\mu, \beta, \sigma^2) p(\mu, \beta, \sigma^2 | \varpi)$$

$$\propto \prod_{i=1}^{n} N\ (y_i | \mu + \sum\nolimits_{j=1}^{P} x_{ij}\beta_j, \sigma^2\ ) \prod_{j=1}^{p} p\ \left(\beta_j\ |\omega\right) P(\rho^2)$$

where

$$P(\mu, \beta, \sigma^2 | y, \omega)$$

is the posterior density of the model unknowns

$$\{\mu, \beta, \sigma^2\}.$$

given the data (y) and hyper parameters. Also,

$$(\omega), p(y|\mu, \beta, \sigma^2) = \prod_{i=1}^{n} N(y_i\ |\mu + \sum_{j=1}^{p} x_{ij}\, \beta_j, \sigma^2)$$

is the conditional density of the data given the unknowns, which for continuous traits are commonly independent normal densities with mean

$$E(y_i|\mu, \beta, \sigma^2) = \mu + \sum_{j=1}^{p} x_{ij}\beta_j$$

and variance

$$VAR\ (y_i|\mu, \beta, \sigma^2) = \sigma^2.$$

The quantity

$$P(\mu, \beta, \sigma^2|\omega) \propto \prod_{j=1}^{p} p\ \left(\beta_j\ |\omega\right) P(\sigma^2)$$

is the joint prior density of model unknowns, including the intercept (μ) , which is commonly assigned a flat prior, marker effects {βj}, which are commonly assigned IID informative priors, and the residual variance (σ^2), which is commonly assigned a scaled-inverse chi-square prior with degree of freedom d.f. and scale parameter S, which is

$$p(\sigma^2) = x^{-2}(\sigma^2\ |\mathrm{d.f.}, S).$$

Here we use a parameterization

$$E(\sigma^2|\mathrm{d.f.}, \mathrm{S}) = (\mathrm{d.f.}\times \mathrm{S})/(\mathrm{d.f.} - 2)$$

(de los Campus 2013).

In Bayes B, a point of mass of zero and a scaled-t density slab priors are used to induce a combination of variable selection and shrinkage (Meuwissen et al., 2001). Therefore, we have set the majority of SNP effects to zero a priori, which is in alignment with our purpose of subset selection of most informative SNPs. The model was fitted with 50,000 iteration of Gibbs sampling; the first 25,000 iterations were considered as burn-in and discarded. Next, the 99% confidence interval for each SNP was calculated, and the 100 most SNPs with largest estimated contributions to the variance were selected for further model development.

4. Genome-Enabled selection using a novel genetic swarm fuzzy expert system

As noted earlier, GS) is becoming increasingly popular in animal and plant breeding and has recently gained attention in human genetics (de los Campus et al., 2013). Many different methods, including Bayesian regression (Meuwissen et al., 2001; Habier et al., 2011), GBLUP (VanRaden, 2008), and machine learning methods (Long et al., 2009) have been proposed. However, there have been no attempts to date to use soft computing techniques, such as fuzzy expert systems. Fuzzy expert systems implement nonlinear mapping from the input space to the output space and classifies observations into different categories. This approach can also be used for breeding value prediction using neuro-fuzzy models (Shahinfar et al., 2012).

4.1 Fuzzy Logic:

Fuzzy logic is a type of multi-valued logic, dealing with reasoning that is uncertain rather than fixed and precise. Against traditional binary sets, which variables take either true or false values, in fuzzy logic variables are associated with a truth value that ranges between 0 and 1. Fuzzy logic has been developed to handle the concept of partial truth, in which the truth value may fall between completely true and completely false (Novák et al., 1999).

It is an approach to compute based on "degrees of truth" rather than the usual Boolean logic on which modern computers are based (Zadeh, 1965). Fuzzy logic includes 0 and 1 as extreme cases of truth, and it may be better to see fuzzy logic as the way reasoning really works and binary or Boolean logic is simply a simplified special case of it (Rouse, 2006)

4.1.1 Fuzziness vs probability: Fuzziness describes the ambiguity of an event. What it measure is the degree to which an event happens, not whether it happens or not. On the other hand, randomness describes the uncertainty of an event's occurrence. Randomness describes how frequently an event occurs or not and how you can bet on it. In another words, "whether an event occurs is "random" and the degree it occurs is "fuzzy""(Kosko, 1990).

4.1.2 Fuzzy expert systems: A fuzzy expert system is simply an expert system which uses a set of if-then rules and respective membership functions to reason about data. The general form of a rule in a fuzzy expert system is as follow:

$$R_i\colon if\ x_{q1}\ is\ A_{i1}\ and\ \ldots..and\ x_{qn}\ is\ A_{in}\ then\ class\ is\ C_i$$

where $A_{i1} \ldots \ldots A_{in}$ are antecedent fuzzy sets of the inputs $x_{q1} \ldots\ldots x_{qn}$ and C_i is the output class label. A set of such rules forms the rule base for the fuzzy expert system in which qualitative reasoning to infer the results is performed.. Fuzzy if-then rules along with membership functions form the core part of a fuzzy inference system.

Fuzzy inference is the process of the mapping from a set of given input to an output using fuzzy logic. This mapping then provides a basis which can be used in decisions making or pattern recognition.Different types of fuzzy reasoning have been developed. Depending on the types of fuzzy reasoning and knowledge base (fuzzy if-then rules) used, most fuzzy inference systems (FIS) would be classified into two types:

Mamdani: This was among the first control systems built using fuzzy set theory. It was formed as an attempt to control a steam engine and boiler combination by synthesizing a set of linguistic control rules obtained from experienced human operators (Mamdani, 1977). The overall fuzzy output is obtained by applying "max" operation to the fuzzy outputs (each of which is derived by applying "min" operation to the firing strength and the output membership function of each rule). Different methods have been proposed to calculate the final crisp output based on the overall fuzzy output. These could be centroid of area, bisector of area, mean of maxima, and maximum criterion, to name a few.

Takagi-Sugeno: In Takagi and Sugeno's fuzzy if-then rules (Takagi and Sugeno, 1983)a linear combination of input variables plus a constant term. The final output of the system would be the weighted average of all rule's output. Most of the differences come from the specification of the consequent part (monotonically non-decreasing or bell-shaped membership functions, or crisp function) and also the de-fuzzification (weighted average, centroid of area, etc.) schemes (Jang, 1993).

FISs are also known as fuzzy-rule-based systems as well. Generally an FIS is contain the following functional parts:

1) Knowledge base, which consists of a rule base containing a set of fuzzy if-then rules and a database that defines the membership functions of the fuzzy sets used in the fuzzy rules;

2) Fuzzification interface, which transforms the crisp inputs into degrees of belongingness to each of the linguistic values;

3) Decision-making unit, which performs the inference operations on the rules, by combining the membership values on the premise part to get the firing strength (weight) of each rule and generatingthe qualified consequence (either fuzzy or crisp) of each rule depending on the firing strength,

4) Defuzzification interface, which transform the fuzzy results of the inference into crisp outputs by aggregating the qualified consequences.

Mamdani and Sugeno FIS (Jang, 1993; Pal, 1999) were used in this research, and their performances were compared against each other as well as with that of GBLUP, RF, OLS, and SVM.

In the model developed herein, a set of input SNPs were compared against the antecedent component of each if-then rule, and in the case of a match the response (output) was obtained through the fuzzy implication operation. The response of each rule is weighted by the extent to which the rule is fired. Next, the responses of fired rules for a particular input are combined to obtain the final class prediction of that input set.

In the development of FIS, knowledge acquisition is the most critical part. It can be done practically, by using an expert's knowledge through a questionnaire or interview, or alternatively by appllying data mining tools to existing databases. In the current study, because of the large number of SNPs, there would be an infinite number of possible rules, and therefore no experts exist to define a complete rule set and the optimum membership functions for individual SNPs that are included in antecedents of the rules.

One or more machine learning methods can be used to tackle this problem. The hybrid proposed fuzzy system uses a genetic algorithm (GA) to come up with the optimum set of if-then rules and particle swarm optimization (PSO) to tune the membership function points.

5. Genetic-Swarm Fuzzy algorithm:

The schematic structure of the implemented algorithm is shown in Figure 5.2. As shown in the figure, the fitness of the current rule set and corresponding membership functions (MFs) are evaluated during each iteration. Then, during the GA run, a higher fitness rule set is created and selected by GA operations. Next, in the PSO run, the optimal points of MF are obtained. The new rule set and MFs will be used to create a new FIS. The algorithm was implemented in Matlab (R2013a, Student version). Mamdani and Sugeno fuzzy inference systems with minimum, maximum, production, and center of gravity approaches to determine the crisp output were developed. The algorithm is designed such that it has the memory of the past population and previous MFs, so at each point it is able to compare the current rule set, MFs, and parameters with those fitted previously and use them to classify observations in the testing set.

5.1 Representation of the rule set:

Each rule of the rule set was represented by an integer string. Each rule has two parts and is shown conceptually in Figure 5.3. The, S_1, S_2,…S_n represent SNP_1 to SNP_n from the input file and were coded as 0, 1, 2, or 3 based on estimated SNP effect and minor allele frequency. A score of 0 means no impact on the trait, whereas scores of 1, 2 or 3 correspond to low, medium, or high impact on the trait, respectively. A single binary digit represents the output class of the rule and can take values of 1, 2, or 3 to represent poor, medium, or high for the output variable. The third part of each rule is a floating point number as a default weight, which can take 0.66 or 0.99 as the degree of belief to the rule.

