Making sense of sensor data:
detecting clinical mastitis in automatic milking systems

Claudia Kamphuis

2010

Making sense of sensor data: detecting clinical mastitis in automatic milking systems
Claudia Kamphuis

Dissertation Faculty of Veterinary Medicine, Utrecht University, the Netherlands
- with summaries in English and Dutch -

ISBN: 978-90-393-5439-1

Cover: Anjolieke Dertien, Faculty of Veterinary Medicine, Utrecht University
Picture: Lely Industries N.V., Maassluis, the Netherlands
Lay-out: Harry Otter, Faculty of Veterinary Medicine, Utrecht University
Printed by: Ridderprint, Ridderkerk, the Netherlands

Making sense of sensor data:
detecting clinical mastitis in automatic milking systems

Het doorgronden van sensor informatie:
detecteren van klinische mastitis in automatische melksystemen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. J.C. Stoof, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op dinsdag 23 november 2010 des middags te 2.30 uur

door

Claudia Kamphuis

geboren op 17 april 1979 te Oldenzaal

Promotor: Prof. dr. ir. J.A.P. Heesterbeek

Co-promotoren: Dr. ir. H. Hogeveen
Dr. ir. H. Mollenhorst

Printing of this thesis was financially supported by Lely Industries N.V. (Maassluis, the Netherlands).
This research is supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Program of the Ministry of Economic Affairs (project number 7274).

Ie goat wiezer van ’t weark, as an ’t weark

Contents

Chapter 1

General introduction

A new way of milking dairy cattle

In 1992, Prolion (Bottema, 1992) and Lely Industries N.V. (Maassluis, the Netherlands) launched the first commercially available automatic milking (AM) systems. With this launch, a whole new concept of milking dairy cows was introduced: from this point onwards the presence of a dairy farmer during the milking process was becoming history as it became possible to milk dairy cattle fully automatically. Since the introduction of these first AM systems, the adoption of these systems took a flight nationally and internationally. The adoption of AM systems particularly took place in areas with a high intensity of dairy production (De Koning, 2010), and there is a growing interest in the implementation of AM in other dairy systems, e.g. pasture-based systems in New-Zealand (Jago et al., 2004). To illustrate the adoption of AM systems: in 2010, 8,000 farms in 25 countries worldwide milk their dairy cattle with an AM system, with the Netherlands having the largest number of farms (n=2,000, approximately) that milk automatically (De Koning, 2010). These figures suggest that these systems are a viable and, in many aspects, advantageous alternative to traditional milking parlors. These advantages are not economically based, as investing in an AM system is not cost-effective (Bijl et al., 2007; Wade et al., 2004). Instead, the advantages that seemingly outweigh the extra (mainly fixed) costs of AM, are the more flexible working hours and improved social life for the dairy farmer (Mathijs, 2004), and a potential increase in milk production due to an increased milking frequency (Svennersten-Sjaunja and Pettersson, 2008; Wade et al., 2004).

Monitoring animal health and milk quality when milking automatically

Because with AM no dairy farmer is present during the milking process, there is no obligatory contact anymore between the farmer and the dairy cows twice or three times daily (Hogeveen and Ouweltjes, 2003). When milking automatically, the whole nature and organization of farm labor changes and dairy farmers become more managers that deal with all kinds of controlling activities, than that they are milkers (De Koning, 2010). As a consequence, an AM system requires a completely different management system for milking, feeding, cow traffic, cow behavior and grazing, but also for safe-guarding milk quality and animal health (Svennersten-Sjaunja and Pettersson, 2008). To control and monitor milk quality and animal health, the AM system has to take over the hands and, even more challenging, the eyes of the dairy farmer (Hogeveen and Ouweltjes, 2003). This control of milk quality is also required by law. In Europe, the legislation dealing with milk production (Regulation (EC) No. 853/2004) states the following:

Milk from each animal is checked for organoleptic or physico-chemical abnormalities by the milker or a method achieving similar results and that milk presenting such abnormalities is not used for human consumption.

Electronic devices or sensors are the tools that need to take over the human visual inspection for abnormality. In order to develop sensors to detect abnormal milk, however, a definition of abnormal milk was needed. After an international workshop held in Denmark, Rasmussen (2004) defined abnormal milk as milk differing from normal milk with respect to color or homogeneity. Most abnormal milk is caused by clinical mastitis (CM) (Hogeveen and Ouweltjes, 2002). In those situations where it is known beforehand that milk is unsuitable for human consumption, e.g., colostrum or milk from a cow treated with antibiotics, it is a simple step for the dairy farmer to program the AM system to separate this milk from the bulk tank for the appropriate number of days. For unforeseen situations, e.g., when a cow suffers from CM, this becomes more difficult. Although mastitis is one of the most frequent and costly diseases in dairy cattle (Halasa et al., 2007), with CM being responsible for a large proportion of the total costs of mastitis (Huijps et al., 2008), still a cow suffering from CM is a rare event on most dairy farms. An overall incidence rate of 33.8 CM cases per 100 cow-years was reported for the Netherlands (van den Borne et al., 2010). Approximately 0.32% of all cow milkings in a year will show abnormalities due to CM (De Koning, 2010). This low prevalence of CM makes a management-by-exception approach necessary (Hogeveen and Ouweltjes, 2003), which means that a dairy farmer should only give attention to those cows that need it. In the case of CM, current AM systems apply this management-by-exception approach by generating mastitis alert lists. These lists, produced by the CM detection model available as part of the AM system, report cows that need a visual health check as they deviate from 'normal' for one reason or another. It is the responsibility of the dairy farmer to check the cows listed on the mastitis alert lists visually for having CM. If a CM case is confirmed, the dairy farmer needs to take the appropriate steps, including the programming of withdrawal of milk by the AM system and to start antibiotic treatment if necessary. The effect of starting with AM on milk quality has been studied extensively (e.g., De Koning et al., 2004; Klungel et al., 2000; Rasmussen et al., 2001). Although the bulk tank somatic cell count increases during and just after the transition period from conventional milking (in a parlor) to AM, within a year it decreases again to levels similar to conventional farms milking twice a day (De Koning et al., 2004). As the somatic cell count (SCC) is the most widely applied parameter used as udder health indicator, the result found by De Koning et al. (2004) may indicate that dairy farmers using AM are capable of managing udder health on their farms. However, Dohmen et al. (2010) studied udder health in relation to hygiene at 151 Dutch

farms that used AM for a longer period of time (more than one year), and concluded that udder health is still a major point of concern on farms with AM. This observation pleads for the need for improved CM detection models in order to further improve udder health on farms milking automatically.

How CM detection models work

In-line and on-line sensors measure more or less continuously all kinds of milk characteristics and milking features at cow or quarter milking level during the milking process. In-line sensors measure milk characteristics in the milk tubes, whereas on-line sensors use a milk sample taken from the milk stream that is used by the sensor that is an integrated part of the AM system. Currently, the AM systems mostly use sensors that determine the electrical conductivity, color, or SCC of milk as indicators for udder health (Viguier et al., 2009). Since the introduction of the AM system, detection of CM mainly relied on measuring the electrical conductivity of milk in-line. Sensors measuring the color of milk or the SCC were added to most AM systems more recently.

The electrical conductivity is a measure of resistance of a particular material to an electric current. In milk, the electrical conductivity is determined by the concentration of ions. When pathogens enter the udder, this causes damage to the blood-milk barrier, the tight junctions and the ion-pumping systems. Due to these destructions, the ion concentration of milk changes and this in return causes the electrical conductivity of milk to increase (Hogeveen and Ouweltjes, 2002; Nielen et al., 1992). It is this increase in electrical conductivity, measured by a relatively cheap and easy to implement sensor (Hogeveen and Ouweltjes, 2002; Mottram et al., 2007), that can be used as indicator for CM. The electrical conductivity of milk is also affected by other factors than pathogens invading the udder (Nielen et al., 1992), such as temperature, the fat content of milk and the fraction of milk, e.g., foremilk, that is used for its determination. This will make it difficult to use electrical conductivity as sole criterion for CM detection. It is also probably the main reason why early research on CM detection models evolved from univariate models (Sheldrake and Hoare, 1981) to multivariate models that added milk yield and temperature (De Mol et al., 1997; Nielen et al., 1995a), and further into sophisticated algorithms including neural networks (e.g., Nielen et al., 1995b), Kalman filters (Chagunda et al., 2006; De Mol et al., 1999), and fuzzy logic algorithms (e.g., Cavero et al., 2006).

The first color sensors were based on the principle of reflection of light generated by a light emitting diode (LED) (Espada and Vijverberg, 2002): the basic idea is that different wavelengths (responding to the colors red, green, and blue)

are reflected equally by a white medium (e.g., milk). If the color of milk changes due to CM, wavelengths will be reflected differently. Recently, Song et al. (2010) reported results from a study that used color sensors based on LED light transmittance for the detection of CM. Espada and Vijverberg (2002) showed that milk from quarters suffering from CM have lower color values and concluded that color sensors were a promising tool for the detection of CM. The usefulness of color sensors to detect blood in milk was confirmed by Rasmussen and Bjerring (2005), as a deviating color of milk due to even very low concentrations of blood in milk was detected very reasonably.

The SCC is strongly related to the influx of polymorphonuclear leucocytes from the blood into the udder initiated by the cow to stop pathogens invading the udder (Kremer et al., 1990). High levels of SCC are associated with higher probabilities of having CM (Steeneveld et al., 2008). In general, SCC data is available from monthly or 6-weekly test days only, but a new sensor has been developed that determines the SCC on-line (CellSense, Sensortec Ltd., Hamilton, New Zealand). In essence, this sensor is an automated Californian Mastitis Test and it determines the SCC based on viscosity measurements (Whyte et al., 2004). First studies on the implementation and potential of this on-line sensor are promising (e.g., Mollenhorst et al., 2010).

Clinical mastitis detection models: what is required

In order to describe what currently available models are lacking, it is necessary to explain first how detection models are judged for their performance in general. Models for the detection of CM can be seen as diagnostic tests as they have to classify cows or quarters as being diseased or healthy, in this case having CM or not. In general, diagnostic tests are evaluated by two epidemiological terms, the sensitivity (Se) and specificity (Sp) (Dohoo et al., 2009). Suppose the outcome of a diagnostic test is a binary classification (yes/no) for a quarter to suffer from CM. The Se refers to the probability that a quarter milking with CM will be classified by the test as a quarter milking with CM (a true positive test result). The Sp refers to the probability that a healthy quarter milking will be classified as healthy (a true negative test result). The Se and Sp are interdependent and they can be varied by changing the threshold for the outcome value of the diagnostic test, e.g., the SCC per mL of milk. If this threshold is increased (e.g., from 200,000 cells/mL to 500,000 cells/mL), fewer quarters will be classified as having CM and the Se will decrease. At the same time, more quarters will be classified as being healthy, causing the Sp to increase. So, if the Se of a test decreases, an increase in Sp can be expected and *vice versa*. In addition to the epidemiological approach for

detection performance evaluation, it is also important to use evaluation parameters that are understandable for the (non-scientific) end-users of these models. Sherlock et al. (2008) propose the 'success rate' as a more useful evaluation statistic for a non-scientific audience. The 'success rate', an alternate name for the positive predictive value, represents the proportion of alerts generated by the model that were correct. The downside of the 'success rate' is that it will vary with the prevalence of CM. This will make the 'success rate' less appropriate for comparisons between farms as the prevalence of CM differs between farms. Sherlock et al. (2008) also suggest the number of false alerts per 1,000 cow milkings as evaluation measure for non-scientists, which is a measure that is practical and comprehensible for dairy farmers and that is less sensitive to the prevalence of CM.

Table 1 summarizes study characteristics and results from studies on CM detection models using sensor data. This table only includes results from peer-reviewed studies that used in- or on-line sensors for the development of CM detection models. Therefore, some recent studies are not included in this table as these studies used cow information (Kramer et al., 2009), laboratory determined L-Lactate dehydrogenase (Friggens et al., 2007), laboratory determination of milk samples (Mottram et al. 2007), or they were not peer-reviewed (Song et al., 2010). The number of studies included in Table 1 makes clear that a lot of research on automated CM detection models has been done. It also makes clear that comparing results between the different studies is not valuable, as the study characteristics differ too much. The CM detection models developed by the studies listed are not applied in practice at all or they are still in an experimental stage working on a few number of commercially producing dairy farms. The main reason for this lack of acceptance is that the models don't fulfill some predetermined requirements. A CM detection model, when applied in an AM situation, should detect (preferably) all cows with CM, it should alert for CM when the infection becomes visible, and it should be able to deal with field data. In addition, although this is not a requirement, it would be valuable for the dairy farmer if a CM detection model presents also information about the mastitis causing pathogen involved. These requirements are described in more detail below.

Table 1. Study characteristics of peer-reviewed published studies conducted from 1990 onwards that used sensor information (in- and on-line) for the detection of clinical mastitis (CM). Characteristics included are the number of farms used for data collection, the definition of non-cases and cases, sensors used, the applied methodology, the time window applied for classification, the sensitivity (Se), and the specificity (Sp)

Paper	Farms (n)	Definition non-cases (n)	Definition cases (n)	Sensors	Methodology	Time window	Se (%)	Sp (%)
Maatje et al. (1992)	1 research farm	Based on bacteriological culturing and SCC[1] (200)	CM based on bacteriological culturing and SCC (25)	EC[2]	Moving average and threshold	14d	100	-
Nielen et al. (1995a)	1 research farm	Based on bacteriological culturing and SCC (25)	CM based on observing abnormal milk (31)	EC, milk yield, milk temperature	Artificial neural network	0d[3]	84.0	97.0
Nielen et al. (1995b)	1 research farm	Based on bacteriological culturing and SCC (17 for training; 13 for testing)	CM based on observing abnormal milk or signs of inflammation (13 for training; 13 for testing)	EC, milk yield, milk temperature	Artificial neural network	1d	77.0	69.0
De Mol et al. (1997)	2 research farms	-- (6,495 milkings)	CM based on clinical signs (52 cases)	EC, milk yield, milk temperature	Time-series with Kalman filter	17d	90[4]	98.2[5]
De Mol and Ouweltjes (2001)[6]	1 research farm	Based on never having CM, bacteriological results, and SCC (29,033 milkings)	CM based on clinical signs (48 cases)	EC, milk yield, milk temperature	Time-series with Kalman filter	7d	100[4, 7]	95.1[5, 7]
De Mol and Woldt (2001)[8]	1 research farm	Based on never having CM, bacteriological results, and SCC (29,033 milkings)	CM based on clinical signs (48 cases)	EC, milk yield, milk temperature	Fuzzy logic	7d	100	99.8
De Mol et al. (2001)	4 semi-research farms	Based on not having CM in the collection period, SCC and times milked (299,842 milkings)	CM based on visual observation (95 cases)	EC, milk yield, milk temperature	Time-series with Kalman filter	4d	67[4, 7]	97.9[5]

Paper	Farms (n)	Definition non-cases (n)	Definition cases (n)	Sensors	Methodology	Time window	Se (%)	SP (%)
							(continued)	
Norberg et al. (2004)	1 research farm	Based on bacteriological culturing and having no treatment for CM by veterinarian (1,353)	CM based on treatment by veterinarian after observing clinical signs by staff members (275)	EC	Discriminant function analysis	0d[3]	47.9	91.9
Cavero et al. (2006)	1 research farm	Based on not being treated for CM (109,690 healthy days for training; 51,588 healthy days for testing)	CM based on treatment (651 days of mastitis for training; 348 days of mastitis for testing)	EC, milk yield, milk flow	Fuzzy logic	Day of treatment, plus 2d prior and 2d after treatment	92.9	93.9
Claycomb et al. (2009)	1 for training 1 for testing	--	CM as clots on filter (23 in test set)	EC	Threshold	4d/2d	83	99.8[9]
Mollenhorst et al. (2010)	3 commercial farms	Based on visual normal milk (3,172 quarter milkings)	CM based on visual observation of abnormal milk (19 quarter milkings)	EC, SCC	Threshold	0d[3]	47.4	99.0
Sun et al. (2010)	1 research farm	Based on SCC and not being treated for CM (3,235 quarter milkings)	CM based on visual observation by farm staff or SCC (895 quarter milkings)	EC, milk yield	Artificial neural network	0d[3]	86.9	91.4

[1] Somatic cell count
[2] Electrical conductivity
[3] Based on one milking
[4] Calculated for a mastitis case (cow level)
[5] Calculated for a mastitis-free milking using only cows that never had mastitis
[6] Based on a model developed for conventional milking and adapted for an automatic milking system
[7] Records with indeterminable (e.g., due to missing values) were excluded
[8] A fuzzy logic was used to classify alerts generated by an earlier developed model (De Mol and Ouweltjes, 2000) in order to decrease the number of false positive alerts, not to increase the sensitivity of the detection model
[9] Approximately, using formula: false alert rate ≈ 10 * (100-Specificity) (Sherlock et al., 2008)

Detecting cows with CM

The first requirement of a CM detection model is that it should detect as many cows with CM as possible (preferably all) and put these on the mastitis alert list, especially the cows with a severe CM infection (Hogeveen and Ouweltjes, 2002). This requires a high Se for the detection model. So far, there is no national or international agreement on the required level of Se of a CM detection model when implemented in practice. A Se of >70% and a Sp of >99% are proposed by Mein and Rasmussen (2008) and by Annex C of Standard ISO/FDIS 20966 (Automatic milking installations – requirements and testing) of the International Standard Organization (ISO, 2007). If a model has to perform similarly as human observation of CM, as stated by the EU legislation regarding milk production (Regulation (EC) No. 853/2004), a Se level of 80% should be the target level (Hillerton, 2000). On the other hand, the Sp of a CM detection model is important as well, as dairy farmers want to have as few cows as possible reported on the mastitis alert lists unnecessarily (false positive alerts), even if this means that some cows suffering truly from CM will be missed by the detection model (Claycomb et al., 2009). Currently, AM systems have CM detection models that alert for CM with a Se of 36.8% and a Sp of 97.9% (Mollenhorst and Hogeveen, 2008, unpublished data). These figures make clear that there is a strong need to improve current CM detection models.

Time windows

The second requirement of a CM detection model applies to the time window in which a model has to alert for CM. The concept of using time windows is explained by Sherlock et al. (2008). Observations of CM as well as alerts for CM by a detection model are points in time. Time windows can be applied to both events, although in most conducted studies time windows are only applied to an alert for CM by a detection model. The time window is then used to formulate a time episode in which an alert by the CM detection model is a valid one. Figure 1 explains the application of time windows to CM alerts by a detection model graphically, and its effect on the false positive, false negative, true positive, and true negative alerts in an AM setting.

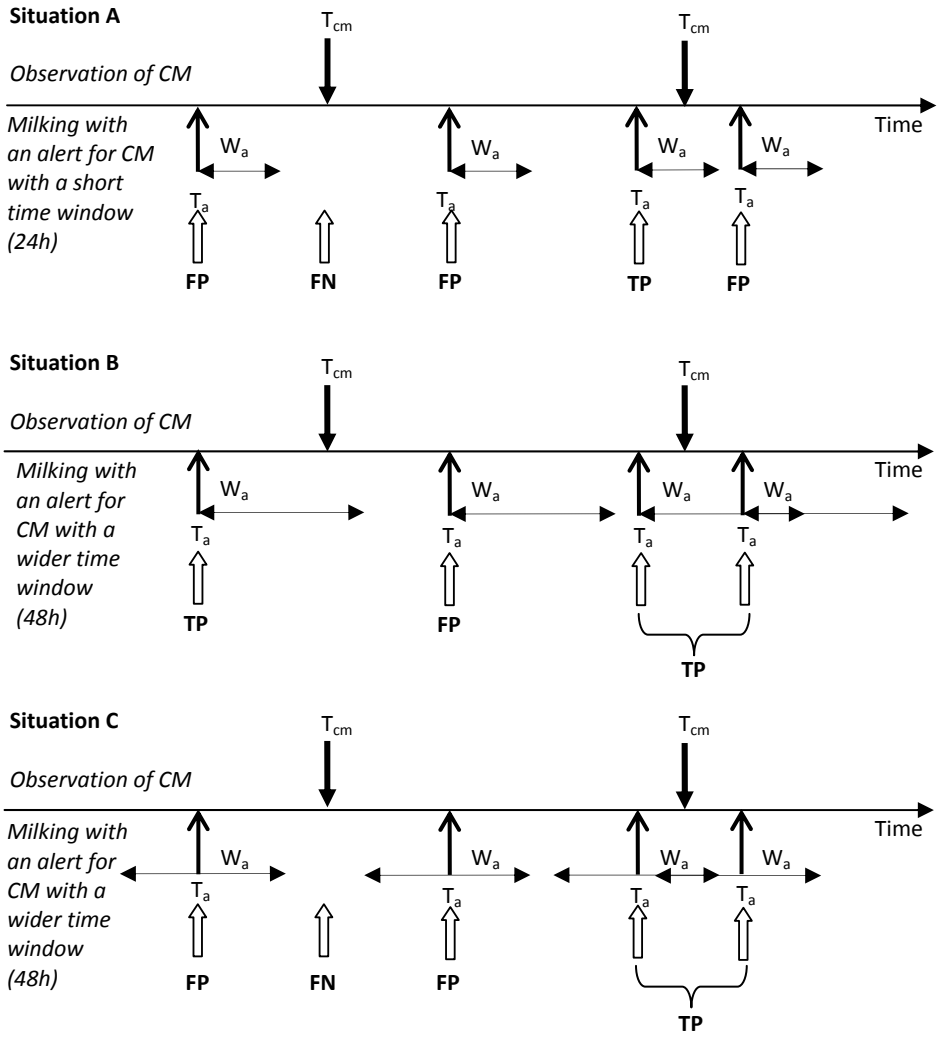


Figure 1. Applying different time windows (24h, situation A; 48h, situation B; 24h before and after an alert, situation C) to alerts (black arrows pointing up) for clinical mastitis (CM; black arrows pointing down) and its effect on the false positive (FP), false negative (FN), and true positive (TP) alerts (white arrows pointing up). An FP alert occurs when an alert for CM (T_a) extended with a time window (W_a) has no observation of CM (T_{cm}) falling into that time window. An FN alert occurs when there is a T_{cm} without any overlapping W_a. A TP alerts occurs when there is a T_a with an extended time window W_a, in which a T_{cm} falls. When two W_a overlap each other (see Situation B and C where a T_a is extended with 48h) these alerts are labeled as one TP (in both situations B and C), FP, or FN

In an AM situation, dairy farmers can detect CM visually at any time of the day, and mostly these times do not coincide with a milking (this in contrast to dairy farmers that milk conventionally in a milking parlor). The CM detection model, however, alerts at the time a cow is being milked by the AM system. Figure 1 depicts three situations; the first uses a short time window in which an alert by the detection model is valid (24h; situation A), the second uses a wider time window of 48h (Situation B), and the latter uses a total time window of 48h but split in 24h before *and* after an alert for CM (Situation C). These numbers are merely examples, but they show clearly the effect of using different lengths of time windows, as well as the effect of applying time windows only after an alert, or both before *and* after an alert. In situation A, the detection model has three false positive alerts, there is just one false positive alert in situation B, and two in situation C. Figure 1 explains visually that when time windows become too small, too many true CM cases will be missed (situation A has one true positive alert, situation B has two true positive alerts). When it is too wide, more CM cases will be accounted for as true positive alerts, but the model will lose its practical ability as dairy farmers will perceive these alerts as false positive (as no signs of mastitis are visible yet). Also, when time windows are applied before *and* after a CM alert (situation C), it is possible that an alert given 24h after a CM observation may be perceived as too late by dairy farmers. In past research on automated CM detection models, time windows have been varied largely (see also Table 1). Some of these time windows were so wide, e.g., a time window of 17 days by De Mol et al. (1997), that practical usefulness, especially in an AM setting, was lost. What the right time window would be for a practical AM setting is arbitrary. However, it should not be too wide as it is of great importance that CM is detected shortly after clinical signs appear in order to eliminate the disease and to prevent recurrence (Hillerton and Kliem, 2002).

Handling field data

Being able to handle field data is a third requirement of a CM detection model. And although this third seems logical at first, most studies reporting on CM detection models lack this essential requirement of being capable to deal with field data. Reasons for not fulfilling this requirement are:

1. Studies were conducted with data from one (research) farm, and included a small number of CM cases (see also Table 1). A lot of studies did develop and validate a model using different data sets, but data for both sets still came from the same farm (e.g., Sun et al., 2010), or data for validation were simulated (Chagunda et al., 2006). These approaches of using one farm for training and testing may result in a model that detects CM at high levels of Se and Sp, but this does not predict the performance on data of a new farm.

The risk of including a small number of CM cases (e.g., 13 CM cases were used by Nielen et al. (1995b)) is that these cases may not represent all variation in CM characteristics (in terms of clinical signs and in sensor measurement patterns), causing a decrease in performance when the model is applied on other farms.

2. Studies used strict definitions for healthy cows or quarters and those that suffered from CM, where SCC, bacteriological culturing, and visual observations are used to define 'healthy' and 'diseased'. This strict definition has been used for cases and non-cases in datasets used for training and for datasets used for validation (e.g., Nielen et al., 1995a; Norberg et al., 2004). For example, Friggens et al. (2007) introduced an interesting new approach of presenting a mastitis *risk*, rather than presenting a binary classification. The model showed high levels of Se and Sp, but performance was based on a validation set that included only highly selected cases and non-cases for CM. However, as a consequence of their strict selection criteria, cows and quarters with a less clear health status were excluded. This does not coincide with daily practice on a dairy farm. For example, cows may be challenged by pathogens invading the udder, but this infection may not (yet) have reached a clinical level. These pathogens cause changes in milk composition that may be detectable by sensors, but not yet by the human eye. When a CM detection model alerts for these sub-clinically infected cows or quarters, dairy farmers will perceive these alerts as false positive ones and they will judge the CM detection model as being inaccurate.
3. Studies excluded complete records with sensor measurement errors or missing data (e.g., De Mol et al., 2001) because the methodology used to develop a CM model was not able to deal with such data. However, real farm sensor data are noisy due to missing values and the need for calibrations of sensors to guarantee a proper functioning and monitoring. Therefore, a CM detection model should be able to deal with these noisy data as well.

Pathogen prediction

Currently available CM detection models do alert for CM, but they fail to give any information about the mastitis causal pathogen. And although this might not be an essential requirement for a CM detection model, having information about the pathogen involved could be beneficial for several reasons. It may increase the cure rate of cows suffering from CM by applying a more appropriate antibiotic treatment (Hillerton and Kliem, 2002; McDougall et al., 2007), and it may limit the use of antibiotics and reduce the risk of antibiotic resistance (Hillerton and Kliem, 2002; Keefe et al., 2010). This information could be a prediction of the mastitis causal pathogen, or of a pathogen characteristic, e.g., the Gram-status. Adding this type

of information on the mastitis alert list will provide a dairy farmer with more information about the disease status of a cow, and, therefore, could improve the decision to treat a cow with CM with antibiotics or not. So far, only Espada and Vijverberg (2002) reported on the potential of using sensor measurements available at an AM system to predict the mastitis causal pathogen or Gram-status. However, that study included a very limited number of cows, and, therefore, further research is necessary to confirm their conclusion.

So, in summary, a CM detection model that will be applied in an AM setting has to fulfill the following requirements: it has to detect at least the severe cases of CM with a high level of Se (>70%) and Sp (>99%); it has to meet these detection requirements using a small time window; and it has to deal with information or data that is noisy and incomplete. In order to improve the detection of CM, additional information about the pathogen involved would be beneficial by improving antibiotic treatment protocols and increasing cure rates, but also by decreasing the risk of resistance occurring due to inappropriate antibiotic use.

Possible solutions

Three main routes can be distinguished in order to improve the needs of currently available CM detection models:

1. Using other algorithms.
   The performance of a CM detection model is dependent of the data used for development, the algorithms used for preprocessing the collected data, and the algorithms used for classification. Past research differs in all these factors, as is summarized in Table 1. Approaches include the use of simple thresholds (e.g., Mollenhorst et al., 2010), moving averages (e.g., Maatje et al., 1992), neural networks (e.g., Sun et al., 2010), time series models with Kalman filters (e.g., Friggens et al., 2007), and fuzzy logic (e.g., Cavero et al., 2006). An algorithm that has not been used so far for CM detection is decision tree induction. This technique, just as neural networks, originates from the scientific field of data mining. In data mining the aim is to discover information from large amounts of data, often to improve a decision making process (Mitchell, 1999). Decision tree induction is a commonly used technique for classification problems (Quinlan, 1986). It is believed to be able to deal with data that is noisy, imbalanced, and incomplete, as field data regarding CM detection is.
2. Adding information from other sensors.
   Studies mentioned in Table 1 in general show higher detection performances for models that included more information than just sensor measurements of

the electrical conductivity. This suggests that adding information from other sensors is valuable in improving the detection of CM. These other sensors may be already existing ones (e.g., sensor determining milk yield or milk temperature as in Nielen et al., 1995), but they also include newer and more complex sensors such as sensors determining the SCC (e.g., Mollenhorst et al., 2010) or the L-Lactate dehydrogenase (e.g., Chagunda et al., 2006).

3. Adding non-sensor information.
   CM detection models can also be based on other information than information collected by sensors available at AM systems. For example, Kramer et al. (2009) used information such as dry matter intake and information about preliminary diseases as input data for a CM detection model. Cow information, e.g., parity, lactation stage and history of CM, has also been used for mastitis pathogen prediction (e.g., Heald et al., 2000; Steeneveld et al., 2009). Combining sensor information with other information sources may very well be able to further improve automated CM detection models (Mottram, 1997).

Objectives and scope of this thesis

The main objective of this thesis is to improve the detection of CM using sensor data readily available from AM systems by applying the three main routes (applying other algorithms, adding information from other sensors, and adding non-sensor information) as described in the previous paragraph. To reach this goal, several studies have been carried out (see also Figure 2). The main body of this thesis embraces the improvement of CM detection models by applying new algorithms for data preprocessing and classification. As a first step, a new concept of data preprocessing was introduced (*Chapter 2*), where new and potential predictive variables were developed using a combination of visual exploration of sensor measurement patterns and a data flow diagram. *Chapter 3* explores the use of a new algorithm (decision tree induction) for the development of a CM detection model. The knowledge gathered in this study is used in *Chapter 4*, where a new CM detection model is developed and validated using decision tree induction and sensor data from 12 AM systems collected at nine dairy farms. This CM detection model used a short time window and included quarter milkings with a less clear mastitis status. *Chapter 5* explores the potential of sensor measurements and decision tree induction to predict Gram-status of CM mastitis causal pathogens, and in this way being (even) more supportive for decisions by dairy farmers that use AM systems.

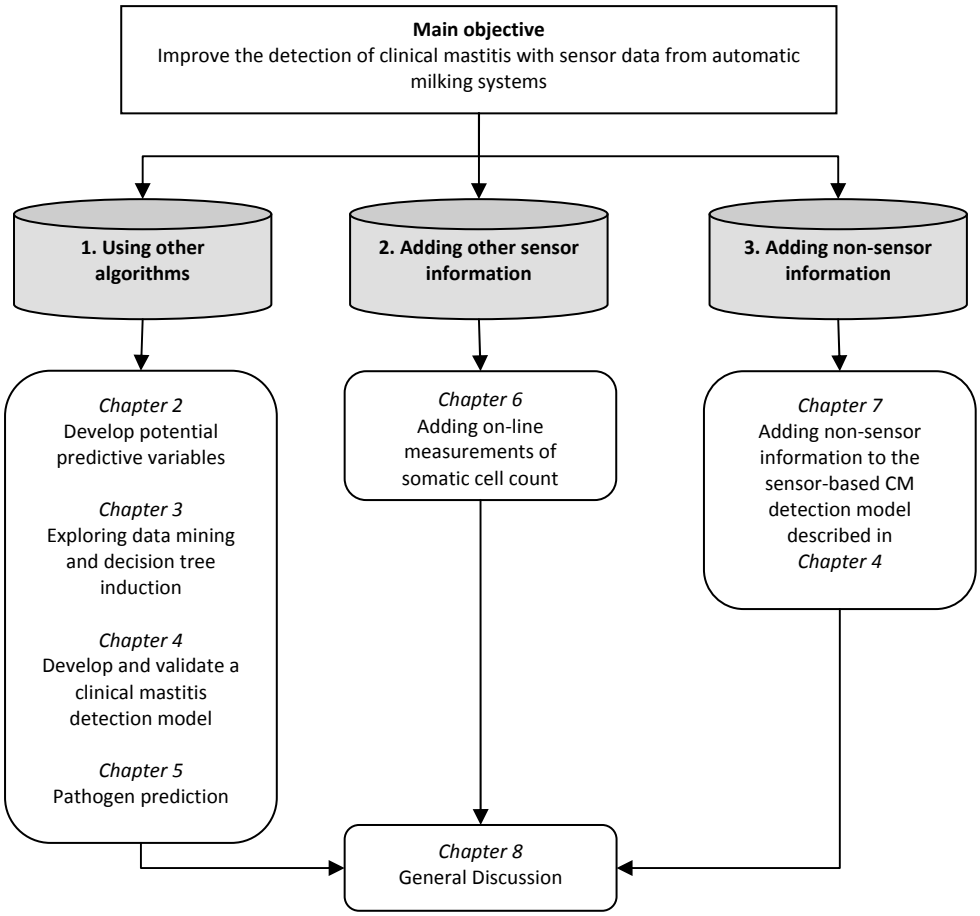


Figure 2. Graphic representation of main objective (improving the detection of clinical mastitis), the three approaches investigated to reach this goal, and their relationship with the chapters in this thesis

Chapter 6 describes the potential additional value of on-line measurements of SCC at cow level, when added to the information of electrical conductivity at the quarter level, for the automated detection of CM. *Chapter 7* adds external non-sensor information to the model developed in *Chapter 4*, to study whether a further improvement in detection performance is possible. In the general discussion (*Chapter 8*) the approaches used to improve the detection of CM and the results found in this thesis are integrated and discussed. Findings from this thesis are used

to formulate future steps to bring the scientific and practical world one step closer to the development of a perfect CM detection model.

References

Bijl, R., S.R. Kooistra, and H. Hogeveen. 2007. The profitability of automatic milking on Dutch dairy farms. J. Dairy Sci. 90(1):239-248.

Bottema, J. 1992. Automatic milking: reality. Pages 63-72 in Proceedings of the international symposium on prospects for automatic milking. A.H. Ipema, A.C. Lippus, J.H.M. Metz, and W. Rossing (eds). Puduc scientific publishers, Wageningen, the Netherlands.

Brandt, M., A. Haeussermann, and E. Hartung. 2010. Invited review: Technical solutions for analysis of milk constituents and abnormal milk. J. Dairy Sci. 93(2):427-436.

Cavero, D., K.-H. Tölle, C. Buxadé, and J. Krieter. 2006. Mastitis detection in dairy cows by application of fuzzy logic. Livest. Sci. 105(1-3):207-213.

Chagunda, M.G.G., N.C. Friggens, M.D. Rasmussen, and T. Larsen. 2006. A model for detection of individual cow mastitis based on an indicator measured in milk. J. Dairy Sci. 89(8):2980-2998.

Claycomb, R.W., P.T. Johnstone, G.A. Mein, and R.A. Sherlock. 2009. An automated in-line clinical mastitis detection system using measurement of conductivity from foremilk of individual udder quarters. New Zealand Veterinary Journal. 57(4):208-214.

De Koning, C.J.A.M. 2010. Automatic milking - common practice on dairy farms. Pages 52-67 in Proceedings of the first North American Conference on Precision Dairy Management. Canada, Toronto.

De Koning, C.J.A.M., B. Slaghuis, and Y. van der Vorst. 2004. Milk quality on farms with an automatic milking system. Pages 311-320 in Automatic Milking: A better understanding. A. Meijering, H. Hogeveen, and C.J.A.M. De Koning (eds). Wageningen Academic Publishers, Wageningen, the Netherlands.

De Mol, R.M., A. Keen, G.H. Kroeze, and J.M.F.H. Achten. 1999. Description of a detection model for oestrus and diseases in dairy cattle based on time series analysis combined with a Kalman filter. Computers and Electronics in Agriculture. 22(2-3):171-185.

De Mol, R.M., G.H. Kroeze, J.M.F.H. Achten, K. Maatje, and W. Rossing. 1997. Results of a multivariate approach to automated oestrus and mastitis detection. Livest. Prod. Sci. 48(3):219-227.

De Mol, R.M., and W. Ouweltjes. 2001. Detection model for mastitis in cows milked in an automatic milking system. Prev. Vet. Med. 49(1-2):71-82.

De Mol, R.M., W. Ouweltjes, G.H. Kroeze, and M.M.W.B. Hendriks. 2001. Detection of estrus and mastitis: Field performance of a model. Applied Engineering in Agriculture. 17(3):399-407.

De Mol, R.M., and W.E. Woldt. 2001. Application of fuzzy logic in automated cow status monitoring. J. Dairy Sci. 84(2):400-410.

Dohmen, W., F. Neijenhuis, and H. Hogeveen. 2010. Relationship between udder health and hygiene on farms with an automatic milking system. J. Dairy Sci. 93(9):4019-4033.

Dohoo, I., W. Martin, and H. Stryhn. 2009. Veterinary Epidemiology Research. 2nd edition. S.M. McPike (ed.), VER Inc., Charlottetwon, Prince Edward Island, Canada.

Espada, E., and H. Vijverberg. 2002. Milk colour analysis as a tool for the detection of abnormal milk. Pages 28-38 in First North American Conference on Robotic Milking, Toronto, Canada. Wageningen Pers, Wageningen, the Netherlands.

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

Halasa, T., K. Huijps, O. Østerås, and H. Hogeveen. 2007. Economic effects of bovine mastitis and mastitis management: A review. Veterinary Quarterly. 29(1):18-31.

Heald, C.W., T. Kim, W.M. Sischo, J.B. Cooper, and D.R. Wolfgang. 2000. A computerized mastitis decision aid using farm-based records: An artificial neural network approach. J. Dairy Sci. 83(4):711-720.

Hillerton, J.E. 2000. Detecting mastitis cow-side. Pages 48-53 in National Mastitis Council 39th Annual Meeting. Madison, WI.

Hillerton, J.E., and K.E. Kliem. 2002. Effective treatment of Streptococcus uberis clinical mastitis to minimize the use of antibiotics. J. Dairy Sci. 85(4):1009-1014.

Hogeveen, H., and W. Ouweltjes. 2002. Mastitis therapy and control: Automatic on-line detection of abnormal milk. Pages 1735-1740 in Encyclopedia of dairy sciences. R. Roginski, J.W. Fuquay, and P.F. Fox (eds). Academic Press, London, UK.

Hogeveen, H., and W. Ouweltjes. 2003. Sensors and management support in high-technology milking. J. Anim. Sci. 81(Suppl. 3):1-10.
Huijps, K., T.J.G.M. Lam, and H. Hogeveen. 2008. Costs of mastitis: facts and perception. J. Dairy. Res. 75(1):113-120.
ISO. 2007. Automatic milking installantions - requirements and testing. ISO/FDIS 20966.
Jago, J., K. Bright, P. Copeman, K. Davis, A.K. Jackson, I. Ohnstad, R. Wieliczko, and M.W. Woolford. 2004. Remote Automatic Selection of Cows for Milking in a Pasture-based Automatic Milking System. Proceedings of the New Zealand Society of Animal Production. 64:241-245.
Keefe, G., J. McCarron, K. MacDonald, and M. Cameron. 2010. The scientific basis for using on-farm culture systems. Pages 141-149 in National Mastitis Council Annual Meeting Proceedings. Alberquerque, New Mexico.
Klungel, G.H., B.A. Slaghuis, and H. Hogeveen. 2000. The effect of the introduction of automatic milking systems on milk quality. J. Dairy Sci. 83:1998-2003.
Kramer, E., D. Cavero, E. Stamer, and J. Krieter. 2009. Mastitis and Lameness detection in dairy cows by application of fuzzy logic. Livest. Sci. 125:92-96.
Kremer, W.D.J., E.N. Noordhuizen-Stassen, and J.A.C.M. Lohuis. 1990. Host defense and bovine coliform mastitis - A review. The Veterinary Quarterly. 12(2):103-113.
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.
Mathijs, M. 2004. Socio-economic aspects of automatic milking. Pages 46-55 in Automatic Milking: A better understanding. A. Meijering, H. Hogeveen, and C. J. A. M. De Koning (eds). Wageningen Academic Publishers, Wageningen.
McDougall, S., K.E. Agnew, R. Cursons, X.X. Hou, and C.R.W. Compton. 2007. Parenteral treatment of clinical mastitis with tylosin base or penethamate hydriodide in dairy cattle. J. Dairy Sci. 90(2):779-789.
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 Control - 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):31-36.
Mollenhorst, H., P.P.J. van der Tol, and H. Hogeveen. 2010. Somatic Cell Count Assessment at Quarter or Cow Milking Level. J. Dairy Sci. 93(7):3358-3364.
Mottram, T. 1997. Automatic monitoring of the health and metabolic status of dairy cows. Livest. Prod. Sci. 48(3):209-217.
Mottram, T., A. Rudnitskaya, A. Legin, J.L.Fitzpatrick, and P.D. Eckersall. 2007. Evaluation of a novel chemical sensor system to detect clinical mastitis in bovine milk. Biosensors & Bioelectronics. 22(11):2689-2693.
Nielen, M., H. Deluyker, Y.H. Schukken, and A. Brand. 1992. Electrical-Conductivity of Milk - Measurement, Modifiers, and Meta Analysis of Mastitis Detection Performance. J. Dairy Sci. 75(2):606-614.
Nielen, M., Y.H. Schukken, A. Brand, S. Haring, and R.T. Ferwerda-van Zonneveld. 1995a. Comparison of Analysis Techniques for Online Detection of Clinical Mastitis. J. Dairy Sci. 78(5):1050-1061.
Nielen, M., M.H. Spigt, Y.H. Schukken, H.A. Deluyker, K. Maatje, and A. Brand. 1995b. Application of A Neural-Network to Analyze Online Milking Parlor Data for the Detection of Clinical Mastitis in Dairy-Cows. Prev. Vet. Med. 22(1-2):15-28.
Norberg, E., H. Hogeveen, I.R. Korsgaard, N.C. Friggens, K.H.M.N. Sloth, and P. Løvendahl. 2004. Electrical conductivity of milk: Ability to predict mastitis status. J. Dairy Sci. 87(4):1099-1107.
Quinlan, J.R. 1986. Induction of Decision Trees. Machine Learning. 1:81-106.
Rasmussen, M.D. 2004. Detection and separation of abnormal milk in automatic milking systems. Pages 189-197 in Automatic Milking: A better understanding. A. Meijering, H. Hogeveen, and C.J.A.M. De Koning (eds.). Wageningen Academic Publishers, Wageningen, The Netherlands.
Rasmussen, M.D., and M. Bjerring. 2005. Visual scoring of milk mixed with blood. J. Dairy. Res. 72(3):257-263.
Rasmussen, M.D., J.Y. Blom, L.A.H. Nielsen, and P. Justesen. 2001. Udder health of cows milked automatically. Livest. Prod. Sci. 72(1-2):147-156.
Sheldrake, R.F., and J.T. Hoare. 1981. The detection of mastitis in individual quarters using electrical conductivity or somatic cell concentration. New Zealand Veterinary Journal. 29:211-213.
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. Pages 275-282 in Mastitis control - from science to practice. T.J.G.M. Lam (ed.). Wageningen Acadamic Publishers, Wageningen, the Netherlands.

Song, X., S. Zhuang, and P.P.J. van der Tol. 2010. New model to detect clinical mastitis in Astronaut A3 Next^{tm} milking robot. Pages 474-480 in Mastitis research into practice: Proceedings of the 5th IDF mastitis conference. J.E. Hillerton (ed). VetLearn, Wellington, New Zealand.

Steeneveld, W., H. Hogeveen, H.W. Barkema, J. van den Broek, and R.B.M. Huirne. 2008. The influence of cow factors on the incidence of clinical mastitis in dairy cows. J. Dairy Sci. 91(4):1391-1402.

Steeneveld, W., L.C. van der Gaag, H.W. Barkema, and H. Hogeveen. 2009. Providing probability distributions for the causal pathogen of clinical mastitis using naive Bayesian networks. J. Dairy Sci. 92(6):2598-2609.

Sun, Z., S. Samarasinghe, and J. Jago. 2010. Detection of mastitis and its stage of progression by automatic milking systems using artificial neural networks. J. Dairy. Res. 77:168-175.

Svennersten-Sjaunja, K.M., and G. Pettersson. 2008. Pros and cons of automatic milking in Europe. J. Anim. Sci. 86(13):37-46.

Van den Borne, B.H.P., G. van Schaik, T.J. Lam, and M. Nielen. 2010. Variation in herd level mastitis indicators between primi- and multiparae in Dutch dairy herds. Preventive Veterinary Medicine. 96(1-2):49-55.

Viguier, C., S. Arora, N. Gilmartin, K. Welbeck, and R. O'Kennedy. 2009. Mastitis detection: current trends and future perspectives. Trends in Biotechnology. 27(8):486-493.

Wade, K.M., M.A.P.M. van Asseldonk, P.B.M. Berentsen, W. Ouweltjes, and H. Hogeveen. 2004. Economic efficiency of automatic milking systems with specific emphasis on increases in milk production. Pages 62-67 in Automatic Milking: A Better Understanding. A. Meijering, H. Hogeveen, and C.J.A.M. de Koning (eds). Wageningen Academic Publishers, Wageningen, the Netherlands.

Whyte, D.S., R.G. Orchard, P.S. Cross, T. Frietsch, R.W. Claycomb, and G.A. Mein. 2004. An on-line somatic cell count sensor. Pages 235-240 in Automatic Milking: A better understanding. A. Meijering, H. Hogeveen, and C.J.A.M. de Koning (eds.). Wageningen Academic Publishers, Wageningen, the Netherlands.

