Data Mining to Detect Clinical Mastitis with Automatic Milking

C Kamphuis[1,] H Mollenhorst[1] JAP Heesterbeek[1] and H Hogeveen[1,2]*

[1] Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, the Netherlands, [2]Chairgroup of Business economics, Wageningen University, the Netherlands

Abstract

Our objective was to use data mining to develop and validate a detection model for clinical mastitis (CM) using sensor data collected at nine Dutch dairy herds milking automatically. Sensor data was available for almost 3.5 million quarter milkings (QM) from 1,109 cows; 348 QM with CM were observed by the participating farmers.

Data was divided into a training and a test set, stratified at the cow level. For model building, QM with CM (n = 243) from the training set were taken together with 24,987 QM with a somatic cell count less than 200,000 cells/ml on a milk production test day from cows that never exceeded this threshold during all test days within parity and that were never visually checked by the farmers for CM. The model used decision tree (DT) induction as base classifier, with and without bagging and boosting techniques. Both bagging and boosting techniques work by building models using the base classifier on various samples of the training data. For validation two test sets were created. The first included 105 QM with CM and 13,313 QM without CM, using the same selection as for the training. This test set (Test_GreyOut) excluded the large pool of QM that have a less clear mastitis status. The second test set included the same 105 QM with CM but this dataset included QM with a less clear mastitis status (Test_GreyIn): for negative examples, those QM that were not scored as having CM by the farmers were labeled as negative for CM. From this large sample (n = 1,146,544), a random sample of 50,000 QM was selected. Sensitivity levels were computed at fixed SP levels, and the transformed partial area under the curve (pAUC) was calculated for specificity values of 97% or more to evaluate performance. To visualize performance of the detection models for specificity values of 97% or more, receiver operating characteristic (ROC) curves were constructed.

When using the Test_GreyOut set, the transformed pAUC increased from 0.713 when using the base classifier alone, to 0.787 when combined with boosting, to 0.800 when combined with bagging. At a specificity of 99%, sensitivity was 43.5% for the base classifier, 60.0% when combined with boosting and 61% when combined with bagging. When testing on the TestInGrey set, pAUC values were lower, but still increased when bagging and boosting techniques were used: values increased from 0.643 when using the base classifier alone, to 0.677 when combined with boosting, to 0.702 when combined with bagging. At a specificity of 99%, sensitivity was 24.7% for the base classifier, 30.5% when combined with boosting and 35.2% when combined with bagging. These results were obtained using very narrow time-windows. It is therefore concluded that models developed by DT induction are promising for future implementation.

Keywords: mastitis detection, robotic milking, sensor information, decision-tree induction

Introduction

Detection of clinical mastitis (CM) is always important, independent of whether cows are being milked in a conventional milking parlour, or by an automatic milking (AM) system. Detection is important to maintain an acceptable standard of milk quality, to initiate an antibiotic treatment when necessary, and to safe-guard the welfare of the herd. As there are no milkers present during the milking process with AM, these systems use several in-line sensors that measure different milk and milking features that can be used to detect cows with CM.

These sensor data are then used by CM detection models to generate mastitis alert lists, reporting those cows and quarters likely to have mastitis. It is up to the famers themselves to check the reported cows and quarters visually for CM at a later stage. It seems that farmers using an AM system are able to detect severe cases of CM, as their bulk tank somatic cell count (SCC) levels are comparable to that of conventional milkers (Van der Vorst et al., 2002). However, there is room for improvement in sensitivity (SE) of current detection models. In addition, famers working with AM systems experience a large number of alerts on the mastitis alert lists that turn out to be false, indicating that specificity (SP) levels need improvement as well.

One of the major difficulties of developing CM detection models is the fact that sensor data itself is noisy and often incomplete. In addition, the low prevalence of CM results in highly imbalanced data, which makes modelling even more difficult. Data mining is a technique which tries to discover new knowledge in large amounts of data. The technique is often used to improve decision making processes (Mitchell, 1999). Decision-tree (DT) induction is a commonly used data mining technique (Quinlan, 1986), often used for classification problems, e.g., whether to grant a credit loan or not. The decision whether a quarter or cow has CM or not is a classification problem as well, and a previous study conducted by Kamphuis et al. (2009) already concluded that DT induction showed potential for developing automated detection models for CM using sensor data from AM systems. Models developed in that specific study showed performances that were comparable with models currently used by AM systems. In that study we suggested to improve performances by including more CM cases, by using a stronger selection for negative examples for CM cases for the training process, and by using other data mining techniques that are believed to improve detection performances, like bagging and boosting techniques.