5.2 Operations of the rule set:

The primary advantage of GA comes from crossing over and selection operations in producing a new state space at each successive generation. Operations like mutation have the tuning role to prevent premature convergence in runs of GA. Several selection and crossing over procedures (all inspired by natural population evolutionary processes) have been suggested by researchers in this field. In the present study, tournament selection and two-point crossing over were used. In tournament selection, an individual will be selected randomly from the population, and the most fit will have a chance to be the predecessor of the next generation of solutions. Although any size of tournament is possible between individuals of the population, we only considered a tournament between a pair of randomly chosen individuals in the present study. It is worth noting that, in this case, each individual is a collection of rules. After the population of the next generation is constructed, two-point crossing over is applied between two randomly selected strings from two selected individuals, based on a roulette wheel followed by the gene cross swap operator (GCSO) introduced by Devaraj and Gunesh (2010) and mutation. In GCSO, two randomly selected SNPs will be swapped between two individuals. Mutation was applied with a very low probability, to add diversity to the population by switching each bit from 0 to either 1, 2, or 3; from1 to either 0 or 2; from 2 to either 1 or 3, or from 3 to either 2 or 0 with uniform probability.

5.3 Representation of MF:

To represent the MF for each SNP, the 99% confidence interval of the SNP's estimated effect from the training data was obtained from BayesB and used as the lowest and highest limit of the MF. This range was arbitrarily divided into three membership functions, named as low (L), medium (M), and high (H), as shown in Figure 5.4. Trapezoidal membership functions were used to represent L and H, and a triangular MF was used to represent M. There are three points to represent each function, and therefore a total of nine membership points are needed for each input SNP. As mentioned above, the first and last points are fixed, because they represent the confidence limits of the estimated SNP effect. The seven intermediate points are optimized during runs of PSO in the algorithm, such that each point has boundaries limited by two other points; P2 has {P1, P9}, P3 has {P2, P9}, P4 has { P2, P3}, P5 has {P4, P9}, P6 has {P5, P9}, P7 has {P5, P6}, and P8 has {P7, P9} as limits (Gunesh et al., 2012).

5.4 Operations for MFs:

Optimum MF points were found during runs of the PSO algorithm. The first and last points of each set of MF were fixed, and the seven intermediate MF points were optimized by standard velocity and position updating rules of a typical PSO algorithm, according the following equations:

$$v_{t+1}^{i} = wv_{t}^{i} + c_1 r_1 \left[\hat{x}_{t}^{i} - \ x_{t}^{i}\right] + c_2 r_2 \left[g_t - x_{t}^{i}\right]$$

In which v_{t+1}^{i} is the velocity of particle *i* in the iteration t+1, and v_{t}^{i} is the velocity of particle *i* in iteration *t*. Parameters $w, c_1,$ and c_2 are user-specified coefficients between 0 and 2, which could be the subject of fine tuning of the algorithm. Parameters $r_1,$ and r_2 are random numbers between 0 and 1 that are regenerated in each iteration to add stochasticity to the process. The $\hat{x}_{t}^{i}$ is the particle best solution up to iteration t, and g_t is the global (swarm) best solution so far. After the velocity for each particle was calculated and checked to be within the boundaries of acceptable velocity, each particle's position was calculated by the equation below and updated consequently:

$$x_{t+1}^{i} = x_{t}^{i} + v_{t+1}^{i}$$

In which x_{t+1}^{i} Is the particle *i*'s position at iteration *t+1*, x_{t}^{i}, is the particle *i*'s position at iteration *t* and v_{t+1}^{i} is the particle *i*'s velocity at iteration *t+1*.

5.5 Fitness Function:

The main core of any iterative learner system is the fitness function. It can be called loss function or cost function, depending on the discipline in which it has been used. Successful application of relative absolute error in previous studies (Shahinfar et al., 2014b) suggests its use as the fitness function because of two additional reasons. First, it measures absolute error, which is not affected by outliers, and second, it considers the relative magnitude of the error as compared with the prediction. Its form is shown below:

$$f = ARG_{\min}(Relative\ Absolute\ Error) = ARG_{min}(\textstyle\sum_{i=1}^{n} \frac{|p_i - a_i|}{|a_i - \bar{a}|})$$

where p_i is predicted value, a_i is actual value, and $\bar{a}$ is the prediction by an arbitrary predictor, in this case the average of actual values (Witten and Frank, 2005). In general, it is always easier to maximize a fitness function, therefore the minimization problem given above can be transformed into a maximization problem as:

$$Fitness = \frac{\theta}{f}$$

in which θ is an amplifier constant.

5.6 Rule coverage and accuracy:

A rule-based classifier uses a set of IF-THEN rules to classify an object. An IF-THEN rule is an expression of the form:

IF: CONDITION THEN: CONSEQUENCE.

For example:

IF SNP_1 is AB, and SNP_2 is AA, and….., and SNP_n is AB THEN DPR is High.

The IF part or the rule antecedent represents the precondition which must be checked from the input and the THEN part is the conclusion upon antecedent. Usually in the rule antecedent, one or more condition exists that is logically ANDed, while the rule consequence is a class prediction. If the condition in a rule antecedent holds for a given input entry, it can be said that the rule antecedent is satisfied and that the rule covers the entry (Han, 2012). Consider two rules, where R1 correctly classifies 38 of the 40 entries it covers, but R2 covers only two entries and classifies both of them correctly. According to this logic, they have 95 and 100% accuracy, respectively. However, although R2 has higher accuracy, it covers only 2 entries, and therefore in rule assessment it is necessary to consider both coverage and accuracy. In the present study, an optimum set of rules was obtained from the training set analyses and assessed by coverage and accuracy jointly:

$$Coverage\ (R) = \frac{N_{covered}}{N_{tested}}$$

$$Accuracy\ (R) = \frac{N_{correct}}{N_{covered}}$$

Where N_{tested} is the total number of testing set entries, $N_{covered}$ is the number of covered entries, and $N_{correct}$ is the number of correctly classified entries among the covered rules. The criteria for coverage, assuming independence between the 100 SNPs in the model, was the percentage of similar SNPs in each test individual with the specific optimum rule being assessed at the time.

6. Results and Discussion:

The coverage and accuracy of each optimal rule were assessed by 6 criteria; accuracies of rules were calculated if that rule could cover 30, 40, 50, 60, 70, or 80% of the SNPs of each test individual. Then, each rule was allowed to classify all entries of the test seting that it covered. The average coverage and accuracy under the aforementioned criteria are shown in Table 5.2 With more conservative criteria, coverage decreased exponentially. In this table, $N_{covered}$% means the percentage of the testing set which the optimum rule under investigation covered, and $N_{correct}$% is the percentage of correctly classified rules that this rule covers, averaged over all optimum rules. No rules satisfied 70% coverage criteria or more. The percentage of $N_{covered}$ was identical between Mamdani and Sugeno, basically because they used the same criteria for calculating $N_{covered}$. However, $N_{correct}$ was totally different and was close to zero in all cases. This may have occurred, because we could not find a proper fitness function. By increasing the coverage constraint, $N_{covered}$ and $N_{correct}$ decreased from 99 and 85% to 93 and 84%, respectively, for Mamdani FIS.

Comparing the performance of the hybrid introduced model with RF, OLS, SVM, GBLUP-100 (GBLUP with only 100 selected SNPs), and GBLUP-17130 (GBLUP with all 17,310 SNPs), Figure 5.5 shows that the hybrid genetic-swarm FIS applied herein had flat performance and predicted the same output value for all individuals in the testing set. Surprisingly, RF, OLS, SVM, GBLUP-100 and GBLUP-17130 were able to catch the trend in DPR despite the low heritability for the trait. The reason for the flat behavior of the genetic-swarm FIS could be the high dimensionality of the rule set, an inappropriate fitness function or a suboptimal defuzzification method. The optimum performance of the genetic-swarm FIS was achieved at generation 100, with population size of 50, a maximum of 500 rules, 70% probability of crossing over, 5% probability of mutation, 70% probability of gene swap, $W_{initial}$ of 0.4, W_{final} of 0.2, C_1 of 1.5, C_2 of 1.5, and C_3 of 0.3.

7. Conclusions

Genetic algorithm and particle swarm optimization were implemented jointly for knowledge acquisition in a hybrid genetic-swarm fuzzy inference system. The algorithm introduced in this paper used a genetic algorithm to optimize rule sets and particle swarm optimization to optimize membership functions, in a sequential manner. Unfortunately, the attempt to develop for a fuzzy inference system for prediction of genetic merit of dairy bulls for daughter fertility based on a selected subset of SNPs failed due to flat behavior of the rule sets and subsequent classifications. Classification performance might be improved by further tuning of model parameters, and if successful the approach developed herein can be used as a decision support system for selection of animals and aiding management decisions.

8. 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. Data for this study were provided by the USDA-ARS Animal Genomics and Improvement Laboratory (Beltsville, MD). Kent Weigel acknowledges partial financial support from the National Association of Animal Breeders (Columbia, MO).

Table 5.1. Description of daughter pregnancy rate phenotypes used in the present study.

	N	Mean	SD	Min	Max
Training	7,000	-0.45	1.87	-9.4	7.1
Testing	1,010	-0.11	1.80	-5.9	5.2
Total	8,010	-0.40	1.86	-9.4	7.1

Table 5.2. Coverage and accuracies of Mamdani and Sugeno fuzzy inference systems according to different criteria for coverage.