Chapter 2

Using sensor data patterns from an automatic milking system to develop predictive variables for classifying clinical mastitis and abnormal milk

C. Kamphuis, D. Pietersma, R.P.P. van der Tol, M. Wiedemann, H. Hogeveen

Published
Computers and Electronics in Agriculture. 2008. 62 (2): 169-181

Abstract

Dairy farmers using automatic milking are able to manage mastitis successfully with the help of mastitis attention lists. These attention lists are generated with mastitis detection models that make use of sensor data obtained throughout each quarter milking. The models tend to be limited to using the maximum or average value of the sensor data pattern, potentially excluding other valuable information. They often put cows on the lists unnecessarily, and their sensitivity for abnormal milk classification is too low for automated separation. Therefore, we analyzed sensor data patterns within quarter milkings in order to identify potentially predictive variables for abnormal milk and clinical mastitis (CM) classification. The data used in this study was obtained at a commercial dairy farm in Germany in September 2002, where a German Simmental farm was milked by a Lely Astronaut system. In total, 3,232 quarter milkings from 63 cows were analyzed; 94 quarter milkings were defined as milk with abnormal homogeneity and 270 as CM. A data flow diagram was developed to systematically describe the steps involved in the transformation of within quarter milking measurements into variables that potentially predict abnormal milk and CM. Three types of pattern descriptors were used: level, variability, and shape. In addition to using the absolute value of the pattern descriptor, the descriptors were considered relative to their expected value based on pattern descriptor values from previous milkings and from other quarters within the same cow milking. Using this method, potentially predictive variables were computed for electrical conductivity, the colors red, green and blue, a combination of color sensors, and milk production. The importance of a variable in predicting abnormal milk and CM was evaluated by computing correlation coefficients as well as information gain ratios. The most important variables came from the sensors for electrical conductivity, blue and green. Variables describing the variability and shape of the measurement patterns were as important as mean and maximum values, and should be included in future modeling. Also variables that are based on absolute values should be considered for future modeling. Results suggest that CM and abnormal milk classification models may include similar predictive variables, but requirements for these models differ resulting in the need for different models. The schematic approach to developing potentially predictive variables will be helpful when exploring the usefulness of new sensors, researching other approaches to estimate expected values, and studying sensor data patterns in general.

Introduction

Since the introduction of automatic milking (AM) in the 1990s, the system has been adopted by a substantial number of dairy farmers. For example, in the Netherlands, 840 farmers (4%) were using AM by 2006 (personal communication: P.J.M. Huijsmans, Organization for quality control of milking equipment (KOM), Lelystad, the Netherlands). Such dairy farmers are able to achieve bulk milk quality levels comparable to those achieved with conventional milking (Helgren and Reinemann, 2003; Van der Vorst et al., 2002). Because AM lacks milkers that can check the milk visually, the AM system uses sensor data to inspect milk quality and to classify milk for CM and normality. So-called 'attention lists' are generated to warn the dairy farmer about cows and quarters likely to have abnormal milk due to mastitis. Current EU legislation dealing with milk production (Regulation (EC) No. 853/2004) requires that milk is checked for abnormalities by the milker or by a method achieving similar results. Additionally, the legislation requires that milk presenting abnormalities is excluded for human consumption (abnormal milk has been defined by Rasmussen (2003) as milk differing from normal milk in homogeneity or color).

In the past, AM systems relied on electrical conductivity (EC) to detect CM and abnormal milk. EC sensor measurements are obtained more or less continuously during each quarter milking, resulting in time-series of data points. From these data patterns, predictive variables are extracted and subsequently used in classification models for abnormal milk or mastitis. Predictive variables describing the EC patterns have involved the maximum value within a quarter milking (e.g., Nielen et al., 1995a; Norberg et al., 2004) and the average of the 20 highest values within a quarter milking (e.g., De Mol et al., 1997; Norberg et al., 2004). In some studies, variables have been compared across milkings of the same quarter (e.g., Lansbergen et al., 1994) or with other quarter milkings (e.g., Maatje et al., 1992) to detect pattern changes possibly due to the occurrence of mastitis. Variables concerning color have also shown themselves to be promising for abnormal milk classification (Espada and Vijverberg, 2002; Ouweltjes and Hogeveen, 2001; Wiedemann and Wendl, 2004); for example, milk with a slightly red color due to (mixing with minute amounts of) blood can be detected using these color sensors (Espada and Vijverberg, 2002; Ouweltjes and Hogeveen, 2001; Rasmussen and Bjerring, 2005).

Although current AM systems can detect mastitis, the classification could be refined in order to reduce the number of cows that are on the attention lists unnecessarily (false positives). It could also be improved with better classification and automated separation of milk with abnormal homogeneity (i.e., clots due to mastitis). After testing several AM systems for their ability to detect milk with abnormal homogeneity, Rasmussen and Bjerring (2004) concluded that the

classification performance of these systems was too low to automate this separation in practice. This may be because descriptive variables were mostly averages and maximum values and mainly concern EC measurements, and so may have missed valuable patterns in the available data (Norberg et al., 2004).

The objective of this study was to identify new, potentially predictive variables based on electrical conductivity and color measured within quarter milkings using an analytical framework, and to evaluate the importance of these variables for the classification of abnormal milk and CM.

Materials and methods

Data acquisition

Data used in this research was obtained in September 2002 as part of a study conducted at a commercial dairy farm in southern Germany. The farm of 65 German Simmental dairy cows was milked automatically with a Lely Astronaut system (Lely Industries N.V., Maassluis, the Netherlands). There were 19 first parity cows, 29 second parity cows and 17 cows were in their third parity or higher. Average milk yield at the time of study was approximately 7,500 kg per cow per year. Cows were milked 2.4 times per day on average.

During two periods of 4 and 10 consecutive days in September 2002, all cows entering the AM system were scored for their quarter milk homogeneity for 24h per day. This visual appearance of quarter milkings was evaluated by seven people, mainly students and milk technicians. All were experienced in dairy farming and confident with the protocol for scoring homogeneity. Approximately 40 mL of the first milk for each quarter milking was collected. Milk was collected after a cow had entered the AM system but before teat cups were attached. Seven categories were used to describe the homogeneity of each quarter milking: (0) visually normal; (1) watery, slightly blue, without flakes; (2) watery, slightly blue, with small flakes; (3) a few clots; (4) many clots; (5) mainly clots and the appearance of milk nearly disappeared, and (6) puss, blood, and slime with a complete loss of milk appearance.

Following the manual collection of milk samples, the AM system continued the normal milking process by attaching the teat cups. The first approximately 30 mL milk per quarter was discarded by the system. After that, average values for each 100 mL of milk were recorded for four in-line sensors (EC, red, green and blue) for each quarter milking. EC was measured in a collection bin with known content between two electrical pins. The colors were measured with a LED sensor combination consisting of three light sources emitting red, green, and blue light, and a receiver to measure the reflected light intensity (Espada and Vijverberg,

2002). The measurements of the color sensors were calibrated automatically by the AM system to achieve farm averages of 500 units. For each sensor, recorded values were visually evaluated with a histogram. Two color sensors (red and green) showed a small cluster of values around 200 units that was clearly below the main cluster of data points. Also the EC sensor showed a small cluster of values below 40 units. These values were thought to be related to malfunctioning of the sensors and were discarded. These excluded values represented less than 0.05% of all the 100 mL measurements for each sensor type. A fifth measurement pattern ('combined') was derived by averaging the measurements of the three color sensors.

A final sensor recorded milk yield for each quarter milking. The quarter milk yield involved only a single value measured at the end of each quarter milking (unlike the within-quarter measurements obtained with the other sensors), and was added as a reference to allow comparison with previous studies. To account for the variability in yield due to variation in the time between two milkings, quarter milk yield was divided by the time since the previous milking and referred to as milk production (mL/h).

The initial dataset involved 65 cows and 11,284 quarter milkings with 100 mL sensor data measurements. Milk homogeneity was evaluated for 29% of the quarter milkings (n=3,232) from 63 cows. More than 90% of the 3,232 quarter milkings that were evaluated for homogeneity appeared visually normal (homogeneity=0; Figure 1).

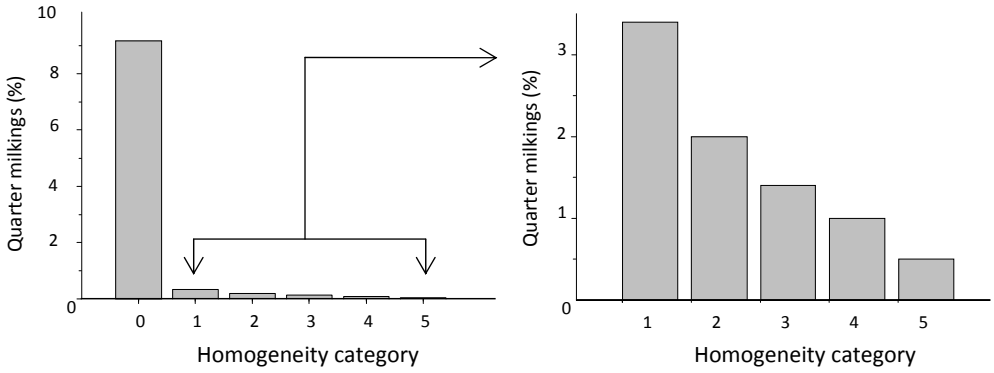


Figure 1. Distribution of quarter milkings over homogeneity categories

The homogeneity scores were used to define both quarter milkings that should be discarded automatically and quarter milkings with CM. Following suggestions by Rasmussen (2005), milk to be discarded automatically was defined as milk containing large flakes or clots, i.e. homogeneity category three or higher involving milk with a few clots or worse. These quarter milkings were referred to as 'abnormal milk'. Quarter milkings with homogeneity category 1 or higher were defined as positive milkings for the classification of CM. Each abnormal or clinical mastitic quarter milking was seen as a separate event, regardless whether the elevated homogeneity categories were assigned to successive quarter milkings or whether more than one quarter was assigned a higher homogeneity score within the same cow milking. There were 94 quarter milkings defined as abnormal milk (from 17 cows) and 270 quarter milkings from 31 of the 65 cows were defined as positive for CM. During the data collection period there were 32 cows with at least one quarter milking having a somatic cell count of more than one million.

Creation of potentially predictive variables to describe measurement patterns
A data flow diagram was developed as an analytical framework to formalize the various data preparation steps in identifying potentially predictive variables within quarter milking sensor measurements (Figure 2). In this diagram, the time series of raw sensor measurements from the AM system are used as input. The output consists of variables that can be used by classification models to detect likely cases of abnormal milk and CM. The first data preparation step involves selecting a time frame to specify the part of the quarter milking to be used for subsequent processing; this might be the entire quarter milking or the first five measurements, or any other time frame. The second step involves calculating so called pattern descriptors to capture specific aspects of the time-series of measurements. Three classes of pattern aspects were defined (level, variability, and shape) to describe general parts of the sensor data pattern. Examples of descriptors are the mean, minimum, and maximum for the level, the standard deviation or range for the variability, and the slope for describing the shape of a pattern. These pattern descriptors can be used as variables directly or they can be stored and used in the estimation of expected values in the third step. The expected value for a pattern descriptor for a particular quarter milking can be computed using the values obtained from other quarters within the same milking (e.g., using the value of the lowest quarter as a reference), using the values obtained from previous milkings of the same quarter, or by using values obtained from a combination of values from other quarters and previous milkings. Finally, the current value of the pattern descriptor can be compared with the expected values to give absolute or relative deviations. These deviations from expected values can be used as potentially

predictive variables, in addition to the absolute values of the pattern descriptors computed in step 2 of the data flow diagram.

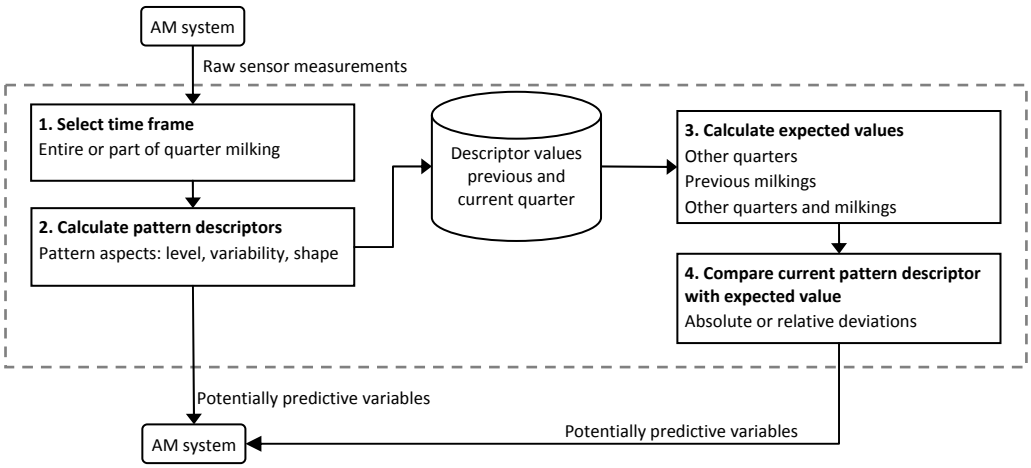


Figure 2. Data flow diagram representing data preparation steps to transform within quarter milking measurements patterns into potentially predictive variables

In this research, five measurement patterns were used: the sensors EC, red, green, blue, and the derived pattern for the combined colors. In the first step, three time frames were selected: the whole milking (Whole), the first 500 mL only (F500), and the last 500 mL only (L500). The second step involved computing 12 different descriptors for each of the measurement patterns (Table 1). This resulted in 180 variables with an absolute value for the pattern descriptor (ABS). In the third step, four different types of expected values were calculated: the average of the two quarters with the lowest values of the current milking (Q2), the value of the previous milking of the same quarter (P1), the mean of the two previous milkings of the same quarter (P2), and the mean of the three previous milkings of the same quarter (P3). The fourth step involved computing the absolute deviation of the pattern descriptor value from the expected values. These steps resulted in 900 variables describing the five measurement patterns. In addition, the absolute value and the four types of deviation from the expected value were calculated for the milk production measurements. These milk production variables were considered as involving the whole milking time frame.

Due to missing values, potentially predictive variables could not always be calculated. Expected values based on the other quarters of the same milking were

set to missing if the value of only one other quarter was available. Expected values based on previous milkings were set to missing if any required variable was missing.

Visual exploration of measurement patterns and their descriptors
To support the identification of potentially predictive variables, patterns of within quarter milking measurements were studied visually using a matrix of panels, each representing a time-series data graph (Figure 3). In this figure, each row of graphs represents a quarter and each column represents a milking. The graphs show 100 mL measurements of EC and the colors red, green, and blue for all four quarters of a cow during four successive milkings. The homogeneity category of the milking is shown with a horizontal line as it involves a single value for the entire quarter milking. It is clear from this figure that each quarter more or less has its own specific sensor pattern. For example, the left front quarter had values for green that were consistently higher than the values for red, and the values for red and blue decreased more rapidly than the values for green during milking. In contrast, EC in this quarter increased slowly during milking.

Table 1. Type and definition of measurement pattern descriptors

Type	Pattern Descriptor	Definition
Level	Mean	Mean of observations
	Maximum	Maximum observation value
	Minimum	Minimum observation value
	First	Value of first observation
Variability	SD	Standard deviation
	CV	Coefficient of variation (100 * SD / Mean)
	Range	Maximum – Minimum
Shape	IncreaseFL	Increase from first to last observation: value last – value first observation
	Slope	IncreaseFL / number of observations
	IncreaseFM	Increase from first to midpoint observation: value at midpoint – value first observation
	IncreaseML	Increase from midpoint to last observation: value last – value at midpoint observation
	DeviationM	Deviation of midpoint observation from expected midpoint: value at midpoint – (First + Last observation) / 2

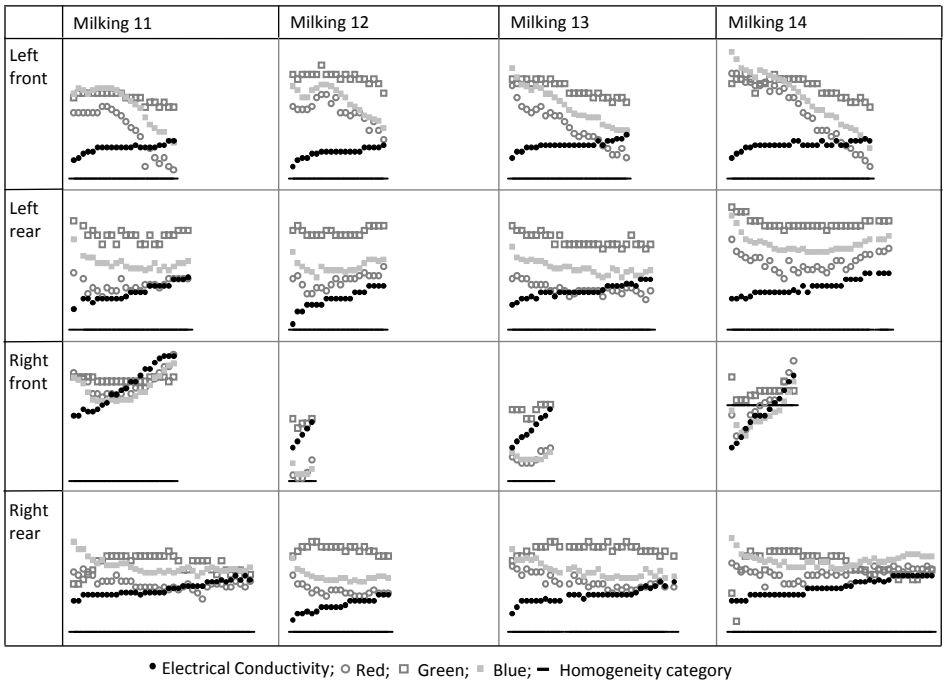


Figure 3. Time-series measurements for electrical conductivity, red, green, and blue, combined with homogeneity category, for all four quarters (from top to bottom) of a cow for four successive milkings (from left to right)

All quarter milkings were evaluated as visually normal (homogeneity=0), except the right front quarter at milking 14, where the milk showed a few clots (homogeneity=3). These visual evaluations were used to explore changes in measurement patterns associated with changes in homogeneity category. For example, in Figure 3, the color red increased in the right front quarter with the homogeneity score of three, whereas there was a decrease in the other three quarters within the same milking. Also, the increase in red was less pronounced in previous milkings and thus the slope could be a pattern descriptor, describing the deviation in patterns properly.

The series of graphs were also used to visually evaluate calculated potentially predictive variables. Figure 4 shows the same cow and quarter milkings as Figure 3, but represents the 100 mL measurements of the EC sensor only (indicated with dotted line), together with the absolute values of three pattern descriptors (the mean, S.D., and slope; one for each descriptor type) and the homogeneity score given to each quarter milking, shown with horizontal lines. The figure indicates that

EC increased during each quarter milking. This increase is more pronounced for the right front quarter and this is already more pronounced in milking 11. The pattern descriptors generally showed higher values for this quarter during all 4 milkings in comparison with values for the other quarters within the same milking. The elevated homogeneity score of the right front quarter at milking 14 was uniquely identified by a combination of a high S.D. and a high slope.

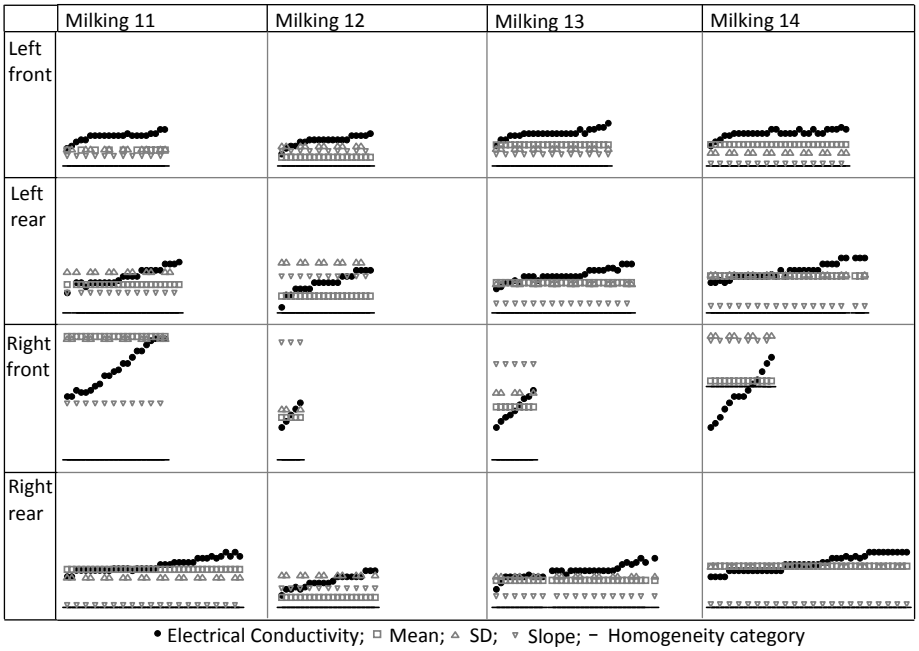


Figure 4. Time-series measurements for electrical conductivity combined with the mean, S.D., and slope of the measurement pattern and homogeneity category, for all four quarters (from top to bottom) of a cow for four successive milkings (from left to right)

Evaluation of variable importance

Two metrics were used to evaluate the importance of the variables for abnormal milk and CM classification; for both, the entire dataset was used. Firstly, Spearman rank correlation coefficients were computed between the potentially predictive variables and the classification variable. For the dependent variable, abnormal milk or CM, respectively the numerical values 0 and 1 were used to represent the

negative and positive classes (Refaat, 2006). Secondly, the information gain ratio (Quinlan, 1986) was used. The gain ratio is based on the change in entropy or information value of a dataset *S* with respect to the dependent variable *Y* after partitioning *S* using the values of the independent variable *X*, and takes into account the number of classes and records per class of *X* (Witten and Frank, 2005). To calculate the gain ratio, the numerical independent variables were made categorical through discretization.

To investigate whether it might be useful to combine potentially predictive variables in future modeling of abnormal milk or CM, Spearman rank correlation coefficients were computed between a subset of potentially predictive variables.

Data preparation was performed using SAS version 9.1 (SAS Institute Inc., Cary, NC). Visualization of within quarter measurements to explore interesting pattern predictors was performed with S-PLUS version 7.0 (Insightful Corp., Seattle, WA). Spearman rank correlation coefficients were computed with SAS and gain ratios were computed with WEKA version 3.4 (Witten and Frank, 2005).

Results

Table 2 shows, for each homogeneity category, mean values for three descriptors (mean, S.D., and slope) of the measurement patterns of EC, red, green, blue, and combined, and the mean values for milk production. These mean values were based on varying numbers of quarter milkings, caused by missing values for some pattern descriptors.

In general, mean values of the pattern descriptors for EC increased with increasing homogeneity category whereas the mean values for the three color sensors, the combined color pattern, and milk production decreased with increasing homogeneity category. Mean values of the pattern descriptors calculated for homogeneity category 1 (watery milk without flakes) deviated from this general observation. For this category, the pattern descriptor mean had high values for the EC, red, blue and combined color measurement patterns. In addition, the slope for the red, blue, and combined measurement patterns was closer to zero for this homogeneity category than for the others and the mean milk production was lower than for homogeneity categories three and four. For EC, the pattern descriptor values were especially high for the fifth homogeneity category (milk with mainly clots). For the color patterns, descriptor means showed especially low values for the two highest categories (milk with many clots and mainly clots). Milk production was very low for the highest homogeneity category. The EC pattern descriptor S.D. increased with homogeneity score while the red pattern descriptor S.D. decreased with increasing homogeneity score.

Table 2. Number of quarter milkings included and means per homogeneity category for three descriptors of the patterns EC, red, green, blue, and combined, and for milk production

			Homogeneity category					
			0 (2936-2962)[a]	1 (103-110)[a]	2 (65-66)[a]	3 (43-45)[a]	4 (31-33)[a]	5 (16)[a]
Pattern	**Descriptor Type**	**Pattern Descriptor**						
EC	Level	Mean	72.0	81.7	75.8	78.2	79.3	86.4
	Variability	SD	3.3	6.6	5.2	6.0	5.4	6.5
	Shape	Slope	0.7	1.4	1.3	1.3	1.6	2.2
Red	Level	Mean	501.6	510.4	504.5	506.3	500.0	497.6
	Variability	SD	8.2	6.7	5.3	6.7	4.6	3.9
	Shape	Slope	-0.9	-0.2	-0.8	-0.6	-0.5	-0.6
Green	Level	Mean	501.1	500.6	500.0	501.1	495.9	492.5
	Variability	Std	3.4	3.2	2.3	2.8	2.7	3.1
	Shape	Slope	-0.2	-0.2	-0.3	-0.3	-0.6	-0.6
Blue	Level	Mean	500.2	511.3	498.2	496.8	471.7	473.1
	Variability	SD	12.6	11.5	9.6	11.7	12.2	9.9
	Shape	Slope	-1.7	-0.5	-1.8	-1.9	-3.3	-2.4
Combined	Level	Mean	500.9	507.5	500.9	501.4	489.2	487.7
	Variability	SD	7.5	6.2	5.3	6.5	6.1	5.2
	Shape	Slope	-0.9	-0.3	-0.9	-0.9	-1.5	-1.2
Milk production			202.2	144.9	159.5	166.0	131.4	95.1

[a] Number of quarter milkings

Table 3 shows the 5% highest correlation coefficients for abnormal milk and CM of all 905 analyzed potentially predictive variables. In general, correlation coefficients were low. The highest correlation coefficients for both abnormal milk (0.18) and CM (0.28) were for the absolute values of the standard deviations of the EC sensor (Table 3). These standard deviations were based on the whole quarter milking. The 5% highest correlations involved mainly absolute values (ABS) of the potential predictive variables, some were based on the deviation from expected values using quarters within the same milking (Q2), and none were based on the deviation from expected values using previous milkings (P1, P2, or P3). The correlation coefficients of the Q2 variables were lower or similar to the correlation coefficients of the corresponding ABS variables.

Table 3. Subset of correlation coefficients (45 highest values) from 905 potentially predictive variables for abnormal milk and clinical mastitis

			Abnormal milk					Clinical mastitis				
Sensor	**Descriptor**	**Pattern**	**ABS**[a]			**Q2**		**ABS**			**Q2**	
	Type	**Descriptor**	**F500**	**L500**	**Whole**	**L500**	**Whole**	**F500**	**L500**	**Whole**	**F500**	**Whole**
EC	Level	Mean	0.13	0.14	0.15			0.18	0.22	0.21	0.14	0.14
		Maximum	0.15	0.16	0.15			0.19	0.25	0.25	0.14	0.14
		Minimum	0.11	0.13	0.11			0.16	0.19	0.16		
		First	0.10	0.12	0.10			0.16	0.19	0.16		
	Variability	SD	0.15	0.14	0.18		0.11	0.14	0.24	0.28		
		CV	0.14	0.13	0.16		0.10		0.21	0.27		0.15
		Range	0.13	0.13	0.13	0.11	0.10		0.21	0.21		
	Shape	IncreaseFL	0.14	0.12	0.13		0.10		0.21	0.21		
		Slope	0.14	0.13	0.17		0.12		0.21	0.25		0.17
		IncreaseFM		0.10	0.11				0.17			
		IncreaseML	0.13	0.10	0.11			0.14	0.18	0.21		
		DeviationM								-0.14		
Red	Level	Mean							0.15			
		Maximum							0.14			
		Minimum							0.15	0.15		
	Variability	Range			-0.10					-0.15		
	Shape	IncreaseFL								0.17		
		IncreaseML								0.14		
Green	Level	Mean	-0.10		-0.10							
		Maximum			-0.12							
Blue	Level	Mean	-0.10									
		Minimum	-0.11									
	Shape	IncreaseFL								0.14		
	Variability	CV	0.10									
Combined	Variability	SD										0.16
	Shape	IncreaseFL								0.14		
Milk production										-0.16		

[a] ABS=absolute value of pattern descriptor; Q2=pattern descriptor relative to mean of other quarters same milking; F500= first 500 mL; L500=last 500 mL; Whole=time frame covering whole milking

For both abnormal milk and CM, correlation coefficients of potential predictive variables describing the variability or shape of sensor data patterns were in the same range as that of variables describing the level of sensor data patterns (Table 3). For both abnormal milk and CM, lower and fewer correlation coefficients were

found for the color sensors than for EC, and all except one were based on absolute values. Whereas for the EC sensor, correlation coefficients for abnormal milk and CM tended to involve the same potential predictive variables, for the color patterns different sets of variables were found. Of the three color sensors, blue and green were important for abnormal milk classification and most of the correlation coefficients described the level of the data pattern. The color sensor red was important for CM, involving level, variability, and shape descriptors. Milk production was negatively correlated with CM but did not appear in the selection of 5% most important correlations for abnormal milk.

The highest gain ratios for abnormal milk (0.28, Table 4) and CM (0.31, Table 5) were found for the variable describing the slope of the color sensor green. For abnormal milk this variable was based on the deviation with the expected value based on the average of the two previous milkings and the last 500 mL as pattern time frame. For CM, this variable was based on the deviation with the expected value based on the previous milking also for the last 500 mL of the milking. Whereas the 5% highest gain ratios mainly involved the blue and EC sensor for abnormal milk, for CM it involved mainly the green sensor. For both abnormal milk and CM, there were gain ratios among the highest 5% for variables based on absolute values or on differences with expected values based on other quarters within the same milking (ABS and Q2). But for both outcome types their values tended to be low. Gain ratios were in general higher for variables based on differences with expected values based on previous milkings. For abnormal milk especially expected values based on two previous milkings showed high gain ratios. For CM, the higher gain ratios were more or less distributed evenly over the three comparison types P1, P2, and P3. For abnormal milk classification, the gain ratios of variables that were based on expected values from previous milkings came from variables describing the level or the variability of a sensor measurement pattern, with the minimum value or the slope as main pattern descriptors. For CM, the gain ratios of variables that were based on expected values from previous milkings involved less often level descriptors. Here, the most common descriptor type was the shape of a sensor data pattern, with the most and also highest gain ratios found for the slope as pattern descriptor. Whereas milk production did appear in the selection of 5% most important gain ratios for abnormal milk with a rather high value (0.25, Table 4), it was absent in the selection of 5% most important gain ratios for CM. The red sensor only appeared in the selection of highest gain ratios found for CM but gain ratios for this sensor were rather low (0.06, Table 5). Only one variable of the green sensor was selected for abnormal milk classification. This variable showed also the highest gain ratio found for this outcome type. For CM, this sensor showed more variables important for classification and all of these showed rather high gain ratios.

Table 4. Subset of gain ratios (45 highest values) from 905 potentially predictive variables for abnormal milk

Sensor	Descr.	Patternn	ABS[a]			Q2			P1			P2			P3		
	Type	Desc.	F500	L500	W	F500	L500	W	F500	L500	W	F500	L500	W	F500	L500	W
EC	Level	Mean			0.03												
		Max													0.03		
		Min							0.19		0.26						
	Var.	SD	0.03		0.03	0.05											
		CV	0.03			0.04											
		Range	0.03			0.05											
	Shape	Incr.FL	0.04			0.04											
		Slope	0.05			0.06							0.05				
		Dev.M	0.05							0.04							
Green	Shape	Slope											0.28				
Blue	Level	Mean	0.09	0.04	0.09			0.05									
		Max	0.04		0.04												
		Min	0.25	0.05	0.04		0.03										
		First	0.06		0.06		0.03										
	Shape	Slope											0.25	0.06		0.12	0.03
		Incr.ML										0.17					
Comb.	Level	Mean	0.03														
		Max	0.03		0.03												
		Min															
		First	0.03		0.03												
	Shape	Slope												0.04			
		Incr.ML															
Milk production								0.03						0.25			

[a] ABS=absolute value of pattern descriptor; Q2=pattern descriptor relative to mean of other quarters same milking; P1=pattern descriptor relative to value previous milking; P2=pattern descriptor relative to mean of previous 2 milkings, P3=pattern descriptor relative to the mean of previous 3 milkings; F500=first 500 mL; L500=last 500 mL; W=time frame covering whole milking

Table 5. Subset of gain ratios (45 highest values) from 905 potentially predictive variables for clinical mastitis

Sensor	Descriptor Type	Pattern Descriptor	ABS[a]		Q2	P1			P2			P3		
			L500	Whole	F500	F500	L500	Whole	F500	L500	Whole	F500	L500	Whole
EC	Variability	SD		0.06	0.06									
		CV			0.07									
		Range			0.06									
	Shape	Slope			0.07									
		DeviationM			0.07		0.09			0.17			0.06	
Red	Variability	Range	0.06											
	Shape	IncreaseFL					0.06							
Green	Level	Minimum										0.25		
		First										0.23	0.15	
	Shape	IncreaseFL										0.08		0.23
		Slope					0.31	0.30	0.22	0.29	0.29	0.11	0.15	
		IncreaseML				0.16		0.18	0.08			0.13		
Blue	Level	Mean	0.06	0.07										
	Variability	Range	0.07				0.29							
	Shape	Slope								0.29	0.07		0.09	
		IncreaseML							0.19			0.09		
Combined	Level	Mean	0.06											
		Minimum	0.10	0.11										
	Variability	Range	0.09											
	Shape	Slope							0.08	0.28	0.27			0.07
		IncreaseML										0.08		

[a] ABS=absolute value of pattern descriptor; Q2=pattern descriptor relative to mean of other quarters same milking; P1=pattern descriptor relative to value previous milking; P2=pattern descriptor relative to mean of previous 2 milkings, P3=pattern descriptor relative to the mean of previous 3 milkings; F500=first 500 mL; L500=last 500 mL; Whole=time frame covering whole milking

A subset of 12 potentially predictive variables with high correlation coefficients or high gain ratios in Tables 3–5 were selected and correlation coefficients between each pair of variables were computed (Table 6). Some pairs of variables showed high correlation coefficients, especially those that belonged to the same sensor and involved the same pattern descriptor type. However, all potentially predictive variables had a very low correlation coefficient with at least one other variable.

Table 6. Correlation coefficients between a subset of potentially predictive variables that showed high correlation coefficients or gain ratios with abnormal milk or clinical mastitis (above diagonal) and the number of records involved (below diagonal)

			Time Frame[a]		Whole	Whole	Whole	Whole	F500	F500	L500	L500	L500	L500	L500	L500
			Comparison Type[b]		ABS	ABS	ABS	P1	P3	P3	P1	P2	P1	P2	P2	P2
			Sensor		EC	EC	EC	EC	Green	Green	Green	Green	Blue	Blue	Comb.	Milk prod.
			Descriptor Type		Level	Var.	Shape	Level	Level	Level	Shape	Shape	Shape	Shape	Shape	
			Pattern Descriptor		Max.	SD	Slope	Min.	Min.	First	Slope	Slope	Range	Slope	Slope	
Time Frame[1]	**Comp. Type[2]**	**Sensor**	**Descr. Type**	**Pattern Descr.**												
Whole	ABS	EC	Level	Max.		0.75	0.44	0.23	0.01	0.00	0.03	0.05	0.00	0.18	0.15	0,19
Whole	ABS	EC	Var.	SD	3212		0.79	-0.04	0.05	0.06	0.03	0.05	-0.01	0.19	0.16	0,10
Whole	ABS	EC	Shape	Slope	3202	3182		-0.16	0.04	0.06	0.05	0.07	-0.04	0.14	0.14	-0,07
Whole	P1	EC	Level	Min.	3109	3092	3081		-0.05	-0.09	0.04	0.03	0.01	0.06	0.05	0,25
F500	P3	Green	Level	Min.	2861	2848	2838	2861		0.91	-0.03	-0.05	0.01	-0.01	-0.02	0,00
F500	P3	Green	Level	First	2775	2762	2775	2775	2775		-0.02	-0.05	0.00	-0.01	-0.02	-0,02
L500	P1	Green	Shape	Slope	3109	3092	3081	3109	2861	2775		0.84	-0.34	0.39	0.56	-0,03
L500	P2	Green	Shape	Slope	2988	2973	2961	2988	2861	2775	2988		-0.31	0.47	0.66	-0,02
L500	P1	Blue	Shape	Range	3109	3092	3081	3109	2861	2775	3109	2988		-0.32	-0.35	0,02
L500	P2	Blue	Shape	Slope	2988	2973	2961	2988	2861	2775	2988	2988	2988		0.94	0,14
L500	P2	Comb.	Shape	Slope	2988	2973	2961	2988	2861	2775	2988	2988	2988	2988		0,10
L500	P2	Milk production			2984	2969	2957	2984	2861	2775	2984	2984	2984	2984	2984	

[a] Time frame: Whole=whole milking; F500=first 500ml of milk; L500=last 500ml of milk

[b] Comparison type: ABS=Absolute value of pattern descriptor; P1=pattern descriptor relative to previous milking; P2=pattern descriptor relative to mean previous 2 milkings; P3=pattern descriptor relative to mean previous 3 milkings

Combinations of variables from different sensors showed generally low correlation coefficients. Within a sensor some variable combinations had low correlation coefficients if they differed in pattern descriptor type and in chosen time frame. For example, with the potentially predictive variables involving the sensor for the color green, the correlation coefficient between the potentially predictive variable describing the first value of the first 500 mL relative to the mean of the previous 3 milkings and the potentially predictive variable describing the slope of the last 500 mL relative to the value of the previous milking was only −0.02. The milk production variable in Table 6 showed low correlations with the other potentially predictive variables, especially with those involving the color sensors.

Discussion

The large proportion of cows with visually abnormal milk within a single month suggests that the herd had a severe mastitis problem. This was confirmed by quarter level milk sampling for somatic cell count determination during the data collection period: 32 of the 65 cows had at least one quarter milking with a somatic cell count of more than one million cells per mL. However, using this farm allowed for the capture of many quarter milkings related to abnormal milk and CM in a relatively short amount of sampling time.

The current study used homogeneity scores to define quarter milkings showing abnormal milk and CM. Although scoring homogeneity is subjective and requires experience (Rasmussen, 2005), results showed a general increase in mean values of EC with increasing homogeneity category. As severity of an udder infection increases with homogeneity category, this result agrees with those of Norberg et al. (2004), where healthy quarters showed a statistically significant lower EC than those subclinically or clinically infected. Values from homogeneity category 1 (watery milk without flakes) showed deviating values from this general pattern. This suggests that the order of homogeneity categories used in the current study may not properly reflect an increase in severity of an udder infection. Based on the EC pattern descriptors and the mean milk production in Table 2, homogeneity category 1 might be better placed between homogeneity categories 3 and 5. Alternatively, based on the pattern descriptors mean and slope of the sensors red and blue, homogeneity category 1 should perhaps be considered as a type of abnormal milk (watery), different from homogeneity categories 2 through 5 (milk with an increasing amount of clots).Watery milk without any flakes has been suggested as an indicator for an upcoming severe clinical *Escherichia coli* infection (Hogeveen et al.,1995). For the development of improved classification models for

abnormal milk and CM it might be useful to consider the detection of watery milk as a classification task separate from the detection of milk with clots.
Using more than one evaluation metric is a conservative and recommended approach to evaluate potentially predictive variables and to reduce the chance that useful variables might be erroneously eliminated (Refaat, 2006). The two evaluation metrics used here indeed selected different variable types as potentially predictive. Whereas correlation coefficients favored EC variables, the gain ratios favored the color sensors blue and green. EC has been used frequently in past research (e.g., De Mol et al., 1997; Maatje et al., 1992; Milner et al., 1996; Norberg et al., 2004), and the color sensor blue was indicated as important for CM classification by Espada and Vijverberg (2002).

The evaluation methods used in the current study selected variables that were based on deviances with expected values (Q2, P1, P2, and P3) as potentially predictive. Other studies have used independent variables based on deviances with expected values, whether based on other quarters within the same milking (e.g., Maatje et al., 1992; Nielen et al., 1992; Pyörälä, 2003;Wiedemann and Wendl, 2004) or the same quarter during previous milkings (e.g., Lansbergen et al., 1994). However, potentially predictive variables that were based on absolute values were emphasized by the correlation coefficients. Sensor data patterns, and thus values for potentially predictive variables, of quarter milkings showing abnormal milk or CM for the first time may strongly deviate from patterns from previous, and visually normal, quarter milkings. But this information may become less useful when several and successive milkings are abnormal or have CM, which was the case in this current study. In these situations, the current abnormal quarter milking will influence the expected value for the next quarter milking strongly and the deviance between actual and expected value will be less pronounced. Therefore, plain absolute values might be more predictive for successive cases of abnormal milk or CM.

Pattern descriptors for EC used in previous studies mainly involved variables describing the level of the pattern, e.g., the maximum value within a quarter milking (e.g., Nielen et al., 1995a; Norberg et al., 2004) and the average of the 20 highest values within a quarter milking (e.g., De Mol et al., 1997; Norberg et al., 2004). Although these pattern descriptors were indicated as potentially predictive in the current study as well, our results also indicate that variables describing the variability or shape of a sensor data pattern seem to be as important as these level based variables. This finding confirms results of a study by Norberg et al. (2004), which concluded that additional information was obtained when variables describing variability were included in the model. Additionally, the current study developed variables that have not been reported as independent variables in previous studies, e.g. DeviationM (the deviation of the actual midpoint observation from the expected

midpoint observation) and Slope. These variables were indicated as potentially predictive as well and thus these new pattern descriptors should be considered in future modeling. Finally, results from Table 6 suggest that combinations of potentially predictive variables that are quite independent of each other can be found for future modeling. In addition to using different sensors, potentially predictive variables involving different types of pattern descriptors, different types of comparisons with previous and other quarters, and focusing on different time frames of the quarter milking seem promising.

In the present study, potentially predictive variables were generated by combining visual exploration of measurement patterns and a data flow diagram. The motivation to use and describe this systematic methodology for generating new variables is made intentionally as it has not been widely reported in previous studies. One of the very few studies that mentioned visual exploration of data before the start of model development, was a study conducted by Nielen et al. (1995b). By visual exploration a more solid reasoning behind the use of certain values of sensor data, such as the maximum or mean, can be made. The methodology used in the current study has been very helpful in developing the large amount of potentially predictive variables and in keeping track of the whole process. It will be easy to implement in future, when new sensors are developed, when sensor data patterns are studied in more detail, or when researching other approaches to estimate expected values. Future research regarding the development of potentially predictive variables from within quarter milking patters may focus on additional combinations of patterns and more robust pattern descriptors. Based on the three color patterns red, green, and blue, new combined patterns can be created that might lead to additional potentially predictive variables. More robust pattern descriptors that are less affected by outlier values might yield improved results. For example, the pattern slope could be calculated through linear regression instead of change from first to last observation.

The current study used one definition for CM and one for abnormal milk. However, with the development of detection models for CM in the future, careful consideration must be paid to why and what kind of CM it is that one wants to detect. There was a large amount of overlap between variables selected for abnormal milk and for CM classification. However, due to different conditions and requirements their classification problems are unlikely to be solved with the same classification model. For example, for abnormal milk detection, the classification model must be able to correctly discard a reasonable proportion of abnormal milk without discarding a large amount of normal milk (Pietersma and Hogeveen, 2004; Rasmussen, 2003). However, a classification model for CM needs to identify those cows and quarters likely to have mastitis for a mastitis attention list. This list can potentially include a moderate number of cows without mastitis (false positives)

(Pietersma and Hogeveen, 2004) but the focus must be on developing a model with a high sensitivity.

Conclusion

The analytical framework – involving four data preparation steps to generate new, potentially predictive variables from within quarter milking measurement patterns – was very helpful, because it made working methodically and maintaining an overview of the entire process relatively easy. This framework can be applied to evaluate additional pattern descriptors and to deal with new sensors or derived measurement patterns. Our results suggest that the patterns from the EC sensor and the blue and green sensors contain the most information for abnormal milk and CM classification. Although a number of the identified potentially predictive variables were based on the deviation from expected values, this study indicates that variables based on the absolute value of the pattern descriptor should be considered as well when developing models. It also shows that variables describing the variability and shape of measurement patterns are as important as variables describing the mean or maximum. Potentially predictive variables that differed in sensor, comparison type and pattern descriptor type showed low correlation coefficients between each other. Classification models for abnormal milk and CM may therefore benefit not only from the use of different sensors, but also from data preparation to extract different predictive variables from the data patterns measured within quarter milkings.

Acknowledgements

We gratefully acknowledge the contributions to the acquisition of the data used in this research by Dr. J. Harms of the Bavarian State Research Centre for Agriculture in Freising-Weihenstephan, Germany. Also gratefully acknowledged is Linda McPhee for her editorial input. This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs.

References

De Mol, R.M., G.H. Kroeze, J.M.F.H. Achten, K. Maatje, and W. Rossing. 1997. Results of a multivariate approach to automated oestrus and mastitis detection. Livest. Prod. Sci. 48(3):219–227.

Espada, E., and H. Vijverberg. 2002. Milk color analysis as a tool for the detection of abnormal milk. Pages 28-38 in First North American Conference on Robotic Milking, Toronto, Canada. Wageningen pers, Wageningen, the Netherlands.

Helgren, J.M., and D.J. Reinemann. 2003. Survey of milk quality on United States dairy farms utilizing automatic milking systems. Page 03-3016 in Proceedings of the ASAE 2003 Annual International Meeting, ASAE, St. Joseph, MI.

Hogeveen, H., E.N. Noordhuizen-Stassen, I Thysen, T van Werven, and T.J.G.M. Lam. 1995. Automated pathogen diagnosis at low somatic cell count farms. Pages s2.47-s2.51 in Proceedings of the Third International Mastitis Seminar, Vol. I. A. Saran, and S. Soback (eds.). Tel Aviv, Israel.

Lansbergen, L.M.T.E., M. Nielen, T.J.G.M. Lam, A. Pengov, Y.H. Schukken, and K. Maatje. 1994. Evaluation of a prototype online electrical-conductivity system for detection of subclinical mastitis. J. Dairy Sci. 77(4):1132–1140.

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.

Milner, P., K.L. Page, A.W. Walton, and J.E. Hillerton. 1996. Detection of clinical mastitis by changes in electrical conductivity of foremilk before visible changes in milk. J. Dairy Sci. 79(1):83–86.