The objective of this current study was to use DT induction to develop and to validate a detection model for CM using sensor data collected during AM that can be used for practical implementation in the nearby future.

Materials and Methods

Data collection and preparation

Data used in the current study was collected at nine Dutch dairy farms using twelve AM systems (Lely Industries N.V., Maassluis, the Netherlands). Farmers were selected for participation based on the primary criterion that they were dealing with mastitis seriously. As a criterion of this, we expected from the farmers that they used the mastitis alert list on a daily basis in a consistent way to select cows being suspicious of CM. In addition, they had to use AM for at least one year. Data collection started at two farms in December 2006. From October 2007 onwards, data was collected at all farms. Data collection continued until March 2009.

A separate remote computer was installed at each AM system to collect raw sensor data. This raw data included average electrical conductivity (EC), red, green, and blue values and estimated yield for each quarter milking (QM), as well as EC, red, green, and blue values measured for each 100 ml of quarter milk produced. Date and time was recorded when a cow entered the AM system, when teat cups were attached and when milk flow started. Quarter milkings where the AM system failed to connect the teat cup, those with missing data for teat cup attachment and 100ml measurements, and those without data at the QM level (sensor measurement averages, start milk flow, teat cup attachment, etc) were deleted from the data set. For the remaining QM, recorded 100ml values for EC, red, green, and blue outside the mean +/- four times the standard deviation were regarded as likely data errors and were set at missing.

A protocol was designed to standardise the visual assessment of the actual mastitis status of the quarters that were checked visually by the farmers, although it was left to farmers themselves to decide which cows and quarters were suspicious enough to check. Their suspicion could be based on the mastitis alert lists, but also on other information, like the somatic cell count determined during the monthly milk production test days or the presence of clots on the milk filter. Although this resulted in slightly different approaches in which quarters were checked per farm (e.g., checking only the quarter mentioned on the mastitis alert list, or checking all four quarters of a particular cow when at least one was mentioned on the alert list), quarters that were checked visually were checked uniformly. Farmers were instructed to visually score the 5th and 6th squirts of milk of all quarters they checked using a clean black paddle as commonly used for the Californian Mastitis Test (without using the reagent normally used when applying this test). When visually normal, the milk was scored with 1. When abnormal, the milk could be scored a 2) watery milk, 3) flakes, 4) clots, 5) serum like milk, or 6) milk with blood. Farmers were instructed to record the cow identification number, quarter, date and time of scoring, and the score assigned to the quarter. Every four to six weeks, farmers were visited to collect data from the remote computer and the scoring forms. During these visits, the scoring forms were discussed with the farmer to ensure a proper use of the protocol. Each visual quarter milk assessment was linked with the sensor data of the most recent milking recorded for that same quarter by the remote computer, within a 24 hour time window prior to the visual assessment time.

Sensor data were combined with data from SCC determination on the milk production test days. The SCC values of a specific cow are combined with all QM recorded by the AM system at a specific test day.

Descriptive variables at the QM level were derived from the 100ml average sensor data as described in Kamphuis et al. (2008): for four sensors (EC, red, green, blue) and two derived measurement patterns (an average value of the red, green, and blue values and milk flow), three time frames, five comparison types, and twelve pattern descriptors were developed. As the remote computers logged additional milk and milking features, three additional measurement patterns (milk flow delay, dead milking time, and milk production) were derived, involving one single value for the whole QM. For these three measurement patterns, the same five comparison types were used to define descriptive variables. A total of 1,065 descriptive variables were used as independent input variables for the CM detection model.

Model development and validation

Sensor data was available from 1,109 cows and almost 3.5 million QM. In total, 1,593 QM were visually checked by the farmers. QM that received a score from 2 through 5 were considered as gold standard positive (n = 348). Data was divided into a training set (containing 2/3 of all data) and a test set, stratified at the cow level, so cows in the training set could not be present in the test set as well and vice versa. The stratification was done in such a way that the number of CM cases in the training set was also about 2/3 of all available CM cases.

The original training set included slightly more than 2.3 million QM from 738 cows. From these QM, 1,034 QM were checked by the farmer of which 243 QM had CM. It was decided to train the models on clear examples of QM with or without CM. Only those QM from cows that never exceeded a SCC level of 200,000cells/ml on the milk production test days within lactation, and in addition were never visually checked by the farmers (over lactations) were selected at first. From these cows, only those QM that were combined with information from the milk production test days, including SCC information, were labelled as gold standard negative QM. By doing so, a large number of QM which we were not able to classify as positive or negative for CM – further referred to as the grey area - were excluded for training. The final training set that was used for model development included a total of 25,230 QM from 430 cows, of which 243 were gold standard positive and 24,987 were gold standard negative.