	Coverage (%)											
	30		40		50		60		70		80	
	Ncover%	Ncorrect%	Ncover%	Ncorrect%	Ncover%	Ncorrect%	Ncover%	Ncorrect%	Ncover%	Ncorrect%	Ncover%	Ncorrect%
Mamdani	99.98	84.51	93.87	84.49	33.46	84.24	0.93	84.04	0.00	0.00	0.00	0.00
Sugeno	99.98	0.00	93.79	0.00	33.47	0.00	0.93	0.00	0.00	0.00	0.00	0.00

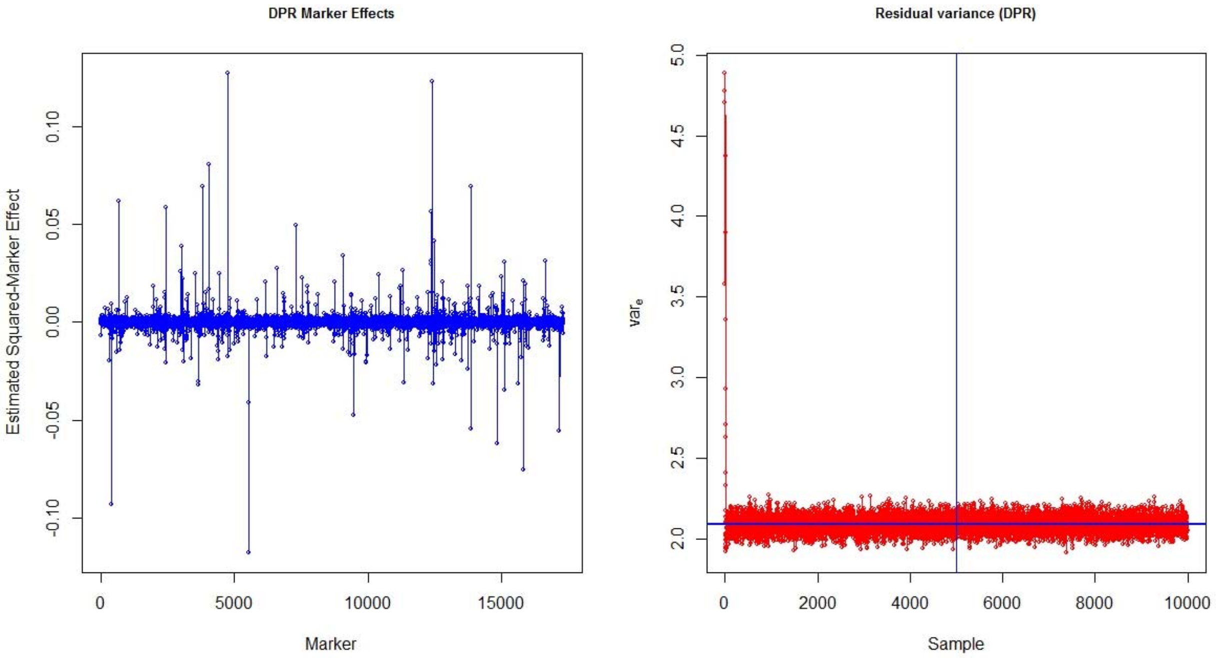

Figure 5.1. Estimated squared marker effects and residual variances for daughter pregnancy rate (DPR), as computed using a Bayes B algorithm in the training set.

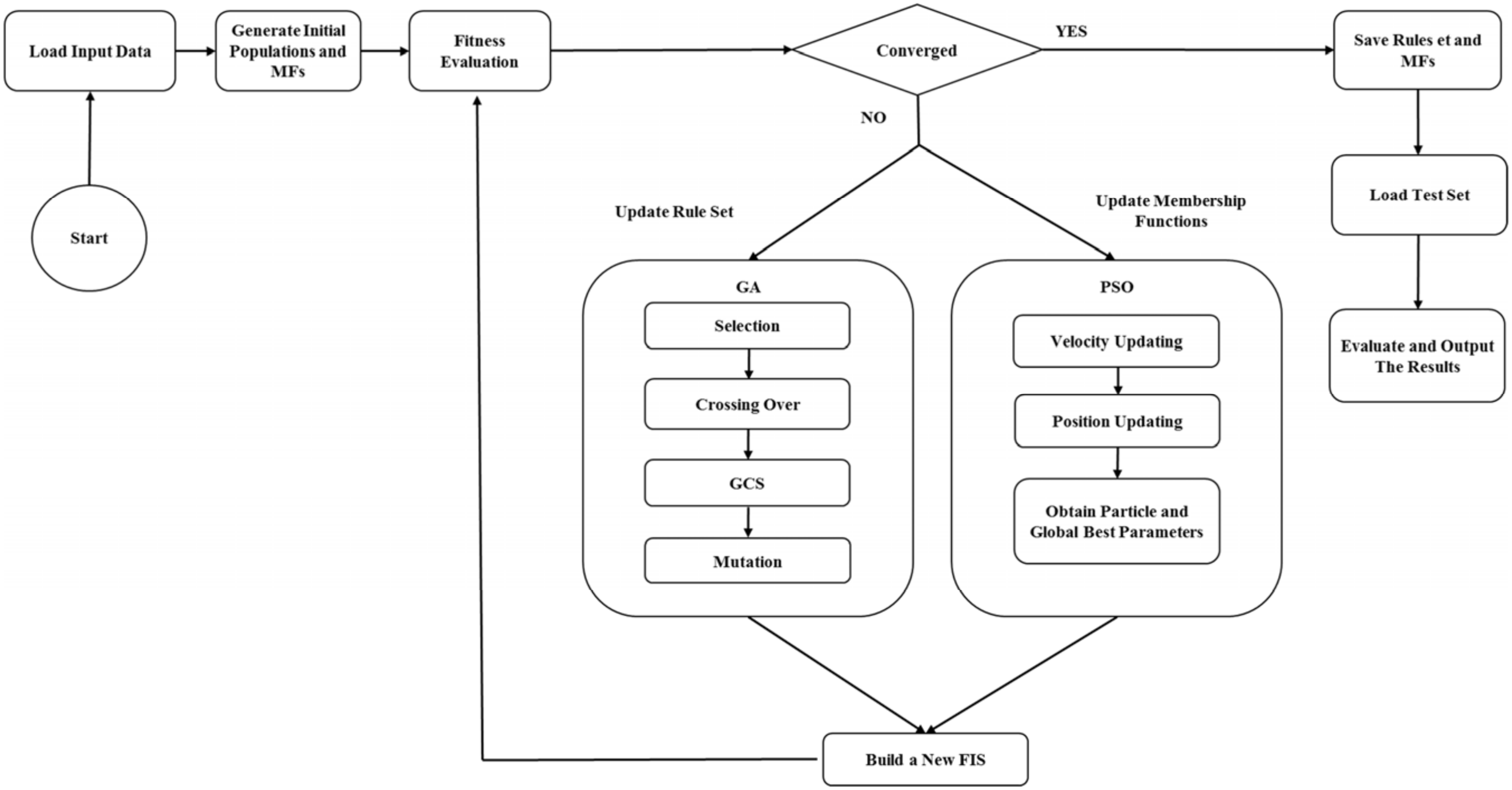

Figure 5.2. Flowchart of the genetic-swarm fuzzy expert system developed in the present study.

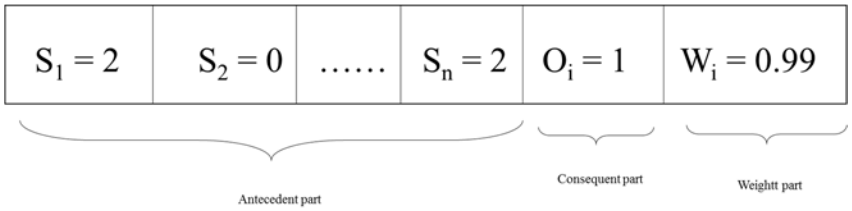

Figure 5.3. Schematic representation of a single rule, where S_1, S_2, .. . S_n refer to specific SNPs, O_i represents the output for a given observation, and W_i represents the corresponding weight.

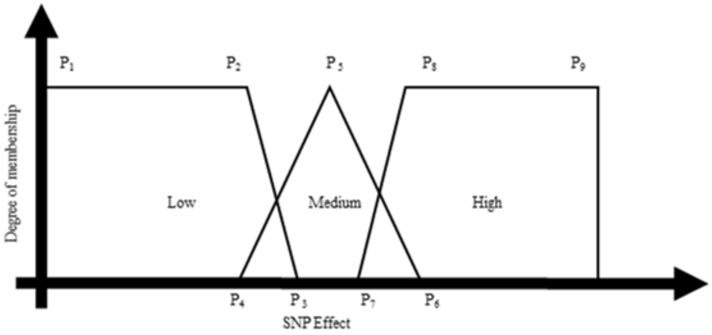

Figure 5.4. Schematic representation of membership functions, as used in the present study, where P_1, P_2, . . ., P_n represent boundary points.

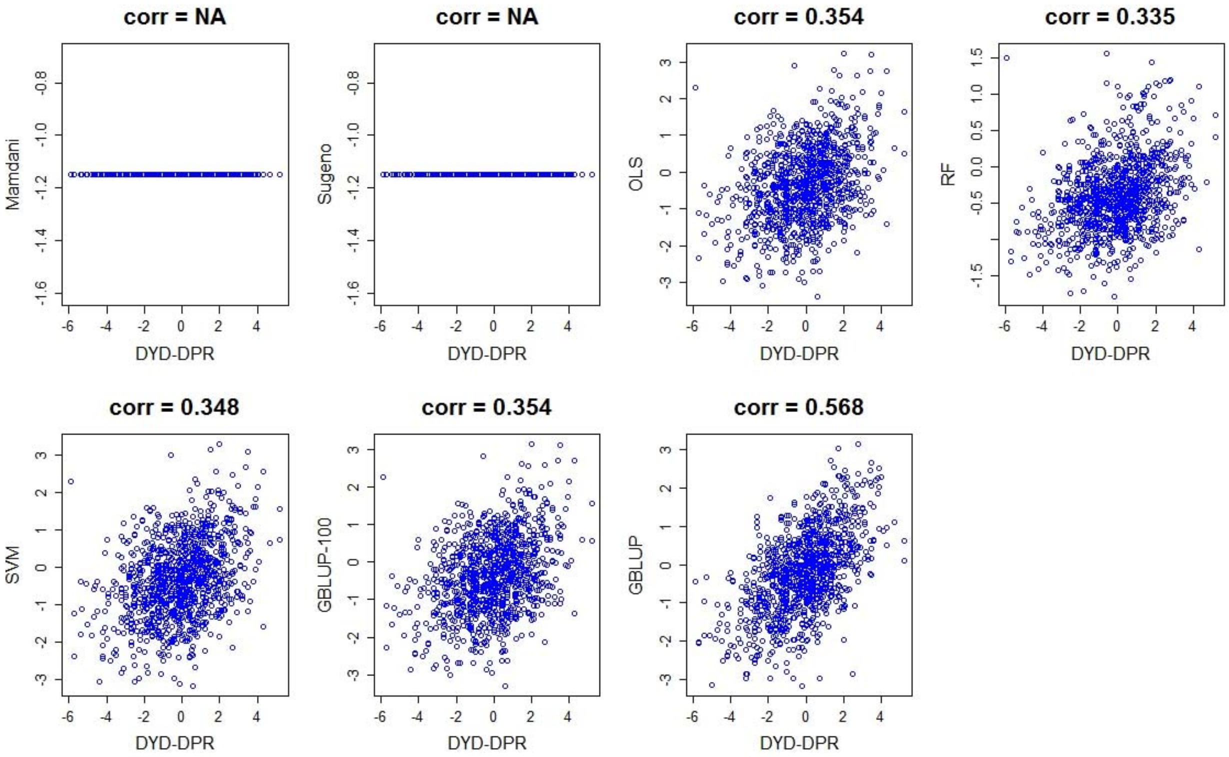

Figure 5.5. Actual vs. predicted phenotypes for 1,010 test individuals using the Mamdani genetic-swarm fuzzy inference system, Sugeno genetic-swarm fuzzy inference system, random forest (RF), ordinary least squares (OLS), support vector machine (SVM), genomic BLUP with 100 SNPs (GBLUP-100), and genomic BLUP with 17,130 SNPs (GBLUP) to predict daughter yield deviation (DYD) for daughter pregnancy rate (DPR) of Holstein sires.