Nielen, M., H. Deluyker, Y.H. Schukken, and A. Brand. 1992. Electrical-conductivity of milk—measurement, modifiers, and meta-analysis of mastitis detection performance. J. Dairy Sci. 75(2):606–614.

Nielen, M., Y.H. Schukken, A. Brand, H.A. Deluyker, and K. Maatje. 1995a. Detection of subclinical mastitis from online milking parlor data. J. Dairy Sci. 78(5):1039–1049.

Nielen, M., Y.H. Schukken, A. Brand, S. Haring, and R.T. Ferwerda-van Zonneveld. 1995b. Comparison of analysis techniques for online detection of clinical mastitis. J. Dairy Sci. 78(5):1050–1061.

Norberg, E., H. Hogeveen, I.R. Korsgaard, N.C. Friggens, K.H.M.N. Sloth, and P. Løvendahl. 2004. Electrical conductivity of milk: ability to predict mastitis status. J. Dairy Sci. 87(4):1099–1107.

Ouweltjes, W., and H. Hogeveen. 2001. Detecting abnormal milk through color measuring. Pages 217-219 in National Mastitis Council Annual Meeting.

Pietersma, D., and H. Hogeveen. 2004. Cost of discarding milk with automatic separation of abnormal milk. Pages 221-227 in Automatic Milking: A Better Understanding. A. Meijering, H. Hogeveen, and C.J.A.M. de Koning (eds.). Wageningen Academic Publishers, Wageningen, the Netherlands.

Pyörälä, S. 2003. Indicators of inflammation in the diagnosis of mastitis. Vet. Res. 34(5):565–578.

Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1:81–106.

Rasmussen, M.D. 2003. Definition of normal and abnormal milk at time of milking: Consequences of definitions of acceptable milk quality for the practical use of automatic milking systems. Deliverable 6, EU project QLK5-2000-31006: Implications of the introduction of automatic milking on dairy farms.

Rasmussen, M.D. 2005. Visual scoring of clots in foremilk. J. Dairy Res. 72(4):406–414.

Rasmussen, M.D., and M. Bjerring. 2004. Definition of normal and abnormal milk at time of milking: abilities of automatic milking systems to detect and separate abnormal milk. Deliverable 7. EU project QLK5-2000-31006: Implications of the introduction of automatic milking on dairy farms.

Rasmussen, M.D., and M. Bjerring. 2005. Visual scoring of milk mixed with blood. J. Dairy Res. 72(3):257–263.

Refaat, M. (ed.). 2006. Data Preparation for Data Mining Using SAS, 1st edition. Morgan Kaufmann, San Fransico.

Van der Vorst, Y., K. Knappstein, and M.D. Rasmussen. 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.

Wiedemann, M., and G. Wendl. 2004. The use of spectral photometry for detection of mastitis milk.Pages 228-234 in Automatic Milking: A better Understanding. A. Meijering, H. Hogeveen, and C.J.A.M. de Koning (eds.). Wageningen Academic Publishers, Wageningen, the Netherlands.

Witten, I.H., and E. Frank (eds.). 2005. Data Mining; Practical Machine Learning Tools and Techniques, 2nd edition. Morgan Kaufmann, San Fransisco.

Chapter 3

Decision tree induction to detect clinical mastitis with automatic milking

C. Kamphuis, H. Mollenhorst, A. Feelders, D. Pietersma, H. Hogeveen

Published
Computers and Electronics in Agriculture. 2010. 70 (1): 60-68

Abstract

This study explored the potential of using decision tree induction to develop models for the detection of clinical mastitis with automatic milking. Sensor data (including electrical conductivity and color) of over 711,000 quarter milkings were collected from December 2006 till August 2007 at six Dutch dairy farms milking automatically. Farmer recordings of quarter milkings with visible signs of mastitis were considered as gold standard positive cases (n=97), quarter milkings that were recorded as being visually normal as gold standard negatives (n=339). Randomly chosen quarter milkings that were not visually checked, that were outside a 2-week range before or after a gold standard positive case and that were not manually or automatically separated were added to end up with 3,000 gold standard negatives. Decision trees, with varying confidence factors and cost matrices to study their effect on performance characteristics, were developed with the probability of having clinical mastitis for each quarter milking as output. Detection performance of decision trees was estimated using 10-fold cross-validation. Evaluated performance characteristics were the sensitivity and specificity, both calculated at a threshold value of 0.50 for the probability estimate for clinical mastitis. The transformed partial area under the curve was used to summarize the diagnostic ability of decision trees within a specified range of interest (specificity ≥97%). Receiver operating characteristic curves visualized all combinations of sensitivity and specificity of decision trees within this range. Results showed that decision trees are easy to interpret when visualized. The lower the confidence factor, the smaller the decision trees: a cost insensitive decision tree with a confidence factor of 0.05 needed only eleven test nodes to classify all 3,097 records with a sensitivity of 23.7% and a specificity of 99.2%. The decision tree with default parameter settings showed a transformed partial area under the curve value of 0.642. By introducing costs for false negative classifications this value increased to 0.648. At a specificity level of 99%, the decision tree with the highest transformed partial area under the curve value showed a sensitivity of 29.8%. Detection performances of the different decision trees were comparable with those of models currently used by automatic milking systems. As it was possible to achieve these results with the use of a rather simple decision tree algorithm, we believe that decision tree induction shows potential for detecting clinical mastitis with automatic milking.

Keywords

Robotic milking, data mining, automatic mastitis detection

Introduction

Automatic milking (AM) systems use several in-line sensors that measure different milk features to detect cows with clinical mastitis (CM). Sensor data are used by detection models to generate mastitis alert lists, reporting those cows and quarters likely to have mastitis. These lists are used by the dairy farmer for further decisions, e.g., to check individual cows for having CM at a later stage. For a CM detection model it is important to identify at least those cows with a severe udder infection. These cows need an antibiotic treatment as soon as possible. However, if a cow with a mild CM (some flakes in the milk) is not immediately detected, this will not be a large problem from a veterinary point of view (Hogeveen and Ouweltjes, 2003). Although the sensitivity of current detection models might be improved, dairy farmers complain especially about the high number of alerts on the mastitis lists that turn out to be false. This increases the work-load of dairy farmers unnecessarily, and moreover the dairy farmer may loose trust in the CM detection model resulting in an improper use of the mastitis alert lists.

In the past, research focused on the electrical conductivity (EC) of milk to detect CM because it offered a relatively simple and inexpensive technique for implementation in-line (Hogeveen and Ouweltjes, 2002; Mottram et al., 2007). During a quarter milking, EC is measured and recorded at specific points in time, e.g., every second or after each 100 mL of milk produced. As a first step in a CM detection model, these in-line point measurements are used to derive EC variables, for example the maximum or the mean value. These derived variables have been compared across milkings of the same quarter (Milner et al., 1996) or with other quarter milkings within the same milking or over milkings (Hovinen et al., 2006; Maatje et al., 1992) to detect pattern changes possibly due to CM. Also time series, Kalman filters or the fuzzy logic theorem were used for CM detection (Cavero et al., 2006; De Mol et al., 1999; De Mol and Ouweltjes, 2001). Most of these models used the maximum or the mean value (or a combination) measured within a cow or quarter milking. Kamphuis et al. (2008) suggested that not only the mean or maximum EC value showed potential for CM detection. Instead, detection performance might be improved by combining different predictive variable types, for example, combining the standard deviation of the EC measurements within a quarter milking with the maximum value of the EC measurements of the first 500 mL of milk. Also, usually models were developed using a highly selected group of positive and negative cases of CM (e.g., Friggens et al., 2007). Others used a very wide time window in which a classification by the model was considered as valid (De Mol et al., 1997; Maatje et al., 1992). Performances of these models were good with high levels of sensitivity and/or specificity. However, what is needed in practice is a CM detection model that is able to detect CM within a very limited time

window (preferably at the milking at which CM occurs) and that classifies the large amount of less clear examples of diseased and healthy quarters correctly as well.

One of the major difficulties of developing CM detection models is the fact that sensor data itself are noisy and often incomplete. In addition, the low prevalence of CM results in highly imbalanced data, which makes modeling even more difficult. Data mining is a technique which tries to discover new knowledge in large amounts of (historical) data. The technique itself is often used to improve the decision making process (Mitchell, 1999). Decision tree (DT) induction is a commonly used data mining technique (Quinlan, 1986), often used for classification problems, e.g., whether a cow suffers from CM or not. For example, DT induction has been used to classify mastitis pathogens by Kim and Heald (1999) using variables that were based on expert knowledge. It is believed, in contrast with more traditional statistical approaches, that DT induction is capable of dealing with data that is noisy, imbalanced and/or incomplete.

The objective of this study was to explore the potential of DT induction to generate models for the detection of CM using sensor data collected at farms milking automatically.

Materials and methods

Data collection

Data used in this study were collected at six commercial dairy farms in the Netherlands using a total of nine Lely Astronaut (version A2; Lely Industries N.V., Maassluis, the Netherlands) AM systems. Table 1 summarizes the start date of collecting data, the number of AM systems, the number of unique cows being milked, and the number of quarter milkings per farm that were collected. Figures in the table are based on raw data, before data preparation was applied. 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 dairy 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 1 year. Data collection started at two farms in November 2006. From March 2007 on, data was collected at all six farms. Data collection was continued until August 2007.

Table 1. Farm characteristics of participating dairy farmers

Farm	Start data collection	Number of AM systems	Number of unique cows milked	Number of quarter milkings
1	November 19th 2006	1	123	95,366
2	November 29th 2006	2	164	137,807
3	January 30th 2007	2	159	138,601
4	January 30th 2007	1	131	106,840
5	March 9th 2007	2	158	152,332
6	March 9th 2007	1	103	80,433
Total		9	838	711,379

To collect sensor data, a separate remote computer was installed and connected to each AM system. The first approximately 30 mL of milk of each quarter was discarded as standard procedure by the AM system without passing the in-line sensors. For each cow, milking data included average EC, red, green, and blue values for each quarter and estimated quarter milk yield. Date and time stamps to indicate when a cow entered the AM system, when teat cups were attached and when milk flow started were recorded as well. Finally, average EC, red, green, and blue values were recorded for each 100 mL of quarter milk produced.

To intervene as little as possible with the daily routine to detect CM, it was left to dairy 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 4–6 weekly milk production recording test days or the inspection of the milk filter for clots. Although this resulted in different approaches in which and what number of 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 having a high somatic cell count), quarters that were checked visually were checked uniformly by using a visual assessment protocol. This protocol instructed dairy farmers to visually score the 5th and 6th squirts of milk of these suspected quarters using a clean black paddle as commonly used for the Californian Mastitis Test (without using the soapy reagent normally used when applying this test). When visually normal, the milk was scored with (1). When abnormal, the milk could be scored as (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 that was checked

visually. Every 4–6 weeks, dairy farmers were visited to collect data from the remote computer and the scoring forms. During these visits, the scoring forms were discussed with the dairy farmer to ensure a proper use of the protocol.

Data preparation

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 24h time window prior to the assessment time. From all quarter milkings available, recorded 100 mL values for EC, the colors red, green, and blue, and the derived measurement pattern milk flow (see below) were plotted. Values outside the main cluster of data points were regarded as likely data errors and were set at missing (6.2% of all sensor measurements). Quarter milkings where the AM system failed to connect the teat cups, quarter milkings with missing data for teat cup attachment and 100 mL measurements, and quarter milkings without any data at the quarter milking level (sensor measurement averages, start milk flow, teat cup attachment, etc.) were completely deleted from the data set (5% of all quarter milkings available).

Descriptive variables at the quarter milking level were derived from the 100 mL average sensor data as described in Kamphuis et al. (2008): for four sensors (EC, red, green, blue) and two derived measurement patterns (combined and milk flow), three time frames, five comparison types, and 12 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 quarter milking. For these three measurement patterns, the same five comparison types were used to define descriptive variables. Exact definitions of sensors, derived measurement patterns, time frames, comparison types, descriptor types, and pattern descriptors are summarized in Appendix A. A total of 1,065 descriptive variables were used as independent input variables for the CM detection model.

Decision tree induction

The DT algorithm used in this study was the J48 as implemented in WEKA. This J48 implements the 8th revision of the commonly known C4.5 algorithm developed by Quinlan (1993). A description of this J48 algorithm can be found in Witten and Frank (2005). In short, the J48 algorithm builds a DT, which is a graphic representation of a divide-and-conquer approach of a classification problem. The DT consists of nodes at which a variable is tested. This variable can be a nominal or a numeric one and in the latter case the test usually determines whether a variable value is greater or less than a predetermined constant resulting in a two-way split. The construction of the 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 recursively 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 – that part of the tree stops developing (Witten and Frank, 2005). From all independent variables that are available, the one chosen to split the data at a node is the variable that produces the purest daughter nodes. The information value is used as basis for evaluating different splits. The entropy, a measure expressing the amount of pureness of a data set, is used to compute this information value (Witten and Frank, 2005).

After a DT has been developed and when it is applied to unseen data, the DT computes probability estimates for each class of the independent variable. Thus, in the current study it computes a probability for each quarter milking to be healthy or to have CM. The threshold used by the DT to classify a quarter milking as having CM is 0.50 for the probability estimate for CM. In addition to the default settings of this J48 algorithm, two important parameters were varied to study their effect on detection performance. First the confidence factor, which is a value that is used by the algorithm to prune developed trees. Pruning of a DT is conducted to avoid over-fitting the model on the records used for modeling. A lower confidence factor results in more pruning (Witten and Frank, 2005). The second parameter that was varied was the cost matrix. By introducing costs it is possible to account for the imbalance of positive and negative cases of CM in the dataset. As J48 is eager to build models with the lowest error rate as possible, the introduction of costs for false negative classifications results in more balance between positive and negative cases of CM and forces the algorithm artificially to put more emphasis on classifying more quarter milkings as positive for CM during the training process. By doing this, it is expected to increase sensitivity and to decrease specificity. Default values of the J48 algorithm are equal costs for false negative and false positive classifications and a confidence factor of 0.25.

Gold standard definition of positive and negative cases of clinical mastitis

A combination of dairy farmer's observations and alerts generated by the AM system was used for gold standard definitions. Sensor data were originally collected from over 800 cows and over 711,000 quarter milkings (Table 1). From the visually checked quarters, 518 could be combined with sensor data as described above. During data preparation, 38,614 quarter milkings were deleted (5%) including 81 quarter milkings with an observation. Of these 81 quarter milkings, 21 had an observation of CM. However, all these quarter milkings were unsuccessful and had no sensor measurements at all to be used for development of descriptive variables. There were 437 visual observations left with AM sensor data available (Table 2).

Gold standard positive (GSpos) cases were quarter milkings that received an observational score of 2 through 5 (n=97). Quarter milkings that were checked and scored a 1 were labeled as gold standard negative (GSneg) cases (n=339). There was one quarter milking that was scored a 6 (milk with blood). As milk with blood does not necessarily imply CM but can also be caused by internal damage in the udder or by teat damage, and because there is already an algorithm available to detect blood in milk that performs well (personal communication, P.P.J. van der Tol, Lely Industries N.V.), this milking was excluded from the dataset.

Table 2. Number of quarter milkings per score assigned to quarter milkings visually checked by dairy farmers

Score	Definition	Quarter milkings visually checked
1	Normal milk	339
2	Watery milk	21
3	Flakes	56
4	Clots	17
5	Serum-like milk	3
6	Milk with blood	1
Total		437

The 339 GSneg cases were quarter milkings that were visually checked for CM by the dairy farmer for any reason, e.g., the quarter appeared on previous mastitis alert lists as well, another quarter of the same cow appeared to have CM and the dairy farmer decided to check the other quarters as well, the quarter showed a high somatic cell count on the monthly milk recording system, or they did appear on the mastitis alert list erroneously. Therefore, it was decided to include more quarter milkings that were labeled as being negative for CM. An additional expected profit by including more GSneg cases was that the J48 algorithm could train a DT model with more variation in the sensor data patterns of negative cases of CM. To select these negative examples, those quarter milkings that were not visually checked were all considered as GSneg cases at first. However, to be more confident that these quarter milkings were indeed truly negative, the GSneg label was removed from all cow milkings 2 weeks prior or post a GSpos case (except for those quarter milkings that received a score of 1), from quarter milkings that were separated manually or automatically, and from quarter milkings where separation details were missing. From this large additional pool of GSneg quarter milkings (n=630,330),

2,661 quarter milkings randomly selected to end up with 3,000 GSneg quarter milkings that were included for further analyses.

Validation

In order to estimate the detection performance of a DT when applied to unseen data, stratified 10-fold cross-validation was applied, which has become a standard method in the field of data mining (Witten and Frank, 2005). Cross-validation involves splitting the data randomly in a fixed number of folds (in this case 10) and the stratification makes sure that each fold has approximately the same distribution of positive and negative cases of CM as the overall dataset. Each fold in turn is used for testing (1/10 of the data), while the remainder (9/10 of the data) is used for training. In the end, every record has been used exactly once for testing (Witten and Frank, 2005).

After a quarter milking was used for testing, each DT model developed used the probability estimate for having CM to classify records into healthy or having CM. The default threshold value of 0.50 for these probability estimates – equivalent to a DT that gives a simple binary classification – was used to classify each quarter milking as being positive or negative for CM. A GSpos quarter milking which was classified by the DT model as being positive for CM was considered a true positive (TP) classification. When the DT classified a GSneg quarter milking as negative for CM, this was considered a true negative (TN) classification. A false positive (FP) classification was a GSneg quarter milking that was classified by the DT as being positive for CM. Finally, a false negative (FN) classification was a GSpos quarter milking that was classified by the DT as being negative for CM. These four classification types were used to compute two performance characteristics to compare the different DT models: first, the sensitivity (Se) as the fraction of GSpos quarter milkings that were classified as being positive for CM (Se (%)=100×TP / (TP + FN)) by the developed DT models. Second, the specificity (Sp) as the fraction of GSneg quarter milkings that were classified as being negative for CM (Sp (%)=100×TN / (TN + FP)).

Receiver operating characteristic (ROC) curves were constructed to visualize the performance of the developed DT models (Detilleux et al., 1999; Witten and Frank, 2005). Data used to create the ROC-curves were based on the probability a record received when it was part of a test fold, using the DT model based on the remaining folds used for training. ROC-curves are graphic representations of the TP rate (or Se) versus the FP rate (or 1-Sp) over the whole range of possible threshold values (Detilleux et al., 1999). To summarize the ROC-curves into a single quantity, the total area under the curve (AUC) is a commonly used measure. However, for the classification problem at hand it is not useful to know a model's detection performance at Sp values that are too low. In other words, it is of no use knowing

the TP rate of a model when you know beforehand that the corresponding FP rate is too high resulting in too many quarters occurring on the mastitis alert list falsely. Therefore, instead of the total AUC, the partial AUC was computed in this study. This partial AUC is restricted to just a relevant portion of the total AUC (Detilleux et al., 1999; Walter, 2005). The interpretation of the partial AUC is the probability that a randomly picked GSpos and a GSneg quarter milking are correctly ranked according to their probability estimate for having CM, conditional on the GSneg probability estimate falling within the restricted range of the curve (Walter, 2005). In this study, the partial AUC is calculated using the trapezoidal rule for an FP 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 using the following formula (McClish, 1989):

$$½ [1 + (A - Min) / (max - min)] \quad (1)$$

where A is the partial AUC as computed by the trapezoidal rule, Min is the area of the triangle area with corners (0.0, 0.0), (0.0, 0.03), and (0.03, 0.03), or in other words the triangle bounded above by the line y=x (representing a random classifier) and below by the horizontal axis. Max is the area of the rectangle with corners (0.0, 0.0), (0.0, 1.0), (0.03, 1.0), and (0.03, 0.0).

Data preparation was done using SAS version 9.1 (SAS Institute Inc., Cary, NC). Development of several DTs with probability estimates for each quarter milking for having CM and the computation of Se 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 to construct ROC-curves in TIBCO S+ version 8.1 (TIBCO Software Inc.).

Results

Figure 1 visualizes a complete cost insensitive DT (equal costs for false negative and false positive classifications) with a confidence factor of 0.05. It is clear that the DT combined tests based on different sensors, variables, and comparisons to end up with a classification for having CM or not. The tree consists of 23 nodes: eleven nodes at which a variable is tested (ovals) and 12 leaf nodes where records are classified as having CM or being healthy (rectangles) based on a threshold of 0.50 for the probability of having CM. The figures between brackets in these rectangular nodes show decimals due to the way the J48 algorithm deals with records with missing values for the variables on which a test at a node applies (Witten and Frank, 2005). The DT starts with a test on the maximum value of EC

measurements in the last 500 mL of a quarter milking when compared with other quarters in the same cow milking (for more details on sensor, variable and comparison type description, see Appendix A).

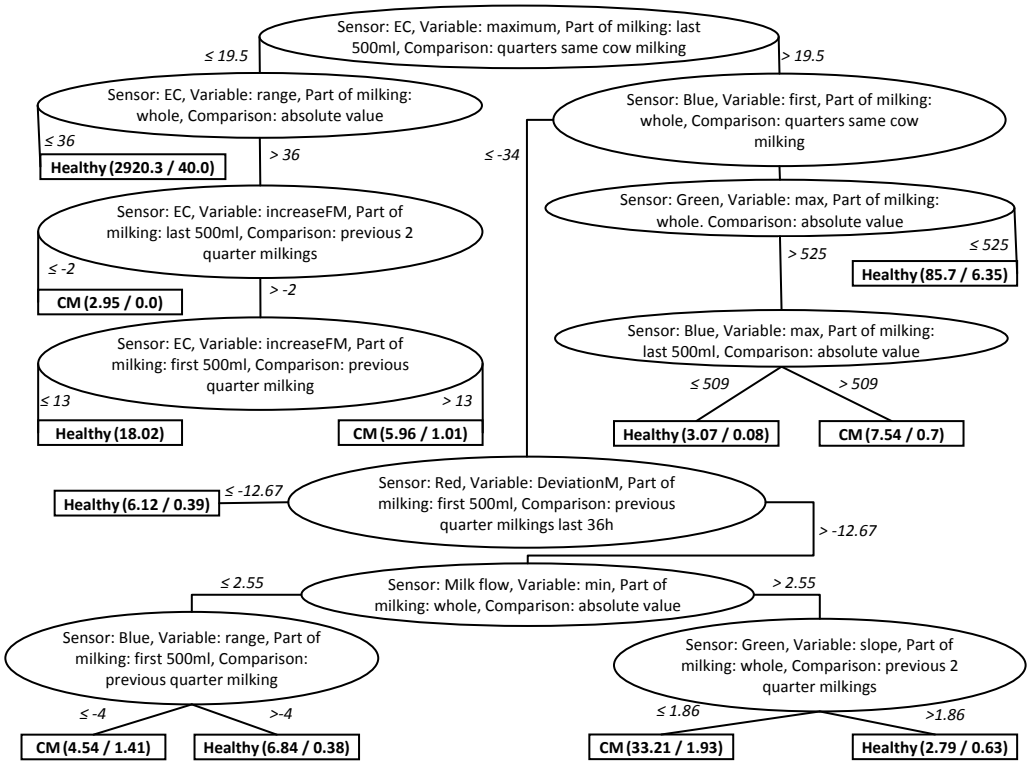


Figure 1. Example of a cost insensitive decision tree (equal costs for false negative and false positive classifications) and a confidence factor of 0.05. The ovals represent nodes at which a variable is tested, the rectangles represent leaf nodes with a classification (being healthy or having clinical mastitis (CM)) for those records reaching this rectangle. The first figure between brackets presented in the rectangles summarize the total number of records reaching this rectangle that are classified as healthy or CM, the second figure represents the number of records that are misclassified by the DT. For a better understanding of variable, sensor and comparison type, see Appendix A

This first test results in two branches ending both in another test node. Records following the left-sided branch are tested against another variable based on EC measurements. This test then results into a leaf node at which 2,920.3 records are classified as being healthy, of which 40.0 records are classified erroneously (false negative). Before records are being classified as having CM, some more tests are applied. For example, at the bottom of the tree, after passing five test nodes, 33.21 records are classified as having CM of which 1.93 are false positive. Figures presented in leaf nodes are used by the DT to calculate probability estimates for CM.

When the DT in Figure 1 is applied to unseen data and encounters a record that has a value of ≤19.5 for the variable used at the root node, and a value ≤36 for the second test node, this record will end up into the leaf node described above. The DT will then assign a probability of CM of 1.37% to this record. This value is calculated as 40.0 divided by 2,920.3. It becomes more complicated when records have missing values for the variables used at test nodes. A record with missing values for variables used at test nodes is proportionally split into pieces – or weights – and the parts are sent down each branch and from there right on down to the leaves of the sub-trees involved. The split into pieces is accomplished by using a numeric weight between zero and one, and the weight for a branch is chosen to be proportional to the number of records going down that branch that do have a value for the variable used at the test node. A split record may be further split at a lower node. Eventually, the various parts – or weights – of the record will reach leaf nodes and the final CM probability estimate for these split records is computed by multiplying the probabilities for CM at the decision nodes by the weight of that record that reaches that specific decision node (Witten and Frank, 2005).

Table 3 summarizes the number of test nodes, the expected detection performance measures (Se and Sp) and the transformed partial AUC of one cost insensitive DT and three cost sensitive trees where costs of false negative classifications are 0.5, 20, or 100 times higher than false positive classifications. For all these trees, confidence factors are varied from 0.05 to 0.35. As expected, trees were smaller as confidence factors decreased. The cost sensitive trees where costs for false negative classifications are 20 or 100 times higher than false positive classifications were larger than the cost insensitive ones, whereas the cost sensitive trees where costs for false negative classifications are 0.5 time higher were smaller. This was expected as well: the DT with costs of 20 or 100 wants to have more certainty before it classifies a quarter milking as being negative which is expressed in more tests before a final classification is made. Similarly, the cost sensitive tree with costs of 0.5 needs less certainty before it classifies a quarter milking as negative; therefore it needs fewer tests to classify all records. The result that cost sensitive trees with costs of 100 are smaller than the cost sensitive trees with costs

of 20 can be explained by the artificially introduced balance between positive and negative cases of CM. By introducing costs of 20 for false negative classifications, one artificially increases the number of GSpos 20 times. In other words, a dataset with 3000 GSneg and 2000 GSpos is created. Similarly, introducing costs of 100 creates a dataset with 3000 GSneg and 10,000 GSpos. In general, the more positive and negative examples in a dataset are in balance, the larger the DT will be. Introducing costs of 20 results in the most balanced dataset in this study and the DT models are therefore the largest trees.

Table 3. Number of test nodes at which a variable is tested (nodes), Sensitivity (Se), Specificity (Sp), and the transformed partial area under the curve (AUC) for one cost insensitive tree (equal costs for false negative and false positive classifications) and three cost sensitive trees (costs for false negative classifications 0.5, 20, or 100 times higher than for false positive classifications). For all tree types, the confidence factor was varied between 0.05 and 0.35. A threshold of 0.50 for the probability estimate was used to classify records for clinical mastitis for calculation of Se and Sp. The transformed partial AUC is calculated for an Sp of 97% or higher

Tree type	Confidence factor	Nodes (n)	Se (%)	Sp (%)	Transformed Partial AUC
Cost Insensitive	0.35	29	32.0	98.4	0.6244
	0.25	26	32.0	98.7	0.6420
	0.15	18	29.9	98.9	0.6417
	0.05	11	23.7	99.2	0.6424
Cost Sensitive (0.5)	0.35	12	21.6	99.4	0.6428
	0.25	7	20.6	99.7	0.6476
	0.15	5	19.6	99.7	0.6401
	0.05	5	5.2	99.8	0.5576
Cost Sensitive (20)	0.35	55	43.3	96.4	0.6061
	0.25	53	44.3	96.3	0.6012
	0.15	51	44.3	96.3	0.5971
	0.05	45	43.3	96.2	0.5859
Cost Sensitive (100)	0.35	42	50.5	94.7	0.6200
	0.25	40	51.5	94.5	0.6192
	0.15	27	52.6	94.0	0.6174
	0.05	26	56.7	93.1	0.6062

The most extreme situation is found for the cost sensitive DT where costs for false negatives classifications are 0.5 and the confidence factor is 0.15 or 0.05. Both these DT models are heavily pruned and need only 5 tests to classify all quarter milkings. Especially the latter tree with a confidence factor of 0.05 shows an extreme low Se (5.2%). The cost insensitive DT with a confidence factor of 0.25 (default settings) resulted in an Se of 32.0% and an Sp of 98.7%. Decreasing the confidence factor from 0.35 to 0.05 in the cost insensitive trees resulted in a decrease in Se of 8.3% whereas the Sp increases slightly with 0.8%. A similar trend is seen in the cost sensitive tree where costs for false negative classifications were 0.5. This decreasing trend for the Se and increasing trend for the Sp was not seen in the cost sensitive trees where costs for false negative classification are 20 or 100. In these trees, the Se remained constant or increased and the Sp decreased as confidence factors decreased. As expected, when costs for false negative classification increased there was an increase in Se together with a decrease in Sp. The DT with default settings showed a transformed partial AUC value of 0.6420. By changing parameters, the transformed partial AUC value increased just slightly to a value of 0.6476 for the cost sensitive tree where costs for false negative classifications are 0.5 and the confidence factor is 0.25. Values of the transformed partial AUC were similar between the cost insensitive trees and the cost sensitive trees where costs for false negative classifications are 0.5, with the exception of the cost sensitive tree with a confidence factor of 0.05. This DT showed the lowest transformed partial AUC (0.5576). In general, values for the transformed partial AUC were lower for the cost sensitive trees where costs for false negative classifications are 20 or 100.

Figure 2 plots the partial ROC-curves of three different trees for the area of interest only — that is an FP rate range of 0.00–0.03 or Sp values of 97% or higher. One DT has the default settings, one has the highest transformed partial AUC value (cost sensitive 0.5, confidence factor 0.25), and there is one DT (cost sensitive 100, confidence factor 0.25) that shows parts of its ROC-curve that lie above the other two ROC-curves in the area of interest. The figure makes clear that, by keeping the confidence factor at the same level (all shown trees have a confidence factor of 0.25) it was possible to vary the detection performance by introducing costs. For example, both the DT with default settings and the cost sensitive tree with cost for false negative classifications of 0.5 have a similar transformed partial AUC value (see Table 3). Figure 2, however, clearly shows that the latter DT especially detects CM at higher TP rates (or Se levels) until an FP rate of 0.013 (or Sp=98.7%). At an FP rate of 0.01 (Sp=of 99%), the Se level of this DT with the highest transformed partial AUC is 29.8%. At the same Sp level, the Se is 21.7% for the DT with default settings and 19.5% for the DT with costs for false negative classifications of 100 and a confidence factor of 0.25. From an FP rate of

0.022 onwards (or Sp ≥ 97.8%), the DT with default settings showed higher TP rates (or Se values). At the maximum FP rate (Sp=97%), the Se of the DT with default settings is 47.4%. Sensitivities for the cost sensitive DT with costs for false negative classifications of 0.5 and for the cost sensitive DT with costs for false negative classifications of 100 were 40.8% and 39.2%, respectively. In between the FP rates of 0.013 and 0.022 (Sp ranging from 97.8 to 98.7%), the DT with costs for false negative classifications of 100 shows the highest Se levels, although the transformed partial AUC value was 0.6192 (Table 3). This indicated that when someone is interested in a specific FP rate range, it might very well be that it is more interesting to choose another DT than the one reported to have the highest transformed partial AUC value within the area of interest.

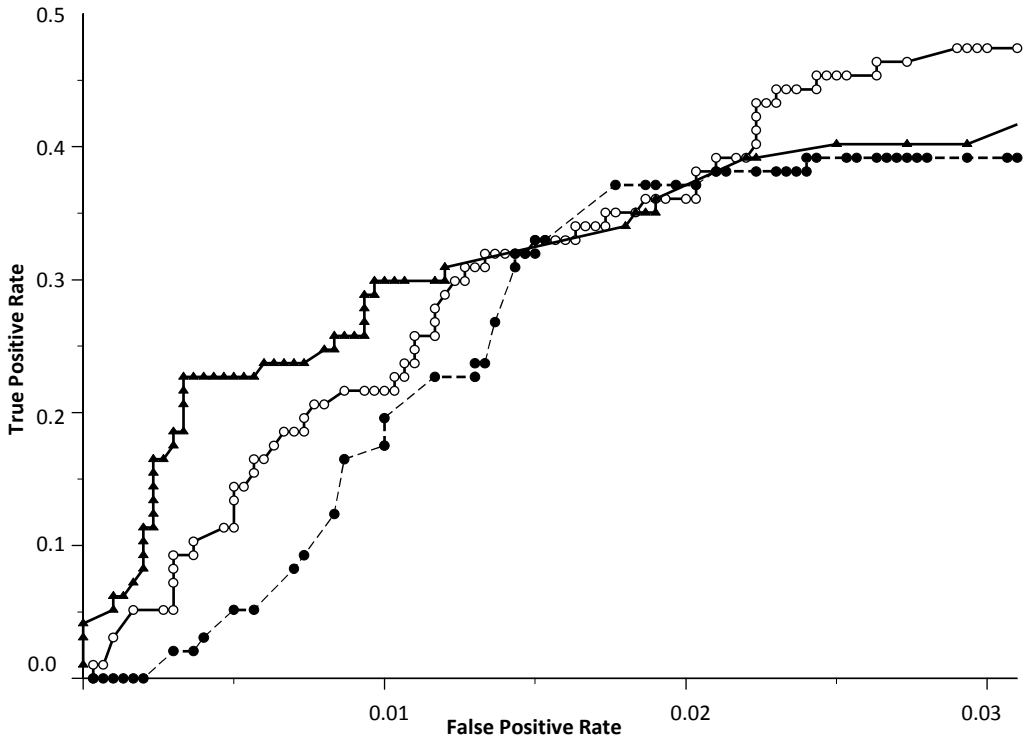


Figure 2. ROC-curves of three decision trees: default settings (○), highest transformed partial AUC (▲, costs for false negative classifications of 0.5 and a confidence factor of 0.25), and a tree with parts of its ROC-curve that lie above the other two ROC-curves in the area of interest (●, costs for false negative classifications of 100 and a confidence factor of 0.25)

Discussion

Figure 1 shows that DT induction reveals an easy to understand model when visualized. This in contrast with some highly sophisticated models using smoothing, fuzzy logic theorems or Kalman filters (e.g., Cavero et al., 2006; De Mol et al., 1999) that are more difficult to interpret. The visualized DT in Figure 1 is the most pruned cost insensitive tree. This DT needed only 11 simple tests to classify all 3,097 records with an Se of 23.7% and an Sp of 99.2% at a threshold value of 0.50 for the probability estimate to classify quarter milkings for having CM or not. These performance estimates are based on a dataset where 1 out of each 30.9 quarter milkings has CM (a prevalence of 3.1%). This prevalence of CM is higher than published in previous studies. A study on CM detection with AM conducted by Mollenhorst and Hogeveen (2008, unpublished data) reported a prevalence of CM of 0.59%, an Se of 36.8%, and an Sp of 97.9%. The DT with the highest transformed partial AUC value in our study showed, at the same Sp level, an Se of 38.1%, indicating that DT induction shows a similar detection performance than models currently used by AM systems.

It is true that previous studies reported higher Se and/or Sp values than found in the current study. However, one has to realize that time windows used were far too wide for practical implications in several studies. There were studies in which classifications 14 days preceding an actual CM case or classifications several days after the actual CM occurred (e.g., De Mol et al., 1997; Maatje et al., 1992) were considered as TP classifications. Of course, the larger the used time window, the better detection performance of models will be (Sherlock et al., 2008). However, 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. One of the few studies with very narrow time windows is conducted by Nielen et al. (1995), where the milking of interest was the milking at which mastitis (defined by visually abnormal milk) was observed (the same short time window as used in the current study). The developed neural network showed an Se of 84% and an Sp of 97%. The Se of the DT with the highest transformed partial AUC in this study was 44% at the same Sp level. However, the neural network as used to predict the results of the same data that were used for model building. This approach will over-estimate the performance of the neural network when applied to truly unseen data. To prevent over-fitting of the DT models developed in the current study, 10-fold cross-validation was applied.

In addition to the used time window, performance of CM detection models largely depends on the chosen gold standard used to define healthy or diseased (CM) quarters. The ideal gold standard would be one where data used to observe CM cases is completely independent from data that is used for model development,

and in addition, that this gold standard would be applied to all quarter milkings to label them correctly in GSneg and GSpos milkings. However, this approach would be far too expensive and time consuming because it would take continuous monitoring of 6 farms for 8 months. Therefore, the current study collected data on a large number of CM observations using a more practical solution—that is, observations of CM were collected by the dairy farmers themselves. Farmers were asked to record observations of each quarter milking they visually checked by using the scoring protocol, but it was left to the dairy farmers themselves to decide which quarters were suspicious enough to check. This suspicion could be based on the mastitis alert lists, but as 82.6% of the analyzed 436 visually checked quarter milkings did not receive an attention by the AM system, there is a strong indication that dairy farmers also used other sources (e.g., clots on the filter) to detect CM cases. This strengthened the assumption that it is a valid approach to use dairy farmer's observations for the gold standard definition.

In the current study, 81 unsuccessful quarter milkings with an observation were excluded during the data preparation steps, of which 21 had an observed CM. It stays subject to debate whether it is fair to remove these unsuccessful quarter milkings from the dataset, as they might contain information about CM very well suitable for CM detection models. However, the goal of this paper was to explore the potential of DT induction to develop CM detection models using sensor measurements. As none of the 81 excluded quarter milkings with an observation had any sensor measurement available, it was considered no problem to exclude these milkings from analyses.

There were also 2,661 quarter milkings included in the analyses of which it was not 100% sure whether they were truly negative for CM. We did this to give the J48 algorithm the opportunity to train a DT model on the whole range of possible variation within the group of quarter milkings that were negative for CM and to include quarter milkings with sensor values that were not considered as abnormal by the currently used CM detection model. The additional negative examples of CM were those quarter milkings of which we were most confident that they were truly negative, as they had to be outside a 2-week range before and after a CM case and they were not separated automatically or manually for any reason. Still, this approach did not rule out that quarter milkings were labeled as negative erroneously. The fact that we did include these quarter milkings might have been the reason why results in this study are much lower than for example a study conducted by Friggens et al. (2007). This study showed very high levels of Se and Sp (92.8% and 97.9%, respectively) but their analysis was based on a very specific dataset — that is, a dataset that included only clear examples of cows being healthy or diseased. Developing a model with such clear examples will result in models that will very well be able to classify unseen data that include similar clear

examples. This is a pitfall in sensor evaluation research in general, as field data never exists of these clear examples alone especially when dealing with data from farms where subclinical mastitis is of high prevalence. It would be of great interest to study their model's performance when applied to unseen real field data.

In general, all DT models showed a transformed partial AUC value higher than 0.5, indicating that all DT models showed better detection performance than a random classifier. Computing the transformed partial AUC showed worthwhile differences between the DT models that would have stayed obscured when presenting performance characteristics that are based on one single threshold value only.

Given the short time window and the chosen gold standard definition used in the current study, our simple DT algorithm showed promising results for the area of interest. This still does not mean that a model with an Sp level of 97% and an Se level of 47.4% (as shown by the DT with default settings in Figure 2) is good enough to be implemented in practice. An Sp of 99% has been suggested by Mein and Rasmussen (2008) to be the value that models should achieve when applied in practice. At this Sp level, our cost sensitive DT with the highest transformed partial AUC value showed an Se of 29.8% (Figure 2). However, detection performance may be improved by including more examples of quarter milkings being positive for CM, or by using more advanced classification techniques within DT induction such as boosting and bagging (Witten and Frank, 2005).

Until now, it has been more or less common practice to evaluate CM detection models based on their Se and Sp levels. Both characteristics are test characteristics and they are independent of the CM prevalence on a farm. However, these evaluation characteristics are not easy to understand, especially not for those that do not use these terms on a daily basis. In addition, for the dairy farmers using AM, it would be of more interest to know the performance of a CM detection model based on the CM prevalence in their own farm. Therefore, it has been suggested by Sherlock et al. (2008) to use the Success Rate and the False Alert Rate per 1,000 cow milkings to evaluate and compare different CM detection models. These characteristics are easier to understand by dairy farmers and, as they are prevalence dependent, they will give a more practical interpretation of the CM detection model on their specific farm.

The developed DT models in the current study are not over-performing and there is still a long way to go before DT models can be applied in practice. However, the goal of this study was to proof the principle of DT induction showing potential for the development of CM detection models. Results of the current study do confirm this potential. Future research has to answer the question whether more advanced classification techniques will improve the ability of DT models to detect CM when applied on unseen field data. Another question that needs further

research is whether it is better to build a model on examples that are clearly positive or negative for CM or to build a model with data including less clear examples of especially negative cases for CM as well.

Conclusion

Detection performances of the different DT models are comparable with detection performance of models currently used by AM systems. These results were achieved with the use of field data, including quarter milkings of which we were not certain whether they were truly negative, the use of a very narrow time window, and a rather simple DT algorithm. Therefore, results suggest that DT induction shows potential for detecting CM using AM sensor data.

Acknowledgements

We gratefully acknowledge the participating dairy farmers for their contribution to the acquisition of the data used in this research. Also greatly acknowledged is Lely Industries N.V. (Maassluis, The Netherlands) for providing addresses of dairy farmers working with AM and their support with the installation of the data acquisition systems at the participating farms. This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs.

Appendix A

Definitions of sensors and their derived measurement patterns, time frames, comparison types, and type and definition of measurement pattern descriptors based on the data flow diagram described in Kamphuis et al. (2008)

Label		Definition
Sensor (100 mL)		**Sensor monitoring and recording values at each 100 mL of milk produced. 100 mL measurements are used to calculate pattern descriptors (see below)**
	EC	Electrical conductivity measured in indexes
	Red	Color red measured in indexes
	Green	Color green measured in indexes
	Blue	Color blue measured in indexes
Derived measurement patterns		**Measurement patterns that are used to compute descriptive variables derived from sensors**
	Combined	Average value from the color sensors red, green, and blue. Representing an 'average color' of the milk. Measurement points are each 100 mL of milk
	Milk flow	Milk production in mL/second within a quarter milking. Measurement points are each 100 mL of milk. Derived from the sensors that estimate yield and the sensor that represents the date and time each time approximately 100 mL of milk was produced
	Milk flow delay	Time between the teat being cleaned and the start of the milk flow. Both are sensors that have a binary outcome and date and time of both sensors are registered when changing from 0 (e.g., milk flow not started) to 1 (milk flow started). The difference between the date and time stamps of these two sensors is the milk flow delay. This involves one value, representing a value for the whole quarter milking
	Dead milking time	Time between teat cup attachment and start of milk flow. Both are sensors that have a binary outcome and data and time of both sensors are registered when changing from 0 (e.g., teat cup is not attached) to 1 (teat cup is attached). The difference between the date and time stamps of these two sensors is the dead milking time. This involves one value, representing a value for the whole quarter milking
	Milk production	The total quarter milk yield (determined by a sensor at the end of a quarter milking) divided by the time since the previous milking. This involves one value, representing a value for the whole quarter milking
	F500	First 500 mL of a quarter milking
	L500	Last 500 mL of a quarter milking
	Whole	All 100 mL measurements of a quarter milking

Label		Definition *(continued)*
Time frame		**Specifying part of a quarter milking for subsequent processing**
Comparison type		**The actual value of a pattern descriptor can be used directly to compute predictive variables but they can also be compared with an expected value to give relative deviations**
	ABS	Absolute values for a pattern descriptor
	Q2[a]	Expected value based on the average of the two quarters with the lowest (or highest) values of the current milking
	P1	Expected value based on the value of the previous milking of the same quarter
	P2	Expected value based on the average of the two previous milkings of the same quarter
	PR36	Expected value based on the average of all previous milkings within 36 hours of the same quarter
Descriptor type and pattern descriptors		**Describing specific aspects of sensor data patterns. Level, variability and shape are descriptor types, describing a general part of a sensor data pattern. Within each of these three pattern descriptor types, pattern descriptors describe parts of a sensor pattern in more detail**
Descriptor Type	**Pattern descriptor**	
Level	*Mean*	Mean of observations within specified time frame
	Maximum	Maximum observation value within specified time frame
	Minimum	Minimum observation value within specified time frame
	First[b]	Value of first observation within specified time frame
Variability	*SD*	Standard deviation
	CV	Coefficient of variation (100 * SD / Mean)
	Range	Maximum – Minimum
Shape	*IncreaseFL*	Increase from first to last observation: value last – value first observation
	Slope	IncreaseFL / number of observations
	IncreaseFM	Increase from first to midpoint observation: value at midpoint – value first observation
	IncreaseML	Increase from midpoint to last observation: value last – value at midpoint observation
	DeviationM	Deviation of midpoint observation from expected midpoint: value at midpoint – (First + Last observation) / 2

[a] As EC, dead milking time, and milk flow delay increase as CM becomes more severe, the pattern descriptors that were based on the Q2 comparison were computed using the two quarters with the lowest values measured within a cow milking. Q2 variables for the other sensors and derived measurement patterns were based on the average values of the two quarters with the highest values measured within a cow milking, as these values decrease with increasing infection severity.
[b] As values for the pattern descriptor 'First' are the same between the time frame 'Whole' and 'F500' for all 6 sensors for which the pattern descriptors is calculated, the variables describing the first measurement for the first 500 mL of milk were excluded.

References

Cavero, D., K.H. Tölle, C. Buxadé, and J. Krieter. 2006. Mastitis detection in dairy cows by application of fuzzy logic. Livestock Science. 105(1–3):207–213.

De Mol, R.M., A. Keen, G.H. Kroeze, and J.M.F.H. Achten. 1999. Description of a detection model for oestrus and diseases in dairy cattle based on time series analysis combined with a Kalman filter. Computers and Electronics in Agriculture. 22(2–3):171–185.