Decision-tree (**DT**) induction was used for model development. A DT is a graphic representation of a divide-and-conquer approach of a classification problem and consists of nodes at which a variable is tested. The construction of a DT can be expressed recursively: a variable is selected to split the data set at the first node (root node). For each possible outcome of the test involved at the node, a branch is made ending in a daughter node. As a next step, the process can be repeated for each branch, using only those records that actually reach the branch. If at any time all records at a node have the same classification – that is, a leaf node is created – that part of the tree stops developing (Witten and Frank, 2005). In a previous study DT induction was already implemented to develop an automated detection model for CM (Kamphuis et al., 2009), using the J48 algorithm as implemented in WEKA (Witten and Frank, 2005). The same algorithm was used in this current study, where it was combined with the data mining techniques of bagging and boosting to study their effect on detection performance. Both boosting and bagging are techniques that can be used with any base classifier, like the J48 algorithm, and both techniques operate by selectively resampling from the training data set to generate derived training sets to which the base classifier is applied (Webb, 2000). The differences between bagging and boosting can be summarized as follows: bagging uses resampling of records in each iteration *t*, it uses the uniform distribution over the records in the training set and in forming the final output or classification, it gives equal weight to each of the individual models build at each iteration *t*. On the contrary, boosting uses reweighting of records in each iteration *t*, it modifies the distribution over records in the training set, and finally, in forming the final output, boosting gives a 'good' model, with less errors, more weight in the final classification (Freund and Schapire, 1996). In this study, the J48 algorithm with default settings was used as base classifier. This base classifier was used alone or in combination with bagging or boosting. The number of iterations for bagging and boosting was set at 10. For boosting the Adaboost.M1 algorithm was used (Witten and Frank, 2005).

The software used for model building and validation showed limitations in the amount of data that could be analysed. It could not cope with the large number of records and all the 1,065 independent variables. In order to develop a model with all the available 25,230 QM in the training set, it was decided to decrease the number of independent variables in a first step. This was done by building 10 DT models, each developed by using the 243 QM with CM from the training set and one out of 10 random samples of size 3,000 from the pool of 24,987 gold standard negatives in the training set. These 10 models used a total of 224 different independent variables. These variables were selected from the training set before the J48 algorithm, with or without bagging and boosting, was used using all the 25,230 QM mentioned above.

The original test set, including approximately 1/3 of all data, was used for validation. It included over 1.1 million QM from 371 cows that did not appear in the training set. A total of 559 QM were visually checked by the farmers of which 105 were recorded as having CM. To validate the developed models, two test sets were created. For the first test set (**Test_GreyOut**), the same selection that was used in the training set to define QM negative for CM was applied. This test set included the 105 gold standard positive QM and 13,418 gold standard negative QM from 217 cows. The second test set (**Test_GreyIn**) included the same 105 gold standard positive QM. However, this time the grey area of QM was included as well: all QM that were not labelled as gold standard positive were labelled as gold standard negatives. From all these QM, a random selection of 50,000 QM was made. This second test set included 50,105 records from 366 cows.

To evaluate model performance on both test sets, SE and SP of the models when applied on the test sets were calculated. Receiver operating characteristic (**ROC**) curves were constructed to visualise the performance of the developed DT models (Detilleux et al., 1999; Witten and Frank, 2005). ROC-curves are graphic representations of the TP rate (or SN) versus the FP rate (or 1-SP) over the whole range of possible threshold values (Detilleux et al., 1999). To summarise the ROC-curves into a single quantity, the partial AUC was computed, which is restricted to just a relevant portion of the total AUC (Detilleux et al., 1999). In this study, the partial AUC is calculated using the trapezoidal rule for a false positive rate ranging from 0.0 to 0.03 (or SP values of 97% or higher). In order to be able to use the same interpretation as the total AUC, the partial AUC values were transformed to the same scale (McClish, 1989).

Data preparation was done using SAS (version 9.1, SAS Institute Inc., Cary, NC). Development of several DT models with probability estimates for each quarter milking for having CM and the computation of SN and SP were done in WEKA version 3.4.8 (Witten and Frank, 2005). Data produced in WEKA were used in SAS to compute transformed partial AUC values and in S+ (version 8.1, TIBCO Sportfire Software Inc.) to construct ROC-curves.