Chapter 6

Short communication: Prediction of Retention Pay-off using a Machine Learning Algorithm

1. ABSTRACT

Replacement decisions have a major impact on dairy farm profitability. Dynamic programming (DP) has been widely studied to find the optimal replacement policies in dairy cattle. However, DP models are computationally intensive and might not be practical for daily decision-making. Hence, the ability of applying machine learning on a pre-run DP model to provide fast and accurate predictions of non-linear and inter-correlated variables makes it an ideal methodology. Milk class (1 to 5), lactation number (1 to 9), month in milk (1 to 20), and month of pregnancy (0 to 9) were used to describe all cows in a herd in a DP model. Twenty-seven scenarios based on all combinations of 3 levels (base, 20% above, and 20% below) of milk production, milk price, and replacement cost were solved with the DP model, resulting in a dataset of 122,716 records, each with a calculated retention pay-off (RPO). Then, a machine learning model tree (MT) algorithm was used to mimic the evaluated RPO with DP. The correlation coefficient factor was used to observe the concordance of RPO evaluated by DP and RPO predicted by the model tree. The obtained correlation coefficient was 0.991, with a corresponding value of 0.11 relative absolute error. At least 100 instances were required per model constraint resulting in 204 total equations (models). When these models were used for binary classification of positive and negative RPO, error rates were 1% false negatives and 9% false positives. Applying this trained model from simulated data for prediction of RPO for 102 actual replacement records from UW-Madison dairy herd resulted in a 0.994 correlation with 0.10 relative absolute error rate. Overall results showed that MT has a potential to be used in conjunction with DP to assist farmers in their replacement decisions.

2. INTRODUCTION

Assessing the value of cows in a herd is an important step for making replacement and other critical decisions such as individual cow breeding or treatment (Cabrera, 2012). Different methods have been developed and used to evaluate cow values in dairy herds, including dynamic stochastic simulation and optimization (Kalantari and Cabrera, 2012). Dynamic programming (**DP**), which is an optimization technique, has been widely studied over the past several decades (van Arendonk, 1985; de Vries, 2004). Dynamic programming promises the optimal replacement policy, considering several biological and economic factors (Kalantari and Cabrera, 2012), by evaluating the value of keeping or replacing individual cows. These values are used for calculating retention pay-off values (**RPO**), which can be interpreted as the expected profit from keeping the cow compared with immediate replacement. This value could be used to rank animals and help farmers to make their final replacement decisions (De Vries, 2004). RPO of a cow is changing throughout the lactation and it follows a curve similar to a lactation curve. Magnitude of RPO is mainly determined by input parameters of the model including replacement cost, carcass price, and milk production level of the herd (Kalantari et al., 2010). A DP model can be described by three main characteristics first stage length, which is the time between 2 consecutive decisions; second state which is used to describe a cow in the model; and third transition probabilities, which are probabilities that are used to weight the values in the model. In this study, stage length was set at 1-mo, and 3 state variables were used to describe the cow in the model. These variables included milk class (1 to 5), lactation number (l to 9), month in milk (1 to 20) and pregnancy status (0 (non-pregnant) or 1 to 9). Probability of pregnancy, abortion, and involuntary culling were used as transition probabilities. The full details of the DP model and its formulation are explained in Kalantari and Cabrera (2012).

Three different levels of milk production, milk price, and replacement costs, for the base and 20% above and 20% below the base, were used to create an extensive data set to estimate different ranges of RPO. The total number of records in this data set was 122,716. These 3 variables were used, because different studies have shown that these variables have the greatest impact on replacement decisions (van Arendonk, 1985; Kalantari et al., 2010). Despite extensive research on DP models and their acknowledged value in finding optimal replacement decisions, this method has not been widely applied for decision-making on commercial farms. One reason could be attributed to DP's complexity, which makes it difficult to both implement and conceptualize (Burt, 1982). The other reason that could affect on-farm decision making, is that DP is not a user friendly model and needs many input parameters. However it seems that farmers have not enough trust to use black-boxed concepts such as "CowVal" provided by DairyComp305, a commercial dairy farm management software, to use it practically in their culling decisions. Thus, the ability of model tree (MT) algorithm to provide fast and accurate predictions of non-linear and inter-correlated variables makes it an ideal methodology to be used together with a DP model in order to mimic the behavior of DP and give a clear interpretation of models and results. In addition, the MT model would abstract the complexity of DP model, and it reduces the number of input parameters that could easily be available from any on-farm database management software. Model trees are a type of decision trees used for numeric prediction. They are similar to decision trees because they use a divide and conquer approach to partition the multi-dimensional state space of the problem (Quinlan, 1992). The MT predicts the values for test instances by the linear model (LM) stored in each leaf based on the attributes of instances that reach that specific leaf (Figure 6.1). The MT and alternating decision trees have been used in assessment of reproductive performance of dairy herds (Caraviello et al., 2006; Schefers et al., 2009).

In this study, the authors developed a user-friendly tool to predict RPO values based on MT, in which the DP remains as a computational core to calculate RPO whenever the input parameters change. Then, the MT can be re-trained based on the new RPO values and used as a terminal predictor of RPO on the farm. Therefore, a Java program was developed using Weka API (Hall et al., 2009), to implement the MT model. An executable version of the program is available online at the University of Wisconsin-Madison Dairy Management Website (DairyMGT.info) which can be used to evaluate animals on commercial farms.

The aforementioned data set from the DP model was partitioned randomly, with two-thirds as a training set and one-third as a testing set for MT. Furthermore, in order to test the trained MT model in a practical situation, 102 records from voluntary culls were extracted from the Allenstein Dairy herd at the UW-Madison Integrated dairy Facility(Madison, Arlington, and Marshfield, WI), and these were used to perform an independent test for MT.

Relative absolute error was used as the error criterion to assess performance of the MT because of 2 main reasons. First, it measures absolute error, which is not affected by outliers. Second, it considers the relative magnitude of the error compared with the prediction.

$$Relative\ Absolute\ Error = \sum_{i=1}^{n} \frac{|p_i - a_i|}{|a_i - \bar{a}|}$$

Where p_i is predicted value, a_i is actual value and $\bar{a}$ is the prediction by an arbitrary predictor, in this case the average of actual values (Witten and Frank, 2005).

In developing a MT there is always two conflicting constraints, number of rules (number of LM stored in the leaves of the tree) and number of instances per LM which is the number of cases (records) that will reach an arbitrary LM. A trade off exist between these two constraints. Increasing the number of rules is equal to increase in accuracy of the model on training set but, decrease in number of instances per LM and therefore loss of generality of the model and poor performance on the new unseen datasets. Hence finding an optimum point in between these two constraints are critical. Performance of MT in terms of number of rules, correlation coefficient, and relative absolute error with respect to the minimum numbers of instances per linear model as an indicating constraint are shown in Table 6.1. There was a noticeable trend in which increasing the minimum numbers of instances per linear model resulted in a decrease in the number of rules and correlation coefficient, but an increase in the error criterion. Predicted (MT) vs. actual (DP) RPO for corresponding 6 scenarios from Table 6.1 are graphically presented in Figure 6.2.

Considering the results from Table 6.1, it seems that a minimum of 100 instances per LM would be a logical constraint because the number of rules dropped considerably to one-tenth in comparison with one instance per model (2,447 vs. 204); and, at the same time, the correlation coefficient and relative absolute error did not change substantially.

Actual and predicted RPOs divided into 10 equal intervals called deciles. Distributions of predicted vs. actual deciles of RPO are presented in Table 6.2. Numbers on the diagonal show the instances that MT predicted the RPO in the same decile as DP. Numbers above the diagonal show the cases of RPO that were cases of RPO that were underestimated by MT, and numbers below the diagonal are the cases that were overestimated. In general, the MT resulted in 76.5% correctly predicted RPOs, 12% overestimated RPOs and 11.5% underestimated RPOs.

Since the sign of RPO is more important than its magnitude (Groenendaal et al, 2004), a binary evaluation of RPO prediction by MT was considered. The confusion matrix in the case of binary classification of RPO is presented in Table 6.3. Classification of RPO into positive and negative cases showed an error rate of 1% false negative (FN) cases and 9% false positive (FP) cases, which indicates the advantage of the method in identifying negative cases. Here, FN are the terminal mistakes in which cows that are predicted to have a negative RPO are culled from the herd erroneously, whereas FP are mistakes that can be corrected by identifying the animal as a cull in the next replacement decision round . However the cost of FN and FP will depend on the magnitude of the RPO in which if a specific FN case has a positive RPO just a little above the zero the related cost would be zero and vice versa.