De Mol, R.M., G.H.Kroeze, J.M.F.H Achten, K. Maatje, and W. Rossing. 1997. Results of a multivariate approach to automated oestrus and mastitis detection. Livest. Prod. Sci. 48(3):219–227.

De Mol, R.M., and W. Ouweltjes. 2001. Detection model for mastitis in cows milked in an automatic milking system. Prev. Vet. Med. 49(1–2):71–82.

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

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

Hogeveen, H., and W. Ouweltjes. 2002. Mastitis therapy and control: automatic online detection of abnormal milk. Pages 1735-1740 in Encyclopedia of Dairy Sciences. R. Roginski, J.W. Fuquay, and P.F. Fox (eds.). Academic Press, London, United Kingdom.

Hogeveen, H., and W. Ouweltjes. 2003. Sensors and management support in high technology milking. J. Dairy Sci. 86(5): 1–10.

Hovinen, M., A.-M. Aisla, and S. Pyörälä. 2006. Accuracy and reliability of mastitis detection with electrical conductivity and milk colour measurement in automatic milking. Acta Agriculturae Scandinavica Section A. 56(3):121–127.

Kamphuis, C., D. Pietersma, R.P.P. van der, 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(2):169–181.

Kim, T., and C.W. Heald. 1999. Inducing inference rules for the classification of bovine mastitis. Computers and Electronics in Agriculture. 23(1):27–42.

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 Control - From science to practice. T.J.G.M. Lam (ed.). Wageningen Academic Publishers, Wageningen, the Netherlands.

Milner, P., K.L. Page, A.W. Walton, and J.E. Hillerton. 1996. Detection of clinical mastitis by changes in electrical conductivity of foremilk before visible changes in milk. J. Dairy Sci. 79(1):83–86.

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

Mottram, T., A. Rudnitskaya, A. Legin, J.L. Fitzpatrick, and P.D. Eckersall. 2007. Evaluation of a novel chemical sensor system to detect clinical mastitis in bovine milk. Biosensors & Bioelectronics. 22(11):2689–2693.

Nielen, M., Y.H. Schukken, A. Brand, S. Haring, and R.T. Ferwerda-van Zonneveld. 1995. Comparison of Analysis Techniques for Online Detection of Clinical mastitis. J. Dairy Sci. 78(5):1050–1061.

Quinlan, J.R. 1986. Induction of decision trees. Machine Learning. 1:81–106.

Quinlan, J.R. (ed.). 1993. C4.5 Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

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. Pages 275-282 in Mastitis Control—From Science to Practice. T.J.G.M Lam (ed.), Wageningen Acadamic Publishers, Wageningen, the Netherlands.

Walter, S.D. 2005. The partial area under the summary ROC curve. Statistics in Medicine. 24(13):2025–2040.

Witten, I.H., and E. Frank (eds.). 2005. Data Mining; Practical Machine Learning Tools and Techniques, 2nd edition. Morgan Kaufmann Publishers. San Fransisco.

Chapter 4

Detection of clinical mastitis with sensor data from automatic milking systems is improved by using decision tree induction

C. Kamphuis, H. Mollenhorst, J.A.P. Heesterbeek, H. Hogeveen

Published
Journal of Dairy Science. 2010. 93 (8): 3616-3627

Abstract

The objective was to develop and validate a clinical mastitis (CM) detection model by means of decision tree induction. For dairy farmers milking with an automatic milking (AM) system, it is desirable that the detection model has a high level of sensitivity (Se), especially for more severe cases of CM, at a very high specificity (Sp). In addition, an alert for CM should be generated preferably at the quarter milking (QM) at which the CM infection is visible for the first time. Data were collected from nine Dutch dairy farms milking automatically during a 2.5-yr period. Data included sensor data (electrical conductivity, color, and yield) at the QM level and visual observations of quarters with CM recorded by the dairy farmers. Visual observations of quarters with CM were combined with sensor data of the most recent automatic milking recorded for that same quarter, within a 24h time window before the visual assessment time. Sensor data of 3.5 million QM were collected, of which 348 QM were combined with a CM observation. Data were divided into a training set, including two-thirds of all data, and a test set. Cows in the training set were not included in the test set and vice versa. A decision tree model was trained using only clear examples of healthy (n=24,717) or diseased (n=243) QM. The model was tested on 105 QM with CM and a random sample of 50,000 QM without CM. While keeping the Se at a level comparable to that of models currently used by AM systems, the decision tree model was able to decrease the number of false positive alerts by more than 50%. At an Sp of 99%, 40% of the CM cases were detected. Sixty-four percent of the severe CM cases were detected and only 12.5% of the CM that were scored as watery milk. The Se increased considerably from 40% to 66.7% when the time window increased from less than 24h before the CM observation, to a time window from 24h before to 24h after the CM observation. Even at very wide time windows, however, it was impossible to reach an Se of 100%. This indicates the inability to detect all CM cases based on sensor data alone. Sensitivity levels varied largely when the decision tree was validated per farm. This trend was confirmed when decision trees were trained using data from eight farms and tested on data from the ninth farm. This indicates that when using the decision tree as a generic CM detection model in practice, some farms will continue having difficulties in detecting CM using mastitis alert lists, whereas others will perform well.

Keywords

Clinical mastitis detection, robotic milking, sensor data, decision tree induction

Introduction

Mastitis is one of the most frequent and costly diseases in dairy cows (e.g., Halasa et al., 2007). Detecting clinical mastitis (CM) is important to maintain an acceptable level of milk quality, to initiate an antibiotic treatment when necessary, and to safeguard the welfare of a dairy farm (Kamphuis et al., 2008b). When using automatic milking (AM) systems, dairy farmers are not present during the milking process to check the cow and the milk visually for CM. Instead, they detect cases of CM by means of several information sources. The mastitis alert list, noting those cows and quarters that are likely to have CM, is one of these information sources. These lists are the output of a CM detection model available at the AM system. As input for such a detection model, the AM system uses in-line sensor information, with electrical conductivity being the trait most commonly measured in-line. Although dairy farmers who use AM systems are able to manage udder health sufficiently (De Koning et al., 2004), udder health is under pressure on farms with AM systems (Dohmen et al., 2010). One way to improve udder health on farms with an AM system is to improve the sensitivity (Se) of the currently available mastitis detection models. In addition, specificity (Sp) levels need improvement, as dairy farmers are eager to minimize the additional labor of checking large numbers of false positive alerts, even at the expense of missing some true cases of CM (Claycomb et al., 2009). A level of 99% for the Sp and a minimum level of 70% for the Se have been suggested as goal for CM detection models when applied in practice (Mein and Rasmussen, 2008).

Several models to detect mastitis using sensor information have been developed in the past and some studies have reported high levels of Se and Sp. However, few of these models are implemented in a practical AM setting. There are several reasons why these models may not be applied in practice, such as the wide time windows used in some studies and the fact that some models were trained and validated using only clear cases of healthy and mastitic quarters or cows. Some of the earlier studies developing a mastitis detection model based on electrical conductivity information used very long time windows; for example, De Mol et al. (1997) used a time window of 10 d before the actual CM case until 7 d after. Using wide time windows will result in models showing a good detection performance (Sherlock et al., 2008), but dairy farmers need a CM alert within a very limited period before or only at the milking when CM occurs (Hogeveen and Ouweltjes, 2003). Friggens et al. (2007) developed and validated their CM detection models with highly selected data including only those cows that clearly had CM and those that were clearly healthy. Such validation, however, will overestimate the detection performance of the model because field data include data from cows (or quarter)

milkings that have a less clear mastitis status as well. A CM detection model should be able to deal with this so-called gray area.

One of the major problems of developing a CM detection model using sensor data is that the data are often noisy and incomplete. In addition, the low prevalence of CM results in highly imbalanced data, which makes modeling even more difficult. These problems make it worthwhile to look at tools other than the traditionally used statistical approaches to develop a detection model. One of these potential tools is data mining, which is the process of finding new and potentially useful knowledge from existing large databases (Mitchell, 1999). Decision tree induction (Quinlan, 1986) is a commonly used data mining technique that is often used for classification problems. Kamphuis et al. (2010a) reported a simple decision tree that was capable of detecting CM with a similar performance compared with detection models currently used by AM systems. They suggested improving detection performance by using more CM cases and using quarter milkings (QM) that clearly had CM or were healthy in the training process. Also, combining decision tree induction with bagging and boosting techniques was expected to improve detection performance.

The objective of this study was to develop and validate a CM detection model by means of decision tree induction that meets the specific requirements when the model is applied in an AM setting. One such specific requirement is that the model should detect at least severe cases of CM at a very high level of specificity. At the same time, it should be able to deal with QM with a less clear mastitis status, because most QM fall into this category in practice. Finally, a CM model should generate alerts for CM within a limited period around or preferably before a CM case.

Materials and methods

Data collection and preparation

Data were collected at nine commercial dairy farms in the Netherlands using a total of 12 Lely Astronaut AM systems [version A2 (n=10) or A3 (n=2); Lely Industries N.V., Maassluis, the Netherlands]. Farmers were selected for participation based on the primary criterion that they were dealing with CM detection seriously. As a prerequisite, dairy farmers were expected to use the mastitis alert list daily and consistently to detect cows suspected of having CM. In addition, dairy farmers had to have at least 1 yr of experience using an AM system. Data collection began at two farms in November 2006. From October 2007 onward, data were collected on all nine farms. Data collection continued until March 2009.

Data from the AM systems were collected by installing remote computers on each AM system. These remote computers stored all data available during a cow milking. Data comprised the cow's identification number, and date and time stamps of the cow's entrance, teat cup attachments, and start of milk flow. All raw sensor measurements were stored as well, after discarding the first 30 mL (approximately) of quarter milk. Raw sensor measurements included average electrical conductivity, red, green, and blue color values for each quarter milking and for each 100 mL of quarter milk produced, and an estimation of quarter milk yield. Quarter milkings where the AM system failed to connect the teat cup, those with missing data for teat cup attachment and 100 mL measurements, and those without data at the QM level (e.g., sensor measurement averages, start milk flow, teat cup attachment) were deleted from the data set. For the remaining QM, recorded 100 mL values for electrical conductivity, and red, green, and blue that were outside the mean ± 4 SD were regarded as likely data errors and were set as missing. Sensor data were combined with somatic cell count (SCC) data collected as part of the dairy herd improvement program. Because SCC was determined at the cow milking level, all QM within a cow milking were assigned the SCC value determined on a specific test-day for a specific cow.

Cows that raised suspicion of CM according to the criteria of the dairy farmer were checked visually for having CM. The dairy farmer's suspicion could be based on the mastitis alert lists, but also on other information such as SCC data or the presence of clots on the milk filter. This approach resulted in different procedures per farm to check quarters for CM. By introducing a scoring protocol, however, the assessment of the actual mastitis status of visually checked quarters was standardized. The scoring protocol instructed dairy farmers to visually score the fifth and sixth squirts of milk of all quarters they checked using a clean black paddle as commonly used for the California Mastitis Test (without using the reagent normally used when applying this test). When visually normal, the milk was scored as 1. When abnormal, the milk could be scored as follows: 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, and the score assigned. Every 4 to 6 wk, dairy farmers were visited to collect data from the remote computer and the scoring forms. During these visits, the scoring forms were discussed with the dairy farmer to ensure proper use of the scoring protocol.

To combine visual observations of CM and sensor data, each visual quarter milk assessment was linked to sensor data from the most recent quarter milking, within a 24h time window before the assessment time, recorded for that same quarter by the remote computer (Figure 1, situation A).

Data from the four sensors (electrical conductivity, and the red, green, and blue color sensors), two derived measurement patterns (average color value and milk

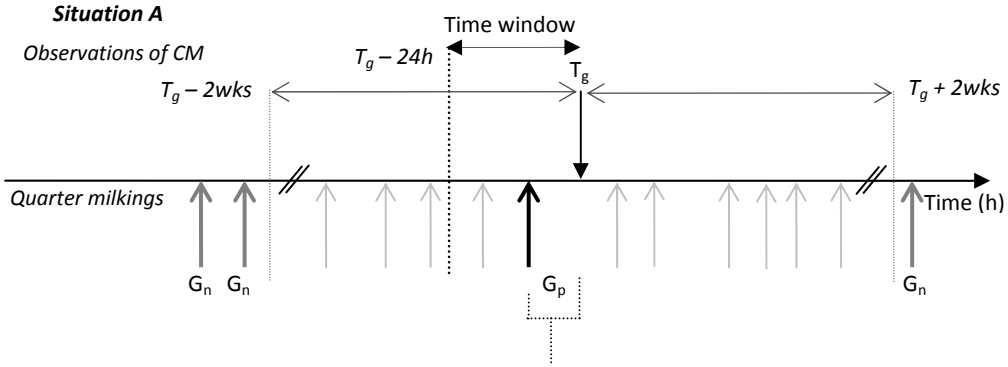
flow), and three measurement patterns involving one single value for a whole QM (milk flow delay, dead milking time, and milk production) were used to define descriptive variables (n=1,065) as described in Kamphuis et al. (2008a, 2010a). These independent variables described characteristics (level, variability, and shape) of the sensor measurement patterns from each QM.

Model development

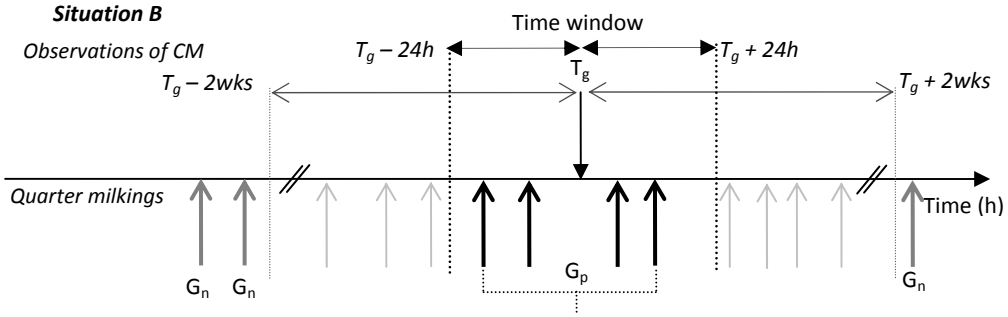
Sensor data were available from 1,109 individual cows and almost 3.5 million QM. Participating dairy farmers checked 1,593 QM visually for CM, of which 348 QM received a score from 2 through 5. These were considered as gold standard positive QM and included 74 QM with watery milk, 217 QM with flakes, 50 QM with clots, and 7 serum-like QM. These 348 QM with CM out of almost 3.5 million QM give a rough estimate of 0.36 CM cases per 365 cow-days at risk. There were also 6 QM containing blood (score 6). These were excluded from the data set because milk with blood does not necessarily imply CM but could be caused by damage in the udder or the teat. In addition, algorithms are already available to accurately detect milk with blood (Rasmussen and Bjerring, 2005). The data were divided into a training set (including two-thirds of all QM) and a test set. Cows in the training set could not be present in the test set and vice versa, and the number of CM cases in the training set was about two-thirds of all available QM with CM.

The training set included slightly more than 2.3 million QM from 738 cows. From these QM, 1,028 QM were combined with a visual observation. Only QM with SCC information from cows that never exceeded a SCC level of 200,000 cells/mL within the lactation and were never visually checked by the dairy farmers during the study period were selected and labeled as gold standard negative. The final training set used for model development comprised 24,960 QM from 404 cows, of which 243 QM were labeled gold standard positive and 24,717 QM were labeled gold standard negative.

Before the CM detection model was developed, the 200 most informative independent variables were selected based on the information gain ratio (Quinlan, 1986; Witten and Frank, 2005; Kamphuis et al., 2008a). This method selects those variables that best discern between positive and negative cases of CM.



Combining an observation of a quarter with CM (T_g) with sensor data from the last milking prior to the observation recorded by the automatic milking system for that same quarter (black arrow pointing up). This first milking had to be within a 24h time window prior to the observation and was considered gold standard positive (G_p). This labeling of G_p quarter milkings was similar for the training and the test set. For validation, quarter milkings within a two-week period around T_g were excluded from the data set (light gray arrows pointing up). Quarter milkings outside this two-week period (dark gray arrows pointing up) were considered gold standard negative (G_n).



Example of implementing a time window of 24h prior and 24h after the observation of CM (T_g) to label quarter milkings (black arrows pointing up) as gold standard positive (G_p). Each of the G_p quarter milkings receives a probability estimate for having CM by the decision tree. Other quarter milkings inside a two-week period but outside the chosen time window are excluded from the data set (light gray arrows pointing up). Quarter milkings outside a two-week period around a T_g (dark gray arrows pointing up) are considered gold standard negative (G_n). The probabilities for CM of the G_p quarter milkings are compared to a threshold that results in a specificity of 99% or 97.9%. If this probability for CM exceeds this chosen threshold, that quarter milking receives an alert for CM. If the number of alerts is ≥1 in the specified time window, then this time window is counted as a true positive alert in calculating sensitivity. If no alerts are assigned, than this time window is considered a false negative alert in calculating sensitivity.

Figure 1. Combining an observation of clinical mastitis (CM) to a quarter milking recorded by the automatic milking system (situation A) and selecting quarter milkings to compute sensitivity when time windows are applied (situation B)

To develop a CM detection model, decision tree induction was applied to the training set, consisting of the 200 selected independent variables and 24,960 QM. A decision tree 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 decision tree can be expressed recursively: based on its information gain ratio an independent variable is selected to split a data set at the first node. For each possible outcome of the test involved at that node, a branch is made ending in a daughter node. Next, 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 part of the tree stops developing (Witten and Frank, 2005). An example of a simplified and hypothetical decision tree is presented in Figure 2. A record in a test set will eventually flow into one of the classification nodes (colored gray in Figure 2), and the record will receive that classification output (CM or healthy).

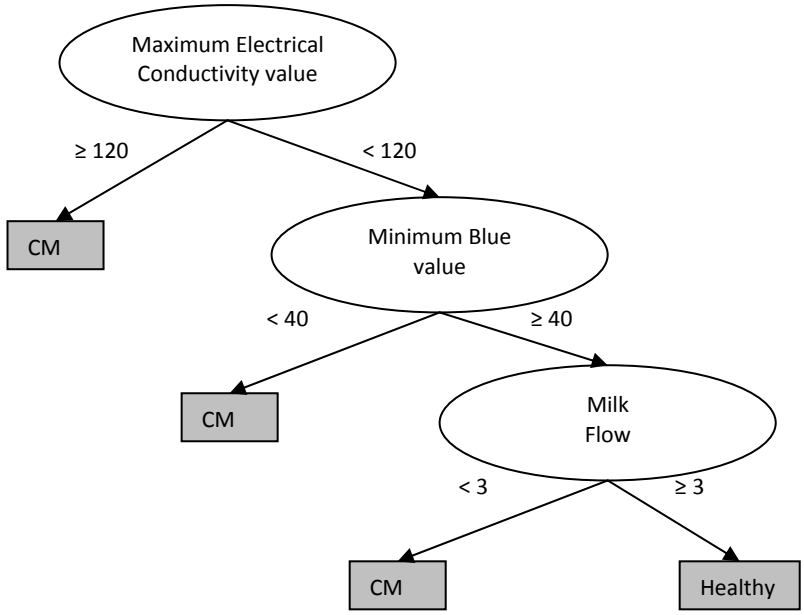


Figure 2. Example of a simplified and hypothetical decision tree using three independent variables based on sensor measurements to classify records into healthy or diseased (clinical mastitis, CM) quarter milkings

To build a decision tree, the J48 algorithm as implemented in WEKA (Witten and Frank, 2005) with default settings was used. This J48 algorithm was used on its own, but also in combination with bagging and boosting to study their effect on detection performance. Bagging and boosting are data mining techniques that can be used with any classification algorithm, such as the J48 algorithm, and both techniques operate by selectively resampling from the training set to generate derived training sets to which the basic algorithm is applied (Webb, 2000). In bagging, records are resampled in each iteration *t*, giving equal weight to each record in the training set. In forming the final classification, bagging also gives equal weight to each of the individual decision tree models built at each iteration *t*. Boosting, on the other hand, uses reweighting of records in each iteration *t*, where records wrongly classified in the previous iteration $t - 1$ receive more weight in the current iteration *t*. Finally, in forming the final classification, boosting gives a "good" model (one with fewer errors in the classification of records in the derived training sets) more weight than "bad" models (Freund and Schapire, 1996). In the current study, the number of iterations for bagging and boosting was set at 10. The Adaboost.M1 and the bagging algorithm in WEKA were used (Witten and Frank, 2005).

Validation

The decision tree developed was validated using data from the test set. This test set 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 dairy farmers, of which 105 QM were labeled as gold standard positive. From this test set, QM at the cow milking level that lay within a 2-wk period around a QM with CM were deleted from the data set, including those QM without CM that appeared in the same cow milking that had a QM with CM. This selection resulted in the exclusion of 24,161 QM (2.1% of all QM in the test set). All other remaining QM were labeled as gold standard negative. This gold standard labeling procedure is represented by situation A in Figure 1. From the large number of gold standard negative QM, a random sample of 50,000 QM was selected. The 50,105 records in the final test set used for validation came from 368 individual cows.

The decision tree model provided probability estimates for having CM for each QM in the test set. These probability estimates were compared with a threshold value to classify a QM as "healthy" or "CM." When the probability estimate exceeded the threshold, this QM was classified as CM. A gold standard positive QM that was classified by the decision tree model as CM was considered a true positive (TP) classification. When the decision tree classified a gold standard negative QM as healthy, this was considered a true negative (TN) classification. A false positive (FP) classification was a gold standard negative QM that was classified as CM, and

finally, a false negative (FN) classification was a gold standard positive QM that was classified as healthy. These four classification types were used to compute the Se as the fraction of gold standard positive QM that were classified as CM [Se (%)=100 × TP / (TP + FN)] by the decision tree, and the Sp as the fraction of gold standard negative QM that were classified as healthy [Sp (%)=100 × TN / (TN + FP)]. To evaluate the detection performance of a decision tree model, the Se was calculated at two fixed Sp levels: an Sp of 99% has been suggested by Mein and Rasmussen (2008) as a level that mastitis detection models should achieve when applied in practice. The second level of Sp was 97.9%, which is an Sp level found by H. Mollenhorst and H. Hogeveen (2008, unpublished data), who studied the performance of CM detection models used currently by the AM system. As a result, the thresholds, at which the probability estimates for having CM were compared with, were 0.4424 for an Sp of 99% and 0.2828 for an Sp of 97.9%.

Receiver operating characteristic curves are graphic representations of the TP rate (or Se) versus the FP rate (or 1 – Sp) over the whole range of possible threshold values (Detilleux et al., 1999). To summarize the receiver operating characteristic curve into a single quantity, the area under the curve (AUC) is a commonly used measure. However, for the classification problem at hand, total AUC is not an informative measure, because detection performance at high FP rates (resulting in too many FP alerts on the mastitis alert list) is of no interest. Therefore, the partial AUC (pAUC) was computed as a third measure to evaluate detection performance. The pAUC is restricted to just a relevant portion of the total AUC (Detilleux et al., 1999) and was calculated using the trapezoidal rule for Sp values of ≥99% and ≥97.9%. To be able to use the same interpretation as the total AUC, the pAUC values were transformed to the same original scale that is used for the total AUC (McClish, 1989).

Exploring time windows and farm-specific performances

As described above, detection performance was based on the labeling procedure depicted in situation A in Figure 1. This approach resulted in a very narrow time window (<24h) in which an alert by the decision tree was considered TP or FN. The application of time windows has been described by Sherlock et al. (2008). Their suggestions were used as the basis for evaluating Se of the decision tree in the current study, when time windows were enlarged. The implementation of wider time windows is shown in situation B in Figure 1, where an example is presented of a time window of 24h before to 24h after an observation of CM. The main difference with situation A (Figure 1) is that not all QM within a 2-wk period were excluded and that not just one QM received a gold standard positive label. Instead, all QM of the affected quarter within the time window of 24h before to 24h after a CM observation were labeled gold standard positive, and all these QM received a

probability estimate for having CM by the decision tree. Again, these probability estimates were compared with the same threshold values that resulted, for situation A (Figure 1), in an Sp of 99% or 97.9%. If the probability estimate exceeded the chosen threshold, that specific QM received an alert for CM. If the number of alerts within the specified time window of 24h before or after a CM observation was ≥1, then this time window was considered a TP alert for computing Se. If none of the selected QM received an alert, this time window was considered an FN alert in computing Se.

The decision tree was trained and tested using QM from all nine farms; data between the training and test sets were thus correlated to some degree. Analyses were conducted to study the detection performance of a CM decision tree model applied to truly unseen data; that is, to data from a farm that was not used for training at all. Therefore, decision tree models were trained, using the same approach as described above, with data from eight farms and validated with data from the ninth farm. This resulted in nine different decision tree models that were validated on nine farms that were not used for training. When a decision tree was trained, all QM with CM of the eight farms were used, and all QM with CM from the ninth farm were included in the test set.

Data preparation was done using SAS (version 9.1, SAS Institute Inc., Cary, NC). Development of several decision tree models and probability estimates for having CM for each quarter milking were done in WEKA (version 3.4.8; Witten and Frank, 2005). Data produced in WEKA were used in SAS to retrieve the Se at the fixed Sp levels and to compute the transformed pAUC values. Decision tree models were tested to determine whether Se levels, at the two fixed Sp levels, were statistically significantly different using the McNemar's test in SAS.

Results

Table 1 shows the results of the stratification procedure to obtain QM used for training and testing. Because some cows had more than one CM case, the stratification at the cow level caused the CM cases not to be distributed strictly by two-thirds for training and one-third for testing. The number of CM cases varied between farms, with a maximum of 81 CM cases observed at farm 2 and a minimum of five CM cases at farm 3. The random sample of 50,000 gold standard negative QM were selected proportionally to the total number of gold standard negative QM per farm in the test set. The number of gold standard negative QM varied largely between farms. Farms 8 and 9 both had two AM systems and data were collected over a similar period (two full years), but farm 8 had substantially more gold standard negative QM than farm 9.

Table 1. Number of automatic milking (AM) systems per farm, and number of unique cows, quarter milkings (QM), and clinical mastitis (CM) cases included in the training and test set per farm

Farm	AM systems (n)	Training set			Test set		
		Cows (n)	QM (n)	CM (n)	Cows (n)	QM (n)	CM (n)
1	1	56	3,259	51	41	7,281	26
2	1	40	1,237	60	30	3,305	21
3	1[a]	28	1,568	3	27	2,840	2
4	1[a]	41	1,221	5	38	2,779	2
5	1	36	3,475	9	28	4,069	5
6	1	40	2,165	44	31	3,825	14
7	2	43	3,178	15	58	8,810	11
8	2	62	6,294	19	67	10,430	9
9	2	58	2,563	37	48	6,766	15
Total	12	404	24,960	243	368	50,105	105

[a] Farms with an A3 Lely Astronaut automatic milking system (Lely Industries N.V., Maassluis, the Netherlands). All other farms used an A2 Lely Astronaut automatic milking system

Table 2 shows the detection performance of three decision tree models when validated on all 50,105 QM of the test set. All three models used the same 200 most informative independent variables, but differed in training procedure. The model that used the J48 algorithm alone showed an Se of 24.7% at an Sp of 99% and an Se of 43.8% at an Sp of 97.9%. The pAUC was 0.556 for Sp levels of ≥99% and 0.612 for Sp levels of ≥97.9%. When combining the J48 algorithm with boosting, there was an increase in both Se and pAUC at both Sp levels but this increase was only statistically significant at an Sp of 99% (P=0.003), and not at an Sp of 97.9% (P=0.532). Values for Se and pAUC were highest (Se of 40.0% and 57.1% for the two fixed Sp levels of 99% and 97.9%, respectively, and pAUC of 0.639 and 0.693 for Sp levels of ≥ 99% and ≥97.9%, respectively) for the decision tree model that combined the J48 algorithm with bagging. The Se of this decision tree model was statistically significantly different from that of the decision tree that used the J48 alone at both Sp levels (P=0.0002 at an Sp of 99%, and P=0.004 at an Sp of 97.9%). The difference in Se between the decision trees that used either bagging or boosting in combination with the J48 algorithm was statistically significant (P < 0.0001) at an Sp of 97.9%. Therefore, the decision tree that

combined the J48 algorithm with bagging was chosen as the best model and is referred to as the base decision tree hereafter. All other results are based on this base decision tree. When applying the base decision tree to farm-specific data in the test set, it became clear that detection performance varied largely between farms. Farms 1, 6, and 8 showed a high Se at an Sp of 97.9% (also reflected by a high pAUC for Sp ≥97.9%). At a higher Sp (99%), the base decision tree still performed well on farms 6 and 8. The base decision tree did not perform well on farms 2, 7, and 9. Performance of the base decision tree was not informative for farms 3, 4 and 5 because of the very low number of CM cases in the test set for these farms (see also Table 1).

Table 2. Sensitivity (Se) at a specificity (Sp) of 99% (Se^{99}) and a Sp of 97.9% ($Se^{97.9}$), and the transformed partial area under the curve for Sp levels ≥ 99% ($pAUC^{99}$) and ≥ 97.9% ($pAUC^{97.9}$) of three decision tree models[1]

Decision tree Model	Se^{99}	$pAUC^{99}$	$Se^{97.9}$	$pAUC^{97.9}$
Using all test data (50,105 quarter milkings)				
J48	24.7[a]	0.556	43.8[a]	0.612
J48 with boosting	38.1[b]	0.636	52.4[ab]	0.681
J48 with bagging	40.0[bc]	0.639	57.1[c]	0.693
Using J48 with bagging on farm specific test data				
Farm				
1	38.5	0.624	76.9	0.70
2	23.8	0.577	23.8	0.596
3	0.0	0.498	0.0	0.495
4	50.0	0.613	50.0	0.682
5	60.0	0.759	60.0	0.779
6	71.4	0.846	71.4	0.851
7	18.2	0.560	36.4	0.592
8	66.7	0.783	77.8	0.822
9	13.3	0.564	46.7	0.602

[a–c] Values within the same column with different superscripts are statistically significantly different ($P < 0.05$)

[1] One model used the J48 algorithm alone, one combined J48 with boosting, and one combined J48 with bagging. Detection performance values are computed using all test data (50,105 quarter milkings; 105 with clinical mastitis) and per farm

The test set included 105 QM with CM, of which 16 were scored as watery milk, 72 contained flakes, 14 contained clots, and three were serum-like. Table 3 summarizes which CM cases were detected by the base decision tree at the two

fixed Sp levels. The base decision tree detected 42 QM with CM at an Sp of 99% and 60 QM with CM at an Sp of 97.9%, coinciding with the Se levels of 40.0% and 57.1% reported earlier (Table 2). The base decision tree had problems detecting watery milk, with only 12.5% of these cases detected at an Sp of 99% and 37.5% at an Sp of 97.9%. Most CM cases were observed as milk containing flakes (68.8%), and at an Sp of 97.9% the base decision tree alerted for more than half of these cases. There were 14 more severe CM cases, indicated by clots in the milk. At an Sp of 99%, the model detected 64.3% of these cases. This percentage increased to 78.6% at an Sp of 97.9%.

Table 3. The number of quarter milkings with clinical mastitis detected by the base decision tree model at two fixed levels of specificity (99% and 97.9%), stratified by the score the observed clinical mastitis received by the participating dairy farmers

Specificity level (%)	Number of clinical mastitis observations receiving score (n)				Total (n)
	2) Watery Milk	3) Flakes	4) Clots	5) Serum-like Milk	
99	2	30	9	1	42
	(4.8)[1]	(71.4)	(21.4)	(2.38)	
	(12.5)[1]	(41.7)	(64.3)	(33.3)	(40.0)
97.9	6	41	11	2	60
	(10.0)	(68.3)	(18.3)	(3.3)	
	(37.5)	(56.9)	(78.6)	(66.7)	(57.1)
Total	16	72	14	3	105
	(15.2)	(68.8)	(13.3)	(2.9)	

[1] First value in parentheses indicates the percentage from the row total; the second value in parentheses indicates the percentage from the column total

Table 4 shows the Se at the two fixed Sp levels of the base decision tree when different time windows were applied. The first row in the table shows the Se of the base decision tree when applied to the time window that was used to combine CM observations with sensor data from the AM systems. A considerable increase in Se was found when a time window of 24h before or after a CM observation was applied: Se increased to 66.7% at an Sp of 99% and to 79% at an Sp of 97.9%. Results showed that Se increased as the selected time windows were widened, and that not all CM were detected by the base decision tree, even when very wide time windows were applied: slightly more than 10% of the CM cases were not detected when a time window of 10d (or 240h) before to 7d (or 168h) after the CM observation was applied at an Sp of 97.9%.

Table 4. Sensitivities (Se) of the base decision tree at a specificity of 99% (Se^{99}) and at a specificity of 97.9% ($Se^{97.9}$) using different time windows

Used time window (h)		Detection Performance	
Before CM observation	After CM mastitis observation	Se^{99}	$Se^{97.9}$
<24[1]	0	40.0	57.1
24	24	66.7	79.0
48[2]	24	69.5	81.0
96[3]	48	73.3	82.9
96	72	75.2	85.7
120[4]	0	70.5	81.0
240[5]	168	84.8	89.5

[1] Based on the single quarter milking that was combined with the observation score of clinical mastitis. This single quarter milking was the first quarter milking prior to the observation of clinical mastitis, but had to be within a 24h period before the observation
[2] Time window applied in Sherlock et al. (2008)
[3] Time window applied in Claycomb et al. (2009)
[4] Time window applied in Kramer et al. (2009)
[5] Time window applied in De Mol et al. (1997)

Results in Table 5 report the Se and pAUC of nine decision trees built on data from eight farms and tested on data from the ninth farm. As in Table 2, the Se varied largely between farms and, for some farms, there was a large difference in Se at the two fixed levels of Sp. Again, detection performance was not very informative for farms 3 and 4 because of the small numbers of QM with CM used for testing. The Se deteriorated for some farms (e.g., farms 6 and 8) at both Sp levels compared with the Se levels reported in Table 2, which were based on a decision tree built with data from all nine herds. For example, the Se of farm 6 decreases with 18% at an Sp of 99%, and a decrease in Se of 13.5% is seen for farm 8 at an Sp of 97.9%. In addition, there was a considerable decrease in pAUC, with a maximum decrease in pAUC of 0.150. On the other hand, farm nine showed an increase in Se of 21.3% at an Sp of 99% when the decision tree was trained on data from eight other farms. The increase was, however, less pronounced in the pAUC that increased by 0.043. In this specific farm, Se did not increase at an Sp of 97.9%.

Table 5. Total number of quarter milkings (QM) and the number of QM with clinical mastitis (CM) used for training decision tree models, total number of QM and the number of QM with CM used for validating the decision tree models, the sensitivity at a specificity of 99% (Se^{99}) and a specificity of 97.9% ($Se^{97.9}$), and the transformed partial area under the curve for specificity levels ≥ 99% ($pAUC^{99}$) and ≥ 97.9% ($pAUC^{97.9}$)[1]

Decision tree model validated using data from farm	Number of QM (with CM) for training	Number of QM (with CM) for validation	Se^{99}	$pAUC^{99}$	$Se^{97.9}$	$pAUC^{97.9}$
1	21,780 (271)	7,332 (77)	41.6	0.654	51.9	0.692
2	23,807 (267)	3,365 (81)	17.3	0.561	24.7	0.583
3	23,495 (343)	2,843 (5)	0.0	0.498	0.0	0.495
4	23,842 (341)	2,784 (7)	57.1	0.599	57.1	0.694
5	21,585 (334)	4,078 (14)	64.3	0.744	64.3	0.783
6	22,886 (290)	3,869 (58)	53.4	0.696	62.0	0.741
7	21,876 (322)	8,825 (26)	26.9	0.579	30.8	0.612
8	18,762 (320)	10,449 (28)	60.7	0.673	64.3	0.748
9	22,487 (296)	6,803 (52)	34.6	0.609	44.2	0.650

[1] All decision trees are trained with data from eight farms using the J48 algorithm combined with bagging (200 best independent variables were selected using information gain ratio) and validated with data from the ninth farm

Discussion

In recent years, several studies on automated detection of CM have been published. Most of these studies collected detailed information on a relatively small number of CM cases, and data were mostly collected at one (research) farm. In the current study, a CM detection model was developed using sensor and observational data collected at nine commercially operating Dutch dairy farms milking automatically, which is a unique approach in the field of CM research. The approach was chosen deliberately to collect data on a large number of CM cases (n=348) in a relatively short period, without intervening with the daily working routine of the dairy farmers involved. A rough estimate of the CM incidence of the current study (0.36 CM cases per 365 cow-days at risk) was similar to the CM incidence rate of 0.31 per 365 cow-days at risk found in a survey of Dutch dairy farmers (Van den Borne et al., 2008). This confirmed the idea that the CM observations by the participating dairy farmers in the current study were reliable and representative for the true CM status in their farms.

The base decision tree differed statistically ($P < 0.05$) from the decision tree model that applied the J48 algorithm alone at both Sp levels (97.9 and 99%) and from the decision tree model that combined the J48 algorithm with boosting at an Sp of 97.9%. When compared with previously conducted studies on CM detection models that report both an Se and an Sp, the base decision tree showed improved detection performance. For example, a recent study conducted by Kramer et al. (2009) reported on a CM detection model with an Sp of 92% and an Se of 75% using a time window of 4d before the CM observation and the day of observation. The base decision tree developed in the current study reached an Se of 92.4% at the same Sp when this time window of 5d was applied. Norberg et al. (2004) applied a stricter time window (the CM detection model had to alert for CM at the milking at which a CM was observed) and reported an Se of 43% at an Sp of 93% to detect CM. The time window was similar to the one used here (<24h), and fixing the Sp at 93%, the base decision tree in the current study showed an Se of 75.2%. The base decision tree showed an Se of 57.1% at an Sp of 97.9%. This is an increase in Se of more than 20 percent points compared with detection models currently used by AM systems at the same Sp of 97.9% (H. Mollenhorst and H. Hogeveen; unpublished data, 2008). Farmers are eager to minimize false positive alerts even at the expense of a lower Se (Claycomb et al., 2009). When interpreting results from the dairy farmer's perspective, the base decision tree was able to keep the Se at the same level as currently used models (40% for the base decision tree vs. 36.8%) while decreasing the number of false positive alerts by more than 50% as Sp increased from 97.9% to 99%.

The application of time windows, in which an alert by a detection model is considered a true positive alert, was previously implemented by De Mol et al. (1997). The discussion about the effect of implementing time windows of different lengths on detection performance is more recent. Sherlock et al. (2008) presented some guidelines for performance evaluation when implementing different time windows. They reported improved detection performance when time windows were wider. This trend was seen in the current study as well: the Se from the base decision tree increased by more than 20 percentage points at both Sp levels when the time window increased from <24h before a CM observation to a 24h time window before or after a CM observation (Table 4). Although detection models that use large time windows for their alerts will show higher levels of Se and Sp than models using smaller time windows, there is a risk that these detection models may lose their practical use (e.g., the time window of 10d before to 7d after a CM reported by De Mol et al. (1997)). What time window is useful for a CM detection model applied in an AM setting remains debatable: use a wider time window for early detection, including the risk that visual signs of CM are not there at the time of an alert or use a time window for detection of quarters with clear visible signs of CM with the risk of alerts for cows already being clinically infected for a few days? Keefe et al. (2010) summarized results from studies using on-farm culture results to apply selective antibiotic treatment. They report no differences in clinical or bacteriological cure rates between cows treated with antibiotics directly after a CM occurrence and cows that received a selective antibiotic treatment based on on-farm culture tests with results typically available within 24h after the CM occurrence. This may suggest that alerts by a CM detection model after the CM occurrence itself do not necessarily result in worse cure rates. It seems sensible, however, to assume that time windows should be as narrow as possible, although it is not implied that the narrow time window of <24h used in the current study is the correct one.

Validation on a data set that includes only QM that are clearly healthy or that clearly suffer from CM as done by Friggens et al. (2007) will give an overestimated performance. They reported high levels of Se and Sp, but their detection model would probably not reach this level of performance in practice because in reality not all QM are clearly healthy or clearly have CM. The effect of including or excluding QM with a less clear CM status has been studied by Kamphuis et al. (2010b). They reported a decrease in Se of 25 percentage points when validation was applied to a test set that included QM with a less clear mastitis status, compared with a test set that used only QM that were clearly healthy or CM. The base decision tree from the current study was trained using QM that were clearly healthy or clearly suffered from CM. In doing so, the amount of noise in the training set was decreased, which in return was expected to ease the training process for the decision tree algorithm.

On the other hand, validation was applied to data that mimicked practice as closely as possible, because a large number of QM with a less clear mastitis status were kept in the validation set. By selecting a random sample of 50,000 QM, it was expected that all potential variation of sensor measurements of these QM with a less clear mastitis status was captured and that validation on such a test set would better reflect detection performance of models when they are implemented in practice.

Another point of debate is the necessary Se and Sp levels a CM detection model should have when it is implemented in practice. For visual observation of CM or abnormal milk, several levels for Se have been suggested. Hillerton (2000) reported an Se of 80% for visual inspection of milk for CM. Rasmussen (2005) suggested an Se of 70% as a reasonable level to detect abnormal milk during foremilking using a filter with pore size of 0.1 mm. The International Organization for Standardization (ISO, 2007) suggests an Se of 70% for automatic milking installations. This Se level is reported as the lower confidence interval value using an 80% confidence interval. When applying this lower confidence interval value in combination with an Sp level of 99%, suggested by Mein and Rasmussen (2008) as minimum level for practical use, the base decision tree should reach an Se of 75% at an Sp of 99%. Using the short time window used in the current study (<24h), however, the base decision tree showed an Se of 40% at an Sp of 99%. To meet the requirements mentioned, it was necessary to widen the time interval to 4d before to 3d after a CM observation (Table 4). The question remains whether this wide time window is still useful for a practical AM setting.

Results showed that the base decision tree performed well on some farms (e.g., farms 1, 6, and 8; Table 2) but poorly on others (e.g., farms 2 and 7; Table 2). Based on our experience when collecting data during the farm visits, the idea arose that dairy farmers from these "bad" farms were more chaotic in their daily working routines than dairy farmers from "good" farms. They did detect CM cases, but it seemed that they detected these cases by using information sources other than the mastitis alert lists and that they did not use the information presented on the mastitis alert list in a systematic and consistent way to detect CM. On the other hand, dairy farmers from "good" farms seemed to be well organized and seemed to have developed a strict protocol of using the information presented on the mastitis alert list, without using a lot of additional information. Including data from "bad" farms in the training set may introduce unwanted noise, which makes it more difficult for the decision tree algorithm to develop a good detection model. Similarly, including data from "good" farms in the training set may be very useful in developing a generic CM detection model. These ideas were reflected by results reported in Table 5: "bad" farms showed a low detection performance for models that were trained with uncorrelated data, suggesting that these dairy farmers

indeed detected the CM cases by means other than the mastitis alert lists. "Good" farms showed lower Se levels of the models trained on uncorrelated data compared with models based on partly correlated data, potentially indicating a loss of useful information when data from these farms were excluded for training. This observation raised the question of whether it would be better to train farm specific models for these "good" farms. For example, when building a farm-specific model for farm 1 (using only data from that farm), the farm-specific decision tree showed an Se of 76.9% at an Sp of 97.9% (data not shown). This is a similar Se compared with that of the base decision tree model, but the farm-specific decision tree was trained with only 55 QM with CM, about one-fourth of the QM with CM on which the base decision tree was based. It is expected that performance would increase if the farm-specific model will be trained using more CM cases, which suggests that a farm-specific model would be very useful for these "good" farms.

Claycomb et al. (2009) indicated that it should be possible for dairy farmers to vary thresholds of future detection models to adapt them to their own specific preferences for detecting TP cases or minimizing FP alerts. This is an interesting idea, and it would be a useful addition for dairy farmers compared with current detection models where the options to customize are minimal. It should be realized, however, that this would require the automatic milking industry to explain to dairy farmers, as end-users of the detection model, the effects of changing these thresholds and the interrelationship between Se and Sp. Suppose the dairy farmer of farm 1 (Table 5) changes the threshold from the decision tree such that Sp decreases from 99 to 97.9%. At the same time, Se would increase from 41.6 to 51.9%. In absolute numbers, the model would now detect 8 additional CM cases at the expense of 80 extra FP alerts. The dairy farmer might be satisfied with this change in FP and TP alerts. For farm 8, however, this change in threshold values would increase the number of FP alerts by more than 100, but just 1 additional CM case would be detected. This would probably be a much less attractive change in TP and FP alerts. A customizable CM detection model is an attractive idea, but it should be applied carefully by dairy farmers and might not be an attractive additional option for all dairy farmers.

The base decision tree detected reasonably well CM cases that were observed as having flakes or clots (Table 3) when using the narrow time window of <24h: more than 56% of CM cases with flakes and almost 80% of CM cases with clots were detected at an Sp of 97.9%. At an Sp of 99%, 64% of all CM cases with clots were detected. Unfortunately, the base decision tree was not able to detect all severe cases of CM (score 4), and the detection of watery milk was very poor, with only 12.5% detection at an Sp of 99%. This indicates that using sensor data (electrical conductivity, color, and milk flow) alone is not enough to detect all CM

cases, and other information sources are necessary to guarantee that all cases are found by dairy farmers.

To evaluate detection performance, the Se was computed at two fixed Sp levels (99% and 97.9%). These evaluators are an estimate of the true detection performance of the developed decision trees because not all QM and not all alerts on the mastitis alert list were checked visually. Because of this lack of continuous monitoring of all QM for CM, alerts are generated by the decision tree model that may have been TP for CM but are now considered FP or in alerts that may have been FN but are considered now TN. Consequently, the number of TP and FN alerts may have been underestimated and the number of FP and TN alerts may have been overestimated. However, the choice to exclude QM within a 2-wk period around a CM observation (resulting in an exclusion of 2.1% of all QM in the test set) was made to prevent mislabeling of QM that already had CM but that were not detected earlier by the dairy farmers or mislabeling of QM with subclinical mastitis as gold standard negative. In addition, the selection procedure for dairy farmers required them to use the mastitis alert lists daily and consistently. By taking these measures, the number of missed true positive alerts was kept as low as possible. This problem was also encountered by Steeneveld et al. (2010). They conducted a study in which cow information was combined with alerts from the mastitis alert list to judge whether these alerts were true positives or false positives. Just as in the current study, they used the selection criterion that farm managers worked seriously in detecting CM cases. In addition, they noted that it is likely that missed cases of CM cured spontaneously or that they were detected later when infection had deteriorated (Steeneveld et al., 2010), two points that are applicable to the current study as well.