Results

Figure 1 plots the ROC-curves of the DT models validated on the Test_GreyOut set. The curves are shown for a maximum false positive fraction of 0.03 – that is for SP values of 97% or higher. When the J48 is combined with bagging or boosting, it is clear that there is an improved detection performance: both curves lie above the curve of the DT model using the J48 alone during the whole false positive rate range of interest. The figure looked similar (not shown) when models were validated on the Test_GreyIn set: the DT models showed similar curves, and the bagging and boosting models were higher on the y-axis than the model using J48 alone during the whole range of interest. However, all curves had lower true positive rate values, or in other words, lower SE values. In addition, the differences between the curves were less pronounced.

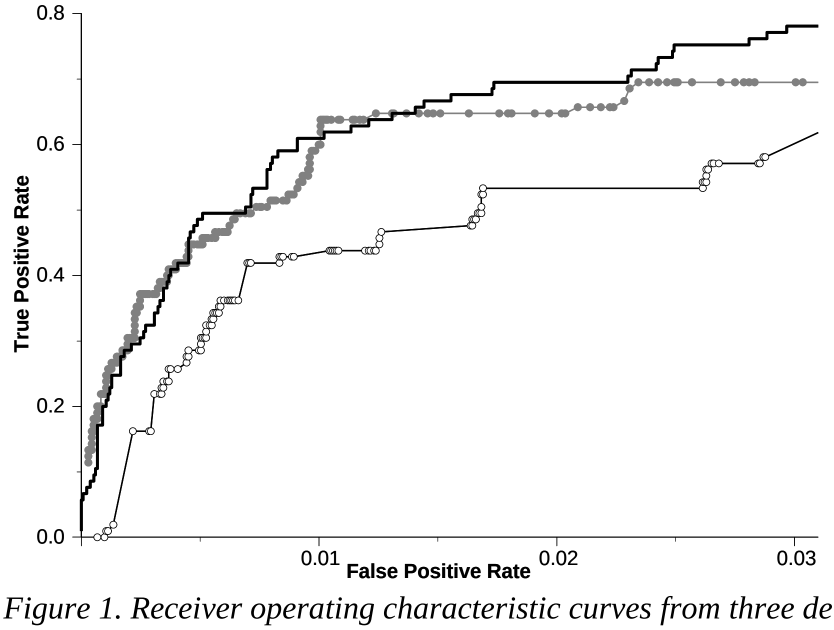

Figure 1. Receiver operating characteristic curves from three detection models validated on a test set excluding the grey area (J48 = —○—); J48 with boosting = —●—; J48 with bagging = ——), for a false positive rate range of 0-0.03 (or Specificity levels of 97% or higher).

Table 1 shows the SE and SP of the different DT models when validated on the two test sets. These values are based on a cut-off value of 0.50 for the probability estimate for having CM. SE values range from 28.6 to 31.4% and are similar for both test sets. This could be expected, as both test sets contain the same gold standard positive QM. SP values range from 98.6 to 99.7%, where SP for the Test_GreyIn set are all lower than those for the Test_GreyOut. For both test sets, SE and SP levels were higher when the J48 base classifier was combined with bagging and boosting techniques. At a SP level of 97%, the SE is 78% when the J48 is combined with bagging and when the grey area of QM is excluded for validation. When the grey area is included, SE decreased to 61%. SE levels decrease considerably when a SP level of 99% is chosen: the SE is 24.7% when the J48 is used alone and 35.2% when the J48 is combined with bagging, and when the grey area of QM is included for validation. Table 1 also summarizes the transformed pAUC values that ranged from 0.643 to 0.800. All pAUC values were higher for the Test_GreyOut set. Although there are differences in the pAUC values between models when tested on the Test_GreyIn set, the differences are larger when the models are tested on the Test_GreyOut set.