The RPO prediction using MT is an efficient method for assessing the value of keeping or culling specific cows and helping farmers make economical culling decisions. A correlation coefficient of 0.994 with 0.10 relative absolute error rate was obtained by applying the trained model for prediction to 102 actual voluntary culling records from the UW-Madison dairy herd, which indicates its accuracy and relevance in real-life replacement decisions.

Table 6.1. Performance of MT[1] in prediction of RPO[2] with different constraints (minimum numbers of instances per LM[3]). Suggested model constraint and its performance are in bold.

Minimum Numbers of Instances per LM	Number of Rules	Correlation Coefficient	Relative Absolute Error %
1	2447	0.998	5.08
100	**204**	**0.991**	**10.93**
200	152	0.985	14.15
400	89	0.976	17.91
800	47	0.966	22.34
1600	28	0.936	30.60

[1] Model tree
[2] Retention Pay-off
[3] Linear model

Table 6.2. Predicted (MT[1]) vs. actual (DP[2]) deciles of RPO[3] for 41,723 test cases. Numbers on the diagonal show the instances that MT predicted the RPO in the same decile as DP predicted. Above the diagonal show the cases of RPOs that were underestimated by MT, and below the diagonal are the cases that were overestimated by MT in compare with DP.

		Actual Deciles									
		1	2	3	4	5	6	7	8	9	10
Predicted Deciles	1	**3,776**	375	21	0	0	0	0	0	0	0
	2	385	**3,257**	491	34	5	0	0	0	0	0
	3	10	528	**3,057**	535	30	10	2	0	0	0
	4	1	12	584	**2,893**	650	31	1	0	0	0
	5	0	0	18	689	**2,816**	636	11	2	0	0
	6	0	0	1	20	653	**2,891**	576	30	1	0
	7	0	0	0	1	17	576	**3,004**	545	29	0
	8	0	0	0	0	1	23	550	**3,101**	487	10
	9	0	0	0	0	0	5	27	485	**3,316**	339
	10	0	0	0	0	0	0	1	9	339	**3,826**

[1] Model tree

[2] Dynamic programming

[3] Retention Pay-off

Table 6.3. Confusion matrix for percentage of binary prediction of RPO[1] for 41,723 test cases.

		Actual (Dynamic Programming)	
		Positive	Negative
Predicted (Model Tree)	Positive	0.91	0.09
	Negative	0.01	0.99

[1] Retention Pay-off

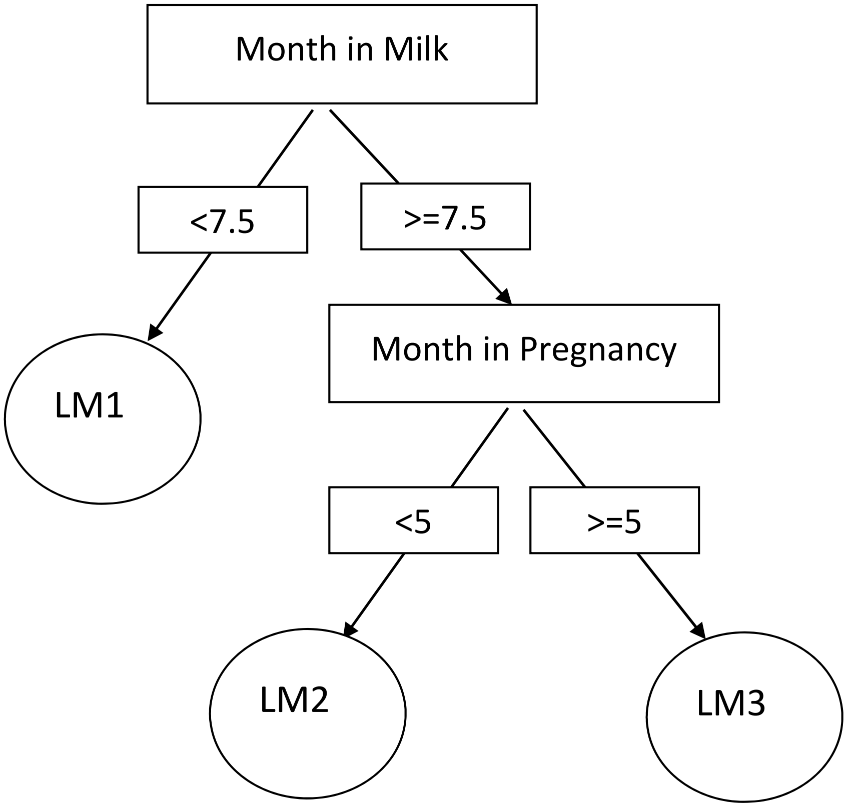

Figure 6.1: Schematic illustration of Model Tree algorithm with two explanatory variables (month in milk and month in pregnancy) and three linear models in leaves. For example if month in milk is <7.5 the value of retention pay-off (RPO) can be predicted by linear model 1 (LM1), but if it is >=7.5 then month in pregnancy need to be considered too. If month in pregnancy is <5, linear model 2 (LM2) and if month in pregnancy is >=7.5 linear model 3 (LM3) will be used. An example of LM2 can be RPO = + 0.1025 * month in milk + 2.5118 * month in pregnancy + 80.5674.

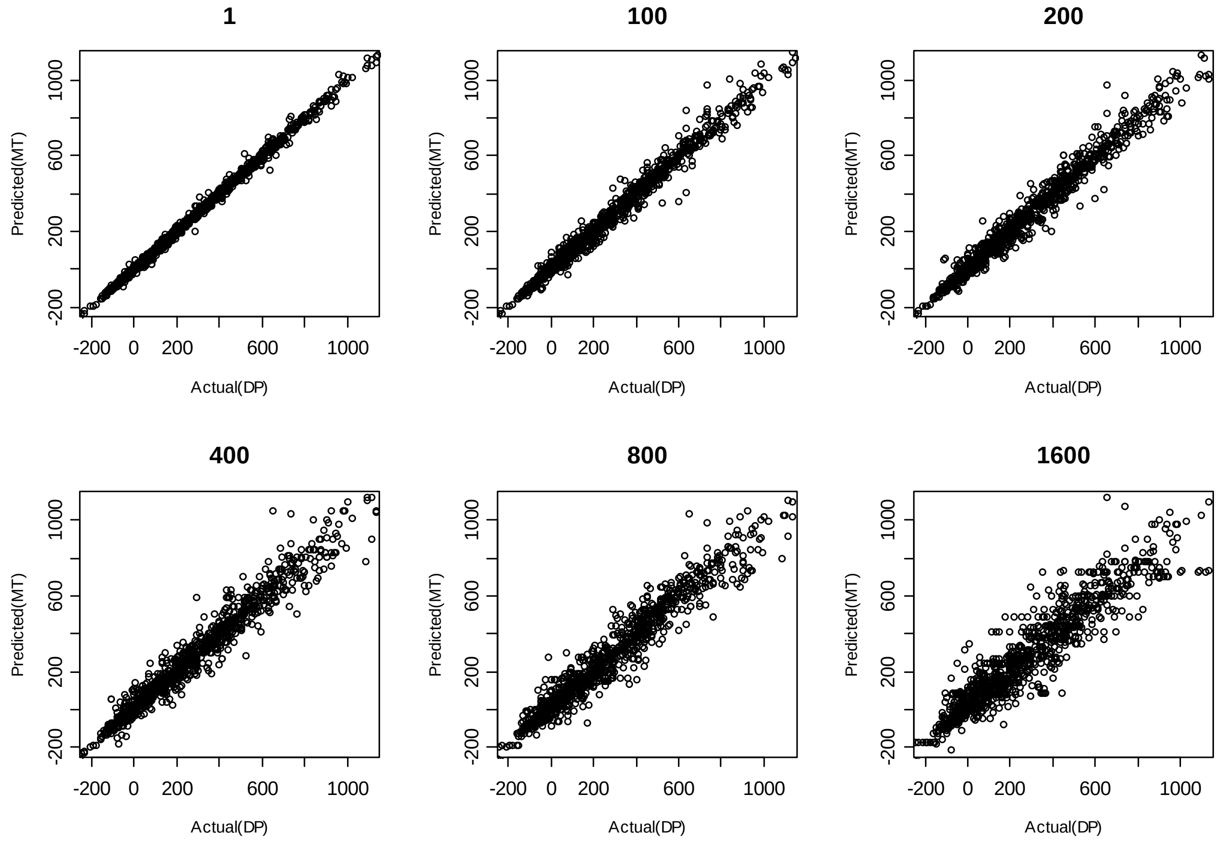

Figure 6.2. Actual (predicted by dynamic programming; DP) vs. Predicted (predicted by machine learning model tree; MT) values of retention pay-off for 6 different scenarios with the test set showed in Table 6.1.

Chapter 7

Overall Conclusions and Suggestions

The objectives of first chapter of this study were to find the best performing models to predict the outcomes of insemination events using a variety of explanatory variables. The advantages of machine learning methods for predicting the outcome of insemination events with inter-correlated and often missing explanatory variables were demonstrated.

The machine learning algorithms considered in this study (Naïve Bayes, Bayesian network, decision tree, bagging, and random forest) were effective in predicting pregnant versus non-pregnant cows at the time of insemination. Among the algorithms considered in this research, RF was significantly better in terms of classification accuracy (72.3% and 73.6% for primiparous and multiparous cows, respectively) and area under the receiver operating characteristic (ROC) curve (75.6% and 73.6%, respectively).

Evaluation of features by gain ratio showed that the incidence of ketosis, mastitis, retained placenta, and lameness for primiparous cows were among the most important explanatory variables, whereas the incidence of mastitis, displaced abomasum, and retained placenta were among the most important explanatory variables for multiparous cows. Overall, results of this study suggest that, although prediction of the insemination outcomes 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.