Conclusion

Decision tree induction made it possible to maintain Se at about the same level as models currently used by AM systems but to decrease the number of false positive alerts by more than 50%. It was also possible to increase Se from 36.8 to 57.1% when the Sp level was kept at the level of models currently used by AM systems (97.9%). The base decision tree is expected to be suitable for use in practice for the following reasons: data from nine commercially producing dairy farms were used for model development and validation, alerts for CM were given within a very short time window (<24h), and finally, validation was conducted on a test set that included QM with a less clear mastitis status, mimicking practice as closely as possible.

Acknowledgements

We gratefully acknowledge the participating dairy farmers for their contribution to the acquisition of the data used in this research. Also gratefully acknowledged is Lely Industries N.V. (Maassluis, the Netherlands) for providing addresses of dairy farmers working with AM systems and their support with the installation of the data acquisition systems at the participating farms. This research is supported by the Dutch Technology Foundation STW, Applied Science division of NOW, and the Technology Program of the Ministry of Economic Affairs.

References

Claycomb, R.W., P.T. Johnstone, G.A. Mein, and R.A. Sherlock. 2009. An automated in-line clinical mastitis detection system using measurement of conductivity from foremilk of individual udder quarters. N. Z. Vet. J. 57(4):208–214.

De Koning, K., B. Slaghuis, and Y. van der Vorst. 2004. Milk quality on farms with an automatic milking system. Pages 311 - 320 in Automatic Milking - a better understanding. A. Meijering, H. Hogeveen, and C.J.A.M De Koning (eds.). Wageningen Academic Publishers, Wageningen, the Netherlands.

De Mol, R.M., G.H. Kroeze, J.M.F.H. Achten, K. Maatje, and W. Rossing. 1997. Results of a multivariate approach to automated oestrus and mastitis detection. Livest. Prod. Sci. 48(3):219–227.

Detilleux, J., J. Arendt, F. Lomba, and P. Leroy. 1999. Methods for estimating areas under receiver-operating characteristic curves: Illustration with somatic-cell scores in subclinical intramammary infections. Prev. Vet. Med. 41(2-3):75–88.

Dohmen, W., F. Neijenhuis, and H. Hogeveen. 2010. Relationship between udder health and hygiene on farms with an automatic milking system. J. Dairy Sci. 93(9):4019-4033.

Freund, Y., and R.E. Schapire. 1996. Experiments with a new boosting algorithm. Pages 148–156 in Proceedings of the Thirteenth International Conference of Machine Learning. L. Saitta (ed.). Morgan Kaufmann Publishers, San Francisco, CA.

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

Halasa, T., K. Huijps, O. Østerås, and H. Hogeveen. 2007. Economic effects of bovine mastitis and mastitis management: A review. Vet. Q. 29:18–31.

Hillerton, J.E. 2000. Detecting mastitis cow-side. Pages 48–53 in National Mastitis Council 39th Annual Meeting. Atlanta, GA.

Hogeveen, H., and W. Ouweltjes. 2003. Sensors and management support in high-technology milking. J. Anim. Sci. 81(Suppl. 3):1–10.

ISO. 2007. Automatic milking systems—requirements and testing. Annex C: Example of methods of evaluating detection systems for milk deemed as abnormal due to blood or to changes in homogeneity. ISO 20966:2007. International Organization for Standardization, Geneva, Switzerland.

Kamphuis, C., H. Mollenhorst, A.J. Feelders, D. Pietersma, and H. Hogeveen. 2010a. Decision-tree induction to detect clinical mastitis with automatic milking. Comput. Electron. Agric. 70(1):60–68.

Kamphuis, C., H. Mollenhorst, J.A.P. Heesterbeek, and H. Hogeveen. 2010b. Data mining to detect clinical mastitis with automatic milking. Pages 568–572 in Proceedings of the 5th IDF Mastitis Conf. J.E. Hillerton (ed.). The New Zealand Veterinary Association Foundation for Continuing Education Inc., Wellington, New Zealand.

Kamphuis, C., D. Pietersma, R..P.P van der Tol, M. Wiedemann, and H. Hogeveen. 2008a. Using sensor data patterns from an automatic milking system to develop predictive variables for classifying clinical mastitis and abnormal milk. Comput. Electron. Agric. 62(2):169–181.

Kamphuis, C., R. Sherlock, J. Jago, G. Mein, and H. Hogeveen. 2008b. Automatic detection of clinical mastitis is improved by in-line monitoring of somatic cell count. J. Dairy Sci. 91(12):4560–4570.

Keefe, G., J. McCarron, K. MacDonald, and M. Cameron. 2010. The scientific bases for using on-farm culture systems. Pages 141–148 in Proceedings of the 49th Annual Meeting of the National Mastitis Council. Albuquerque, NM.

Kramer, E., D. Cavero, E. Stamer, and J. Krieter. 2009. Mastitis and lameness detection in dairy cows by application of fuzzy logic. Livest. Sci. 125(1):92–96.

McClish, D. K. 1989. Analyzing a portion of the ROC curve. Med. Decis. 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 Control— 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. Commun. ACM 42(11):30–36.

Norberg, E., H. Hogeveen, I.R. Korsgaard, N.C. Friggens, K.H.M.N. Sloth, and P. Løvendahl. 2004. Electrical conductivity of milk: Ability to predict mastitis status. J. Dairy Sci. 87(4):1099–1107.

Quinlan, J.R. 1986. Induction of decision trees. Mach. Learn. 1:81– 106.

Rasmussen, M.D. 2005. Visual scoring of clots in foremilk. J. Dairy Res. 72:406–414.

Rasmussen, M.D., and M. Bjerring. 2005. Visual scoring of milk mixed with blood. J. Dairy Res. 72(3):257–263.

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. Pages 275–282 in Mastitis Control—From Science to Practice. T.J.G.M. Lam (ed.). Wageningen Academic Publishers, Wageningen, the Netherlands.

Steeneveld, W., L.C. van der. Gaag, W. Ouweltjes, H. Mollenhorst, and H. Hogeveen. 2010. Discriminating between true positive and false positive clinical mastitis alerts from automatic milking systems. J. Dairy Sci. 93(6):2559–2568.

Van den Borne, B.H.P., G. Van Schaik, T.J.G.M. Lam, and M. Nielen. 2008. A comparison of the occurrence of mastitis in Dutch primi- and multiparous cows. Pages 55–62 in Mastitis Control— From Science to Practice. T.J.G.M. Lam (ed.). Wageningen Academic Publishers, Wageningen, the Netherlands.

Webb, G.I. 2000. Multiboosting: A technique for combining boosting and bagging. Mach. Learn. 40:159–196.

Witten, I.H., and E. Frank (eds.). 2005. Data Mining; Practical Machine Learning Tools and Techniques. 2nd edition. Morgan Kaufmann Publishers, San Francisco, CA.

Chapter 5

Sensor measurements revealed; predicting the Gram-status of clinical mastitis causal pathogens

C. Kamphuis, H. Mollenhorst, H. Hogeveen

Submitted

Abstract

This study explored whether decision tree induction was able to predict the Gram-status of clinical mastitis (CM) causal pathogens, using in-line sensor measurements from automatic milking systems. Data consisted of 140 bacteriological cultured CM cases with sensor measurements of electrical conductivity, colors red, green, and blue and milk yield. In total, 110 CM cases were classified as Gram-positive cases and 30 were classified as Gram-negative. A decision tree was trained with two thirds of all data (n=96 CM cases) and validated with the remaining one third (n=44 CM cases), where the distribution of the Gram-status of the CM cases in the training set and in the test set were similar. Results showed that the decision tree used three variables, based on the electrical conductivity and the color sensor blue, to predict the Gram-status of the CM causal pathogen. This decision tree had an accuracy of 90.6% and a kappa value of 0.76 based on data in the training set. When validated with data from the test set, accuracy dropped to 54.5% and the kappa value to -0.20. Although the developed decision tree for Gram-status prediction was not a robust one, results from the training set indicate a potential for sensor measurements as tool for Gram-status prediction. When only those CM cases were considered with extreme high probability estimates for their Gram-status (either positive or negative), 74% of all 96 records in the training set could be classified with an accuracy of 97.1%. Predicting the CM causal pathogen itself was not possible; the decision tree had an accuracy of 52% and a kappa of 0.0, based on data in the training set.

Keywords

Gram-status prediction, sensor measurement patterns, decision tree induction

Introduction

Mastitis is one of the most frequent and costly diseases in the dairy industry (Halasa et al., 2007; Viguier et al., 2009), with clinical mastitis (CM) being responsible for a large proportion of the total costs of mastitis (Huijps et al., 2008). The disease has negative effects on farm economics due to production losses, treatment costs, and costs of culling (Huijps et al., 2008; McDougall et al., 2007; Viguier et al., 2009). It is important that cows that suffer from CM are detected during the early onset of the disease to eliminate the infection and to prevent recurrence, and that the appropriate type and amount of antibiotic is applied by the correct route and for the correct time to increase cure rate (Hillerton and Kliem, 2002). When milking fully automatically, CM is detected using two diagnostic tests: the first is the CM detection model on the automatic milking system itself. This model uses sensor measurements as input and gives a CM alert as output. These CM alerts appear on a mastitis alert list to warn the dairy farmer for cows that need attention ("management by exception", Hogeveen and Ouweltjes, 2003). The second test involves the visual confirmation of CM and is conducted by the dairy farmer, who checks the alerts from the mastitis alert list he or she thinks necessary. If a CM case is visually confirmed, it is the responsibility of the dairy farmer to decide on antibiotic treatment. Initially, the choice of antibiotic treatment is based in absence of any knowledge about the CM causal pathogen, as the whole process from taking a milk sample, culturing it, and determining the pathogens usually takes three days. Choosing an inappropriate treatment protocol or an ineffective antibiotic may be reasons that cure rates of 75%, as suggested at the introduction of commercial mastitis injectors, are not met in practice (Erskine et al., 2003). In order to make a more informed decision on which antibiotic to use, it could be beneficial for dairy farmers to have information about the CM causal pathogen at the same time a cow is listed on the mastitis alert list.

In the past, studies were conducted to use several information sources to provide or to predict the CM causal pathogen or the Gram-status of the pathogen involved. For example, Godden et al. (2007) and MacDonald et al. (2010) presented the potential of on-farm culture systems in order to replace bacteriological culturing that is done in the laboratory. Despite attractive benefits, major disadvantages of this approach are the need of experienced personnel and good on-farm laboratory conditions in order to achieve a proper accuracy. Instead of replacing the laboratory to the farm itself, other studies focused on using other information sources to predict CM causal pathogens or Gram-status; expert knowledge was used by Jones et al. (1990) and Kim and Heald (1999), cow information by Milne et al. (2003) and Steeneveld et al. (2009), cow data and farm management practices by Heald et al. (2000), and somatic cell count patterns by

De Haas et al. (2004). These studies showed varying predictive performances of the classification models developed, or concluded that collection of the necessary information would take too long for an appropriate prediction at cow-level. Kamphuis et al. (2008) reported a difference in sensor measurement patterns of the electrical conductivity and the colors red, green, and blue between healthy quarters and quarters infected with CM, but they did not check for different patterns between different pathogens. Espada and Vijverberg (2002) expected color sensor measurements to be useful for abnormal milk detection and potentially also for pathogen prediction. However, that study was based on a small data set, collected during one month at two farms, with only 6 cows showing abnormal milk in one quarter.

Kamphuis et al. (2010b) developed a CM detection model using decision tree induction. Sensor measurements of the electrical conductivity and colors red, green, and blue were used as input for this detection model. The final model showed a good performance (a sensitivity of 40% at a specificity of 99%), considering the narrow time window in which the model should alert for CM and the inclusion of quarter milkings with a less clear mastitis status. Decision tree induction could very well be used to extract knowledge from sensor data, amongst that of color sensors, for Gram-status prediction. The main objective of this study is to explore whether sensor measurements from automatic milking systems can be used by decision tree induction to predict the Gram-status of the CM causal pathogen. In addition, the potential of predicting the actual CM causal pathogen itself is evaluated as well.

Materials and methods

Data collection

The collection of data used for this study has been described in detail by Kamphuis et al. (2010b). In short, raw sensor data and observations of CM were collected at nine commercial Dutch dairy farms milking automatically (version A2 (n=10) or A3 (n=2); Lely Astronauts N.V., Maassluis, the Netherlands) from November 2006 until March 2009. Raw sensor data of the electrical conductivity, the colors red, green, and blue, and an estimation of quarter milk yield were collected by connecting a remote computer to each of the 12 automatic milking systems. From these raw sensor measurements 1,065 potentially descriptive variables were developed using a data flow diagram (Kamphuis et al., 2008; 2010a). These variables described characteristics (level, variability, and shape) of sensor measurements patterns from each quarter milking. Cows that raised suspicion of being infected with CM, according to the own criteria of the participating dairy farmer, were checked

visually for having CM. The dairy farmers' suspicion could be based on the mastitis alert lists, but also on other information, like the SCC data or the presence of clots on the milk filter. This approach resulted in different procedures per farm to check quarters for CM. By introducing a scoring protocol, however, the assessment of the actual mastitis status and the procedure to collect milk samples for bacteriological culturing was standardized. The scoring protocol instructed dairy farmers to visually score the 5^{th} and 6^{th} 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 2) watery milk, 3) flakes, 4) clots, 5) serum-like milk, or 6) milk with blood. If a dairy farmer decided the CM infection was severe enough to start an antibiotic treatment, they were asked to first take two milk samples for bacteriological culturing and to store these milk samples in their refrigerator. Furthermore, they were instructed to record the cow's identification number, quarter, date and time, and the mastitis score assigned to the visually checked quarter. Every 4 to 6 weeks, dairy farmers were visited to collect sensor data stored by the remote computer connected to each automatic milking system, the scoring forms, and the milk samples if any. Milk samples were bacteriological cultured by the Veterinary Microbiological Diagnostic Centre (VMDC, Faculty of Veterinary Medicine, Utrecht University) according to the standards of the National Mastitis Council (Harmon et al., 1990).

Data preparation

In order to combine visual observations of CM and sensor data, each visual quarter milk assessment was linked with sensor data from the most recent quarter milking, within a 24h time window prior to the assessment time, recorded for that same quarter by the remote computer (for a more detailed description, see Kamphuis et al., 2010b).

By the end of the data collection period, there were 2,222 quarters checked visually for having CM, of which 1,593 could be combined with sensor measurements within a 24h time window. Quarters that received a score from 2 through 5 (n=348) were considered as quarters with CM. These included 74 quarters with watery milk, 217 quarters with flakes, 50 quarters with clots and 7 serum-like quarters. There were 6 quarters containing blood (score 6). These were excluded from further analyses, as milk with blood does not necessarily imply CM, but can also be caused by damage in the udder or of the teat. From the 348 CM cases, 243 quarters were sampled for bacteriological culturing, with the majority (89.3%) being sampled twice. In this study only culture results of major CM causal pathogens (*Escherichia coli, Staphylococcus aureus, Staphylococcus dysgalactiae, Streptococcus uberis, and Klebsiella SPP*) were included for further analyses. If a

culture showed a combination of major CM causal pathogens (e.g., an *S. aureus* in combination with an *S. dysgalactiae*), the quarter was excluded. Combinations of one major CM causal pathogen with a 'mixed culture', 'no growth', or '*Bacillus SPP*' within one milk sample or over milk samples of the same quarter were included. In total, 140 quarters with CM were included for further analyses; all of them were labeled according to the major CM causal pathogen cultured. This resulted in 26 quarters infected with *E. coli*, 4 with *Klebsiella SPP*, 22 with *S. aureus*, 72 with *S. uberis*, and 16 with *S. dysgalactiae*. Quarters infected with *E. coli* or *Klebsiella SPP* were classified as Gram-negative CM cases (n=30), and all others were classified as Gram-positive CM cases (n=110). Table 1 summarizes the distribution of the major CM causal pathogens isolated from bacteriological culturing of the 140 CM cases, their Gram-status, and their mastitis score recorded by the participating dairy farmers.

Table 1. Bacteriological culturing results of 140 clinical mastitis cases with sensor measurements and with at least one milk sample being bacteriological cultured, their Gram-status, and their mastitis score as recorded by the participating dairy farmers

Major pathogen isolated	Gram-status	Score assigned to quarter by dairy farmer				Total (n)
		Watery (n)	Flakes (n)	Clots (n)	Serum-like (n)	
E. coli	Negative	6	13	6	1	26
Klebsiella SPP	Negative	2	1	1	-	4
S. aureus	Positive	2	19	1	-	22
S. dysgalactiae	Positive	-	14	2	-	16
S. uberis	Positive	5	40	22	5	72
Total		15	87	32	6	140

Model development and validation for Gram-status prediction

In order to develop a decision tree model that predicts the Gram-status of a CM causal pathogen, a training set and a test set were constructed. Two thirds of all data was selected for training and the remaining third for testing, where quarters were stratified according to their Gram-status. This stratification resulted in a training set with 21 Gram-negative and 75 Gram-positive CM cases. The test set included nine Gram-negative and 35 Gram-positive CM cases.

All CM cases in the training set had 1,065 independent variables based on sensor measurements. To prevent overfitting of the decision tree, first all independent variables that had an information gain ratio higher than 0.01 were selected. The information gain ratio of an independent variable *X* is based on the change in information value (or entropy) of a dataset *S* with respect to the dependent variable *Y*, after partitioning *S* using the values of independent variable *X*. The information gain ratio takes into account the number of classes and records per class of *X* (Witten and Frank, 2005). Table 2 lists the 16 independent variables in the training set that met this information gain ratio requirement.

Table 2. Abbreviations, full descriptions, and the information gain ratio value of the 16 independent variables that were selected from the training set for predicting Gram-status, and one independent variable (abbreviation and full description) that was selected from the training set for predicting the actual CM causal pathogen. The selected variables all had an information gain ratio higher than 0.01

Variable name (abbreviated)	Variable description	Information gain ratio
Predicting Gram-status		
wcdtmax	Maximum value of the electrical conductivity of a whole quarter milking	0.2923
fcdtmax	Maximum value of the electrical conductivity of the first 500 mL of quarter milk produced	0.1855
lcdtmax	Maximum value of the electrical conductivity of the last 500 mL of quarter milk produced	0.1693
lcdtmean	Mean value of the electrical conductivity of the last 500 mL of quarter milk produced	0.1693
wcdtrangeP2	The range (maximum – minimum value) of the electrical conductivity of the whole quarter milking compared with the mean value for the range of the two previous milkings for that same quarter	0.1511
fcdtmiddevPR36	The deviation of the electrical conductivity between the actual midpoint observation and the expected midpoint observation from the first 500 mL of quarter milk produced compared with the mean value for this variable of all milkings within the previous 36h for that same quarter	0.1419
wblueincrFLQ2	The increase from the first to the last sensor measurement of the whole quarter milking of the color sensor blue compared with the mean value of the increase of two other quarters within the same cow milking that showed the highest values for this increase	0.1075

Variable name (abbreviated)	Variable description	Information gain ratio
Predicting Gram-status		*(continued)*
startmilkflowdelayP2	Time difference between teat cup attachments and start of milk flow, compared with the mean time difference of the two previous milkings of that same quarter	0.1008
lredmiddevP2	The deviation of the color sensor red between the actual midpoint observation and the expected midpoint observation from the last 500 mL of quarter milk produced compared with the mean value for this variable of the previous two milkings for that same quarter	0.0732
wcdtrange	The range (maximum – minimum value) of the electrical conductivity of the whole quarter milking	0.0675
lcdtincrFM	The increase from the first to the middle sensor measurement of the electrical conductivity of the last 500 mL of quarter milking produced	0.0603
lcdtincrFL	The increase from the first to the last sensor measurement of the electrical conductivity of the last 500 mL of quarter milking produced	0.0464
lcdtslope	The increase from the first to the last sensor measurement of the electrical conductivity of the last 500 mL of quarter milk produced / the number of sensor measurement recordings in this last 500 mL of quarter milk produced	0.0464
lcdtmin	The minimum value of the electrical conductivity of the last 500 mL of quarter milk produced	0.0401
wcdtslope	The increase from the first to the last sensor measurement of the electrical conductivity of the whole quarter milking / the number of sensor measurement recordings in quarter milking	0.0313
wcdtincrFL	The increase from the first to the last sensor measurement of the electrical conductivity of the whole quarter milking	0.0260
Predicting CM causal pathogen		
lcombincrFMP2	The increase from the first to the middle sensor measurement of the mean color value (average of red, green, and blue sensor measurements) of the last 500 mL of quarter milk produced compared with the mean value for the increase from the first to the middle sensor measurements of the previous two milkings for that same quarter	0.1390

Decision tree induction was applied to the training set. A decision tree is a graphic representation of a divide-and-conquer approach of a classification problem and consists of nodes at which a variable is tested. Based on its information gain ratio, an independent variable is selected to split a data set at the first node. For each possible outcome of the test involved at that node, a branch is made ending in a daughter node. Next, the process can be repeated for each branch, using only those records that actually reach that branch. If at any time all records at a node have the same classification, that part of the tree stops developing (Kamphuis et al., 2010b; Witten and Frank, 2005). To build a decision tree, the J48 algorithm as implemented in WEKA (Witten and Frank, 2005) with default settings was used. As the number of Gram-negative CM cases was much lower than Gram-positive CM cases (21 vs. 75, respectively), a cost matrix was introduced to account for this imbalance (Kamphuis et al., 2010a). To prevent the selection of correlated variables in the development of a Gram-status predicting decision tree, first 16 univariate decision trees (including the cost matrix) were developed; all 16 variables listed in Table 1 were used one by one to split the training data according to their Gram-status. To select the best univariate decision tree, the kappa value was used. The kappa value is a test of agreement between two tests, in absence of a golden standard, which takes into account the agreement due to chance between these two tests (Dohoo et al., 2009). The decision tree resulting in the highest kappa value was selected, and a forward selection procedure started including the variable chosen in the first selection round and all 15 remaining variables added one by one. This procedure was continued until the kappa value no longer improved. The decision tree that was developed at that point was used to predict the Gram-status of the CM cases in the test set, where the output was a probability estimate for a quarter to have a CM infection caused by a Gram-positive or a Gram-negative pathogen. Kappa value and accuracy of this test set were evaluated as well.

Model development and validation for pathogen prediction

Decision tree induction was also applied to sensor data to develop a model that predicts the CM causal pathogen itself. However, as the number of CM cases classified with *Klebsiella SPP* was very limited (n=4), these CM cases were excluded from the training set (n=3) and test set (n=1). The remaining 93 records in the training set were used to select independent variables with an information gain ratio higher than 0.01. This selection resulted in one variable that fulfilled this requirement (see Table 2). Therefore, a further forward selection procedure was not performed. A final decision tree was trained using a cost matrix that balanced the pathogen ratio in this training set and the single independent variable that had a gain ratio higher than 0.01. Also this decision tree was evaluated with data from

the test set, containing 43 CM cases, with the kappa value and the accuracy as evaluation measures.

The data mining software WEKA (Witten and Frank, 2005) was used to select variables based on their information gain ratio, to perform the forward selection of independent variables, and to develop the final decision tree. The PROC FREQ procedure in SAS (version 9.1, SAS Institute Inc., Cary, NC) was used for computing kappa values and accuracies.

Results

Table 1 shows the distribution of the major pathogens isolated from bacteriological culturing of the 140 CM cases, their Gram-status, and their mastitis score recorded by the participating dairy farmers. The pathogen isolated mostly was S. *uberis*. Most quarters were scored as having flakes or clots in their milk (85%). Compared to Gram-positive CM cases, Gram-negative CM cases were often scored as having watery milk (6.4% *vs*. 26.7% for Gram-positive and Gram-negative pathogens, respectively).

Table 2 lists 17 independent variables with an information gain ratio higher than 0.01. The first 16 of these listed variables were selected for predicting the Gram-status, with information gain ratios ranging from 0.0260 to 0.2923. Most of these variables are based on the sensor measuring the electrical conductivity, although the sensors measuring the colors red and blue also seem informative (based on their information gain ratio value). Selected variables are based on absolute values, but also on comparisons with previous milkings or on other quarters within the same cow milking.

Figure 1 shows which variables are selected by the final decision tree (with the highest kappa value) to predict the Gram-status of the CM causal pathogen. The accuracy of this final model based on data from the training set was 90.6%, and the kappa value was 0.76. This final decision tree was constructed after three forward selection rounds, and used two independent variables based on the electrical conductivity, and one based on the color sensor blue. All three selected variables were used more than once, and in total the decision tree uses seven test nodes (ovals in Figure 1) in classifying all 96 CM cases. The three independent variables used by this decision tree were not the variables with the highest information gain ratios. The numbers between brackets in the end nodes (rectangles in Figure 1) are used to calculate the probability estimates for a CM case to be infected with a Gram-positive or a Gram-negative CM causal pathogen.

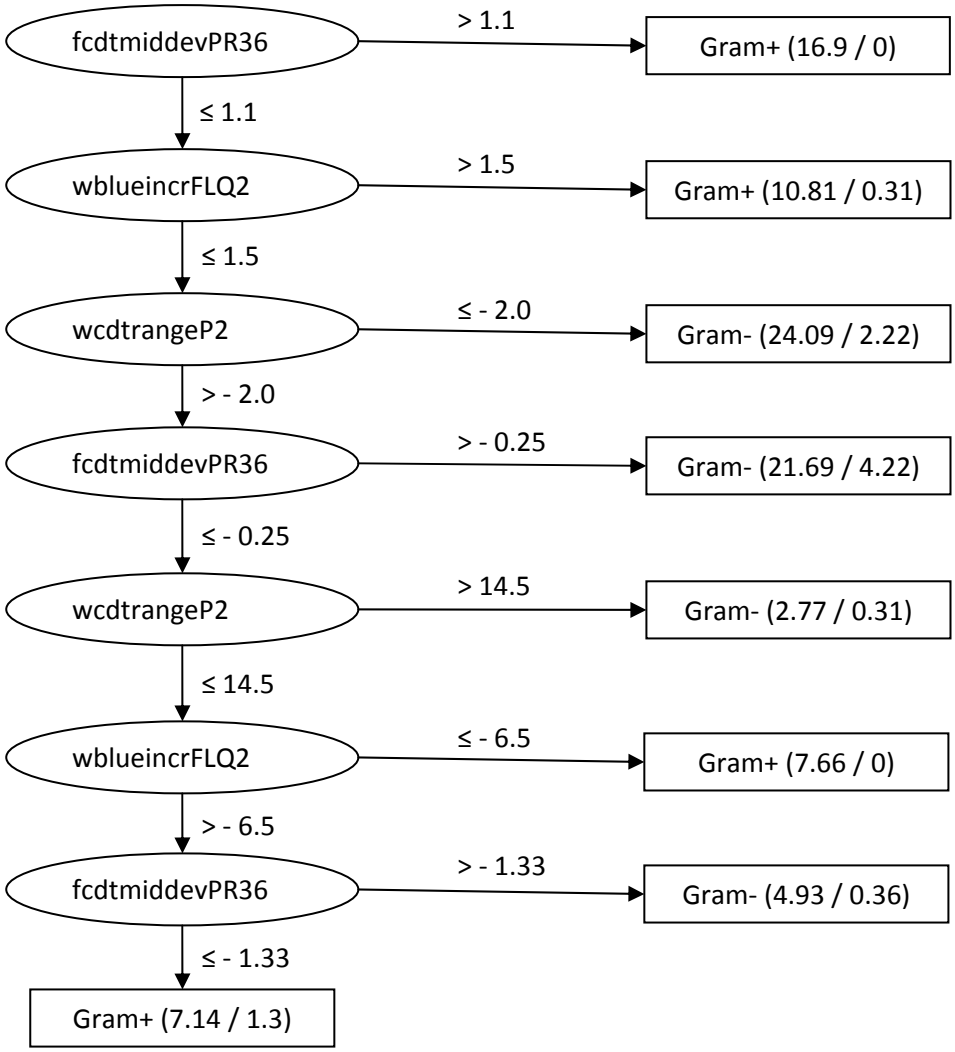


Figure 1. Final decision tree using three different independent variables to classify 96 quarters in the training set for to their Gram-status. The ovals represent test nodes at which a variable is tested (variable names are explained in Table 2). The rectangles represent end nodes at which a prediction is made (being infected with a Gram-positive (Gram+) or Gram-negative (Gram-) clinical mastitis causal pathogen) for those records reaching this rectangle. The first figure between brackets presented in these rectangles summarizes the total number of records reaching this rectangle; the second figure represents the records that are misclassified by the decision tree. The selected variables had a gain ratio >0.01

The first number between brackets indicates the number of CM cases that flow into this end node; the second value indicates the number of records that are misclassified by the tree. The figures in brackets in these end nodes show decimals due to the way the J48 algorithm deals with records with missing values for the variables on which a test at a node applies (Witten and Frank, 2005). For example, if a CM case ends up in the second end node from the top, it will be classified by the decision tree as being infected with a Gram-positive CM causal pathogen. This end node contains 10.81 CM cases, of which 0.31 are misclassified (and thus were labeled as Gram-negative CM cases). When a CM case ends up in this end node, the decision tree will assign a probability estimate of 97.1% that this CM case is infected with a Gram-positive CM causal pathogen. This percentage is calculated as 100*((10.81 – 0.31) / 10.81). This CM case will receive a probability estimate of 2.9% by the decision tree to be a Gram-negative CM case.

Figure 2 presents the number of Gram-negative CM cases (white area within a bar) or Gram-positive CM cases (grey area within a bar) per class of probability estimate in the training set. These probability estimate classes are estimates for a quarter to be a Gram-positive CM case. However, a low probability estimate for a Gram-positive CM case coincides with a high probability estimate for a Gram-negative CM case. The higher the class of probability estimate, the more accurate the prediction for a Gram-positive CM case gets. But also, the lower the class of probability estimate, the more accurate the prediction for a Gram-negative CM case gets. There were 65 quarters out of the 96 (67.7%) that received a probability estimate of > 0.60 for being a Gram-positive CM case. All of these CM cases were indeed Gram-positive cases. An extreme low probability estimate for a Gram-positive CM case (≤ 0.10), which is thus similar to an extreme high probability estimates for a Gram-negative CM case (> 0.90), also shows an accurate prediction: 12 out of the 96 CM cases (12.5%) received such a high probability estimate for a Gram-negative CM case, and 83.3% of these 12 CM cases were indeed Gram-negative cases.

The accuracy of the final decision tree for Gram-status prediction dropped to 54.5% when applied to the test set. The kappa value decreased to -0.20. Figure 3 presents the number of Gram-negative CM cases (white area within a bar) and the number of Gram-positive CM cases (grey area within a bar) per class of probability estimate in the test set. Again, the probability estimate classes are estimates for a quarter to be a Gram-positive CM case. The figure explains visually the decrease in both accuracy and kappa value. First of all, even at extremely high probabilities (e.g., >0.90) for a Gram-positive CM case, still 7 out of the 27 CM cases (25.9%) were Gram-negative. And at extremely low probabilities (e.g., ≤ 0.10) for Gram-positive CM cases (which is similar to an extreme high probability of > 0.90 for a Gram-negative CM case), still 4 out of the 5 CM cases (80%) are Gram-positive.

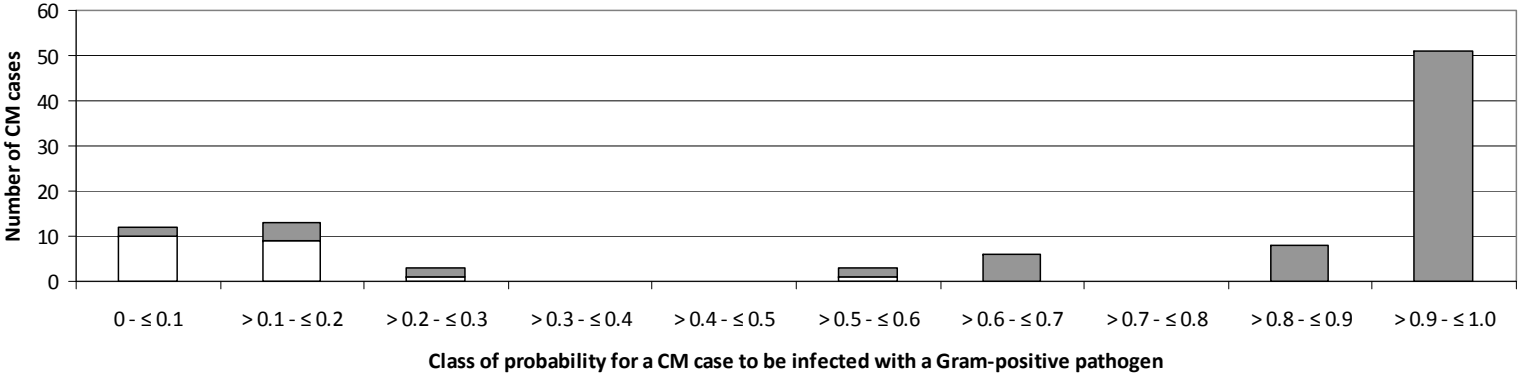


Figure 2. The number of clinical mastitis (CM) cases in the training set infected with a Gram-positive (grey area within a bar) or a Gram-negative (white area within a bar) CM causal pathogen (y-axis) per class of probability estimate for a CM case to be infected with a Gram-positive pathogen (x-axis). In total, 96 CM cases were included in the training set

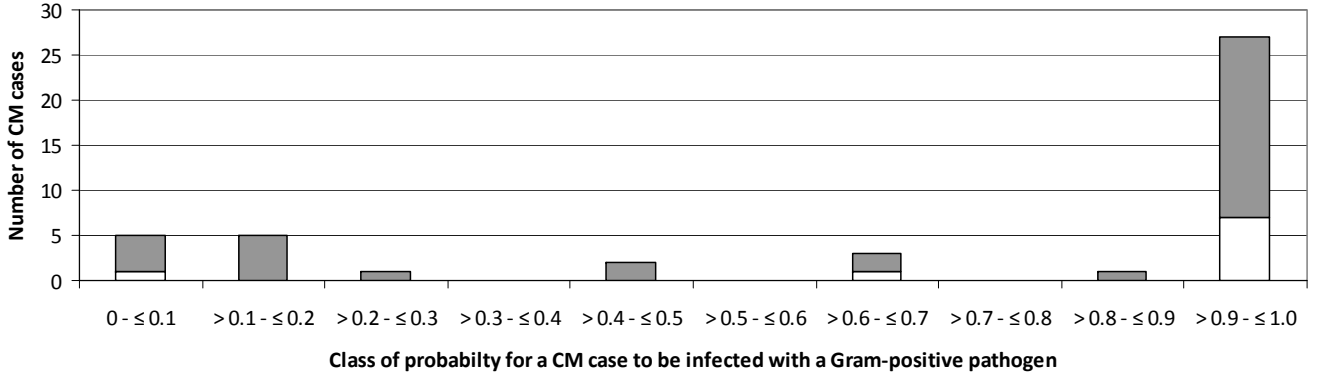


Figure 3. The number of clinical mastitis (CM) cases in the test set infected with a Gram-positive (grey area within a bar) or a Gram-negative (white area within a bar) CM causal pathogen (y-axis) per class of probability estimate for a CM case to be infected with a Gram-positive pathogen (x-axis). In total, 44 CM cases were included in the test set

There was only one independent variable that fulfilled the requirement of having an information gain ratio higher than 0.01 for predicting the actual CM causal pathogen (Table 2). This variable averaged the three color sensor measurements and made a comparison with previous milkings. The information gain ratio of this variable was 0.1390. However, the application of the cost matrix that was used to account for the imbalance between CM pathogens in the training set, made the final decision tree such that it did not use this variable at all. Instead, the model simply predicted all CM cases to be *S. uberis* CM cases. Based on the 93 records in the training set, the accuracy of this model was 52.7% and the kappa value was 0.0. When applied on the test set, accuracy was 53.4% and the kappa value stayed 0.0.

Discussion

Only 16 out of 1,065 potentially predictive variables had an information gain ratio higher than 0.01 for predicting the Gram-status of a CM case (Table 2). This indicates that all other variables had an information gain ratio value so low, that these had no discriminative power to predict the Gram-status. The electrical conductivity has been used often for the development of sensor based CM detection models, but it has not been suggested to be useful for pathogen detection before. Most of the listed variables in Table 2 are based on the sensor measuring the electrical conductivity, indicating that this information seems to be useful for Gram-status prediction. Also the final decision tree for predicting the Gram-status of CM causal pathogens used two out of three variables that were based on electrical conductivity measurements. The third one was based on the color sensor blue. The selected variables characterized the variability or shape of sensor measurements, and they were based on comparisons with previous quarter milkings or with other quarters in the same cow milking. These results coincide with results from the first study on the potential of color sensors for pathogen prediction by Espada and Vijverberg (2002), which served as basis for the study described in this paper. Although that study included just a very small number of cases (6 cows with abnormal milk in one quarter), they concluded that abnormalities in milk (e.g., clots) caused a significant change of milk color in comparison with milk from other quarters in the same milking or with previous milkings. The findings of the current study are also in agreement with results from Kamphuis et al. (2008). They concluded that sensor data from the electrical conductivity and the color sensors blue and green contained the most information for abnormal milk or CM classification, and that variables based on the variability or shape (e.g., the range

or increase) of sensor measurement patterns could be as predictive as variables based on the level of sensor measurement patterns (e.g., the mean value).

The three independent variables the final decision tree selected (Figure 1) were not the three variables with the highest information gain ratios (see Table 2). The combination of these three variables, however, resulted in the decision tree (Figure 1) with the highest kappa value (0.76). This kappa value can be interpreted as a substantial agreement (Dohoo et al., 2009) between the classifications of the decision tree and the labeling of quarters according to their bacteriological culturing results as Gram-positive or negative CM cases. The accuracy of this final decision tree was 90.6%. For comparison, a simple model that would classify CM cases based on prevalence would have an accuracy of 66.7%, and a model that would classify all CM cases as Gram-positive cases would have an accuracy of 78.1%. These results suggest that the three sensor based variables used by the decision tree were able to discriminate between Gram-positive and Gram-negative CM cases.

Figure 2 shows the range of probability estimates that the final decision tree assigned to the CM cases in the training set. This indicates that the decision tree is more confident for some quarters to be Gram-positive CM cases than for others. Quarters with an extreme high probability (e.g., > 0.90) for being a Gram-negative CM case (and thus having a low probability estimate to be a Gram-positive CM case) are in 83.3% indeed Gram-negative. Also, quarters receiving a probability estimate of > 0.60 to be a Gram-positive case were indeed all Gram-positive CM cases. Results from Figure 2 also indicate that Steeneveld et al. (2009) and Van der Gaag (2009) made sensible suggestions that presenting a probability distribution for the causal CM pathogen would be more informative for dairy farmers as it reveals the uncertainty involved with a binary or multiclass classification. Both studies discuss the presentation of a stratified accuracy, a measure which is based upon different strata of the dataset under study. When this idea is applied to the current study, it would be possible to present to dairy farmers only those quarters with extreme high probability estimates for being a Gram-positive CM case (e.g., > 0.80) or a Gram-negative CM case (e.g., > 0.90, which is similar as a probability of ≤ 0.10 for being a Gram-positive CM case). By doing this, still 74% from the 96 quarters will have a probability estimate presented to the dairy farmer, with a stratified accuracy of 97.2% (results not shown).

The decision tree from Figure 1 was also applied to unseen data to validate its performance. Results showed a dramatic drop in kappa value to -0.20, indicating a poor agreement (Dohoo et al., 2009) between the classifications of the decision tree and the labeling according to the bacteriological culturing. Also the accuracy decreased considerably to 54.4%. This indicates that the three selected variables based on data in the training set which were unable to predict the Gram-status of

the CM cases in the test set and that the trained decision tree was not a robust one. The drop in kappa value was expected when the variables with a gain ratio higher than 0.01 in the test set were evaluated (results not shown); just one out of 18 variables with a gain ratio higher than 0.01 in the test set, also showed a gain ratio higher than 0.01 in the training set (wcdtrangeP2, see Table 2). The accuracy of 54.5% is just slightly higher than the accuracy of 48% that White et al. (1986) observed for inexperienced clinicians when predicting the Gram-status of a mastitis infection. The accuracy of the test set in this study was much lower than the accuracy of 73% reported by Steeneveld et al. (2009) and 79% by Milne et al. (2003). These studies were based on much more CM cases used for analyses with 573 CM cases for Milne et al. (2003) and 3,833 CM cases for Steeneveld et al. (2009). Using more CM cases for model development will certainly improve robustness of a model and probably also the accuracy and kappa value. In this study, however, there were 140 CM cases with results from bacteriological culturing and sensor data. This is a large amount in the field of CM detection using sensor data from automatic milking systems as most studies used less CM cases, e.g., 19 CM cases by Mollenhorst et al. (2010) and 36 CM cases by Song et al. (2010). Still, the 96 CM cases used for training the decision tree seemed not enough to capture all sensor measurement characteristics of Gram-positive or Gram-negative CM cases to develop a robust decision tree model.

Decision tree induction was not able to build a model for CM pathogen prediction using sensor measurements from automatic milking systems: it simply classified all CM cases as *S. uberis* cases, resulting in a model with an accuracy of 52.6% or 53.5% (equal to the proportion of *S. uberis* cases in the training set or test set, respectively), resulting in a kappa of 0.0. Results from this study suggest that sensor measurements do not have enough discriminative power to predict the CM causal pathogen. This may indicate that other information sources perhaps are more suitable for pathogen prediction. For example, Steeneveld et al. (2009) used cow information to predict the CM causal pathogen. Their model was able to predict the CM causal pathogen with an overall accuracy of 52%, and a stratified accuracy of 89% for 4% of all CM cases. However, analyses of more CM cases with more CM causal pathogens per category, is necessary to confirm this lack of discriminative power of sensor measurements for CM pathogen prediction using decision tree induction.

Although the current study included only 140 CM cases and that a robust decision tree for Gram-status prediction could not be build, results from this study suggest that sensor measurements are useful for predicting the Gram-status of CM causal pathogens. If such a model becomes applicable in practice, this could affect the dairy farmer's choice on antibiotic treatment. For example, a dairy farmer in practice being confronted with a quarter with an extreme high probability estimate

for being a Gram-negative CM case (e.g., > 0.90), might better start a supportive treatment to relieve the cow of systemic signs with fluids and other supportive care than to start an antibiotic treatment (Erskine et al., 2003; Pyörälä and Pyörälä, 1998). If a reduction of antibiotics used in the dairy industry is induced, the real or perceived concerns about antibiotic use as a human health hazard by the regulatory sector and the milk consuming public may be reduced. Future work should include more CM cases to develop a more robust decision tree for Gram-status prediction.

Conclusion

A decision tree for predicting the Gram-status of CM causal pathogens showed a kappa of 0.76 and an accuracy of 90.6% for the training set. The kappa value and accuracy decreased to -0.20 and 54.5%, respectively, when the decision tree was validated with data from the test set. These figures indicate that the developed decision tree was not a robust one, but that there is potential in using in-line sensor measurements from automatic milking systems to differentiate between Gram-positive and Gram-negative CM pathogens.

Acknowledgements

We acknowledge the participating dairy farmers for their contribution to the acquisition of the data used in this research. Also acknowledged is Lely Industries N.V., Maassluis, The Netherlands) for providing addresses of dairy farmers working with automatic milking systems and their support with the installation of the data acquisition systems at the participating farms. This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs.

References

De Haas, Y., RF. Veerkamp, H.W. Barkema, Y.T. Gröhn, and Y.-H. Schukken. 2004. Associations between pathogen-specific cases of clinical mastitis and somatic cell count patterns. J. Dairy Sci. 87(1):95-105.

Dohoo, I., W. Martin, and H. Stryhn. 2009. Veterinary Epidemiology Research. 2nd edition. S.M. McPike (ed.). VER Inc., Charlottetwon, Prince Edward Island, Canada.

Erskine, R., S. Wagner, and F. DeGraves. 2003. Mastitis therapy and pharmacology. Veterinary Clinics of the North America / Food Animal Practice. 19:109-138.

Espada, E., and H. Vijverberg. 2002. Milk colour analysis as a tool for the detection of abnormal milk. Pages 28-38 in First North American Conference on Robotic Milking. Wageningen Pers, Wageningen, the Netherlands.

Godden, S., A. Lago, R. Bey, K. Leslie, P. Ruegg, and R. Dingwell. 2007. Use of on-farm culture systems in mastitis control programs. Pages 1-9 in NMC Regional Meeting Proceedings. Visalia, California.

Halasa, T., K. Huijps, O. Østerås, and H. Hogeveen. 2007. Economic effects of bovine mastitis and mastitis management: A review. Veterinary Quarterly. 29(1):18-31.

Harmon, R.J., R.J. Eberhart, D.E. Jasper, B.E. Langlois, and R.A. Wilson. 1990. Microbiological Procedures for the Diagnosis of Bovine Udder Infection. National Mastitis Council, Inc. Arlington, VA.

Heald, C.W., T. Kim, W.M. Sischo, J.B. Cooper, and D.R. Wolfgang. 2000. A computerized mastitis decision aid using farm-based records: An artificial neural network approach. J. Dairy Sci. 83(4):711-720.

Hillerton, J.E., and K.E. Kliem. 2002. Effective treatment of *Streptococcus uberis* clinical mastitis to minimize the use of antibiotics. J. Dairy Sci. 85(4):1009-1014.

Hogeveen, H., and W. Ouweltjes. 2003. Sensors and management support in high-technology milking. J. Anim. Sci. 81(Suppl. 3):1-10.