Table 1: Sensitivity, Specificity, and transformed partial area under the curve (pAUC) for the three clinical mastitis detection models (J48 alone, J48 combined with Boosting, and J48 combined with Bagging) when validated on two different test sets (one excluding the grey area (Test_GreyOut) and the other including the grey area (Test_GreyIn)), as well as the Sensitivity of the three clinical mastitis detection models when Specificity was fixed (two levels: 97% and 99%)

	Test_GreyOut			**Test_GreyIn**		
	J48	**J48 – Boosting**	**J48 – Bagging**	**J48**	**J48 – Boosting**	**J48 – Bagging**
Sensitivity (%)	28.6	32.4	31.4	28.6	32.4	31.4
Specificity (%)	99.5	99.8	99.7	98.6	98.9	99.1
pAUC[*]	0.713	0.787	0.800	0.643	0.677	0.702
Sensitivity (%) at a Specificity level of 97%	60.0	69.5	78.0	47	54.4	61.0
Sensitivity (%) at a Specificity level of 99%	43.5	60.0	61.0	24.7	30.5	35.2

[*] Transformed partial AUC values are computed for specificity levels of 97% or higher;

Discussion

The current study used DT induction, with and without bagging and boosting techniques, to develop and validate CM detection models using on-farm collected (sensor) data. One of the major concerns in any model-developing study is the gold standard definition, and discussions about the true gold standard are still lively (Mein and Rasmussen, 2008). The ideal situation would be a truly independent observation of all CM cases – that is, data used for gold standard definitions of CM is independent of data used by the model – and that all QM were checked for CM. However, this approach would be far too expensive and time consuming because, when applied to this study, it would take continuous monitoring of nine herds for two years. Therefore, the current study collected data on CM cases using a more practical approach, which are CM cases as observed by farmers themselves. Farmers were asked to record observations of each QM they visually checked by using the scoring protocol, but it was left to the farmers themselves to decide which quarters were suspicious enough to check.

This suspicion could be based on the mastitis alert lists, but as 76% of all the 1,593 visually checked QM did not receive an attention by the AM system (result not shown), there is a strong indication that farmers also used other sources (e.g., clots on the filter) to detect CM cases. Therefore, it seems fair to conclude that the gold standard used for CM definitions, is largely independent from data used for model development.

Bagging and boosting are both techniques that have been used in the field of data mining to improve detection performance. Quinlan (1996) reported results of applying both techniques to a DT algorithm and concluded that both approaches substantially improved the predictive accuracy. The reported improvement was subscribed to the fact that error rates of boosted and bagged classifiers are lower than those of single classifiers. This explanation was also provided for the boosting algorithm by Freund and Schapire (1996). The improved predictive accuracy of bagging and boosting is confirmed by the pAUC values in the current study as well. The transformed pAUC values for bagging and boosting were higher than when the base classifier J48 was used alone, independent whether the models were validated on data including or excluding the grey area of QM. Models were also robust as a second random sample of 50,000 QM (including the grey area) that were labelled as negative for CM was used for validation, and detection performances of the DT models were similar.

Earlier studies also found high levels of SE and SP for mastitis detection models, but results were based on wide time windows in which an alert by the model was considered as true positive (Maatje et al., 1992), or results were based on a highly selected validation set including only clear examples of cows being healthy or having CM. For example, a study by Friggens et al. (2007) tested a LDH-based CM detection model on a dataset including less than 1% of the full dataset, with a strict selection of negative cases of CM and only severe cases with CM that required veterinary treatments. Using wide time windows will result in models showing a good detection performance (Sherlock *et al.*, 2008), but at the same time these models will lose their goal as in practice an automatic CM detection model should generate an alert within a very limited period of time before or only at the milking when CM occurs. Developing models with only clear examples of healthy and diseased animals will also result in methods that will only be able to classify data that include similar clear examples. Unfortunately, it is a fact that only a small part of real field data exists of these clear examples. The effect of including the large amount of less clear examples in a validation set becomes very clear in the current study. Whereas the J48 combined with bagging showed a SE of 78% when validated on the Test_GreyOut set, the SE decreased considerably to 61% when the model was validated on the Test_GreyIn set, indicating that the large grey area, where much of the field data can be found, is just not easily classified by the detection model. Still, the validation on the Test_GreyIn set shows higher SE levels compared to results found in a study by Mollenhorst and Hogeveen (2008). This latter study was conducted to evaluate detection performances of models currently available at AM systems. They showed a SE level of 36.8% at a SP level of 97.9%. The DT combined with the bagging technique in the present current study reached a SE of 49.5% at the same SP level when validated on the Test_GreyIn set, and this is an improvement. In addition, a SP level of 99% has been suggested by Mein et al. (2008) as a level that should be met by detection models when applied in practice. In the current study, the DT combined with bagging showed an SE of 35.2% when tested on the Test_GreyIn set and one could argue that this model still not over performing. However, it remains a matter of discussion whether this SE level will be perceived as being too low by farmers when it is applied in practice. After all, when this 35.2% of detected QM with CM includes a high proportion of the more severe cases of CM (e.g., milk with clots or serum-like milk in the current study), then two major demands of an automated CM detection model are fulfilled with this DT model. That is, at least the more severe cases of CM are detected using a very narrow time window, while at the same time, the number of false positive alerts is at a level that has been suggested as being of practical use.