When semen is inexpensive and technician costs are modest, the common practice is to simply inseminate all eligible cows and ignore the costs associated with breeding lowly fertile cows with a history of health and reproductive problems. Therefore, using the best prediction model for the outcome of an insemination event along with cost-sensitive evaluation and lift chart analysis showed the economic differences between in various breeding scenarios. Decision support tools developed herein 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 utilize gender-enhanced semen or expensive semen from high merit sires. Lift chart analysis can be used to stratify the pool of eligible cows into those that are likely to conceive and those that should not be inseminated at a given point in time. When coupled with cost-sensitive evaluation of misclassified and correctly classified insemination events, the approach provides a potentially powerful tool optimizing the reproductive management strategy of a given farm under that farm's economic conditions. Implementation of such a decision-making tool would likely require the assistance of a professional reproductive management consultant, but perhaps key features of this approach could be incorporated into existing dairy herd management software programs in the future. The program developed here is ready for commercial use and is going to be available to farmers through a farm management software, which is under developement.

The third part of this dissertation focused on genomic prediction of the merit of dairy sires for daughter pregnancy rate, proposing a novel hybrid fuzzy inference system. Although the results of this portion of the study were unfavorable, further tuning of the parameters and investigation of the reasons for poor model performance are needed to achieve acceptable prediction accuracy.

The fourth part of this study was focused on developing a user-friendly tool to predict RPO values based on MT, in which dynamic programming remains as a computational core to calculate RPO whenever the input parameters change. Then, the MT can be re-trained based on the new RPO values and used as a terminal predictor of RPO on the farm. Therefore, a Java program was developed using Weka API (Hall et al., 2009), to implement the MT model. An executable version of the program is available online at the University of Wisconsin-Madison Dairy Management Website (DairyMGT.info), which can be used to evaluate animals on commercial farms.

In general, a limitation of this study was the shortage of high quality health records on farms for use in the predictive model of insemination outcomes. In addition, very few studies have investigated the cost and benefits associated with pregnancy, abortion, insemination and, the most challenging, the forgone pregnancy. These values are critical to make optimum decisions regarding breeding policies on commercial farms. Therefore, more research on these issues is needed in order to create a successful and realistic decision making tool. Moreover, because of fluctuating costs and profits associated with reproduction, having a fast and cost effective method to estimate these values is a necessary and worthy investment.

Understanding the cause-effect relationship between explanatory variables associated with reproductive performance in general, and specifically the outcomes of insemination events, can enhance model performance for prediction of those traits. It is likely that the use of an influence diagram, Bayesian networks or structural equation models may be helpful in this regard.

Lastly, considerable uncertainty exists regarding the state of nature with regard to additive, dominance, and epistatic effects of genes affecting complex traits such as female fertility. Methods of inference in such an uncertain space must be compatible with the problem. Fuzzy logic has proven itself in handling imprecision and nonlinearity in many industrial and medical applications. However, utilization of its potential in animal breeding and genetics has not been extensively explored, and as one of the first attempts to utilize fuzzy inference systems in genomic prediction, our results were not appealing. Additional work is needed to determine if such systems have a place in animal breeding and dairy farm management.

References:

Adamczyk, K., K. Molenda, J. Szarek, and G. Skrzynski. 2005. Prediction of bulls' slaughter value from growth data using artificial neural network. *J Cent Euro Agric 6133–142*.

Andersson, L. 1988. Subclinical ketosis in dairy cows. *Vet. Clin. North Am. Food Anim. Pract.* 4:233–251.

Bahl, V., S. Lin, N. Xu, B. Davis, Y. Wang, and P. Talbot. 2012. Comparison of electronic cigarette refill fluid cytotoxicity using embryonic and adult models. *Reprod. Toxicol.* 34:529–537. doi:10.1016/j.reprotox.2012.08.001.

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.

Blondin, J. 2009. Particle Swarm Optimization: A tutorial (September 4, 2009).

Breiman, L. 1996. Bagging predictors. *Mach. Learn. 24*. 24.

Breiman, L. 2001. Random Forests. *Mach. Learn. 45*. 45.

Brieman L. 1994. Bagging predictors. Department of Statistics, University of California Berkeley, CA, USA.

Britt, J.H. 1985. Enhanced reproduction and its economic implications. *J Dairy Sci*. 68:1585–1592.

Browning, B.L., and Z. Yu. 2009. Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. *Am. J. Hum. Genet.* 85:847–861. doi:10.1016/j.ajhg.2009.11.004.

Burt, O. R. 1982. Dynamic programming: Has its day arrived? *West J Agric Econ*. 7:381:394.

Butler, W.R., and R.D. Smith. 1989. Interrelationships Between Energy Balance and Postpartum Reproductive Function in Dairy Cattle. *J. Dairy Sci.* 72:767–783. doi:10.3168/jds.S0022-0302(89)79169-4.

Cabrera, V.E. 2012. A simple formulation and solution to the replacement problem: A practical tool to assess the economic cow value, the value of a new pregnancy, and the cost of a pregnancy loss. *J Dairy Sci*. 95:4683–4698.

Cabrera, V.E. 2014. Economics of fertility in high-yielding dairy cows on confined TMR systems. *animal*. 8:211–221. doi:10.1017/S1751731114000512.

Caraviello, D.Z., K.A. Weigel, P.M. Fricke, M.C. Wiltbank, M.J. Florent, N.B. Cook, K.V. Nordlund, N.R. Zwald, and C.L. Rawson. 2006. Survey of Management Practices on

Reproductive Performance of Dairy Cattle on Large US Commercial Farms. *J. Dairy Sci.* 89:4723–4735. doi:10.3168/jds.S0022-0302(06)72522-X.

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* in Proc. 9th European Conference on Artificial Intelligence. Stockholm, Sweden.

Chapinal, N., M.A.G. von Keyserlingk, R.L.A. Cerri, K. Ito, S.J. LeBlanc, and D.M. Weary. 2013. Short communication: Herd-level reproductive performance and its relationship with lameness and leg injuries in freestall dairy herds in the northeastern United States. *J. Dairy Sci.* 96:7066–7072. doi:10.3168/jds.2013-6967.

Chebel, R.C., J.E.P. Santos, J.P. Reynolds, R.L.A. Cerri, S.O. Juchem, and M. Overton. 2004. Factors affecting conception rate after artificial insemination and pregnancy loss in lactating dairy cows. *Anim. Reprod. Sci.* 84:239–255. doi:10.1016/j.anireprosci.2003.12.012.

Chen, S.-M., and F.-M. Tsai. 2008. Generating fuzzy rules from training instances for fuzzy classification systems. *Expert Syst. Appl.* 35:611–621. doi:10.1016/j.eswa.2007.07.013.

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 892473–2478.*

De los Campos, G., D. Gianola, G.J.M. Rosa, K.A. Weigel, and J. Crossa. 2010. genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet Res.* 92:295–308.

De los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P.L. Calus. 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics.* 193:327–345. doi:10.1534/genetics.112.143313.

De los Campos, G., and P. Perez Rodriguez. 2012. BGLR R-Package.

De Jarnette, 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 943477–3483.*

De Vries, A. 2004. Economics of Delayed Replacement When Cow Performance is Seasonal*. *J. Dairy Sci.* 87:2947–2958. doi:10.3168/jds.S0022-0302(04)73426-8.

De Vries, A. 2006. Economic Value of Pregnancy in Dairy Cattle1. *J. Dairy Sci.* 89:3876–3885. doi:10.3168/jds.S0022-0302(06)72430-4.

De Vries, A., and C.A. Risco. 2005. Trends and seasonality of reproductive performance in Florida and Georgia dairy herds from 1976 to 2002. *J. Dairy Sci.* 88:3155–3165.

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. doi:10.3168/jds.S0022-0302(00)74856-9.

Domecq, J.J., A.L. Skidmore, J.W. Lloyd, and J.B. Kaneene. 1997. Relationship Between Body Condition Scores and Conception at First Artificial Insemination in a Large Dairy Herd of High Yielding Holstein Cows1. *J. Dairy Sci.* 80:113–120. doi:10.3168/jds.S0022-0302(97)75918-6.

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

Duffield, T.F., K.D. Lissemore, B.W. McBride, and K.E. Leslie. 2009. Impact of hyperketonemia in early lactation dairy cows on health and production. *J. Dairy Sci.* 92:571–580. doi:10.3168/jds.2008-1507.

Elkan, C. 2001. The Foundations of Cost-Sensitive Learning. Morgan Kaufmann, Seattle, WA, USA. 973–978.

Elkjær, K., R. Labouriau, M.-L. Ancker, H. Gustafsson, and H. Callesen. 2013. Short communication: Large-scale study on effects of metritis on reproduction in Danish Holstein cows. *J. Dairy Sci.* 96:372–377. doi:10.3168/jds.2012-5584.

Ferguson, J.D., and D.T. Galligan. 2011. The value of pregnancy diagnosis – a revisit to an old art. *In* PTheriogenology Annual Conference Symposium, August 9–13. Milwaukee, USA.

Földi, J., M. Kulcsár, A. Pécsi, B. Huyghe, C. de Sa, J.A.C.M. Lohuis, P. Cox, and G. Huszenicza. 2006. Bacterial complications of postpartum uterine involution in cattle. *Anim. Reprod. Sci.* 96:265–281. doi:10.1016/j.anireprosci.2006.08.006.

Fourichon, C., H. Seegers, and X. Malher. 2000. Effect of disease on reproduction in the dairy cow: a meta-analysis. *Theriogenology*. 53:1729–1759. doi:10.1016/S0093-691X(00)00311-3.

Galvao, K.N., P. Federico, De Vries, and G. Schuenemann. 2013. Economic comparison of reproductive programs for dairy herds using estrus detection, timed artificial insemination, or a combination. *J. Dairy Sci.* 96:2681–2693.

Ganesh Kumar, P., T. Aruldoss Albert Victoire, P. Renukadevi, and D. Devaraj. 2012. Design of fuzzy expert system for microarray data classification using a novel Genetic Swarm Algorithm. *Expert Syst. Appl.* 39:1811–1821. doi:10.1016/j.eswa.2011.08.069.

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 874123–4131*.

Garverick, H.A., M.N. Harris, R. Vogel-Bluel, J.D. Sampson, J. Bader, W.R. Lamberson, J.N. Spain, M.C. Lucy, and R.S. Youngquist. 2013. Concentrations of nonesterified fatty acids and glucose in blood of periparturient dairy cows are indicative of pregnancy success at first insemination. *J Dairy Sci*. 96:181–188.