Huijps, K., T.J.G.M. Lam, and H. Hogeveen. 2008. Costs of mastitis: facts and perception. J. Dairy. Res. 75(1):113-120.

Jones, G.F., and G.E. Ward. 1990. Evaluation of a Scheme for Predicting the Gram-Staining Reaction of Organisms Causal Bovine Mastitis. Journal of the American Veterinary Medical Association. 196(4):597-599.

Kamphuis, C., H. Mollenhorst, A.J. Feelders, D. Pietersma, and H. Hogeveen. 2010a. Decision-tree induction to detect clinical mastitis with automatic milking. Computers and Electronics in Agriculture. 70(1):60-68.

Kamphuis, C., H. Mollenhorst, J.A.P. Heesterbeek, and H. Hogeveen. 2010b. Detection of clinical mastitis with sensor data from automatic milking systems is improved by using decision-tree induction. Journal of Dairy Science. 93 (8): 3616-3627.

Kamphuis, C., D. Pietersma, R. van der 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.

Kim, T., and C.W. Heald. 1999. Inducing inference rules for the classification of bovine mastitis. Computers and Electronics in Agriculture. 23(1):27-42.

MacDonald, K., D. Poole, and A. Muckle. 2010. Preliminary assessment of the accuracy of on-farm diagnosis of clinical mastitis using 3M Petrifilms compared to standard microbiology. Pages 481-482 in Mastitis research into practice: Proceedings of the 5th IDF mastitis conference. J.E. Hillerton (ed). VetLearn, Wellington, New Zealand.

McDougall, S., K.E. Agnew, R. Cursons, X.X. Hou, and C.R.W. Compton. 2007. Parenteral treatment of clinical mastitis with tylosin base or penethamate hydriodide in dairy cattle. J. Dairy Sci. 90(2):779-789.

Milne, M.H., A.M. Biggs, J.L. Fitzpatrick, G.T. Innocent, and D.C. Barrett. 2003. Use of clinical information to predict the characteristics of bacteria isolated from clinical cases of bovine mastitis. Veterinary Record. 152(20):615-617.

Mollenhorst, H., P.P.J. van der Tol, and H. Hogeveen. 2010. Somatic Cell Count Assessment at Quarter or Cow Milking Level. J. Dairy Sci. 93:3358-3364.

Pyörälä, S. H. K., and E. O. Pyörälä. 1998. Efficacy of parenteral administration of three antimicrobial agents in treatment of clinical mastitis in lactating cows: 487 cases (1989-1995). Journal of the American Veterinary Medical Association. 212(3):407-412.

Song, X., S. Zhuang, and P. P. J. van der Tol. 2010. New model to detect clinical mastitis in Astronaut A3 Next^{tm} milking robot. Pages 474-480 in Mastitis research into practice: Proceedings of the 5th IDF mastitis conference. J.E. Hillerton (ed). VetLearn, Wellington, New Zealand.

Steeneveld, W., L.C. Van der Gaag, H.W. Barkema, and H. Hogeveen. 2009. Providing probability distributions for the causal pathogen of clinical mastitis using naive Bayesian networks. J. Dairy Sci. 92(6):2598-2609.

Van der Gaag, L.C., S. Renooij, W. Steeneveld, and H. Hogeveen. 2009. When in Doubt ... Be Indecisive. Pages 518-529 in Proceedings of the 10th European Conference on Symbolic and quantitative approaches to reasoning with uncertainty. C. Sossai and G. Chemello (eds). Springer Verlag, Berlin / Heidelberg.

White, M.E., L.T. Glickman, F.G. Barnes-Pallesen, E.S. Stem III, P. Dinsmore, M.S. Powers, P. Powers, M.C. Smith, and D. Jasko. 1986. Accuracy of Clinicians in Predicting the Bacterial Cause of Clinical Bovine Mastitis. Canadian Veterinary Journal. 27:218-220.

Witten, I.H., and E. Frank (eds.). 2005. Data Mining; Practical Machine Learning Tools and Techniques. 2nd edition. Morgan Kaufmann Publishers, San Fransisco.

Chapter 6

Automated detection of clinical mastitis is improved by on-line monitoring of somatic cell count

C. Kamphuis, R. Sherlock, J. Jago, G. Mein, H. Hogeveen

Published
Journal of Dairy Science. 2008. 91 (12): 4560-4570

Abstract

This study explored the potential value of on-line composite somatic cell count (ISCC) sensing as a sole criterion or in combination with quarter-based electrical conductivity (EC) of milk, for automatic detection of clinical mastitis (CM) during automatic milking. Data generated from a New Zealand research herd of about 200 cows milked by two automatic milking systems during the 2006–2007 milking season included EC, ISCC, monthly laboratory-determined SCC, and observed cases of CM that were treated with antibiotics. Milk samples for ISCC and laboratory-determined SCC were taken sequentially at the end of a cow milking. Both samples were derived from a composite cow milking obtained from the bottom of the milk receiver. Different time windows were defined in which true positive, false negative, and false positive alerts were determined. Quarters suspected of having CM were visually checked and, if CM was confirmed, sampled for bacteriological culturing and treated with an antibiotic treatment. These treated quarters were considered as gold-standard positives for comparing CM detection models. Alert thresholds were adjusted to achieve a sensitivity of 80% in 3 detection models: using ISCC alone, EC alone, or a combination of these. The success rate (also known as the positive predictive value) and the false alert rate (number of false positive alerts per 1,000 cow milkings) were used to evaluate detection performance. Normalized ISCC estimates were highly correlated with normalized laboratory-determined SCC measurements (r=0.82) for SCC measurements $>200\times10^3$ cells/mL. Using EC alone as a detection tool resulted in a range of 6.9 to 11.0% for success rate, and a range of 4.7 to 7.8 for the false alert rate. Values for the ISCC model were better than the model using EC alone with 12.7 to 15.6% for the success rate and 2.9 to 3.7 for the false alert rate. Combining sensor information to detect CM, by using a fuzzy logic algorithm, produced a 2- to 3-fold increase in the success rate (range 21.9 to 32.0%) and a 2- to 3-fold decrease in the false alert rate (range 1.2 to 2.1) compared with the models using ISCC or EC alone. Results suggest that the performance of a CM detection system improved when ISCC information was added to a detection model using EC information.

Key words

Clinical mastitis detection, automatic milking, on-line somatic cell count, electrical conductivity

Introduction

Detection of clinical mastitis (CM) is important to maintain an acceptable standard of milk quality, to initiate an antibiotic treatment when necessary, and to safeguard the welfare of the herd. Among others, the European and Australasian legislations dealing with milk production state that milk presenting abnormalities or milk from diseased or injured udders must be excluded for human consumption. Detection of abnormal milk should be done by the milker or by methods achieving similar results; for example, electronic mastitis detectors [Regulation (EC) No 853/2004; NZCP1: Code of Practice, Version 5]. Automatic milking (AM) systems today use several types of electronic detectors or sensors and management information systems to assist the herdsman in detecting cows with CM. Because the AM system lacks milkers who check the milk visually, it uses sensor data to collect data about milk quality and to classify milk for CM or normality. Electrical conductivity (EC) is the most widespread method of automatic detection of CM (Mottram et al., 2007); yield and temperature are also commonly used in-line milk sensing methods (Hogeveen and Ouweltjes, 2003; Reinemann and Helgren, 2004). These sensors measure during milking at the cow and quarter levels and this information is used as input for the CM detection models used by the AM systems. The outputs of these models, also generated by the AM system, are at the quarter level and are so-called mastitis alert lists; that is, a list of cows with quarters likely to have mastitis. The herdsman must visually check the cows and quarters identified on these lists to validate the warnings that were generated automatically.

Early research focused on the EC of milk as a detection tool for CM because it offered a relatively simple and inexpensive technique for implementation in-line (Hogeveen and Ouweltjes, 2002; Mottram et al., 2007). The change from univariate models (Maatje et al., 1992) via multivariate logistic regression models including yield and temperature (Nielen et al., 1995) and more sophisticated methods such as Kalman filters and fuzzy logic algorithms (De Mol et al., 1999; De Mol and Woldt, 2001; Cavero et al., 2006) resulted in improved detection performance. Although these models were not implemented in practice, a common complaint by farmers working with AM is the high number of false positive alerts generated (De Mol and Woldt, 2001). For example, a study conducted by Mottram et al. (2007) showed a sensitivity and specificity of a CM detection model at the quarter level using EC alone of 56% and 82%, respectively. This means that 18% of the truly healthy quarters are listed on the mastitis alert list unnecessarily (false positive alerts), and about half of the truly diseased cases will not be listed at all. Unfortunately, it is almost impossible to give an exact figure of an acceptable level of false positive alerts, because this level depends on the opinion of the dairy farmer as the end-user of an AM system. However, all models in which specificity is not 100% will cost

money. In this context, it might not be monetary costs alone, but also the costs of an undesirable additional workload (Hogeveen and Ouweltjes, 2003) due to these false positive alerts.

Along with EC, the laboratory-determined SCC has been used as an indicator or predictor of udder health and milk quality (Peeler et al., 2003; Steeneveld et al., 2008). Dairy producers using AM often use SCC information, usually determined monthly, to validate an alert but the SCC information itself has not been used by the AM system so far to generate alert lists. Recently a new on-line sensor has been developed (CellSense, Sensortec Ltd., Hamilton, New Zealand), which estimates the SCC based on viscosity measurements (Whyte et al., 2004). Initial results presented by Leslie et al. (2007) were promising, showing a correlation coefficient of 0.71 between a sensor-determined "spot-sample" and a laboratory-determined SCC of composite udder milk. To date, however, no information has been published on the potential success of on-line SCC (ISCC) information used by an automatic sensing system for CM detection.

To test the hypothesis that ISCC information estimated from a composite cow milking adds a useful contribution to the detection of CM by decreasing the number of false positive alerts, three detection models were tested. The first model used one simple threshold value for ISCC. The second model was based on software currently used by AM systems and included EC information alone. The third model combined EC and ISCC information using a fuzzy logic algorithm. The three detection models were evaluated by comparing performance parameters.

Materials and Methods

Data collection

Data were generated from a New Zealand research herd (Greenfield Project, DairyNZ, Hamilton, New Zealand), which began AM in 2001. Cows were managed as described by Jago et al. (2004). Dairy cows (80% Holstein-Friesian, remainder Jersey and Jersey-Holstein Friesian crossbred, average age: 6 yr) were seasonally calved after a dry cow period. Cows were milked, on average, 1.3 times per day by two Merlin AM systems (Fullwood Ltd., Shropshire, UK). Cows were kept outside year-round and were mainly pasture fed on 44.5 ha (500 g of concentrates per day was fed through the AM system). The average milk production for the complete 2006–2007 milking season was 3,700 kg/cow, with an average bulk tank SCC of 189×10^3 cells/mL.

Data used in this study were collected from November 22, 2006, until March 31, 2007. Data included EC information recorded continuously during milking at quarter level by the AM system. Raw EC data was pre-processed by the commercial

AM software to correct for drifts and calibration errors of EC sensors. A new sensor (CellSense, Sensortec Ltd.) for estimating ISCC as described by Whyte et al. (2004) was installed on both AM systems. The sensor was connected to the milk receiver and programmed to collect a composite udder milk sample from the bottom of the receiver can at the end of each cow milking for ISCC estimation. Every three to four weeks, composite udder milk samples were taken for laboratory determination of SCC (FSCC) using the Fossomatic technique (CombiFoss 5000, Foss Electric, Hillerød, Denmark). This milk sample, also collected from the bottom of the receiver can, was taken directly after the milk sample for ISCC estimation, before the milk was pumped into the vat. Milk samples were stored in a freezer within 8h after collection into tubes with added preservatives (Bronopol, 10% solution, Spec-Chem Industries Inc., Nanjing, China). After all cows were sampled, milk samples were taken to the laboratory where they were analyzed for FSCC within 3d. Finally, those cows that were suspected of having CM were visually checked. This check usually took place around 12:00h (noon) but the exact time was dependent upon when cows were alerted for CM and when they were drafted from the herd for checking. If CM was confirmed, by showing obvious signs of wateriness or clots in any quarter or other clear signs of CM such as swelling or redness, the quarter was sampled for bacteriological culturing and treated with an antibiotic treatment. Cows were suspected of having CM on the basis of multiple indicators, including the mastitis alert lists generated by software used by the AM system (Fullwood Ltd.). Other triggers to check cows for CM directly or for the farm staff to reach a greater level of alertness to find cows with CM included finding clots on the filter sock, bulk tank SCC information, results of bacteriological culturing (collected at calving, at dry off, and at suspicion of CM), a high ISCC estimation, and results of FSCC determinations.

Data preparation

Data for antibiotic treatments and FSCC results were merged with the EC and ISCC data collected by the AM systems. All data from three cows without any ISCC information were excluded from the data set. Specific cow milkings that were recorded as being a failure milking by the AM system (milk yield <1 L, including milkings in which the AM systems were unable to connect teat-cups to the teats) were excluded as well. Because two separate systems were used to estimate ISCC values (the AM system) and to determine FSCC (a shuttle was used to store milk samples), recordings of date and time between the two systems differed slightly. To be confident that milk samples for ISCC estimations and FSCC determinations were collected concurrently, a maximum time difference between the two date and time stamps was used: when the time difference between FSCC storage and the cow exiting the AM system was >2 minutes, values of FSCC were classified as missing.

Values of ISCC estimations and FSCC determinations were log (base 10) transformed to normalize the data and are referred to as normalized ISCC or normalized FSCC in this article.

A generated alert for CM could be assigned to a cow milking by three detection models using sensor information. The ISCC model used a simple threshold value of ~2.0 × 10^6 cells/mL: cow milkings with a value greater than this threshold were alerted as CM. The second model generated a CM alert based on EC threshold values which included a ~50% increase in the normalized quarter EC maximum value and/or a ~36% increase in the running mean (over five milkings) of the minimum quarter EC for a specific cow. The third model used a fuzzy logic algorithm (McNeil and Freiberger, 1994) to combine ISCC and EC information. This technique has been used and described previously as a detection tool for mastitis by De Mol and Woldt (2001) and Cavero et al. (2006). Fuzzy logic algorithms use linguistic variables whose values are words rather than numbers. In combination with simple IF-THEN rules, a linguistic CM likelihood variable is created. After transforming this linguistic likelihood variable into a numeric one, a simple threshold value for this CM likelihood variable is used to differentiate between normal quarter milkings and those with CM. In short, the fuzzy logic algorithm contains three steps:

1. Fuzzification:
   Numeric input variables are transformed to linguistic ones. For example, the numeric input variable ISCC is transformed to a linguistic one with the memberships low, middle, and high. These three memberships are defined by an empirically defined minimum, maximum, and midpoint value of the input variable ISCC. For example, the low membership value is 1 for values below the minimum, it linearly decreases to 0 until the maximum value is reached, and stays at 0 for values greater than the maximum value. For a particular cow, the ISCC value of 500×10^3 cells/mL may be transformed to membership values 0.0 for low, 0.6 for middle, and 0.4 for high. In linguistic terms, these membership values indicate that the ISCC of this cow is certainly not low, but rather middle and somewhat high.
2. Fuzzy inference:
   The memberships of the linguistic variables are applied in IF-THEN rules. The degree to which each part of the condition has been satisfied for each rule is known by the corresponding grades of the membership functions. A new linguistic variable, the linguistic variable of CM likelihood, is created, which also consists of the three memberships low, middle, and high. Initially, the values of these memberships are all zero, but the membership values of the input linguistic variables are accumulated in this CM likelihood variable after the IF-THEN rules have been applied.

3. Defuzzification:
   After all IF-THEN rules have been applied, the cumulative result of the linguistic CM likelihood variable is transformed back into a single numeric value that is used to compare with a threshold value to discriminate between normal quarter milkings and those with CM.

The current study used five rules applied in the fuzzy logic algorithm to provide a continuous defuzzified value of the CM likelihood variable. The first rule used the ISCC information. Because ISCC was estimated at the cow level and the fuzzy logic was used to detect CM at the quarter level, the assumption was made that all quarters within a cow milking had the same ISCC estimation. This first rule essentially captured the idea that the greater the ISCC value, and thus the lower the value for the low membership and the greater the value for the high membership of the linguistic ISCC variable, the greater the likelihood of CM. The other four rules used (combinations of) derived variables from EC data. Two of these used the same idea as the ISCC variable: an increasing EC contributed to an increased likelihood of CM. The last two rules, both using combinations of derived EC variables, contributed to an increased likelihood of CM when there was a positive trend of increasing EC values or when there were quarters with outstanding high EC values. A threshold value of ~0.80 was used for the defuzzified CM likelihood variable to distinguish healthy quarter milkings from those with CM.

The time of an alert for CM generated by a model and the time of an observed CM case will usually not coincide in an AM setting. Therefore, a finite-duration time window was associated with both an alert for and an observation of CM, a technique also applied by De Mol and Woldt (2001). The alert and observation were considered coincident when these time windows showed any overlap. Whenever a model assigned an alert to a quarter for having CM after a cow milking, the alert was extended using one of three different time windows (48, 72, and 96h) to create a period in which the alert was valid (Alertperiod). Whenever a quarter was checked and treated with antibiotics for CM, the time window in which this CM observation was valid (Obsperiod) started on the treatment day at 12:00h and ended 24 or 48h later. This Obsperiod starting point was chosen as most cows were drafted from the herd to be checked visually, and treated if necessary, for CM during mid morning. Different time windows for Alertperiod and Obsperiod were used to explore its effect on detection performance of the different models.

If there was any overlap in the Alertperiod and Obsperiod, the alert generated by the model was considered as a true positive (TP) alert. This is illustrated in Figure 1 where a single alert assigned by a model was valid for 48h, and where an observed CM was valid for 24h. The TP alert occurs in Figure 1 in the extreme situation where there is a marginal overlap in time windows, as the actual alert

precedes the actual observation of CM for quite some time. An alert was considered as a false positive (FP) alert when there was an alert but where the Alertperiod passed without an overlapping Obsperiod.

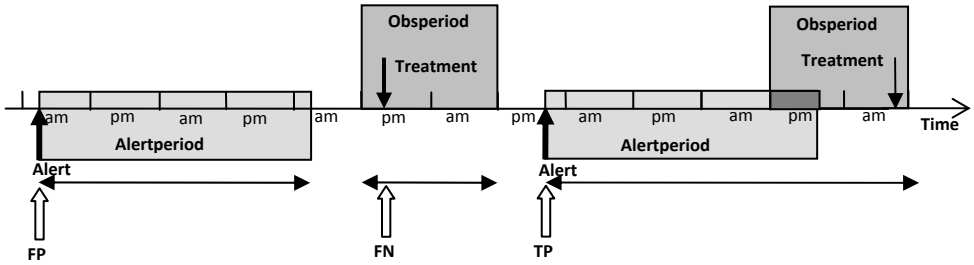


Figure 1. Example of classification of clinical mastitis (CM) alerts: one milking with an automatic milking (AM) system alert extended by 48h to define an automatic detected CM period (Alertperiod) without a treatment recording resulting in a false positive (FP) alert; one treatment date to define an observed CM case extended by 24h to define an observed CM period (Obsperiod) without an overlapping Alertperiod resulting in a false negative (FN) alert; one Alertperiod with an overlapping Obsperiod resulting in a true positive (TP) alert

If CM was observed, but there was no overlapping Alertperiod within the Obsperiod, this was considered as a false negative (FN) alert. In those situations where a model alerted a second time for CM within the time window of the first, these consecutive alerts were merged into one Alertperiod, where the period started at the first alert and ended at the second alert extended with the specified time window. This merging procedure was repeated if necessary when more consecutive alerts arose. In these situations, only one TP, FN, or FP alert could be assigned. The general rule covering all situations is that a TP alert occurs in two situations: the first is when the last alert in an Alertperiod is less than the used Alertperiod time window width before the first treatment in an Obsperiod. The second is when the first alert in an Alertperiod is less than an Obsperiod time window width after the last treatment date.

Statistical analyses

Values of ISCC and FSCC were categorized into 6 levels (≤50, 50 ≤ 100, 100 ≤ 200, 200 ≤ 500, 500 ≤ 1,000, and >1,000×10^3 cells/mL), after which a frequency table was created summarizing cow milkings that fell into the same range of SCC as estimated or determined by the two SCC measurements. Pearson correlation

coefficients were computed to see whether information about SCC was similar between ISCC estimations and the more commonly used FSCC determinations, or in other words, to study how precise ISCC estimations were when compared with the traditionally used FSCC determinations. Input values for the correlation coefficients were the normalized ISCC and normalized FSCC values. Spearman rank correlation coefficients were computed between FSCC values $\leq 200\times10^3$ cells/mL and ISCC estimations, and between FSCC values $>200\times10^3$ cells/mL and ISCC estimations. For those records with both FSCC and ISCC values, Pearson correlation coefficients were also computed between normalized ISCC values and the normalized maximum EC value within a cow milking to determine whether ISCC contained the same information as retrieved from EC measurements. Spearman rank correlation coefficients were computed between low values of ISCC ($\leq 200\times10^3$ cells/mL) and EC, and high values of ISCC ($>200\times10^3$ cells/mL) and EC.

As illustrated in Figure 1, the counts of occurrences of TP, FN, and FP are readily computed. These counts were used to calculate three performance characteristics. First, the sensitivity (Se) of the detection model was computed, which gives a measure of the fraction of all true cases of CM that were actually detected by the model. The Se (%) is calculated as (100 × TP alerts/ (TP alerts + FN alerts)). Alert thresholds were adjusted to achieve an Se of 80% in all three detection models for each different time window. So, the above-mentioned threshold value for the detection model using ISCC alone (which was $\sim2.0\times10^6$ cells/mL for all time windows) resulted each time in an Se of about 80%. To compare the three detection models, two other performance characteristics associated with the specificity were computed as suggested by Sherlock et al. (2008): the success rate, representing the proportion of quarter milkings that are true cases of CM from all the alerts generated by the model, which is also known as the positive predictive value and calculated as (100 × TP alerts/ (TP alerts + FP alerts)), and the number of FP alerts per 1,000 cow milkings (the false alert rate).

Data preparation and computations of Pearson and Spearman rank correlation coefficients were performed using SAS version 9.1 (SAS Institute Inc., Cary, NC). Model building and performance analyses were accomplished using Python programming language (www. python.org). Figures were made in S-PLUS (version 7.0, Insightful Corp., Seattle, WA).

Results

After data preprocessing, 194 cows and 27,719 cow milkings remained for analyses. The dataset included a total of 20 quarters from 13 cows that received an antibiotic treatment for CM. Two cows were treated for two different quarters on the same day. From the 20 quarter cases that were visually checked and confirmed with CM, seven quarters were infected with *Streptococcus uberis*, five with *Staphylococcus aureus*, two with *Escherichia coli*, one with CNS, one with other pathogens, and two quarters showed no growth on the plate. From two quarters no milk samples were available for bacteriological culturing.

Correlation

As FSCC was determined for all cows each three to four weeks, only a limited number of cow milkings (n=456) showed information for both ISCC and FSCC within the specified time limit of two minutes. Table 1 shows a frequency table of cow milkings falling into the same range of SCC as determined by ISCC estimation or by FSCC determination, where ISCC and FSCC values were categorized into 6 groups. Of the 283 cow milkings with a FSCC of ≤200x10^3 cells/mL, 267 (94%) also showed an ISCC of ≤200×10^3 cells/mL. There were 98 cow milkings (37%) where the ISCC and the FSCC fell into the same category below 200×10^3 cells/mL. In just one case, a cow milking was estimated to have an ISCC of more than 1,000×10^3 cells/mL, whereas the FSCC was determined to be between 50×10^3 and 100×10^3 cells/mL.

Table 1. Frequency table of cow milkings by somatic cell count (SCC) class as determined by an on-line SCC (ISCC) estimation or by a laboratory SCC determination (FSCC)

	FSCC (x 10^3 cells/mL)						
ISCC (x 10^3 cells/mL)	≤50	50 – ≤ 100	100 – ≤ 200	200 – ≤ 500	500 – ≤ 1,000	>1,000	Total
≤50	12	62	42	3	0	0	119
50 - ≤ 100	6	22	45	8	0	0	81
100 - ≤ 200	3	11	64	27	0	1	106
200 - ≤ 500	1	3	8	68	13	2	95
500 - ≤1,000	2	1	0	6	17	4	30
>1,000	0	1	0	1	2	21	25
Total	24	100	159	113	32	28	456

Another 173 cow milkings had an FSCC of >200×10^3 cells/mL, of which 134 (77%) also showed an ISCC >200×10^3 cells/mL. There were 106 cow milkings (79%) where ISCC and FSCC fell into the same category >200 × 10^3 cells/mL.

Figure 2 plots the normalized ISCC values against the normalized FSCC values to visualize the precision of ISCC estimations when compared with the generally used FSCC determinations. The dotted lines indicate normalized ISCC and normalized FSCC values of 5.3 on the log (base 10) scale, which corresponds with estimated ISCC or determined FSCC values of 200×10^3 cells/mL.

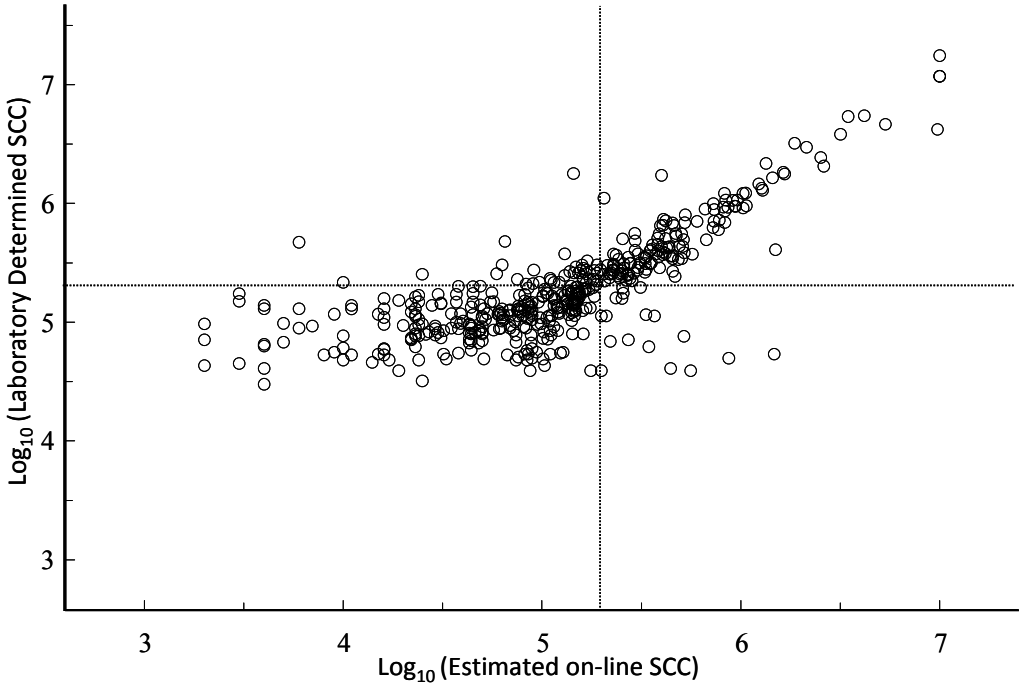


Figure 2. Normalized on-line somatic cell count (SCC) estimations against normalized laboratory determined SCC measurements

Figure 2 confirms the findings presented in Table 1 that there is more variation in ISCC values at lower values of FSCC. There is a more linear relationship between ISCC and FSCC at greater values of FSCC. This was also reflected in the calculated correlation coefficients: the Pearson correlation between the normalized ISCC and the normalized FSCC values over the whole range of data points was 0.76 (P <

0.001). Spearman rank correlation coefficients for lower normalized values of FSCC (≤5.3 on the log10 scale) and normalized ISCC was 0.33 ($P < 0.001$), whereas the correlation coefficient for greater normalized values of FSCC (>5.3 on the log10 scale) and normalized ISCC was 0.82 ($P < 0.001$).

Figure 3 plots the normalized ISCC values against the normalized maximum EC values within a cow milking. The dotted line indicates a normalized ISCC value of 5.3 on the log10 scale, which represents an estimated ISCC value of 200×10^3 cells/mL. Correlation coefficients between the normalized ISCC and the normalized maximum quarter EC value were computed to study whether ISCC contains the same information as retrieved from EC measurements. The Pearson correlation over the whole range of data points was 0.48 ($P < 0.001$). For normalized ISCC values of ≤5.3 (on the log10 scale) the Spearman rank correlation was 0.25 ($P < 0.001$; n=306) and for greater normalized values of ISCC, the correlation was estimated at 0.33 ($P < 0.001$; n=150).

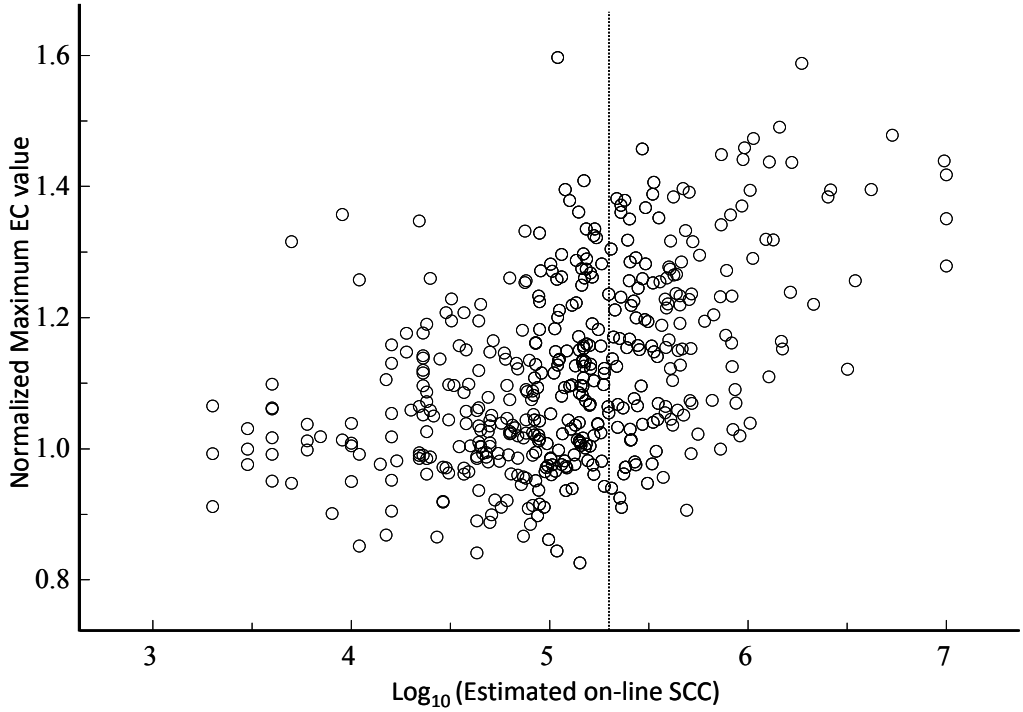


Figure 3. Normalized on-line somatic cell count (SCC) estimations against normalized maximum electrical conductivity (EC) values

Performance analyses

Table 2 shows the performance measures for the three different detection models using different time windows to define TP, FP, and FN alerts. The Se for the model using ISCC information alone was slightly greater than 80% because ISCC is measured at the cow level, resulting in merging two quarter cases into one in those two situations where cows were treated in two different quarters on the same day. Looking at the models using ISCC or EC information alone, the success rate decreased as the analyzed time windows became wider, the range being only 2.9% for ISCC and 4.1% for EC. The false alert rate increased for both models with wider time windows, but this increase seems minor with a range of 0.8 for ISCC and 3.1 for EC.

When applying the fuzzy logic algorithm to combine EC and ISCC information there was an improvement in the success rate and the false alert rate. All analyzed time windows showed an almost three-fold increase of the success rate and a three-fold decrease of the false alert rate at the same time when compared with the EC model, and a two-fold increase of the success rate and a two-fold increase of the false alert rate when compared with the ISCC model. The success rate increased from 11.0% or 15.6% for the models using EC or ISCC alone in the widest time window to 32.0% using the model with EC and ISCC information, and from 6.9% or 12.7% to 21.9% when the narrowest time window was analyzed. For the false alert rate, figures decreased from 4.7 or 2.9 FP alerts per 1,000 cow milkings when the models using EC or ISCC alone and the largest time windows was analyzed to 1.2 FP alerts per 1,000 cow milkings with the model using EC and ISCC, and from 6.9 or 3.7 to 2.1 FP alerts per 1,000 cow milkings when the narrowest time window was analyzed. The model using ISCC information alone showed less variation in the success rate or the false alert rate than the models using EC alone or where EC and ISCC information was combined. The results for the 96h– 48h time window and the 96h–24h time window for the ISCC model were the same. A closer look at the results revealed that in all CM cases, alerts were given by the model preceding the observed CM case, which made the detection performances independent of the chosen time window for the Obsperiod.

Table 2. Performance statistics using electrical conductivity (EC) and on-line somatic cell count (ISCC) information as a detection tool for clinical mastitis (CM)

		Alerts based on EC			Alerts based on ISCC			Alerts based on EC and ISCC		
Time-window Alertperiod[1] (hours)	Time-window Obsperiod[2] (hours)	Se[3] (%)	Success Rate[4] (%)	False Alert Rate[5]	Se (%)	Success Rate (%)	False Alert Rate	Se (%)	Success Rate (%)	False Alert Rate
96	48	80	11.0	4.7	83.3	15.6	2.9	80	32.0	1.2
96	24	80	8.9	5.9	83.3	15.6	2.9	80	30.2	1.3
72	24	80	8.2	6.4	83.3	13.8	3.4	80	27.1	1.6
48	24	80	6.9	7.8	83.3	12.7	3.7	80	21.9	2.1

[1]Alertperiod=automatically detected period of clinical mastitis: time of alert extended by a specified number of hours
[2]Obsperiod=Observational period for CM: starts at time of first observation of clinical symptoms [nominally 1200h (noon) on the day of inspection] and extended by the specified time window
[3]Se=Sensitivity
[4]Success Rate=also known as the positive predictive value. Representing the fraction of all true cases of CM that were actually detected by the model
[5]False Alert Rate=number of false positive alerts per 1,000 cow milkings

Discussion

One of the few published studies on on-line SCC measurements (Leslie et al., 2007) reported a correlation coefficient between ISCC readings and laboratory determined SCC values of 0.71, which is similar to the overall correlation coefficient we found in the research reported here (r=0.76). The correlation coefficient of 0.82 found in this study between normalized ISCC and normalized FSCC values for cow milkings with an FSCC $>200\times10^3$ cells/mL was of special interest because milk affected by CM has been associated with high levels of laboratory-determined SCC in previous studies (Peeler et al., 2003; Green et al., 2004). These two correlation coefficients (0.76 and 0.82) were therefore considered to be a promising first indication that ISCC contains valuable information for an automatic sensing system for CM.

The correlation coefficient found in the current study between the normalized ISCC values and the normalized maximum EC value within a cow milking was low (r=0.48). This weak correlation can be explained by the different physiological mechanisms that cause a change in EC or SCC when pathogens enter the udder: whereas EC is related to the degree of destruction of the blood-milk barrier due to pathogens entering the udder (Nielen et al., 1992), SCC is strongly related to the influx of polymorphonuclear leucocytes from the blood into the udder initiated by the cow to stop pathogens invading the udder (Kremer et al., 1990). The observation that ISCC and EC were weakly correlated combined with the knowledge that SCC and EC explain different physiological mechanisms supported our idea that combining these two sources might improve detection of CM. This view is confirmed by our results. The models using EC or ISCC alone detected a similar number of cases of CM. Consequently, no difference could be observed in types of pathogens that were detected by the different detection models, which may be caused by the small number of CM cases (results not shown). However, when parameterized for the same Se level, the ISCC model showed fewer FP alerts than the EC model, and when ISCC and EC information was combined, this decrease in FP alerts was even more pronounced. This indicates that the combination of sensor information is better able to distinguish healthy from diseased quarters. This improved result is also in agreement with an earlier study by De Mol et al. (1997), which showed that combining EC data with other sensor information improves CM detection performances compared with using EC alone, and also with more recent studies using the fuzzy logic technique to improve CM detection and to decrease the number of FP alerts (De Mol and Woldt, 2001; Cavero et al., 2006).

Results suggest that detection models including ISCC alone might produce better performance characteristics than the model using EC alone at the target Se level used in the analyses. Both models showed a similar trend of decreasing values for the success rate and increasing values for the false alert rate as time windows

shortened, with the ISCC model showing less variation in performance characteristics than the EC model. The best performance was related to the widest time interval to define a TP case (96h for Alertperiod and 48h for Obsperiod), in which a simple ISCC threshold detected CM with 1.8 fewer FP alerts per 1,000 cow milkings than the model using EC alone. Previous studies used different time windows to define TP alerts, and this difference influences how the results have been, or should be, interpreted. One of the first studies on EC as a detection tool for CM (Maatje et al., 1992) showed that all 25 cases of CM were detected by the model used (Se=100%) and that in 65% of the CM cases, the model alerted before clinical signs of mastitis appeared. However, the average number of days that this model alerted before clinical signs appeared was 14d. Using a model with such a wide time interval to define a TP case will certainly show good detection performances. An automatic sensing system, however, that alerts for CM 14d before the actual clinical signs appear is of no use for farmers working with AM because these farmers do not see the cows and quarters daily, unlike conventional milkers. A similar impractical situation occurs when an automatic detection model generates alerts several days after the actual CM case appeared, as reported by De Mol et al. (1997). Hillerton and Semmens (1999) showed that 19 out of 23 quarters incubated with *Strep. uberis* showed clinical signs of mastitis one day before changes in EC were measured and that antibiotic treatment before clinical signs appeared resulted in good cure rates. They also discussed that an early treatment will not be possible until the Se of a mastitis detection model is sufficient. Hillerton and Semmens (1999) failed to mention what a sufficient Se for mastitis detection would be, but they did stress an important issue for detection models for AM systems. Although it is tempting to use wide time intervals to increase the apparent detection performance of models, an automatic sensing system for CM should, in practice, generate an alert within a very limited period before or only at the milking when CM is present. The practical application of an automatic CM detection model should remain the focus for future research.

The fuzzy logic model in the current study showed a two- to three-fold increase in success rate and a two- to three-fold decrease in the false alert rate, the two performance parameters on which current CM sensing models are often criticized. These results mean that in practice, when applying the different models developed in this study with the most narrow time window, the number of FP alerts during the data collection period (including 27,719 cow milkings) would decrease from 216 FP alerts using the EC model, to 103 FP alerts using the ISCC model, and to 58 FP alerts when ISCC and EC information are combined. However, results must be interpreted while keeping in mind the objective of the study and the limitations of the data set used. First, the parameters of all three models were set to reach a Se of 80%, a percentage that is stated as the average sensitivity of milkers to detect

CM by visual observation (Hillerton, 2000; Rasmussen, 2005). This does not imply that the chosen threshold should be used in future models. The approach, however, was used to explore the potential of adding new on-line information (ISCC) available on AM systems to improve the detection of CM when compared or added to information generally available on AM systems today (EC). This is a different objective than optimizing detection performance of the three models used in this study, for which more data are necessary to validate the models. Furthermore, the limitations of the used data set are two-sided. First, model parameters were not optimized in the current study and optimizing parameters is likely to result in a further improvement of the success rate and the false alert rate of all three detection models, possibly even at greater Se levels. Although the ISCC model has limited options for optimization (because the ISCC is a single-point estimate for a whole cow milking), optimizing this model might be done by choosing another threshold value or by comparing ISCC values with values found in previous milkings. The model using EC alone is a more sophisticated model compared with the model using ISCC alone. It also has more optimizing options because EC is continuously measured during a quarter milking. This creates possibilities to make comparisons with EC values from other quarters within the same milking or with EC values from previous milkings. The model may include other descriptive variables than the commonly used (normalized) mean and maximum EC value (Kamphuis et al., 2008). Also, the fuzzy logic model may improve by several means: other membership threshold values can be defined that affect their contribution to the mastitis likelihood variable. The fuzzy rules themselves can be changed, or a weight could be applied to the different rules. For example, when information about ISCC is considered to be more important than information extracted from EC measurements, the fuzzy rule considering ISCC information can be valued as more important by applying a greater weight to this rule. The effect of this rule to the mastitis likelihood variable will be larger than the rules applying EC information. Finally, the threshold value used to discriminate between normal quarters and quarters with CM can be optimized.

The second limitation of this study is the gold standard definition used. This definition was based on suspicious cases of CM that were visually checked and that showed signs of CM that were severe enough, according to the farm manager and staff of the research farm, to start an antibiotic treatment. This definition is somewhat subjective: it might differ from one staff member to another on the research farm itself, but it will certainly differ between farms as each farmer has a self-defined detection goal. In addition, AM farmers will decide, from all cows for which the AM system alerts, which cows to draft from the farm for a visual check. This qualitative selection will differ between farmers. Because of this selection, it is a valid assumption that the number of quarters with CM in this study will probably

be an underestimation of the real CM situation of the farm. However, it also seems sensible to assume that the observed CM cases included in this study were the more severe ones as they all were treated with antibiotics. It is subject to debate whether an AM system should alert for mild CM infections because these infections do have a greater chance to be cured by the cow herself. The consequence of the underestimation of the true CM status of this farm is that some unobserved quarter cases with CM were, in fact, considered healthy. This will result in an incorrect specificity (Sp) of any detection model, because this measure can only be calculated in those situations where there is certainty that all quarter milkings considered as negative are indeed truly negative. This certainty, however, was not achievable in the current study. Therefore, it was decided to report on two alternative performance measures: the success rate, which represents another measure of erroneous classification (of healthy quarter milkings being classified as CM), and the false alert rate. The false alert rate was chosen as an evaluation measure as it is a more comprehensible measure for the end-users of AM systems: the dairy farmers (Sherlock et al., 2008). As each dairy farmer will have a self-defined CM detection goal it will be an improvement for future detection models that these are flexible so that farmers themselves can decide about the trade-off between Se and the false alert rate.

Furthermore, the gold standard used in this study may also differ from another lively debated definition: that for abnormal milk. A new international standard (International Standardization Organization standard 20966; ISO, 2007) attempts to describe how to detect abnormal milk due to CM and an example of this definition is described in Annex C of the standard. Although abnormal milk and CM seem to be similar, their detection goals do differ. Whereas CM is primarily detected to start an antibiotic treatment, abnormal milk is expected, in the near future, to be separated from the bulk tank automatically. The model requirements for this latter detection goal differ greatly from the one reported here and the potential for using ISCC information should be studied for this problem domain as well.

Although we had to deal with these limitations, the primary goal of this study was to explore whether ISCC information estimated from a composite cow milking would make a useful contribution to the detection of CM by decreasing the number of FP alerts. Our results showed a first and promising attempt to combine ISCC with EC data to decrease FP alerts using the same Se level compared with models using ISCC or EC alone. However, further research is needed to determine the full potential of ISCC information in automatic sensing systems before it can be implemented on a large scale by AM and conventional milking systems. The EC is measured as standard procedure on most, if not all, AM systems today, as it is a relatively simple and inexpensive technique for on-line implementation (Hogeveen and Ouweltjes, 2002; Mottram et al., 2007). Sensor monitoring of EC lacks the

additional costs of reagent fluid and discarding of milk, which are present when estimating SCC on-line. In addition, the ISCC sensor itself is more expensive than the EC sensor. The ISCC has been estimated for all milkings at a composite cow milking level in the current data set, but there are AM systems where it is possible to estimate ISCC at quarter level as well. It would be of great interest to explore several implementation strategies; for example, to estimate ISCC only for those quarters suspected of having CM where EC or other information may be used as a first indicator. Further research should focus on a cost-benefit analysis between these two sensors and explore economically attractive implementation strategies of using ISCC and EC together to detect CM.

Conclusions

The use of ISCC alone showed slightly better capabilities in detecting CM than use of EC alone. Combining ISCC and EC sensor information via a fuzzy logic algorithm to detect treated cases of CM produced a two- to three-fold improvement in the success rate and a two- to three-fold reduction in the false alert rate when the Se was maintained at the same value (80%). Results suggest that the performance of a CM detection system improved when ISCC information was added to a detection model using EC information.

Acknowledgements

We gratefully acknowledge the contributions to the acquisition of the data used in this research by the farm staff. Also gratefully acknowledged is Jacqueline McGowan (DairyNZ Ltd., Hamilton, New Zealand) for providing general farm data and information about on-line SCC data collection, Vicky Farr (Sensortec Ltd., Hamilton, New Zealand) for detailed information on CellSense in practice, and Barbara Dow (DairyNZ Ltd., Hamilton, New Zealand) for helping with data handling and data analysis. This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs, and the New Zealand Foundation for Research, Science and Technology (contract number DRCX0201).

References

Cavero, D., K.-H. Tölle, C. Buxadé, and J. Krieter. 2006. Mastitis detection in dairy cows by application of fuzzy logic. Livest. Sci. 105(1-3):207–213.