Conclusion

The results found in this study show that DT models do have difficulties when validated on a test set that includes the grey area of less clear examples of CM, of which field data is largely composed. Although performance estimates decreased, when compared to results of a test set excluding the grey area, results of the test set including the grey area reflect the performance when the model would be applied in the real world. Still, DT induction seems to be able to reach higher SE levels than models currently available on AM systems. As these results were obtained using very narrow time-windows in which an alert by the model was accounted for as positive, it is concluded that models developed by DT induction are promising for future implementation.

References

Detilleux, J., Arendt, J., Lomba, F., Leroy, P., 1999. Methods for estimating areas under receiver-operating characteristic curves: illustration with somatic-cell scores in subclinical intramammary infections. Preventive Veterinary Medicine. 41 (2-3), 75-88.

Freund, Y., and R. E. Schapire. 1996. Experiments with a new boosting algorithm. Pages 148-156 in Machine Learning: Proceedings of the Thirteenth International Conference. L. Saitta, ed. Morgan Kaufmann Publisers, San Fransisco.

Friggens, N. C., Chagunda, M. G. G., Bjerring, M., Ridder, C., HØjsgaard, S., Larsen, T., 2007. Estimating degree of mastitis from time-series measurements in milk: A test of a model based on lactate dehydrogenase measurements. Journal of Dairy Science. 90 (12), 5415-5427.

Kamphuis, C., H. Mollenhorst, A. J. Feelders, D. Pietersma, and H. Hogeveen. 2009. Decision-tree Induction to detect clinical mastitis with automatic milking. doi:10.1016/j.compag.2009.08.012.

Kamphuis, C., D. Pietersma, R. v. d. Tol, M. Wiedemann, and H. Hogeveen. 2008. Using sensor data patterns from an automatic milking system to develop predictive variables for classifying clinical mastitis and abnormal milk. Computers and Electronics in Agriculture 62:169-181.

Maatje, K., P. J. M. Huijsmans, W. Rossing, and P. H. Hogewerf. 1992. The Efficacy of In-Line Measurement of Quarter Milk Electrical-Conductivity, Milk-Yield and Milk Temperature for the Detection of Clinical and Subclinical Mastitis. Livest. Prod. Sci. 30(3):239-249.

McClish, D. K., 1989. Analyzing a portion of the ROC curve. Medical Decision Making. 9 (3), 190-195.

Mein, G. A., and M. D. Rasmussen. 2008. Performance evaluation of systems for automated monitoring of udder health: would the real gold standard please stand up? Pages 259-266 in Mastitis Conrol - From science to practice. T. J. G. M. Lam, ed. Wageningen Academic Publishers, Wageningen, the Netherlands.

Mitchell, T. M. 1999. Machine learning and data mining. Communications of the Association for Computing Machinery 42(11):30-36.

Mollenhorst, H. and Hogeveen, H. 2008. Detection of changes in homogeneity of milk. Internal report.

Quinlan, J. R. 1986. Induction of Decision Trees. Machine Learning 1:81-106.

Quinlan, J. R. 1996. Bagging, Boosting, and C4.5. Pages 725-730 in Proceedings of the Thirteenth Conference of Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference. AAAI Press / The MIT Press, Portland, Oregon.

Sherlock, R., H. Hogeveen, G. Mein, and M. D. Rasmussen. 2008. Performance evaluation of systems for automated monitoring of udder health: Analytical issues and guidelines. In: Mastitis control - from science to practice. T. J. G. M. Lam (Ed.), Wageningen Acadamic Publishers, Wageningen, The Netherlands. pp. 275-282.

Van der Vorst, Y., Knappstein, K., and Rasmussen, M. D. 2002. Milk quality on farms with an automatic milking system: effects of automatic milking on the quality of produced milk. Deliverable 8.EU project QLK5 -2000-31006: Implications of the introduction of automatic milking on dairy farms.

Webb, G. I. 2000. Multiboosting: a technique for combining boosting and wagging. Machine Learning 40(2):159-196.

Witten, I. H., and E. Frank. 2005. Data Mining; Practical Machine Learning Tools and Techniques. 2 ed. Morgan Kaufmann Publishers, San Fransisco.