Gianola, D., G. de los Campos, W.G. Hill, E. Manfredi, and R. Fernando. 2009. Additive Genetic Variability and the Bayesian Alphabet. *Genetics*. 183:347–363. doi:10.1534/genetics.109.103952.

Gianola, D., R.L. Fernando, and A. Stella. 2006. Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. *Genetics*. 173:1761–1776. doi:10.1534/genetics.105.049510.

Gianola, D., H. Okut, K.A. Weigel, and G.J. Rosa. 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet.* 12:87. doi:10.1186/1471-2156-12-87.

Giordano, J.O., P.M. Fricke, M.C. Wiltbank, and V.E. Cabrera. 2011. An economic decision-making support system for selection of reproductive management programs on dairy farms. *J Dairy Sci*. 94:6216–6232.

Giordano, J.O., A.S. Kalantari, P.M. Fricke, M.C. Wiltbank, and V.E. Cabrera. 2012. A daily herd Markov-chain model to study the reproductive and economic impact of reproductive programs combining timed artificial insemination and estrus detection. *J Dairy Sci*. 95:5442–5460.

Giuliodori, M.J., R.P. Magnasco, D. Becu-Villalobos, I.M. Lacau-Mengido, C.A. Risco, and R.L. de la Sota. 2013. Metritis in dairy cows: Risk factors and reproductive performance. *J. Dairy Sci.* 96:3621–3631. doi:10.3168/jds.2012-5922.

González-Recio, O., and R. Alenda. 2005. Genetic Parameters for Female Fertility Traits and a Fertility Index in Spanish Dairy Cattle. *J. Dairy Sci.* 88:3282–3289. doi:10.3168/jds.S0022-0302(05)73011-3.

González-Recio, O., R. Alenda, Y.M. Chang, K.A. Weigel, and D. Gianola. 2006. Selection for Female Fertility Using Censored Fertility Traits and Investigation of the Relationship with Milk Production. *J. Dairy Sci.* 89:4438–4444. doi:10.3168/jds.S0022-0302(06)72492-4.

González-Recio, O., M.A. Pérez-Cabal, and R. Alenda. 2004. Economic Value of Female Fertility and its Relationship with Profit in Spanish Dairy Cattle. *J. Dairy Sci.* 87:3053–3061. doi:10.3168/jds.S0022-0302(04)73438-4.

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.* 54:69–83. doi:10.1016/j.compag.2006.08.004.

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. doi:10.1016/j.compag.2010.09.001.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*. 12:186. doi:10.1186/1471-2105-12-186.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reuteman, and I.H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor.* 11.

Hamel, L. 2008. Model Assessment with ROC Curves, The Encyclopedia of Data Warehousing and Mining, 2nd Edition.

Han, J. 2012. Data mining concepts and techniques. 3rd ed. Morgan Kaufmann Publishers, Waltham, Mass.

Herdt, T.H. 2000. Ruminant adaptation to negative energy balance Influences on the etiology of ketosis and fatty liver. *Vet Clin North Am Food Anim Pr.* 16:215–230.

Hertl, J.A., Y.T. Gröhn, J.D.G. Leach, D. Bar, G.J. Bennett, R.N. González, B.J. Rauch, F.L. Welcome, L.W. Tauer, and Y.H. Schukken. 2010. Effects of clinical mastitis caused by gram-positive and gram-negative bacteria and other organisms on the probability of conception in New York State Holstein dairy cows. *J. Dairy Sci.* 93:1551–1560. doi:10.3168/jds.2009-2599.

Ho, S.-Y., C.-H. Hsieh, H.-M. Chen, and H.-L. Huang. 2006. Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. *Biosystems*. 85:165–176. doi:10.1016/j.biosystems.2006.01.002.

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

Holland J. H. 1995. Hidden Order : How adaption Builds Complexity. Addison-Wesley.

Huang, C., S. Tsuruta, J.K. Bertrand, I. Misztal, T.J. Lawlor, and J.S. Clay. 2009. Trends for conception rate of Holsteins over time in the southeastern United States. *J. Dairy Sci.* 92:4641–4647. doi:10.3168/jds.2008-1982.

Huang, W., B.W. Kirkpatrick, G.J.M. Rosa, and H. Khatib. 2010. A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. *Anim. Genet.* 41:570–578. doi:10.1111/j.1365-2052.2010.02046.x.

Hudson, C.D., A.J. Bradley, J.E. Breen, and M.J. Green. 2012. Associations between udder health and reproductive performance in United Kingdom dairy cows. *J. Dairy Sci.* 95:3683–3697. doi:10.3168/jds.2011-4629.

J. H. Holland. 1975. Addaption in Natural and Artificial Systems. Unoversity of Michigan Press.

Jang, J.-S.R. 1993. ANFIS : Adap tive-Ne twork-Based Fuzzy Inference System. *IEEE Trans. Syst. Man Cybern.* 23:665–685.

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.

Jordan, E.R. 2003. Effects of Heat Stress on Reproduction. *J. Dairy Sci.* 86:E104–E114. doi:10.3168/jds.S0022-0302(03)74043-0.

Kalantari, A.S., and V.E. Cabrera. 2012. The effect of reproductive performance on the dairy cattle herd value assessed by integrating a daily dynamic programming model with a daily Markov chain model. *J. Dairy Sci.* 95:6160 – 6170. doi:http://dx.doi.org/10.3168/jds.2012-5587.

Kalantari, A.S., H. Mehrabani-Yeganeh, M. Moradi, A.H. Sanders, and A.D. Vries. 2010. Determining the optimum replacement policy for Holstein dairy herds in Iran. *J. Dairy Sci.* 93:2262 – 2270. doi:http://dx.doi.org/10.3168/jds.2009-2765.

Keil, F., and R.A. Wilson. 1999. The MIT Encyclopedia of the Cognitive Sciences. The MIT Press. 37 pp.

Kennedy, J., and R. Eberhart. 1995. Particle swarm optimization. IEEE Press, Piscataway, NJ. 1942–1948.

Khatib, H., R.L. Monson, W. Huang, R. Khatib, V. Schutzkus, H. Khateeb, and J.J. Parrish. 2010. Short communication: Validation of in vitro fertility genes in a Holstein bull population. *J. Dairy Sci.* 93:2244–2249. doi:10.3168/jds.2009-2805.

Khatib, H., R.L. Monson, V. Schutzkus, D.M. Kohl, G.J.M. Rosa, and J.J. Rutledge. 2008. Mutations in the STAT5A Gene Are Associated with Embryonic Survival and Milk Composition in Cattle. *J. Dairy Sci.* 91:784–793. doi:10.3168/jds.2007-0669.

Kononenko, I. 1990. Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. *In* Current Trends in Knowledge Acquisition. IOS. Press, Amsterdam, the Netherlands.

Kontkanen, P., P. Myllymaki, T. Silander, H. Tirri, S.F. Lauderdale, and FL. 1997. Comparing predictive inference methods for discrete domains. Pages 311–318. *In* in Proc. 6th International Workshop on Artificial Intelligence.

Kosko, B. 1990. Fuzziness vs. Probability. *Int. J. General Systems*. 17:211–240.

Kuhn, M.T., and J.L. Hutchison. 2008. Prediction of Dairy Bull Fertility from Field Data: Use of Multiple Services and Identification and Utilization of Factors Affecting Bull Fertility. *J. Dairy Sci.* 91:2481–2492. doi:10.3168/jds.2007-0743.

Kuhn, M.T., J.L. Hutchison, and G.R. Wiggans. 2006. Characterization of Holstein heifer fertility in the United States. *J. Dairy Sci.* 89:4907–4920.

Lacroix, R., Wade, K.M., Kok, R., and Hayes, J.F. 1995. Prediction of cow performance with a connectionist model. *Trans ASAE*. 38:1537–1579.

Langton C. 1995. Artificial Life. MIT Press, MA, USA.

LeBlanc, S. 2005. Overall Reproductive Performance of Canadian Dairy Cows: Challenges We Are Facing. *Adv. Dairy Technol.* 17:137–157.

LeBlanc, S. 2007. Economics of improving reproductive performance in dairy herds. *Adv. Dairy Technol.* 19:201–214.

Lima, F.S., A. De Vries, C.A. Risco, J.E.P. Santos, and W.W. Thatcher. 2010. Economic comparison of natural service and timed artificial insemination breeding programs in dairy cattle. *J. Dairy Sci.* 93:4404–4413. doi:10.3168/jds.2009-2789.

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. doi:10.3168/jds.2008-1029.

Löf, E., H. Gustafsson, and U. Emanuelson. 2007. Associations Between Herd Characteristics and Reproductive Efficiency in Dairy Herds. *J. Dairy Sci.* 90:4897–4907. doi:10.3168/jds.2006-819.

Löf, E., H. Gustafsson, and U. Emanuelson. 2014. Factors influencing the chance of cows being pregnant 30 days after the herd voluntary waiting period. *J. Dairy Sci.* 97:2071 – 2080. doi:http://dx.doi.org/10.3168/jds.2012-5874.

Lomander, H., C. Svensson, C. Hallén-Sandgren, H. Gustafsson, and J. Frössling. 2013. Associations between decreased fertility and management factors, claw health, and somatic cell count in Swedish dairy cows. *J. Dairy Sci.* 96:6315–6323. doi:10.3168/jds.2012-6475.

Long, N., D. Gianola, G.J. 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. doi:10.1186/1297-9686-41-18.

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

Lucy, M.C. 2001. Reproductive Loss in High-Producing Dairy Cattle: Where Will It End? *J. Dairy Sci.* 84:1277–1293. doi:10.3168/jds.S0022-0302(01)70158-0.

Ludwich, T.M., and E.R. Rader. 1967. Diagnosis of early pregnancy in cattle by ovarian analysis. *J Dairy Sci*. 51:74–77.

Mamdani, E.H. 1977. Advances in the linguistic synthesis of fuzzy controllers. *Int. J. Man-Mach. Stud.* 7:1182–1191.