De Mol, R.M., A. Keen, G.H. Kroeze, and J.M.F.H. Achten. 1999. Description of a detection model for oestrus and diseases in dairy cattle based on time series analysis combined with a Kalman filter. Comput. Electron. Agric. 22(2-3):171–185.
De Mol, R.M., G.H. Kroeze, J.M.F.H. Achten, K. Maatje, and W. Rossing. 1997. Results of a multivariate approach to automated oestrus and mastitis detection. Livest. Prod. Sci. 48(3):219–227.
De Mol, R.M., and W.E. Woldt. 2001. Application of fuzzy logic in automated cow status monitoring. J. Dairy Sci. 84(2):400–410.
Green, M.J., L.E. Green, Y.-H. Schukken, A.J. Bradley, E.J. Peeler, H.W. Barkema, Y. de Haas, V.J. Collis, and G.F. Medley. 2004. Somatic cell count distributions during lactation predict clinical mastitis. J. Dairy Sci. 87(5):1256–1264.
Hillerton, J.E. 2000. Detecting mastitis cow-side. Pages 48–53 in National Mastitis Council 39th Annual Meeting, Madison, WI.
Hillerton, J.E., and J.E. Semmens. 1999. Comparison of treatment of mastitis by oxytocin or antibiotics following detection according to changes in milk electrical conductivity prior to visible signs. J. Dairy Sci. 82(1):93–98.
Hogeveen, H., and W. Ouweltjes. 2002. Mastitis Therapy and Control: Automatic on-line detection of abnormal milk. Pages 1735–1740 in Encyclopedia of Dairy Sciences. R. Roginski, J. W. Fuquay, and P. F. Fox (eds.). Academic Press, London, UK.
Hogeveen, H., and W. Ouweltjes. 2003. Sensors and management support in high-technology milking. J. Anim. Sci. 81(Suppl.3):1–10.
ISO. 2007. Automatic milking installations—Requirements and testing. International Standard ISO 20966. ISO, Geneva, Switzerland.
Jago, J.G., K. Bright, P. Copeman, K. Davis, A.K. Jackson, I. Ohnstad, R. Wieliczko, and M. W. Woolford. 2004. Remote automatic selection of cows for milking in a pasture-based automatic milking system. Proc. N.Z. Soc. Anim. Prod. 64:241–245.
Kamphuis, C., D. Pietersma, R. van der 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. Comput. Electron. Agric. 62(2):169–181.
Kremer, W.D.J., E.N. Noordhuizen-Stassen, and J.A.C.M. Lohuis. 1990. Host defense and bovine coliform mastitis—A review. Vet. Q. 12:103–113.
Leslie, K.E., R. Dingwell, L. Yan, A. Bashiri, and P. Johnstone. 2007. An Evaluation of Sensortec Cellsense® for Determining Udder Health Status in Lactating Dairy Cattle. Pages 232–233 in National Mastitis Council 46th Annual Meeting, Verona, WI.
Maatje, K., P.J.M. Huijsmans, W. Rossing, and P.H. Hogewerf. 1992. The efficacy of in-line measurement of quarter milk electricalconductivity, milk-yield and milk temperature for the detection of clinical and subclinical mastitis. Livest. Prod. Sci. 30(3):239–249.
McNeil, D., and P. Freiberger. 1994. Fuzzy Logic: The Discovery of a Revolutionary Computer Technology and How it is Changing Our World. Simon and Schuster, New York, NY.
Mottram, T., A. Rudnitskaya, A. Legin, J.L. Fitzpatrick, and P.D. Eckersall. 2007. Evaluation of a novel chemical sensor system to detect clinical mastitis in bovine milk. Biosens. Bioelectron. 22(11):2689–2693.
Nielen, M., H. Deluyker, Y.-H. Schukken, and A. Brand. 1992. Electrical-conductivity of milk: Measurement, modifiers, and meta analysis of mastitis detection performance. J. Dairy Sci. 75(2):606–614.
Nielen, M., Y.-H. Schukken, A. Brand, H.A. Deluyker, and K. Maatje. 1995. Detection of subclinical mastitis from online milking parlor data. J. Dairy Sci. 78(5):1039–1049.
Peeler, E.J., M.J. Green, J.L. Fitzpatrick, and L.E. Green. 2003. The association between quarter somatic-cell counts and clinical mastitis in three British dairy herds. Prev. Vet. Med. 59(3):169–180.
Rasmussen, M. D. 2005. Visual scoring of clots in foremilk. J. Dairy Res. 72(3):406–414.
Reinemann, D. J., and J. M. Helgren. 2004. Online milk sensing issues for automatic milking. Paper 04–4191 in ASAE/CSAE Meeting. Am. Soc. Agric. Eng., St. Joseph, MI.
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. Pages 275–282 in Mastitis control – From science to practice. T.J.G.M Lam (ed.). Wageningen Academic Publishers, Wageningen, the Netherlands.
Steeneveld, W., H. Hogeveen, H. W. Barkema, J. Van den Broek, and R.B.M. Huirne. 2008. The influence of cow factors on the incidence of clinical mastitis. J. Dairy Sci. 91(4):1391–1402.
Whyte, D. S., R. G. Orchard, P. S. Cross, T. Frietsch, R. W. Claycomb, and G.A.Mein. 2004. An on-line somatic cell count sensor. Pages 235–240 in Automatic Milking: A better understanding. A. Meijering, H. Hogeveen, and C.J.A.M. de Koning (eds.). Wageningen Academic Publishers, Wageningen, the Netherlands.

Chapter 7

The effect of adding cow information to a clinical mastitis detection model using sensor information from automatic milking systems

W. Steeneveld, C. Kamphuis, H. Mollenhorst, H Hogeveen

Abstract

Using a probability of having clinical mastitis (CM) based on sensor measurements given by an automatic milking (AM) system in combination with a probability of having CM based on non-AM cow information was expected to improve CM detection. Based on data of nine Dutch commercial dairy farms, a CM detection model was developed. The model provided for each cow milking the probability that that cow milking was a mastitis one, based on sensor measurements of electrical conductivity, color and milk yield. In total, 47,049 cow milkings, including 99 milkings for which CM was reported, were available with a probability of having CM based on the sensor measurements. To each of these cow milkings, non-AM cow information (parity, day in milk, season, somatic cell count history, and CM history) was added. After adding a probability of having CM based on the non-AM cow information, a posterior probability of having CM was calculated for each cow milking. To evaluate the CM detection performance by adding non-AM cow information, the posterior probabilities were used to calculate the sensitivity and specificity over the whole range of possible threshold probabilities for classification. Receiver operating characteristic (ROC) curves were constructed to visualize the detection performance. To summarize the ROC-curves into a single quantity, the area under the curve (AUC) was computed. The AUC based on the probability of having CM based on sensor measurements only was compared with the AUC based on the posterior probability and thus on the probability of having CM based on sensor measurements and non-AM cow information. The AUC was 83.0 for the CM detection model based on only sensor measurements, while the AUC based on sensor measurements and non-AM cow information was 80.8. Results of the current study show that adding non-AM cow information to sensor information from an AM system did not improve CM detection performance. So, if probabilities based on sensor measurements are available, adding non-AM cow information has no additional value for the detection of CM.

Introduction

Using a probability of having clinical mastitis (CM) based on sensor measurements given by an automatic milking (AM) system in combination with a probability of having CM based on non-AM cow information was expected to improve the detection of CM (Steeneveld et al., 2010a). However, Steeneveld et al. (2010b) reported a minor effect of using non-AM cow information (parity, day in milk (DIM), somatic cell count (SCC) history, and CM history) on making a distinction between true positive and false positive mastitis alerts. This minor effect was explained by the fact that within the group of alerted cows, amongst others the SCC history of the true positive alerts was not significantly different from the false positive alerts. In that specific study, the effect of adding non-AM cow information to only those cow milkings with mastitis alerts was studied. It was not investigated if adding non-AM cow information to all milkings could improve CM detection. Moreover, reported results were based on data from a single research farm, where the proportion of cows at risk for CM differed from commercial Dutch dairy farms (Steeneveld et al., 2010b).

Based on sensor data collected at nine commercial dairy farms in the Netherlands, a new CM detection model has been developed by Kamphuis et al. (2010). That new CM detection model was capable of keeping the sensitivity (Se) at about the same level as models currently used by AM systems, but at the same time to decrease the number of false positive alerts by more than 50%. The model provided a probability for having CM at quarter milking level, based on sensor measurements of electrical conductivity, color and milk yield. The CM detection performance of the described model was highly farm dependent, with Se levels ranging between 0.0% and 71.4% at a fixed specificity (Sp) of 99% for the nine farms.

The objective of the current study is to determine whether CM detection can be improved by adding non-AM cow information to the sensory based CM detection model developed by Kamphuis et al. (2010). Detection performance will be determined for nine farms together and for each farm separately. Detection performance of a model including only sensor information will be compared with a model including both sensor information and non-AM cow information.

Materials and Methods

A CM detection model based on sensor measurements

A detailed description of the development and validation of a new CM detection model, using decision tree induction, is provided by Kamphuis et al. (2010). In

short, sensor measurements of electrical conductivity, color and milk yield were collected for each quarter milking at nine commercial dairy farms in the Netherlands using a total of twelve Lely Astronaut AM sysyems (version A2 (n=10) or A3 (n=2); Lely Industries N.V., Maassluis, the Netherlands). These sensor measurements were used to define descriptive variables (n=1,065) as described in Kamphuis et al. (2008). These independent variables described characteristics (level, variability and shape) of sensor measurement patterns from each quarter milking.

A decision tree was trained using 24,960 quarter milkings from 404 cows. From these quarter milkings, 243 had CM as observed and recorded by the participating dairy famers. All others had a very low likelihood of having CM: these quarter milkings had SCC information (from the milk production test day system) and belonged to cows that never exceeded a SCC level of $200x10^3$ cells/mL within the lactation, and in addition were never visually checked by the dairy famers during the whole study period. The training procedure used decision tree induction as base classifier and combined it with bagging. Bagging is a commonly used data mining technique that is able to improve detection performance of a classification model (Witten and Frank, 2005). The model was validated using data from cows that were not used for training. In addition, the test set also included quarter milkings with a less clear mastitis status: all quarter milkings outside a 2-week range from a CM case were considered as being negative for having CM. A random sample of 50,000 of these quarter milkings was then used for testing. The final test set included 50,105 quarter milkings from 368 cows. From these quarter milkings, 50,000 were negative for CM and 105 quarter milkings had CM. The decision tree model provided probability estimates for having CM for each of the 50,105 quarter mikings in the test set.

The non-AM cow information is at cow level, so it was necessary to transform the probability for having CM at quarter milking level to a probability estimate at cow milking level. To do so, all quarter milkings of the cow were added to the quarter milking originally in the test dataset. These added quarter milkings also received a probability of having CM determined by the CM detection model developed by Kamphuis et al. (2010). A cow level probability for CM ($P_{cowmilk}$) was calculated using the probabilities for CM of the quarter milkings as follows:

$$P_{cowmilk}=P_{Q1}+(1-P_{Q1})*P_{Q2}+(1-P_{Q1})*(1-P_{Q2})*P_{Q3}+(1-P_{Q1})*(1-P_{Q2})*(1-P_{Q3})*P_{Q4} \quad (1)$$

where P_{Q1}, P_{Q2}, P_{Q3} and P_{Q4} are the probabilities of having CM for each of the four quarter milkings of the cow, respectively. Subsequently, per cow milking, one record remained which resulted in a dataset of 47,111 cow milkings, each with a probability of having CM at cow milking level. To all these cow milkings, the non-AM

cow information on parity, month in lactation, the season of the year, the SCC in the previous 30 days, the SCC in the 30 days before the previous 30 days and, for multiparous cows, the geometric mean SCC from all available test-day records from the previous lactation were added. In addition, for each cow milking the accumulated number of CM cases of the cow in the previous 30 days and the accumulated number of CM cases of the cow in the days before the previous 30 days were added. For 62 cow milkings, it was not possible to add non-AM cow information, and therefore these milkings were excluded for further analyses. The final test set, used for analyses in the current study, consisted of 47,049 cow milkings with in total 99 CM cases. The test dataset was constructed for all farms together and for the nine farms separately. Creating the test dataset was performed using SAS version 9.1 (SAS Institute Inc., Cary, NC).

Adding non-AM cow information using naïve Bayesian network

To update the probabilities for having CM based on the CM detection model developed by Kamphuis et al. (2010) with non-AM cow information, a naive Bayesian network (NBN) was constructed. An NBN determines conditional probabilities for the feature variables given the possible classes of the class variable (Friedman et al., 1997). In the current study, the variable capturing whether a cow had CM or not at a particular day in milk (DIM) was used as the class variable. The variables describing non-AM cow information were feature variables. To construct the NBN, a dataset including each DIM for all cows present on the nine included farms but that were not present in the test set of Kamphuis et al. (2010) was used. That dataset was established using data provided by the Dutch national milk production recording system (CRV, Arnhem, the Netherlands), including information from the monthly milk recording. For each DIM, the non-AM cow information feature variables (parity, season of the year, etc) were added. The final set for constructing the NBN consisted of 260,046 DIM, including 261 DIM for which CM was recorded. The descriptive statistics of this dataset used for constructing the NBN are presented in Table 1.

NBN based on the whole dataset and for the nine farms separately were constructed. Creating the dataset for constructing the NBN was performed using SAS. Constructing the NBN, which included estimating conditional probabilities, was done by using the Bayesian-network editing package Dazzle (Schrage et al., 2005).

Table 1. Number of clinical mastitis (CM) cases (CM=1) and non-CM cases (CM=0) for different levels of cow information in the dataset used for constructing the naive Bayesian network. The *P*-value indicates whether the distribution over the levels for a cow information variable is statistically different between CM and non-CM cases

Cow information	**CM=1 (n=261)**	**CM=0 (n=259,785)**	***P*-value**
Parity			<0.0001
1	34 (13%)	97,242 (37%)	
2	88 (34%)	72,139 (28%)	
3	47 (18%)	36,348 (14%)	
≥ 4	92 (35%)	54,056 (21%)	
Day in milk			0.025
1-30	29 (11%)	23,800 (9%)	
31-60	34 (13%)	23,646 (9%)	
61-90	22 (9%)	23,431 (9%)	
91-120	32 (12%)	23,338 (9%)	
121-150	24 (9%)	22,603 (9%)	
151-180	26 (10%)	21,629 (8%)	
181-210	26 (10%)	20,415 (8%)	
≥211	68 (26%)	100,923 (39%)	
Season			0.012
January – March	69 (26%)	72,618 (28%)	
April – June	54 (21%)	63,983 (25%)	
July – September	82 (31%)	54,904 (21%)	
October – December	56 (22%)	68,280 (26%)	
SCC in last 30 days			<0.0001
<200x10^3 cells/mL	89 (34%)	162,439 (63%)	
≥200x10^3 cells/mL	130 (50%)	53,436 (21%)	
SCC before last 30 days			<0.0001
<200x10^3 cells/mL	102 (39%)	163,399 (63%)	
≥200x10^3 cells/mL	105 (40%)	50,025 (19%)	
Mean SCC previous lactation			<0.0001
<200x10^3 cells/mL	88 (34%)	73,607 (28%)	
≥200x10^3 cells/mL	62 (24%)	30,485 (12%)	
# CM cases in last 30 days			0.169
0	236 (90%)	253,580 (98%)	
1	20 (8%)	5,906 (2%)	
2	5 (2%)	299 (0%)	
# CM cases before last 30 days			0.335
0	179 (69%)	228,676 (88%)	
1	53 (20%)	22,358 (9%)	
2	29 (11%)	8,751 (3%)	

Validation

For each cow milking in the test set, the constructed NBN was used to calculate a posterior probability of the cow milking being positive for CM (P_{cow}), based on the prior probability of a cow having CM ($P_{cowmilk}$), as calculated with formula 1, and the added non-AM cow information. For computing the posterior probability P_{cow} ($c \mid f_1, \ldots, f_n$) of the output c given levels $f_1,\ldots,f_n$ for its n feature variables, the model uses

$$P_{cow}(c \mid f_1,\ldots,f_n) = \frac{\prod_{i=1}^{n} P(f_i \mid c_1) * P_{cowmilk}(c_1)}{\sum_{j=1}^{2} \prod_{i=1}^{n} P(f_i \mid c_j) * P_{cowmilk}(c_j)} \quad (2)$$

where $P_{cowmilk}(c_1)$ is the prior probability of cow milking being positive. The probabilities $P(f_i \mid c_1)$ are the conditional probabilities of finding the level f_i for the i^{th} feature variable given that milking is positive for CM. The conditional probabilities $P(f_i \mid c_j)$ for all levels for the i^{th} selected feature have already been estimated from the training data upon constructing the NBN and therefore are readily available in the NBN for the computation of the posterior probabilities using formula (2). Computing the posterior probabilities for each cow milking was done using Microsoft Excel.

To evaluate the CM detection performance by adding non-AM cow information, the posterior probabilities (P_{cow}) obtained for the validation dataset were used to calculate the Se and Sp over the whole range of possible threshold probabilities for classification. Receiver operating characteristics (ROC) curves were constructed to visualize the performance. To summarize the ROC-curves into a single quantity, the area under the curve (AUC) was computed (e.g., Dohoo et al., 2003). The AUC based on P_{cow} was compared with the AUC based on $P_{cowmilk}$. The difference between Se levels at two fixed Sp levels was tested for significant difference using McNemar's test in SAS. Also at two fixed Se levels the Sp levels were compared. Validation was performed for all farms together and for each of the nine farms separately.

Results and Discussion

For construction of the NBN, the dataset as described in Table 1 was used. The non-AM cow information variables on parity, DIM, season and SCC history variables were found to be statistically significantly different between CM cases and non-CM cases. For instance, while 50% of the CM cases were from cows with an SCC in the

last 30 days above 200x10^3 cells/mL, the same information was found in only 21% of the non-CM cases (Table 1, $P<0.0001$).

Despite the significant difference in having CM and not having CM for different non-AM cow information variables (Table 1), adding these variables to the sensor measurements given by the AM system did not improve CM detection. Based on data of all nine farms, the AUC was 83.0 for the CM detection model based on only sensor measurements, while the AUC based on sensor measurements and non-AM cow information was 80.8 (Table 2). For the farm-specific CM detection models, the AUC ranged between 63.6 and 89.4 when non-AM information was added. Except for farm 4, these AUC-values were lower than the AUC-values for the CM detection model based on only sensor measurements (Table 2). Previously it was reported that adding non-AM cow information to alerted cows did not improve the ability to discriminate between false positive and false positive mastitis alerts (Steeneveld et al., 2010b). Results of the current study showed that adding non-AM cow information to all cow milkings also did not improve CM detection performance. So, if probabilities based on sensor measurements are available, adding non-AM cow information has no additional value for the detection of CM.

Table 2. Area under the curve (AUC) based on a clinical mastitis (CM) detection model including sensor measurements ($P_{cowmilk}$), and for a CM detection model including sensor measurements and non-AM cow information (P_{cow})

Herd[1]	# cow milkings	# CM cases	AUC ($P_{cowmilk}$)	AUC (P_{cow})
all	47,049	99	83.0	80.8
1	6,834	26	92.3	89.0
2	3,104	16	72.8	70.4
3	3,822	5	79.3	69.1
4	3,614	14	87.9	89.4
5	8,284	11	73.4	63.6
6	9,816	9	82.1	81.9
7	6,364	14	79.3	74.6

[1] For two farms the AUC was not calculated because they had only two CM cases in the test set

The CM detection model based on sensor measurements alone showed an Se of 26.3% at an Sp of 99% and an Se of 37.4% at an Sp of 97.9% (Table 3). The CM detection model based on sensor measurements and non-AM cow information showed lower Se values at both Sp levels. Also for the farm-specific CM detection models, Se values were lower when non-AM cow information was added at a fixed

Sp of 99%. It must be noted that the farm-specific CM detection models were based on a small number of CM cases, for instance, for farm 3 only 5 CM cases were present in the test set. At a fixed Sp of 99%, the Se of 26.3% is much lower than the Se of 40% reported by Kamphuis et al. (2010). This can be explained by the fact that the current study is performed at cow milking level, while the study of Kamphuis et al. (2010) was performed at quarter milking level, which makes comparison of Sp and Se values very difficult. For instance, reaching a Sp of 99% at cow milking level is more difficult than reaching a Sp of 99% at quarter milking level. Subsequently, comparison of the Se at a fixed Sp level does not give a good indication.

Table 3. Sensitivity (Se) at a specificity of 99% (Se^{Sp99}) and at a specificity of 97.9% ($Se^{Sp97.9}$) for the clinical mastitis (CM) detection model including sensor measurements ($P_{cowmilk}$) and the CM detection model including sensor measurements and non-AM cow information (P_{cow})

Herd[1]	Se^{Sp99} ($P_{cowmilk}$)	Se^{Sp99} (P_{cow})	*P*-value[2]	$Se^{Sp97.9}$ ($P_{cowmilk}$)	$Se^{Sp97.9}$ (P_{cow})	*P*-value[2]
all	26.3	20.2	0.157	37.4	29.3	0.033
1	34.6	23.1	0.257	38.5	30.8	0.414
2	12.5	12.5	1	12.5	12.5	1
3	60	40	0.317	60	40	0.317
4	71.4	28.5	0.014	78.6	50	0.046
5	9.1	0	0.317	9.1	18.2	0.317
6	44.4	33.3	0.564	66.6	55.5	0.317
7	14.3	14.3	1	14.3	21.4	0.317

[1] For two farms the AUC was not calculated because they had only two CM cases in the test set

[2] Indicates whether the Se is statistically significantly different between the CM detection model including sensor measurements ($P_{cowmilk}$) and the CM detection model including both sensor measurements and non-AM cow information (P_{cow})

By adding non-AM cow information, the Sp values at fixed Se-levels did also not improve (Table 4). For instance, based on sensor measurements only, the Sp was 95.7% at a fixed Se-level of 50%, while the Sp decreased to 94.5% when non-AM cow information was added ($P<0.0001$). Also for all farm specific CM detection models, the Sp at a fixed Se of 50%, were significantly lower when non-AM cow information was added. For 2 farms, the Sp at a fixed Se of 70% were higher ($P<0.0001$). The reason for significant differences in Sp between CM detection models including only sensor information and CM detection models including both

sensor information and non-AM cow information was the large amount of milkings without CM (n=46,950).

Table 4. Specificity (Sp) at a sensitivity of 50% (Sp^{Se50}) and at a sensitivity of 70% (Sp^{Se70}) for the clinical mastitis (CM) detection model including sensor measurements ($P_{cowmilk}$) and the CM detection model including sensor measurements and non-AM cow information (P_{cow})

Herd[1]	Sp^{Se50} ($P_{cowmilk}$)	Sp^{Se50} (P_{cow})	*P*-value[2]	Sp^{Se70} ($P_{cowmilk}$)	Sp^{Se70} (P_{cow})	*P*-value[2]
all	95.7	94.5	<0.0001	91.6	89.6	<0.0001
1	96.2	94.5	<0.0001	93.4	90.0	<0.0001
2	87.2	81.5	<0.0001	69.5	75.6	<0.0001
3	99.2	81.2	<0.0001	91.7	49.0	<0.0001
4	99.8	98.6	<0.0001	99.1	94.9	<0.0001
5	94.5	56.5	<0.0001	89.1	43.7	<0.0001
6	98.7	98.5	0.0196	95.2	94.4	0.0001
7	93.6	91.4	<0.0001	87.5	89.3	<0.0001

[1] For two farms the AUC was not calculated because they had only two CM cases in the test set
[2] Indicates whether the Sp is statistically significantly different between the CM detection model including sensor measurements ($P_{cowmilk}$) and the CM detection model including both sensor measurements and non-AM cow information (P_{cow})

Besides the CM detection model based on the described NBN, an NBN was constructed using an SCC threshold of $500x10^3$ cells/mL instead of an SCC threshold of $200x10^3$ cells/mL for the three SCC history variables. In addition, also an NBN was developed using backward elimination of feature variables. This resulted in an NBN without the feature variable 'season of the year'. The AUC of the CM detection model based on an SCC threshold of $500x10^3$ cells/mL and the AUC of the CM detection model developed with the backward NBN were lower than the AUC-values presented in Table 2 (data not shown). This indicates that adding non-AM cow information with different constructed NBN did also not improve CM detection performance.

To develop an NBN, also forward selection of feature variables was performed. This resulted, however, in an NBN without any feature variable. This indicates that each of the feature variables on non-AM cow information was not able to increase the AUC of the NBN under construction. So, this denotes that, although the statistically significant differences between CM cases and non-CM cases in non-AM cow information (Table 1), the variables on non-AM cow information did not have

enough discriminative power to distinguish between cow milkings having CM and those that do not have CM.

Conclusions

The results of the current study show that, in presence of sensor information, detection of CM cannot be improved by adding non-AM cow information. Test characteristics of the CM detection model including sensor measurements and non-AM cow information showed lower performance than the CM detection model including sensor measurements alone.

Acknowledgements

We gratefully acknowledge the participating dairy famers for their contribution to the acquisition of the data used in this research. Also gratefully acknowledged is Lely Industries N.V (Maassluis, The Netherlands) for their support with the installation of the data acquisition systems at the participating farms. This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs.

References

Dohoo, I., W. Martin, and H. Stryhn. 2003. Veterinary epidemiologic research. AVC Inc., Charlottetown, Prince Edward Island, Canada

Friedman, N., D. Geiger, and M. Goldszmidt. 1997. Bayesian network classifiers. Machine Learning 29:131-163.

Kamphuis, C., D. Pietersma, R. van der 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(2):169-181.

Kamphuis, C., H. Mollenhorst, J. A. P Heesterbeek, and H. Hogeveen. 2010. Detection of clinical mastitis with sensor data from automatic milking systems is improved by using decision tree induction. J. Dairy Sci. 93(8):3616-3627.

Schrage, M. M., A. Van IJzendoorn, and L. C. Van der Gaag. 2005.Haskell ready to Dazzle the real world. Pages 17-26 in Proceedings of the 2005 ACM SIGPLAN workshop on Haskell, ACM Press, New York.

Steeneveld, W., L. C. van der Gaag, H. W. Barkema, and H. Hogeveen. 2010a. Simplify the interpretation of alert lists for clinical mastitis in automatic milking systems. Computers and Electronics in Agriculture 71(1):50-56.

Steeneveld, W., L. C. van der Gaag, W. Ouweltjes, H. Mollenhorst, and H. Hogeveen. 2010b. Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems. J. Dairy Sci. 93(6):2559-2568.

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

Chapter 8

General discussion

The main objective of this thesis was to improve the detection of clinical mastitis (CM) with sensor data from automatic milking (AM) systems. In *Chapter 1*, three main requirements were defined for a CM detection model if it is to be applied in practice: (1) it has to detect at least severe cases of CM with high sensitivity (Se) and specificity (Sp) levels, (2) it has to detect CM within a small time window, and (3) it should be able to deal with noisy, incomplete, and imbalanced field data. A fourth, but less urgent requirement considers the ability of a CM detection model to provide information about the CM causal pathogen. To develop a CM detection model that meets all these requirements, three main routes were explored throughout this thesis: (1) using other algorithms for both data preprocessing and classification (*Chapters 2, 3, 4,* and *7*), (2) adding information from other sensors (*Chapter 5*), and (3) adding non-sensor information (*Chapter 6*).

In this chapter, the approaches used to improve the detection of CM and the results found in this thesis are integrated and discussed. The first section of this chapter discusses each specific requirement in a separate subparagraph. The second section will discuss each main route explored separately. A third section briefly discusses a major issue found throughout this thesis that requires attention for future research in the field of CM detection. Finally, this chapter will present a number of conclusions and recommendations that will hopefully bring us one step closer to a CM detection model perfect for both science and dairy farming.

Requirements of a future CM detection model

Chapter 1 introduced four requirements that will make a CM detection model successful in practice. This section discusses which steps were taken in this thesis in order to fulfill these requirements, and the main results found.

Detecting cows with CM with high levels of sensitivity and specificity

Hogeveen and Ouweltjes (2002) suggested that a CM detection model should detect as many cows with CM as possible. Preferably all cows with CM should be detected, but at least those cows with a severe CM infection. Since then, several suggestions regarding the required detection performance of a CM detection model have been made in literature (*Chapter 1*). Focusing on one of the most recent suggestions by Mein and Rasmussen (2008), a CM detection model should have a Se of >70% in combination with a Sp of >99%. With an Se of 40% at an Sp of 99%, the base decision tree validated as described in *Chapter 4* seemed unable to reach their suggested detection performance. However, that decision tree was validated using the strict and narrow time window of <24h, whereas Mein and Rasmussen (2008) did not specify any requirements about the time window in

which their suggested detection performance should be reached. When applying the proposed time window for use in practice of Sherlock et al. (2008), being 48 hours for a CM alert and 24h for a CM observation, the Se of the base decision tree was 69.5% at an Sp of 99% (*Chapter 4*). Putting the results in that perspective, there is a strong and positive indication that decision tree induction is able to use sensor measurements for developing a CM detection model that reaches the recommended detection performance rates, even when noisy field data were included in the test set.

Although the detection requirements suggested by Mein and Rasmussen (2008) are in sight, the requirement of detecting at least all severe cases of CM (Hogeveen and Ouweltjes, 2002) is not met by the base decision tree. Results from *Chapter 4* showed that this decision tree was unable to detect all severe cases of CM (defined as those that were scored as having clots in the fifth or sixth squirt of milk): 64% of these severe cases were detected by the base decision tree at an Sp of 99%. In addition, the detection of watery milk was poor, with only 12.5% of all watery quarter milkings being detected at an Sp of 99%. It is impossible to say whether these results deviate from other studies regarding CM detection, as this information has not been provided previously.

Time windows

The discussion about the effect of time windows on detection performance rates started with the introduction of time windows (Sherlock et al., 2008), and it is expected that this discussion will receive much more attention in future. Although Sherlock et al. (2008) proposed a 48h time window for an alert by a CM detection model and a 24h time window for an observation of CM, a very narrow time window of <24h before an observation of CM (*Chapter 3* and *4*) was used in this thesis. This narrow time window was used as it was felt that in practice a time window should be as small as possible. Detection performances improved when time windows became wider (*Chapter 4* and *6*), a trend confirmed by Sherlock et al. (2008). However, it was hard to compare the results found in this thesis with previously reported results, due to the large variation in applied time windows (see also Table 1, *Chapter 1*) in the different studies. An international agreement on an appropriate and practically relevant time window to validate CM detection models is therefore necessary. Is it useful to apply a wide time window for early detection, including the risk that signs of mastitis are not visible yet? Or do we want to apply a time window for detection of quarters with severe CM mastitis with the risk of receiving the alert too late for a high cure rate after an antibiotic treatment? Treating mild or moderate CM cases 24h after detection, instead of treating them immediately after observation, does not influence cure rate (Keefe et al., 2010) or CM recurrence, somatic cell count (SCC) levels, milk production and culling (Lago et al., 2009).

These findings suggest that a model alerting for mastitis within 24h after the infection becomes clinical is still sensible. Using a time window of 48h, defined as 24h before a CM observation to 24h after a CM observation, resulted in a decision tree model with an Se of 66.7% at an Sp of 99% (*Chapter 4*). This is an increase of more than 25 percent points compared to the base decision tree that was only allowed to alert for CM before the actual CM observation. In addition, detection performance requirements as suggested by Mein and Rasmussen (2008) are almost met when this wider time window is applied.

Handling field data

It is essential that data used for developing and validating a CM detection model show similarities with real farm data. *Chapter 1* already indicated three reasons why most studies on CM detection models lack this requirement. These three reasons will be discussed in more detail below:

1. Number of (research) farms and CM cases included for data analyses. Usually, a limited number of farms and a small number of CM cases are included in studies regarding CM detection. For example, Maatje et al. (1992), Nielen et al. (1995a,b), and Norberg et al. (2004) all used one (research) farm to collect sensor and CM data (see also Table 1, *Chapter 1*). Others collected data at more farms (e.g., Mollenhorst et al., 2010; Song et al., 2010) but during a limited time period. Both approaches result in the inclusion of a limited number of CM cases. Most studies used different data sets for model development and validation, but the data sets still came from the same farm (e.g., Cavero et al., 2006), or simulated data were used for testing logic and robustness of the model (Chagunda et al., 2006). By using just one or a limited number of (research) farms, the collected data are probably more detailed and of better quality than data collected at more and commercially producing dairy farms. This detailed data may be an advantage for model development. The major disadvantage is, however, that the small number of CM cases included may not represent all variation in CM characteristics (in terms of clinical signs and sensor data), causing a drop in detection performance of the developed model when applied to other farms. This means that a CM detection model developed and tested on one farm or a small number of farms may show a disappointing detection performance when applied to others (a trend also seen in the farm specific analyses in *Chapter 4*). This thesis used a unique approach in the field of CM detection models, as it collected data from nine commercially producing dairy farms (*Chapter 4, 5,* and *7*) with a total of 12 AM systems for a period of two years without intervening with the daily working routine of the participating dairy

farmers. A disadvantage of this approach is the risk of collecting less detailed data of probably a lower quality and with large dependence on the motivation of the participating dairy farmers. This resulted in a lack of information about visual abnormalities around the period of a CM observation. It could very well be that the AM system alerted for a cow to have CM but that the dairy farmer decided not to check this cow. Or that a dairy farmer checked a true positive alert by the CM detection model quite a while after the alert, with the 'risk' that the cow cured spontaneously and thus showing no clinical signs anymore. Such a situation caused the dairy farmer to score the milk as normal milk, and thus as a false positive alert. This lack of information around these CM periods make it difficult, if not impossible, to give a true estimate about the detection performance of the CM model developed in *Chapter 4*. However, the disadvantage of this approach of data collection is outweighed by one major advantage: data used for model development and validation in this thesis consisted of field data, with all its noise and incompleteness, making it possible to generalize obtained results to practice. Two other important advantages are that the inclusion of nine commercial dairy farms resulted in the collection of a large number of CM cases (n=348) within a relatively short period of time, and the possibility to analyze farm specific detection performances.

2. Definitions used for CM cases and non-cases (gold standard). In the past, strict inclusion criteria for CM cases and non-cases were applied for both model development and validation. Data of SCC, bacteriological culturing, and visual observations were used to define CM cases and non-cases (e.g., Nielen et al., 1995a; Norberg et al., 2004; Friggens et al. 2007). These strict definitions exclude a large part of the noise present in sensor and CM data, by excluding cows and quarters with a less clear health status. Over the years, a lot of time has been spent on defining a 'true' gold standard applicable for an AM setting. Rasmussen et al. (2005) suggested that for evaluation and testing of AM systems to detect abnormal milk, the abnormal milk should be defined using both visual appearance and the CMT score of foremilk, where a higher CMT score indicates a higher SCC of the milk. This approach is used to make a clear distinction between normal and abnormal milk for test purposes, and excludes the milkings with a less clear status (e.g., visually abnormal milk with a low SCC, or milk with a high SCC without visual abnormalities) from the gold standard status. It also makes it easier for a detection model to show better detection performances. Mollenhorst et al. (2010) applied this gold standard based on CMT and visual appearance and confirmed that detection performance was better when validated using this gold standard compared to a gold standard based on

visual appearance alone. Claycomb et al. (2009) used six different ways for defining a CM case based on published or hypothetical 'gold standard' reference methods. Again, different gold standards resulted in different detection performances. These differences in performances are a perfect illustration that it is unjustified to compare detection performances of different CM detection models without taking the used gold standard into consideration. One of the applied gold standards by Claycomb et al. (2009) was based on observed CM cases requiring antibiotic treatment, which was indicated as a 'best-practice' gold standard. In addition, they concluded that repeated observations of clots in milk at consecutive milkings was a more justified gold standard definition than one that is based on clot observations at a single cow milking. Mein and Rasmussen (2008) made a first attempt in defining an internationally accepted definition of a gold standard for abnormal milk. They suggested that the classification of the 'true' CM status needs to be confirmed either by linking clot observation with a raised CMT score (or raised SCC) in quarter foremilk, or by repeated clot observation at consecutive milkings. However, getting this visual information for consecutive milkings is extremely time consuming and expensive, especially when information on a substantial number of CM cases is wanted. Their suggestion using clot observations at consecutive milkings as gold standard could not be used in this thesis, as not all quarter milkings were visually checked by the dairy farmers. Therefore, another gold standard was used that reflected reality as much as possible: the training process applied a strict gold standard to define CM cases and non-cases (*Chapter 4*), where farmer's observations, based on a scoring protocol which instructed them to score the fifth and sixth squirts of milk, were used to define CM cases. This definition does not use a bacteriological culturing to confirm presence of CM causal pathogens and, therefore, may be considered as incorrect according to the strict definition of a CM infection. On the other hand, it is a practical definition, as dairy farmers judge a cow's health based on visual abnormalities in milk. Non-cases in the training set were quarter milkings with a very high likelihood to be truly negative of CM. This strict definition of cases and non-cases was applied to generate a clear contrast between these two groups, which was expected to ease the training process. Validation, however, was applied to data that mimicked practice as closely as possible, because quarter milkings with a less clear CM status were kept in the validation set. By selecting a random sample of 50,000 quarter milkings, it was expected that all potential variation of sensor measurements of these quarter milkings with a less clear mastitis status was captured and that

validation on such a test set would better reflect detection performance of the CM detection model when they are implemented in practice.

3. Dealing with missing values.
   Sensor data collected in the field are noisy and incomplete by definition. Therefore, a CM detection model should be able to deal with this noisy data as well. Excluding records with missing values or erroneous measurements as done by De Mol et al. (2001) does not reflect reality, and thus is probably not the solution for future research. Regression analysis is not able to handle missing values, which makes this method less appropriate for developing a CM detection model with sensor data. Threshold values or moving averages have a similar disadvantage. *Chapter 1* already indicated that more complex algorithms were used to develop CM detection models, all with their own advantages and disadvantages. For example, neural networks (Nielen et al., 1995a,b; Sun et al., 2010) are often referred to as black-box models. These algorithms try to discover and utilize new knowledge present in the data. A disadvantage of neural networks is that it is difficult to interpret the model in traditional analytical terms due to the hidden layers in these systems. Fuzzy logic (De Mol and Woldt, 2001; Cavero et al., 2006), on the other hand, is an easy to understand algorithm, as shown in *Chapter 6*. The method uses linguistic variables to classify fuzzy or approximate outcomes. For example, CM is not a binary outcome, but it varies between a mild to a severe systemic infection. This makes CM an ideal outcome variable to be used in a fuzzy logic model. The method can deal with missing values, by simply not activating the fuzzy rule involving the variable with the missing value. But as long as there are enough other fuzzy rules, it is likely that a sensible result is returned. A disadvantage of fuzzy logic is that it is heuristic in nature, which means that it is not searching for new knowledge in data itself but that the users of this algorithm have to find out the 'truth' and the fuzzy rules themselves. It uses knowledge already available by experts in the field of CM detection. The essential memberships, the membership values, and the IF-THEN rule blocks are all predetermined by experts. Finally, time-series analyses in combination with Kalman filters are a method to develop cow specific models, by updating parameters of the underlying models with new information by the Kalman filter each time a cow is milked (e.g., De Mol and Ouweltjes, 2001; Cavero et al., 2006). This method is useful for detecting short term changes in sensor data, possibly indicating an acute CM infection. On the other hand, the approach makes the models less appropriate to detect chronically infected cows. In addition, the models are no longer valid if new measurements widely deviate because the cow is ill or in estrus (De Mol et al., 1999), and this type of algorithm has difficulties with

measurement errors and missing values. In this thesis, data mining was used for developing a CM detection model using sensor data from AM systems. Data mining, also referred to as 'knowledge discovery from databases', has been used in areas such as analyzing medical outcomes and detecting credit card fraud (Mitchell, 1999). It includes a number of techniques, of which neural networks is one of many. A classification technique that is applied widely in the field of data mining is decision tree induction (Quinlan, 1986). This algorithm is used in this thesis to classify quarter milkings as being healthy or as having CM. Decision tree induction has been used in the dairy industry to classify cows in estrus (Mitchell, 1996), to classify CM causal pathogens (Kim and Heald, 1999), and to interpret lactation curves (Pietersma et al., 2002). An advantage of decision tree induction is that the models can be graphically displayed which make them easy to interpret. In addition, decision trees are flexible in the sense that new information can be added quite easily, and they are believed to be capable of dealing with data that is noisy, incomplete, and imbalanced. The application of decision tree induction is further discussed in second section of this chapter ('Main routes explored to improve CM detection').

Predicting pathogens

Although not an essential requirement, it would be beneficial to have information about the pathogen causing the CM infection. Adding this type of information on the mastitis alert list will provide a dairy farmer more information about the disease status of a cow, and could improve the decision to treat a cow with antibiotics or not. Information about the pathogen involved with the CM infection could include the Gram-status, but other information may be useful as well (e.g., major or minor pathogen, high or low cure rate). Espada and Vijverberg (2002) were the first to explore color sensor data for CM prediction and reported the potential of using sensor data for pathogen prediction. However, their study included a limited amount of data (six cows with abnormal milk). *Chapter 5* confirmed the earlier results reported by Espada and Vijverberg (2002). That study included 140 CM cases with sensor data and pathogen information, of which 96 CM cases were used for training a decision tree. Sensor data (including the color sensor blue) were able to differentiate between Gram-positive and Gram-negative CM cases, and thus that sensor data can be used for Gram-status prediction of the CM causal pathogen. The decision tree was unable to predict the actual CM causal pathogen, possibly due to the low number of CM cases per group of pathogens.

Main routes explored to improve CM detection

Chapter 1 introduced three main routes which can be used to improve the detection of CM using sensor data from AM systems. This section discusses how this thesis explored these three routes and what main results were found.

Using other algorithms for preprocessing and classification

1. Algorithms for data preprocessing.
   Chapter 1 described that sensors have to take over the eyes and hands of a milker during the milking process. Whichever sensor is used, whether it measures the electrical conductivity, SCC, color, or any other CM indicator, it delivers data. Many sensors deliver data in large amounts, as they measure continuously during the milking process. These point measurements do not provide any information itself, but they need to get preprocessed before any information can be generated from it. Extracting predictive variables from these point estimates, to be used subsequently by the CM detection model, is one of these preprocessing steps. In the past, predictive variables based on electrical conductivity measurements mainly involved the maximum value within a quarter milking (e.g., Norberg et al., 2004) or the average of the 20 highest values within a quarter milking (e.g., De Mol et al., 1997; Norberg et al., 2004). Cases of CM were then identified by comparing the values of these variables with the rolling average for electrical conductivity of the same quarter (e.g., Milner et al., 1996), with other quarter milkings within the same milking (e.g., Hovinen et al., 2006), a combination of these two comparisons (e.g., Maatje et al., 1992), or with a threshold value (e.g., Mollenhorst et al., 2010). Little literature-based evidence was available to answer the question whether also other predictive variables could be useful for describing patterns in sensor data. Therefore, a data flow diagram was developed (*Chapter 2*). Applying this analytical framework resulted in the insight that not only variables that described the level of sensor measurement patterns (e.g., mean or maximum values) explained differences in sensor patterns. Instead, also variables that described the variability (e.g., standard deviation) or the shape (e.g., slope) of a sensor measurement pattern could be useful in differentiating between healthy and diseased quarter milkings. In addition, absolute values of predictive variables could be as informative as variables based on comparisons with other quarters in the same cow milking or with previous quarter milkings. The study described in *Chapter 2* also revealed that variables never used before, e.g., the slope of sensor measurement patterns or the deviation of the actual midpoint observation from the expected midpoint observation, had potential

predictive value for CM detection. This observation was confirmed in a subsequent study (*Chapter 3*) where a straightforward decision tree model combined information from different sensors, including variables that described the shape, variability and level of the sensor measurement patterns. Variables were based on comparisons with other quarters or with previous quarter milkings or on simple absolute values. A second data preparation step consisted of combining CM observations with sensor data recorded by the AM systems. In this thesis, it was chosen to combine a CM observation with sensor data prior to the CM assessment time. This approach has led to the situation that some CM observations were combined with sensor data prior to the CM assessment time, while sensor data were also available closer to and after the CM observation. In other words, it could be that a CM observation is combined with sensor data e.g., 10h prior to the CM assessment time, while it would have been more appropriate to combine the CM observation with sensor data one hour after the CM assessment time.

2. Algorithms for data classification

The development of potential predictive variables was a first, but very important step in the preprocessing of raw sensor data. After the development of a huge amount of predictive variables (n=1,065, *Chapter 3* and *4*) little other preprocessing was conducted. This was done deliberately to be sure that data used for model development and validation in this thesis showed as much similarity with field data as possible. This lack of other data preprocessing steps, therefore, resulted in the inclusion of quarter milkings with missing values due to malfunctioning of the sensors. The collected field data were highly imbalanced due to the very low prevalence of CM at quarter milking level. As explained earlier in this chapter, it was decided to apply decision tree induction as this technique was believed to be capable of dealing with data that is noisy, incomplete and imbalanced. This thesis proved the method to be capable of dealing with real farm data (*Chapter 3,* and *4*). Regarding the ability to deal with noisy data, the base decision tree described in *Chapter 4* showed an Se of 40% at an Sp of 99%. This performance was reached with the application of an extreme small time window and the inclusion of data that were hardly preprocessed to eliminate the noise naturally present in sensor data. Regarding the ability to deal with missing values, the decision trees developed in the studies described in *Chapter 3* and *4* did classify quarter milkings with missing values for the variables used at test nodes. This implies that, as long as a quarter milking has values for variables used at other test nodes, a quarter milking with missing values is still likely to be classified correctly. It should be realized that this becomes more difficult with increasing amounts of missing values;

in those situation where a quarter milking has a lot of missing values for variables used at test nodes it will probably get classified as being a healthy quarter milking as that is the majority class present in the data. The application of a cost-matrix to correct for the imbalance between CM cases and non-cases did not markedly improved detection performance compared to a decision tree with default settings (*Chapter 3*). A cost-matrix was therefore not applied in the development of the base decision tree (*Chapter 4*). The prevalence of CM in both the training and test set in *Chapter 4* was extremely low (<1% of all quarter milkings had CM), but even with this low prevalence of CM in the training set, the base decision tree did not simply classify all quarter milkings as healthy. This indicated that there is information available in sensor data that can be used to differentiate between healthy and diseased quarter milkings. Although cost-matrices did not improve detection performance markedly (*Chapter 3*), the application of bagging and boosting techniques did (*Chapter 4*). The Se of a default decision tree without bagging or boosting was 24.7% at an Sp of 99%, whereas the Se increased to 40% in just 10 iterations when bagging was applied. A disadvantage of using bagging is the decrease in the interpretability of the decision tree model.

Adding information from other sensors

Combining sensor information to improve CM detection has been done in previous studies as well (see also Table 1 in *Chapter 1*). For example, De Mol et al. (1997) showed that combining electrical conductivity data with other sensor information (milk yield and temperature) improved the detection performance of a detection model compared to a model that used electrical conductivity data only.