Maraziotis, I.A. 2012. A semi-supervised fuzzy clustering algorithm applied to gene expression data. *Pattern Recognit.* 45:637–648. doi:10.1016/j.patcog.2011.05.007.

McArt, J. a. A., D.V. Nydam, and G.R. Oetzel. 2012. A field trial on the effect of propylene glycol on displaced abomasum, removal from herd, and reproduction in fresh cows diagnosed with subclinical ketosis. *J. Dairy Sci.* 95:2505–2512. doi:10.3168/jds.2011-4908.

McCullock, K., D.L.K. Hoag, J. Parsons, M. Lacy, G.E.S. Jr, and W. Wailes. 2013. Factors affecting the economics of using sexed semen in dairy cattle. *J Dairy Sci*. 96:6366–6377.

Meadows, C., P.J. Rajala-Schultz, and G.S. Frazer. 2005. A spreadsheet-based model demonstrating the nonuniform economic effects of varying reproductive performance in Ohio dairy herds. *J Dairy Sci*. 88:1244–1254.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*. 157:1819–1829.

Mitchell, T.M. 1997. Machine Learning. International Edition. McGraw-Hill, 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. doi:10.3168/jds.2006-574.

Mrode, R. 2005. Linear Models for the Prediction of Animal Breeding Values. 2nd edition. CABI, Wallingford, UK ; Cambridge, MA. 208 pp.

Nebel, R.L., and M.L. McGilliard. 1993. Interactions of High Milk Yield and Reproductive Performance in Dairy Cows. *J. Dairy Sci.* 76:3257–3268. doi:10.3168/jds.S0022-0302(93)77662-6.

Norman, H.D., J.R. Wright, S.M. Hubbard, R.H. Miller, and J.L. Hutchison. 2009. Reproductive status of Holstein and Jersey cows in the United States. *J. Dairy Sci.* 92:3517–3528. doi:10.3168/jds.2008-1768.

Novák, V., I. Perfilieva, and J. Močkoř. 1999. Mathematical principles of fuzzy logic. Kluwer Academic, Dodrecht.

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 914323–4332*.

Olynk, N.J., and C.A. Wolf. 2009. Stochastic economic analysis of dairy cattle artificial insemination reproductive management programs. *J. Dairy Sci.* 92:1290–1299. doi:10.3168/jds.2008-1418.

Oseni, S., I. Misztal, S. Tsuruta, and R. Rekaya. 2003. Seasonality of Days Open in US Holsteins. *J. Dairy Sci.* 86:3718–3725. doi:10.3168/jds.S0022-0302(03)73977-0.

Oseni, S., I. Misztal, S. Tsuruta, and R. Rekaya. 2004a. Genetic Components of Days Open Under Heat Stress. *J. Dairy Sci.* 87:3022–3028. doi:10.3168/jds.S0022-0302(04)73434-7.

Oseni, S., S. Tsuruta, I. Misztal, and R. Rekaya. 2004b. Genetic Parameters for Days Open and Pregnancy Rates in US Holsteins Using Different Editing Criteria. *J. Dairy Sci.* 87:4327–4333. doi:10.3168/jds.S0022-0302(04)73578-X.

Ospina, P.A., D.V. Nydam, T. Stokol, and T.R. Overton. 2010. Association between the proportion of sampled transition cows with increased nonesterified fatty acids and β-hydroxybutyrate and disease incidence, pregnancy rate, and milk production at the herd level. *J. Dairy Sci.* 93:3595–3601. doi:10.3168/jds.2010-3074.

Pal, S.K. 1999. Neuro-fuzzy pattern recognition: methods in soft computing. Wiley, New York. 375 pp.

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. doi:10.3168/jds.S0022-0302(07)71547-3.

Peñagaricano, F., K.A. Weigel, and H. Khatib. 2012. Genome-wide association study identifies candidate markers for bull fertility in Holstein dairy cattle. *Anim. Genet.* 43:65–71. doi:10.1111/j.1365-2052.2012.02350.x.

Philipsson, J., H. Jorjani, J. Jakobsen, E. Hjerpe, F. Forabosco, and F. Fikse. 2007. Breeding for health and fertility in dairy cattle.

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

Purcell, S., B. Neale, L. Thomas, F. MAR, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81.

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

Rechenberg I. 1965. Cybernetic solution path of an experimental problem.

Rouse, M. 2006. Fuzzy logic. http://whatis.techtarget.com.

Russell, S., and P. Norvig. 2002. Artificial intelligence-a modern approach. 3rd ed. Pearson, Upper Sadle River, NJ.

Schaefer, G., and T. Nakashima. 2010. Data Mining of Gene Expression Data by Fuzzy and Hybrid Fuzzy Methods. *IEEE Trans. Inf. Technol. Biomed.* 14:23–29. doi:10.1109/TITB.2009.2033590.

Schaefer, G., M. Závišek, and T. Nakashima. 2009. Thermography based breast cancer analysis using statistical features and fuzzy classification. *Pattern Recognit.* 42:1133–1137. doi:10.1016/j.patcog.2008.08.007.

Seidel, G.E.J. 2003. Economics of selecting for sex: the most important genetic trait. *Theriogenology*. 59:585–598.

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. Methods Med.* 2012. doi:10.1155/2012/127130.

Shahinfar, S., D. Page, J. Guenther, V. Cabrera, P. Fricke, and K. Weigel. 2014. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. *J. Dairy Sci.* 97:731–742. doi:10.3168/jds.2013-6693.

Sharma, A.K., C.J. Wilcox, F.G. Martin, and W.W. Thatcher. 1990. Effects of Stage of Lactation and Pregnancy and Their Interactions on Milk Yield and Constituents. *J. Dairy Sci.* 73:1586–1592. doi:10.3168/jds.S0022-0302(90)78829-7.

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. doi:10.1530/rep.0.1230837.

Shi, Y., and Russell Eberhart. 1998. A modified particle swarm optimizer. 69–73.

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

Takagi, T., and M. Sugeno. 1983. Derivation of fuzzy control rules from human operator's control actions. *In* IFAC Symp. Fuzzy Inform. Knowledge Representation and Decision Analysis. 55–60.

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 946162–6172.*

Tyrell, H.F., and J.T. Reid. 1965. Prediction of the energy value of cow's milk. *J Dairy Sci 481215–1223.*

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

Van Arendonk, J.A.M. 1985. Studies on the replacement policies in dairy cattle. II. Optimum policy and influence of changes in production and prices. *Livest Prod Sci*. 13:101–121.

VanRaden, P. 2008. Efficient methods to compute genomic predictions. *J Dairy Sci*. 91:4414–4423.

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. doi:10.3168/jds.2011-4624.

VanRaden, P.M., A.H. Sanders, M.E. Tooker, R.H. Miller, H.D. Norman, M.T. Kuhn, and G.R. Wiggans. 2004a. Development of a national genetic evaluation for cow fertility. *J. Dairy Sci.* 87:2285–2292.

VanRaden, P.M., A.H. Sanders, M.E. Tooker, R.H. Miller, H.D. Norman, M.T. Kuhn, and G.R. Wiggans. 2004b. Development of a National Genetic Evaluation for Cow Fertility. *J. Dairy Sci.* 87:2285–2292. doi:10.3168/jds.S0022-0302(04)70049-1.

VanRaden, P.M., and G.R. Wiggans. 1991. Derivation, Calculation, and Use of National Animal Model Information. *J. Dairy Sci.* 74:2737–2746. doi:10.3168/jds.S0022-0302(91)78453-1.

Veerkamp, R.F. 1998. Selection for Economic Efficiency of Dairy Cattle Using Information on Live Weight and Feed Intake: A Review. *J. Dairy Sci.* 81:1109–1119. doi:10.3168/jds.S0022-0302(98)75673-5.

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. doi:10.3168/jds.S0022-0302(05)72821-6.

Wang, X., C. Maltecca, R. Tal-Stein, E. Lipkin, and H. Khatib. 2008. Association of Bovine Fibroblast Growth Factor 2 (FGF2) Gene with Milk Fat and Productive Life: An Example of the Ability of the Candidate Pathway Strategy to Identify Quantitative Trait Genes. *J. Dairy Sci.* 91:2475–2480. doi:10.3168/jds.2007-0877.

Washburn, S.P., W.J. Silvia, C.H. Brown, B.T. McDaniel, and A.J. McAllister. 2002. Trends in Reproductive Performance in Southeastern Holstein and Jersey DHI Herds. *J. Dairy Sci.* 85:244–251. doi:10.3168/jds.S0022-0302(02)74073-3.

Weigel, K.A. 2004. Improving the Reproductive Efficiency of Dairy Cattle through Genetic Selection. *J. Dairy Sci.* 87:E86–E92. doi:10.3168/jds.S0022-0302(04)70064-8.

Weigel, K.A. 2011. Haplotyps Affecting Fertility and their impact on dairy cattle breeding programs.

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. Data Mining. 2nd ed. Elsevier, San Francisco, CA.

Woolf, P.J., and Y. Wang. 2000. A Fuzzy Logic Approach to Analyzing Gene Expression Data. *Physiol Genomics*. 3:9–15.

Wright S. 1932. Evolution in Mendelian populations. *Genetics*. 97–159.

Yang, X. Z., R. Lacroix, and K. M., Wade. 1999. Neural detection of mastitis from dairy herd improvement records. *Trans ASAE*. 42:1063–1071.

Yao, C., D.M. Spurlock, L.E. Armentano, C.D. Page, M.J. VandeHaar, D.M. Bickhart, and K.A. Weigel. 2013. Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. *J. Dairy Sci*. 96:6716–6729. doi:10.3168/jds.2012-6237.

Zadeh, L.A. 1965. Fuzzy sets. *Inf. Control*. 8:338–353.

Zadeh, L.A. 1975. The concept of a linguistic variable and its application to approximate reasoning ii. *Information Sciences*. 8:301–357.

Zadeh, L.A. 1978. Fuzzy sets as a basis for theory of possibility. Fuzzy Sets and Systems, ii. *Information Sciences*. 1:3–28.