The detection of CM may also be improved by using better, but already existing, sensors. For example, the CM detection model developed by Song et al. (2010) showed a promising performance with an Se of 83.3% and an Sp of 99.4%, using a new generation of color sensors. These color sensors measure the transmittance of a light emitting diode (LED) rather than the older generation of sensors that measure LED light reflection. Data from these latter sensors were used in the model developed as described in *Chapter 4* of this thesis. The results reported by Song et al. (2010) were obtained with a rather wide time window (4 days), and the number of CM cases used for validating the model was limited (n=36). When the same time window of 4 days was used in combination with the base decision tree developed in *Chapter 4*, the detection performance of the base decision tree, based on data in the test set, was still lower (results not shown) than the performance reported by Song et al. (2010). The application of these newer

color sensors may at least partly explain the difference in detection performance with the performance found in this thesis.

Finally, CM detection models may also benefit from the implementation of new in-line or on-line sensors. These new sensors may add to the detection performance itself, but may also improve the detection of the severe CM cases. Detection of these severe cases was insufficient with currently available sensors as indicated previously in this chapter and in *Chapter 4*. Examples of these new and more complex sensors are biosensors that measure Lactatedehydrogenase (LDH) and N-acetyl-β-glucosaminidase (NAGase) (Brandt et al., 2010; Viguier et al., 2009). Whyte et al. (2004) introduced a new sensor assessing the SCC of milk based on viscosity measurements. It was this sensor that was used in the study described in *Chapter 6* to explore its potential as on-line CM indicator. Results from that study proved that cow level SCC information combined with quarter level electrical conductivity data improved the detection of CM, compared with using either cow level SCC or electrical conductivity alone. Mollenhorst et al. (2010) concluded that a further improvement can be expected by measuring on-line SCC at quarter level rather than at cow level.

Although these new and complex sensors are promising, they also have one disadvantage in common compared to the old-fashioned but simpler sensors; they are expensive to purchase, maintain, and operate. The older sensors (e.g., the one measuring the electrical conductivity) are not that expensive to purchase, and except for maintenance costs, no additional costs are made regarding the measurement itself. The newer sensors, however, cost money each time they are used. They may get marketed with comments such as 'a few cents per test', but a dairy farmer has to realize that a sensor that costs €0.02 per test (these are the estimated costs for measuring SCC on-line at cow level) is probably too expensive to be applied for all cow milkings at a farm using AM. For example, an AM farm milking 70 cows approximately 2.5 times/day, will apply this on-line SCC test 175 times a day, costing €3.50 a day or approximately €1,300 a year just to let this sensor do its work. Costs increase to an approximate €5,000 when the sensor is applied to quarter milking level. These costs consist of discarded milk and reagent fluids used for testing. The costs for sensors using a dipstick technology (e.g., measuring LDH) are probably even more expensive. No literature is available yet about the economic impact of these new and more expensive sensors. Before application on a large scale in practice, it is essential to explore several economically attractive implementation strategies, e.g., the measurement of SCC on-line during the first few weeks after calving, or using it as a CM confirmation test for those cows with a high electrical conductivity. In addition, whichever new sensors are going to be developed in the (near) future, to be useful in practice, these new sensors need to be easy to understand and work with, the costs need to

be sufficiently low, and the sensors themselves need to be robust, resulting in few measurement errors due to malfunctioning of the sensor.

Adding non-sensor information

Mottram (1997) suggested a potential improvement of CM detection by combining information sources. Steeneveld et al. (2010a) suggested that an improved CM detection can be expected when a probability of CM based on sensor information from AM systems is combined with a probability of CM based on non-AM cow information. This non-AM cow information included parity, lactation stage, SCC history, and CM history. It was expected that a ranked-ordered CM mastitis alert list, based on probability estimates of cows suffering from CM, would make it easier for the dairy farmer to decide which cows to check visually and to improve the detection of CM this way.

However, a first indication that this expectation was not correct was reported by Steeneveld et al. (2010b). They reported no additional value of adding non-AM cow information in order to discriminate between false positive and true positive alerts on the mastitis alert list, which were based on sensor information from AM systems alone. That unexpected result of non-AM cow information having no additional value was confirmed by results presented in *Chapter 7* of this thesis. That particular study showed that the performance of a slightly modified CM detection model as described in *Chapter 4* of this thesis was not improved when non-AM cow information (e.g., lactation stage) was added. This unexpected and perhaps disappointing result may be caused by the presence of subclinically infected cows. Subclinical mastitis has several risk factors in common with CM, such as an increased parity (e.g., Busato et al., 2000; Zadoks et al., 2001), and an increased SCC (e.g., Djabri et al., 2002a). Data used for constructing a naïve Bayesian network in *Chapter 7* included a large proportion of cows that had a high SCC in the previous 30 days, and that, therefore, can be suspected of suffering from a subclinical mastitis infection: 53,566 cow milkings had a SCC ≥ 200.000 cells/mL in the previous 30 days, while for only 130 cow milkings (0.2%) CM was reported (Table 1, *Chapter 7*). Adding this SCC information resulted in an increased posterior probability for having CM for all these 53,566 cow milkings, potentially causing the CM detection model to alert for CM based on these higher posterior probabilities. But apparently, these alerted cows stopped an intramammary infection successfully, or they were not infected at all. As a consequence, however, these alerts were considered as misclassifications. This may explain the lower CM detection performance for the model that added the SCC from the monthly milk recording to the AM sensor information, compared with the CM detection model that used the AM sensor information alone.

The observation that adding non-AM cow information, such as the variables based on the SCC obtained from milk recording test days, did not improve CM detection (*Chapter 7*), did not coincide with results reported in *Chapter 6*. In *Chapter 6*, adding SCC information improved CM detection when added to electrical conductivity sensor data. This disagreement between studies can be explained by the different time windows in which the SCC was considered as valid information. Whereas the SCC information used in the study described in *Chapter 7* is sometimes used for a cow milking that occurs four weeks after the SCC measurement during test days, an on-line measurement of SCC as used in the study described in *Chapter 6* applies directly to the cow milking at which the on-line measurement is assessed. Thus, SCC information seems a useful mastitis indicator, as long as it is measured very closely to (or preferably at) the milking under consideration by a CM detection model.

The non-AM cow information used in the study described in *Chapter 7* included more than just variables based on SCC measurement during monthly test days. For example, lactation stage, parity, and season were other variables that were included as independent variables as well. However, these non-AM cow information variables also did not improve the ability to discriminate between milkings with and without CM. This result agrees with a French study (Djarbi et al., 2002b) that concluded that adding parity and lactation stage to a detection model for IMI had a very limited impact on the model's accuracy.

Additionally gained knowledge

One of the major additional lessons learned from this thesis is that the performance of CM detection models cannot be judged, nor compared fairly, without having the essential information about the gold standard and the time window that were used, and the data inclusion criteria. If these differ too much between studies, comparison of detection performance becomes very difficult, if not impossible. It is necessary to have and internationally accepted gold standard, together with an agreement on the width of the time window. Also, agreement on in- and exclusion criteria for data analyses and testing is necessary. Finally, it is necessary to have an international agreement on which detection performances are required, and which performance measures are to be evaluated. It is important to realize that all these necessary agreements on gold standards, time windows, and data inclusion criteria will depend on one basic question: "What is to be detected and who wants to detect it?". Different detection goals require different definitions for gold standards, time windows and detection performance requirements. If the goal is to detect severe cases of CM, a small time window is essential for a timely start of

antibiotic treatment. If the goal is to divert abnormal milk automatically, Se and Sp requirements are totally different from a detection model developed to alert for cows with subclinical mastitis. Detection goals will differ between scientists (who develop models) and dairy farmers (who have to base decisions on the outcome of these models). Whereas scientists are eager to search for the perfect alert (e.g., even the smallest flakes in foremilk are detected), a dairy farmer may be satisfied with a more practical alert: a model alerting for cows that have severe CM is probably enough as these cows need an antibiotic treatment as soon as possible. Cows having only a few flakes in the foremilk (mild CM infection) may be of less interest for dairy farmers, as Keefe et al. (2010) indicated that there was no difference in cure rate for moderate CM cases being treated immediately or 24h after detection. Regarding economics, dairy farmers may be interested in a detection model alerting for subclinical mastitis, as Van den Borne et al. (2010) reported that treatment of contagious subclinical mastitis during the lactation is economically beneficial.

As CM detection models are developed to be applied in practice, it is essential to gain more insight in what dairy farmers want and which requirements they think are important. Based on what they want, an international agreement on gold standard, time windows, and detection performance requirements can then be formulated. A next step is then to set up an extensive and detailed field study for future CM detection research, where data are collected at more commercial farms for a longer period of time in which all quarter milkings are visually checked for abnormalities in milk. This dataset, in combination with the agreements on gold standards, time windows, and detection requirements can then be used by different classification algorithms in order to answer the question whether one CM detection model outperforms another.

Conclusion

So, has the perfect CM detection model for an AM system been developed in this thesis? No, in fact it has not, and as long as there is no agreement on what is to be detected, the perfect CM model may continue to be fiction. However, results from this thesis made us learn some important lessons, and they made it possible to define a number of conclusions:

1. Decision tree induction is able to deal with field data that are noisy, incomplete, and imbalanced. The base decision tree seems to have improved detection performances compared with currently available models. This improvement was obtained while applying a small time window, relevant to

farming practice, and by including quarter milkings with a less clear CM status and missing values.

2. The recommended Se of >70% with an Sp of >99% for CM detection models was almost reached with the use of currently available sensor data and decision tree induction, even with the additional requirements used in this thesis (small time window and similarity with field data).

3. Sensor data show potential to discriminate between Gram-positive and Gram-negative CM causal pathogens. More CM cases with sensor data and bacteriological culturing results are necessary to develop a robust decision tree model.

4. Including information from other sensors is beneficial to improve CM detection models. These other sensors may consist of improved versions of the sensors available at the AM systems included in this thesis, but also new sensor information seems promising as CM indicator; adding on-line measurements of cow level SCC to electrical conductivity measurements improved the detection of CM. However, economically beneficial implementation strategies of these new sensors need to be explored.

5. Adding non-sensor cow information did not improve the detection of CM. Detection models using AM sensor measurements outperformed models using AM sensor information and non-AM cow information.

6. It is not possible to have a fair and sensible comparison between results found in this study with results reported in previous studies, due to the large variation in gold standard definitions, the variation in applied time windows, the different evaluation characteristics used, and the differences in data inclusion criteria and quality of sensors used for analyses.

This thesis confirmed that CM detection models can be improved by adding new or improved sensors or by applying new algorithms for data preprocessing and classification. It showed the importance of developing and validating models with field data from more commercial farms, and the necessity of a data set where all quarter milkings are visually checked for abnormalities or CM. These aspects are needed in order to give a good estimate for the performance of a CM detection model when applied in practice. However, there is still a long way to go in the search for the perfect CM detection model that satisfies both (theoretical) scientists and (practical) dairy farmers. We, as scientists, are only at the beginning of this

interesting and challenging quest for such a model. Essential efforts necessary for taking the next steps are: gaining more insight in the needs of dairy farmers, obtaining international agreements on gold standard definitions, applied time windows, data inclusion criteria, and collecting a data set from more than one commercial dairy farm where all quarter milkings are checked visually for CM.

References

Brandt, M., A. Haeussermann, and E. Hartung. 2010. Invited review: Technical solutions for analysis of milk constituents and abnormal milk. J. Dairy Sci. 93(2):427-436.

Busato, A., P. Trachsel, M. Schällibaum, and J.W. Blum. 2000. Udder health and risk factors for subclinical mastitis in organic dairy farms in Switzerland. Prev. Vet. Med. 44(3-4):205-220.

Cavero, D., K.-H. Tölle, C. Buxadé, and J. Krieter. 2006. Mastitis detection in dairy cows by application of fuzzy logic. Livest. Sci. 105(1-3):207-213.

Chagunda, M.G.G., N.C. Friggens, M.D. Rasmussen, and T. Larsen. 2006. A model for detection of individual cow mastitis based on an indicator measured in milk. J. Dairy Sci. 89(8):2980-2998.

Claycomb, R.W., P.T. Johnstone, G.A. Mein, and R.A. Sherlock. 2009. An automated in-line clinical mastitis detection system using measurement of conductivity from foremilk of individual udder quarters. New Zealand Veterinary Journal. 57(4):208-214.

De Mol, R.M., A. Keen, G.H. Kroeze, and J.M.F.H. Achten. 1999. Description of a detection model for oestrus and diseases in dairy cattle based on time series analysis combined with a Kalman filter. Computers and Electronics in Agriculture. 22(2-3):171-185.

De Mol, R.M., G.H. Kroeze, J.M.F.H. Achten, K. Maatje, and W. Rossing. 1997. Results of a multivariate approach to automated oestrus and mastitis detection. Livest. Prod. Sci. 48 (3):219-227.

De Mol, R.M., and W. Ouweltjes. 2001. Detection model for mastitis in cows milked in an automatic milking system. Prev. Vet. Med. 49(1-2):71-82.

De Mol, R.M., W. Ouweltjes, G.H. Kroeze, and M.M.W.B. Hendriks. 2001. Detection of estrus and mastitis: Field performance of a model. Applied Engineering in Agriculture. 17(3):399-407.

De Mol, R.M., and W.E. Woldt. 2001. Application of fuzzy logic in automated cow status monitoring. J. Dairy Sci. 84(2):400-410.

Djabri, B., N. Bareille, F. Beaudeau, and H. Seegers. 2002a. Quarter milk somatic cell count in infected dairy cows: a meta-analysis. Vet. Res. 33(4):335-357.

Djabri, B. N. Bareille, B. Poutrel, F. Beaudeau, M. Bucelliez, and H. Seegers. 2002b. Accuracy of the detection of intramammary infection using quarter somatic cell count when taking parity and stage of lactation of the dairy cow into account. Anim. Res. 51(2):135-148.

Espada, E., and H. Vijverberg. 2002. Milk colour analysis as a tool for the detection of abnormal milk. Pages 28-38 in First North American Conference on Robotic Milking. Toronto, Canada. Wageningen Pers, Wageningen, the Netherlands.

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

Hogeveen, H., and W. Ouweltjes. 2002. Mastitis therapy and control: Automatic on-line detection of abnormal milk. Pages 1735-1740 in Encyclopedia of dairy sciences. R. Roginski, J.W. Fuquay, and P. F. Fox (eds.). Academic Press, London, UK.

Hovinen, M., A.-M. Aisla, and S Pyörälä. 2006. Accuracy and reliability of mastitis detection with electrical conductivity and milk colour measurement in automatic milking. Acta Agriculturae Scandinavica Section A. 56(3):121–127.

Keefe, G., J. McCarron, K. MacDonald, and M. Cameron. 2010. The scientific basis for using on-farm culture systems. Pages 141-149 in National Mastitis Council Annual Meeting Proceedings. Alberquerque, New Mexico.

Kim T., and C.W. Heald. 1999. Inducing inference rules for the classification of bovine mastitis. Comput. Electron. Agr. 23(1):27-42.

Lago, A., S. Godden, R. Bey, P. Ruegg, and K. Leslie. 2009. Effect of the selective treatment of clinical mastitis based on farm culture results on clinical mastitis recurrence, somatic cell count, milk production, and culling. Pages 150-151 in Proceedings of the Annual Meeting of the National Mastitis Council. Charlotte, North Carolina.

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.

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 Control - From science to practice. T.J.G.M. Lam (ed.). Wageningen Academic Publishers, Wageningen, the Netherlands.

Milner, P., K.L. Page, A.W. Walton, and J.E. Hillerton. 1995. Detection of Clinical Mastitis by Changes in Electronical Conductivity of Foremilk Before Visible Changes in Milk. J. Dairy. Sci. 79:86-86.

Mitchell, R.S., R.A. Sherlock, and L.A. Smith. 1996. An investigation into the use of machine learning for determining oestrus in cows. Comput. Electron. Agr. 15: 195-213.

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

Mollenhorst, H., P.P.J. van der Tol, and H. Hogeveen. 2010. Somatic Cell Count Assessment at Quarter or Cow Milking Level. J. Dairy Sci. 93(7):3358-3364.

Mottram, T. 1997. Automatic monitoring of the health and metabolic status of dairy cows. Livestock. Prod. Sci. 48(3): 209-217.

Nielen, M., Y.-H. Schukken, A. Brand, S. Haring, and R.T. Ferwerda – van Zonneveld. 1995a. Comparison of Analysis Techniques for Online Detection of Clinical Mastitis. J. Dairy Sci. 78(5):1050-1061.

Nielen, M., M.H. Spigt, Y.-H. Schukken, H.A. Deluyker, K. Maatje, and A. Brand. 1995b. Application of a Neural-Network to Analyze Online Milking Parlor Data for the Detection of Clinical Mastitis in Dairy-Cows. Prev. Vet. Med. 22(1-2):15-28.

Norberg, E., H. Hogeveen, I.R. Korsgaard, N. C. Friggens, K. H. M. N. Sloth, and P. Løvendahl. 2004. Electrical conductivity of milk: Ability to predict mastitis status. J. Dairy Sci. 87(4):1099-1107.

Pietersma, D., R. Lacroix, D. Lefebvre, and K.M. Wade. 2002. Decision-tree induction to interpret lactation curves. Canadian Biosystems Engineering. 44:7.1-7.13.

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

Rasmussen, M.D., M. Bjerring, and F. Skjøth. 2005. Visual appearance and CMT score of foremilk of individual quarters in relation to cell count of cows milked automatically. J. Dairy. Res. 72(1): 49-56.

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. Pages 275-282 in Mastitis control - from science to practice. T.J.G.M. Lam (ed). Wageningen Acadamic Publishers, Wageningen, the Netherlands.

Song, X., S. Zhuang, and P.P.J. van der Tol. 2010. New model to detect clinical mastitis in Astronaut A3 Next^{tm} milking robot. Pages 474-480 in Mastitis research into practice: Proceedings of the 5th IDF mastitis conference. J.E. Hillerton (ed). VetLearn, Wellington, New Zealand.

Steeneveld, W., L.C. Van der Gaag, H.W. Barkema, and H. Hogeveen. 2010a. Simplify the interpretation of alert lists for clinical mastitis in automatic milking systems. Comp. Electron. Agr. 71(1): 50-56.

Steeneveld, W., L. C. Van der Gaag, W. Ouweltjes, H. Mollenhorst, and H.Hogeveen. 2010b. Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems. J. Dairy Sci. 93(6):2559-2568.

Sun, Z., S. Samarasinghe, and J. Jago. 2010. Detection of mastitis and its stage of progression by automatic milking systems using artificial neural networks. J. Dairy. Res. 77(2):168-175.

Van den Borne, B.H.P., T. Halasa, G. van Schaik, H. Hogeveen, and M. Nielen. 2010. Bio-economic modeling of lactational antimicrobial treatment of new bovine subclinical intramammary infections caused by contagious pathogens. J. Dairy Sci. 93(9):4034-4044.

Viguier, C., S. Arora., N. Gilmartin, K. Welbeck, and R. O'Kennedy. 2009. Mastitis Detection: current trends and future perspectives. Trends Biotech. 27(8):486-493.

Whyte, D.S., R.G. Orchard, P.S. Cross, T. Frietsch, R.W. Claycomb, and G.A. Mein. 2004. An on-line somatic cell count sensor. Pages 235–240 in Automatic Milking: A better understanding. A. Meijering, H. Hogeveen, and C.J.A.M. de Koning (eds.). Wageningen Academic Publishers, Wageningen, the Netherlands.

Zadoks, R.N., H.G. Allore, H.W. Barkema, O.C. Sampimon, G.J. Wellenberg, Y.T. Gröhn, and Y.-H. Schukken. 2001. Cow- and quarter level risk factors for *Streptococcus uberis* and *Staphylococcus aureus* mastitis. J. Dairy Sci. 84(12):2649-2663.

Summary

The first commercially available automatic milking (AM) systems were introduced in the early 1990s. With AM, the milking process is completely automated, which makes it no longer necessary for the dairy farmer to be physically present during this process. Since the introduction of AM, the adoption of these systems took a flight with 8,000 farms in 25 countries that use AM in 2010. As with conventional milking (in a milking parlor), a farmer that uses AM has to fulfill requirements by law regarding udder health (e.g., Regulation (EC) No. 853/2004). One of these requirements considers the inspection of milk for abnormalities in respect of color and homogeneity. This abnormal milk must be excluded for human consumption. Most abnormalities in milk are caused by clinical mastitis (CM). As there is no obligatory contact during milking between the farmer and the dairy cows with AM, sensors have to be used to replace visual inspection to monitor udder health. Clinical mastitis detection models available at AM systems use the available sensor information as input data to produce mastitis alert lists. These lists report cows that need a visual health check as they deviate, according to the sensor information, from 'normal' for one reason or another. It is the responsibility of the dairy farmer to check these listed cows visually to confirm a CM. Most CM detection models make use of a sensor that measures the electrical conductivity of the milk. In case of a CM, physiological changes in the udder cause the electrical conductivity to increase, and this increase is used as indicator for udder health. More recently, also other sensor information (e.g., color or somatic cell count) is used by detection models to monitor udder health. Current CM detection models used by AM systems show a sensitivity of 36.8%, indicating that 36.8% of all quarter milkings with CM are detected by the CM detection model (true positive alerts). This sensitivity of 36.8% can be improved. At the same time, the specificity of current detection models is 97.9%, indicating that 2.1% of all healthy quarter milkings (without CM) will be listed on the mastitis alert list as well (false positive alerts). This may seem a low percentage at first, but with the knowledge that over 99.5% of all cow milkings are healthy (no CM), then this 2.1% results in a high absolute number of false positive alerts. There is a strong need to improve the specificity of the detection models as dairy farmers especially complain about these high numbers of false positive alerts on the mastitis alert lists. In the past, a lot of research has been conducted in the field of CM detection. Despite this research, most developed models are not, or just experimentally, applied in practice. A reason for this lacking adoption is that these models fail to meet a number of essential requirements for the automated detection of CM. These requirements are: a CM detection model should

- detect at least severe cases of CM with high levels of sensitivity (>70%) and specificity (>99%);
- alert for CM within a short time window at which the CM actually occurs;

- be able to deal with field data. Field data (collected at commercially producing dairy farms rather than research farms) are noisy and incomplete by definition. In addition, the low prevalence of CM at cow milking level causes the data to be highly imbalanced. A model should also be able to deal with the large difference between the number of healthy quarter milkings and the relatively small number of quarter milkings that have CM;
- provide information about the CM causal pathogen. This requirement is less important than the first three requirements. However, such information would be a valuable addition for current CM detection models and can be used as a decision support tool for farmers in their antibiotic treatment decisions.

The objective of this thesis was to improve the detection of CM with sensor data from AM systems. Three different approaches were explored to develop a CM detection model that meets the four requirements described above. These approaches are: applying other algorithms for data preprocessing and data classification, adding information from other sensors to existing detection models, and adding non-sensor information to sensor based CM detection models.

The main body of this thesis focused on improving CM detection by using other algorithms for data preprocessing and for classifying the preprocessed data. The development of any detection model starts with preprocessing collected data. An example of this preprocessing is the transformation from raw sensor measurements (e.g., measured each second during a quarter milking) to a variable that summarizes or describes these raw measurements. Previous studies often averaged the raw electrical conductivity measurements, or they used only the maximum value of the electrical conductivity. However, a clear motivation for using these two variables specifically was lacking. Therefore, this thesis started with the introduction of a data flow diagram to systematically describe the steps involved in the transformation of within quarter milking sensor measurements (electrical conductivity, color, milk yield) into potentially predictive variables used for the detection of abnormal milk and CM (*Chapter 2*). These potentially predictive variables were evaluated for their importance to classify abnormal milk and CM by computing correlation coefficients and information gain ratios. Results of the study described in *Chapter 2* showed that models for abnormal milk or CM detection may benefit from the inclusion of variables that are based on different sensors and that describe different characteristics of the raw sensor measurements. Especially variables based on the electrical conductivity and the colors blue and green were expected to have predictive power for the detection of CM. Also variables that describe other characteristics of raw sensor measurements than the often used mean and maximum value were expected to be useful for a model to detect

abnormal milk and CM. Finally, variables based on absolute sensor measurement values could be as important as variables based on differences with expected values. These expected values can be based on values measured at previous milkings of the same quarter or on values measured at other quarters within the same cow milking.

A next step was to apply a new algorithm for the classification of preprocessed sensor data (*Chapter* 3). This new algorithm, called decision tree induction, uses the preprocessed sensor data (electrical conductivity, color, and milk yield), and should be able to be implemented in practice as well. This study used field data collected at six Dutch dairy farms using AM, and included 3,000 quarter milkings without CM and 97 quarter milkings with CM for the statistical analyses. Due to this small number of quarter milkings, ten fold cross validation was used for the development and validation of several decision trees. Cost matrices were introduced to account for the imbalance of the data, and confidence factors were adapted to prevent over-fitting the decision trees. Results showed that the decision trees were easy to interpret when visualized. As predicted in the study described in *Chapter 2*, the developed decision trees combined information from different sensors, including variables that described the shape, variability and level of sensor measurement patterns. Variables were based on comparisons with expected values (based on other quarters or on previous quarter milkings) or on absolute sensor measurement values. Detection performances of the developed decision trees were comparable with those of detection models currently used by AM systems. As results from this study were obtained using a simple algorithm, field data, and a narrow time window of <24h in which the decision trees could alert for CM, the main conclusion of the study described in *Chapter 3* was that decision tree induction showed potential for detecting CM with sensor data from AM systems. Recommendations were formulated to further improve the detection of CM. These recommendations were: using sensor information of more CM cases, using only those quarter milkings that were clearly healthy or that clearly suffered from CM for the training process, and to apply bagging or boosting techniques. Both techniques use selective resampling of training data to improve classification results.

Chapter 4 describes a study that implemented the recommendations formulated in *Chapter 3* to improve the detection of CM. For this study, field data were collected at nine Dutch dairy farms using AM. A total of 3.5 million quarter milkings with sensor data (electrical conductivity, color, and milk yield) were available, of which 348 had CM according to observations recorded by the participating dairy farmers. Data were divided into a training (including two thirds of all data) and a test set (including the remaining one third of the data). Decision trees were trained, with or

without bagging and boosting techniques, with 243 quarter milkings with CM and 24,717 quarter milkings with a very high likelihood of being healthy. Validation was done using 105 quarter milkings with CM and a random sample of 50,000 quarter milkings without a CM observation (thus including quarter milkings with a less clear mastitis status as well). The decision tree developed in combination with bagging showed the best detection performance; this tree, called the base decision tree, showed a sensitivity of 40% at a specificity of 99%. This is a similar sensitivity of CM cases compared with models currently used by AM systems, but with a more than 50% reduction of the number of false positive alerts. When applying a suggested practical time window (being 48h for a CM alert and 24h for a CM observation), sensitivity increased to 69.5% at a specificity of 99%. This result indicates that decision tree induction is able to get close to the suggested detection performance requirements (being a sensitivity of >70% and a specificity of >99%), but not fulfill these requirements completely. It is expected that the base decision tree described in *Chapter 4* is suitable for use in practice as field data from nine dairy farms was used for model building and validation, alerts for CM were given within a narrow time window of less than 24h, and detection performance was based on a test set including quarter milkings with a less clear mastitis status as well, mimicking practice as closely as possible. It was difficult to compare the results described in *Chapter 4* with previously reported results, due to the large variation in gold standard definition, used data, and applied time windows.

In order to meet the fourth requirement of a CM detection model – that is, providing additional information about the CM causal pathogen – *Chapter 5* describes a study that developed a decision tree for the prediction of the Gram-status of a CM causal pathogen. Whenever a farmer has to choose to start an antibiotic treatment, it can be useful to know the Gram-status of the CM causal pathogen. Cows with a Gram-negative CM may benefit more from a supportive treatment (e.g., fluids and painkillers) to relieve the cow of systemic signs than to start an antibiotic treatment, while cows with a Gram-positive CM can be more easily cured with an antibiotic treatment. For development and validation of a Gram-status predicting decision tree, 140 CM cases with sensor data (electrical conductivity, color, and milk yield) and with bacteriological culturing results were used: 96 CM cases were used to train a decision tree (training set) of which 21 were Gram-negative. The test set included 44 CM cases of which nine were Gram-negative. Results from the training set indicated that sensor measurements could be used as tool for Gram-status prediction; the developed decision tree showed an accuracy of 90.6%, indicating that of all CM cases the decision tree predicts for 90.6% of these cases the correct Gram-status of the CM causal pathogen. From all 96 CM cases in the training set, 74% received a high probability to be infected with

either a Gram-positive or a Gram-negative pathogen by the decision tree. When only these CM cases were selected, the decision tree showed an accuracy of 97.2%. The decision tree, however, was not robust as the accuracy dropped to 54.5% when the model was validated using data in the test set. Predicting the CM causal pathogen itself was not possible; that specific decision tree had an accuracy of 52% based on data in the training set. However, the results do indicate a potential to use in-line sensor measurements from automatic milking systems to differentiate between Gram-positive and Gram-negative CM causal pathogens.

Adding information from other sensors to already existing CM detection models was a second route that was explored to improve detection performance of CM detection models. *Chapter 6* of this thesis describes the potential additional value of including on-line measured somatic cell count at cow level in the detection of CM. Data were collected at one research farm that applied AM (the Greenfield Project, Hamilton, New Zealand). Three detection models were compared; the first model used the electrical conductivity as sole criterion for udder health, the second one used somatic cell count information only, and the third model applied a fuzzy logic algorithm to combine these two information sources. The sensitivity of all three models was fixed at 80%. The false alert rate per 1,000 cow milkings was similar between the first two detection models. However, combining the two information sources resulted in a 2- to 3-fold reduction in false alert rate. It was concluded that performance of a CM detection model improved when on-line somatic cell count information at cow level was added to a model based on electrical conductivity measurements.

The third explored method to improve the detection of CM was to add non-sensor cow information. *Chapter 7* describes a study that adds non-sensor cow information to a slightly modified version of the base decision tree described in *Chapter 4*. This modified decision tree uses sensor data to provide a prior probability estimate for a cow milking to be a mastitic one. By adding non-sensor cow information (parity, month in lactation, season, SCC history, and CM history) the CM probability estimate was then adapted to calculate a posterior probability estimate. Despite expectations, adding non-sensor cow information to a CM detection model based on sensor data did not improve detection performance. On the contrary, the CM detection model based on only sensor information showed the highest area under the receiver operating characteristic curve, indicating a better detection performance than the CM detection model where sensor and non-sensor cow information were combined. The main conclusion of this study was that adding non-sensor cow information had no additional value for the detection of CM when a model based on sensor measurements is available.

In the general discussion (*Chapter 8*) each specific requirement of an automated CM detection model (requirements of detection performance, using time windows, field data, and providing information of the CM causal pathogen) as well as the three approaches investigated (using other algorithms, adding new sensor information, and adding non-sensor information) to improve the detection of CM with AM systems have been integrated and discussed. *Chapter 8* also discusses that, for the future development of a CM detection model, it is extremely important to get more insight in what dairy farmers expect from CM detection models and what it is they want to be alerted for. Finally, this chapter describes the importance of getting an international agreement on gold standard definitions, time windows, and detection performance requirements in order to get a more fair comparison between CM detection models in the future.

Samenvatting

De eerste commercieel beschikbare automatische melksystemen (AMS) werden begin 1990 geïntroduceerd. Bij AMS is het gehele melkproces volledig geautomatiseerd, waardoor het voor melkveehouders niet langer noodzakelijk is om fysiek bij dit proces aanwezig te zijn. Sinds de introductie van AMS is deze manier van melken uitgegroeid tot een wereldwijd geaccepteerde technologie waarbij in 2010 ongeveer 8.000 melkveehouders uit 25 landen gebruik van dit melksysteem. Net als bij het conventioneel melken (in een melkput), heeft een melkveehouder met een AMS de verplichting te voldoen aan een aantal wettelijk geformuleerde eisen met betrekking tot uiergezondheid (bijvoorbeeld Wetgeving (EC) No. 853/2004). Een van deze eisen is dat melk gecontroleerd moet worden op afwijkingen wat betreft kleur en homogeniteit. Deze afwijkende melk mag niet gebruikt worden voor humane consumptie. De meeste afwijkingen in melk worden veroorzaakt door een zichtbare uierontsteking, ofwel een klinische mastitis (KM). Aangezien er bij AMS geen verplicht contact meer is tussen de melkveehouder en zijn melkvee, wordt het monitoren van uiergezondheid bij AMS overgenomen door sensoren om op deze manier aan de wettelijke eisen te voldoen. De AMS gebruiken modellen voor de detectie van KM, die op hun beurt de sensorinformatie gebruiken als input voor het produceren van zogenoemde mastitis attentie lijsten. Deze lijsten vermelden koeien die afwijken van 'normaal' op basis van deze sensor gegevens. Het is vervolgens de verantwoordelijkheid van de melkveehouder om deze koeien vervolgens visueel te controleren op het hebben van KM. Van oudsher maken KM detectiemodellen vooral gebruik van een sensor die de elektrische geleidbaarheid van melk meet. Wanneer een koe een KM heeft, treden er fysiologische veranderingen op in de uier die de elektrische geleidbaarheid van de melk verhogen. Het is deze verhoging in de elektrische geleidbaarheid die gebruikt wordt als voorspeller voor KM. Tegenwoordig worden ook sensoren gebruikt die bijvoorbeeld de kleur en het celgetal van de melk meten om de uiergezondheid te monitoren. Op dit moment hebben detectiemodellen die gebruikt worden door AMS een sensitiviteit van 36,8%. Dit betekent dat 36,8% van alle kwartiermelkingen met KM inderdaad door het detectiemodel opgespoord worden (positieve attenties). Deze sensitiviteit van 36,8% kan verbeterd worden. Tegelijkertijd hebben de detectiemodellen een specificiteit van 97.9%. Dit betekent dat 2,1% van alle gezonde kwartiermelkingen (geen KM) ook op de mastitis attentielijst komen te staan (vals positieve attenties). Dit lijkt op het eerste gezicht een klein percentage, maar wanneer ruim 99,5% van alle koemelkingen gezond is (geen KM), is deze 2,1% toch een groot aantal vals positieve attenties wanneer uitgedrukt in absolute aantallen. Er is een duidelijke behoefte vanuit de praktijk om ook de specificiteit van detectiemodellen te verbeteren omdat melkveehouders vooral klagen over dit grote aantal vals positieve attenties op de mastitis attentie lijsten. Ondanks dat er al veel onderzoek verricht is op het gebied van KM detectiemodellen, worden de

meeste ontwikkelde modellen niet, of slechts experimenteel, toegepast. Dit komt omdat deze modellen niet voldoen aan een aantal essentiële voorwaardes voor de detectie van KM. Deze voorwaardes zijn dat een detectiemodel voor KM

- op zijn minst de ernstige KM gevallen zou moeten detecteren met een hoge sensitiviteit (>70%) en specificiteit (>99%);
- een attentie voor KM moet geven binnen een kort tijdsbestek rondom het daadwerkelijk optreden van een KM;
- moet met velddata kunnen omgaan. Velddata (verzameld op commerciële bedrijven in plaats van op proefbedrijven) zit per definitie vol met ruis en missende waardes. Door de lage prevalentie van KM op koemelkingsniveau is de verdeling tussen gezonde en zieke (met KM) kwartiermelkingen ook nog eens extreem scheef verdeeld. Ook met deze onbalans tussen zieke en gezonde kwartiermelkingen moet een KM detectiemodel kunnen omgaan;
- informatie moeten geven over de bacterie die de KM veroorzaakt. Deze laatste voorwaarde is minder belangrijk dan de eerste drie voorwaardes. Toch zou deze informatie een belangrijke uitbreiding kunnen zijn van de huidige KM detectiemodellen omdat melkveehouders met deze informatie een betere inschatting kunnen maken van de juiste antibiotica behandeling.

Het doel van dit proefschrift is het verbeteren van de detectie van KM met behulp van sensorinformatie van AMS. Er is op drie manieren geprobeerd om een detectie model voor KM te ontwikkelen die voldoet aan de vier hierboven beschreven criteria. De drie manieren zijn: het toepassen van andere modellen voor het voorbereiden en het classificeren van de data, het toevoegen van andere sensorinformatie aan bestaande modellen, en als laatste het toevoegen van niet-sensor informatie aan modellen gebaseerd op sensorinformatie.

In dit proefschrift ligt de nadruk op het verbeteren van de detectie van KM door middel van het toepassen van andere modellen voor zowel de voorbereiding van ruwe data als voor de classificatie van de voorbereidde data. Het ontwikkelen van een model begint altijd bij het voorbereiden van de verzamelde data. Een voorbeeld van deze voorbereiding is de omzetting van ruwe sensormetingen (welke bijvoorbeeld iedere seconde gedurende een kwartiermelking gemeten worden) naar een variabele die deze ruwe sensormetingen omschrijft. Zo werd in voorgaande onderzoeken vaak de gemeten waardes van de elektrische geleidbaarheid tijdens een koe- of kwartiermelking gemiddeld, of werd alleen de maximaal gemeten waarde eruit gefilterd. Het was echter onduidelijk waarom juist deze waardes gekozen werden als voorspellende variabelen. Daarom is dit proefschrift begonnen met de introductie van een gegevensstroomschema om op een systematische manier de stappen te omschrijven die nodig zijn om ruwe sensordata (elektrische

geleidbaarheid en kleur gemeten op kwartiermelkings niveau) om te zetten naar mogelijk voorspellende variabelen voor de detectie van abnormale melk en KM (*Hoofdstuk 2*). De voorspellende waarde van de ontwikkelde variabelen voor de detectie van abnormale melk en KM zijn vervolgens bepaald door het berekenen van correlatie coëfficiënten en 'information gain ratios'. Resultaten van de studie beschreven in *Hoofdstuk* 2 lieten zien dat detectiemodellen voordeel kunnen hebben wanneer variabelen gebruikt worden die gebaseerd zijn op verschillende sensoren en die verschillende karakteristieken van de ruwe sensormetingen beschrijven. Vooral variabelen gebaseerd op de elektrische geleidbaarheid en de kleuren groen en blauw leken een goede voorspellende waarde te hebben voor de detectie van KM. Ook variabelen die andere eigenschappen beschrijven van ruwe sensormetingen zouden, naast het vaak gebruikte gemiddelde of maximum gemeten waarde, een toevoeging kunnen zijn voor detectiemodellen voor abnormale melk en KM. Tot slot leken variabelen die gebaseerd zijn op absoluut gemeten sensorwaardes net zo veel voorspellende waarde te hebben voor de detectie van abnormale melk en KM als variabelen die gebaseerd zijn op een verwachte waarde. Deze verwachte waardes kunnen betrekking hebben op waardes gemeten bij voorgaande melkingen van hetzelfde kwartier of op waardes gemeten bij andere kwartieren binnen dezelfde koemelking.

Een volgende stap was het toepassen van een nieuw model voor de classificatie van de voorbereidde sensordata (*Hoofdstuk 3*). Dit nieuwe model, 'decision tree induction', omvat het hele proces van het ontwikkelen van een beslisboom voor de detectie van KM. Deze beslisboom maakt gebruik van voorbereidde sensorinformatie (elektrische geleidbaarheid, kleur en melkproductie), en zou ook in de praktijk door AMS gebruikt kunnen worden. Voor deze studie is gebruik gemaakt van een dataset bestaande uit gegevens van zes melkveebedrijven, waarbij uiteindelijk 3.000 kwartier melkingen zonder KM en 97 kwartier melkingen met KM gebruikt werden voor de statistische analyses. Vanwege deze kleine aantallen is gebruik gemaakt van tienvoudige kruislingse validatie om beslisbomen te ontwikkelen en te valideren. Verschillende kostmatrices werden toegepast om de onbalans in de data op te heffen. Om te voorkomen dat een beslisboom te veel aangepast is op de data die gebruikt wordt voor de ontwikkeling, en daardoor niet meer algemeen toepasbaar is, werden ook verschillende betrouwbaarheidsfactoren gebruikt bij de analyses. De resultaten toonden aan dat beslisbomen gemakkelijk te interpreteren zijn. Zoals voorspelt door de studie beschreven in *Hoofdstuk 2* van dit proefschrift, gebruikten de ontwikkelde beslisbomen informatie van verschillende sensoren voor de detectie van KM, waarbij variabelen gebruikt werden die de vorm, variabiliteit, en het niveau van sensormetingen omschreven. De beslisbomen maakten gebruik van variabelen die gebaseerd waren op vergelijkingen met

verwachte waardes (gebaseerd op zowel andere kwartieren binnen dezelfde koemelking als op voorgaande melkingen van hetzelfde kwartier) en van variabelen gebaseerd op absoluut gemeten waardes van sensormetingen. De detectievermogens van de ontwikkelde beslisbomen waren vergelijkbaar met die van detectiemodellen die op dit moment door AMS gebruikt worden. Omdat in deze studie gebruik werd gemaakt van een simpel model, velddata, en een tijdsinterval van <24 uur waarin een beslisboom een attentie voor KM moest afgeven, was de belangrijkste conclusie van de studie beschreven in *Hoofdstuk 3* dat beslisbomen heel goed in staat zijn om KM te detecteren met behulp van sensorgegevens van AMS. Een aantal adviezen werden geformuleerd om het detectievermogen van de beslisbomen te verbeteren. Deze adviezen waren: het gebruik maken van meer KM gevallen met sensorinformatie, voor het ontwikkelen van een beslisboom alleen gebruik maken van kwartiermelkingen die duidelijk ziek of gezond zijn, en bij het ontwikkelen van een beslisboom gebruik maken van 'bagging' en 'boosting' technieken. Dit zijn twee technieken die gebruik maken van het selectief herschikken van de data die gebruikt wordt om een model te ontwikkelen om zo het detectievermogen te verbeteren.

Hoofdstuk 4 beschrijft een studie waarbij de adviezen vermeldt in *Hoofdstuk 3* zijn toegepast. Voor deze studie is een dataset gebruikt bestaande uit gegevens van negen melkveebedrijven. In totaal waren er 3,5 miljoen kwartiermelkingen met sensorinformatie (elektrische geleidbaarheid, kleur en melkopbrengst) beschikbaar, waarvan er 348 KM hadden op basis van waarnemingen door de deelnemende melkveehouders. Data werden verdeeld in een training set (2/3 van alle data) en een test set (met de 1/3 data). Voor het ontwikkelen van beslisbomen (met en zonder 'bagging' en 'boosting' technieken) werden uit de training set 24.717 duidelijk gezonde kwartiermelkingen geselecteerd en 243 kwartiermelkingen met KM. Het detectievermogen van de beslisbomen werd getest met behulp van 105 kwartiermelkingen met KM en van een willekeurige steekproef van 50.000 kwartiermelkingen zonder KM waarneming (en dus van kwartiermelkingen met een minder duidelijk omschreven gezondheidsstatus wat betreft mastitis). Het beste detectievermogen (een sensitiviteit van 40% en een specificiteit van 99%) werd behaald door een beslisboom gecombineerd met 'bagging' (verder de basis beslisboom genoemd). Dit resultaat is wat betreft de sensitiviteit vergelijkbaar met detectiemodellen die op dit moment door AMS gebruikt worden, maar de basis beslisboom laat een daling van meer dan 50% zien voor het aantal vals positieve attenties in vergelijking met detectiemodellen die op dit moment door AMS gebruikt worden. Wanneer een tijdsinterval toegepast werd dat in eerder onderzoek werd voorgesteld als een praktisch toepasbaar tijdsinterval (48 uur voor een attentie voor KM en 24 uur voor een observatie van KM), was de sensitiviteit van de basis

beslisboom 69,5% bij een specificiteit van 99%. Dit resultaat toont aan dat de basis beslisboom in de buurt komt van de eisen aan het detectievermogen van een KM detectie model (sensitiviteit van >70% en een specificiteit van >99%), maar er nog niet helemaal aan voldoet. De ontwikkelde basis beslisboom beschreven in *Hoofdstuk 4* zal naar verwachting goed toepasbaar zijn in de praktijk omdat bij de ontwikkeling en de validatie gebruikt is gemaakt van velddata van negen verschillende praktijkbedrijven, de basis beslisboom binnen 24 uur een attentie voor KM moest afgeven, en omdat het detectievermogen van de basis beslisboom getest is op een dataset die praktijkdata zo goed mogelijk nabootste. Het was lastig om de gevonden resultaten beschreven in *Hoofdstuk 4* te vergelijken met voorgaande onderzoeksresultaten omdat deze allemaal verschillen in de gebruikte gouden standaard, in de gebruikte data, en in de gebruikte tijdsintervallen waarin een model een attentie voor KM zou moeten afgeven.

Om aan de vierde voorwaarde voor een KM detectiemodel te voldoen – het verstrekken van informatie over de bacterie die de KM veroorzaakt – wordt in *Hoofdstuk 5* een studie beschreven over het ontwikkelen van een beslisboom die een voorspelling doet van de Gramstatus van de KM veroorzakende bacterie. Wanneer een boer de keuze moet maken om te starten met een antibiotica behandeling, kan het zinvol zijn om vooraf de Gramstatus van de KM veroorzakende bacterie te weten. Koeien met een Gramnegatieve KM hebben misschien meer baat bij een ondersteunende behandeling (bijvoorbeeld het toedienen van pijnstillers of vocht) waardoor de koe zich beter gaat voelen dan dat ze een antibioticabehandeling krijgt om de bacterie te doden. Koeien echter met een Grampositieve KM zijn doorgaans wel beter te behandelen met antibiotica. Voor het ontwikkelen en valideren van een Gramstatus voorspellende beslisboom zijn 140 KM gevallen met sensorinformatie (elektrische geleidbaarheid, kleur en melkopbrengst) en met bacteriologische informatie gebruikt; 96 KM gevallen werden gebruikt voor het ontwikkelen van een beslisboom (training set) waarvan 21 Gramnegatief waren. Voor het testen werden 44 KM gebruikt (test set) waarvan negen Gramnegatief waren. Resultaten op basis van de training set lieten zien dat sensorinformatie gebruikt kan worden om de Gramstatus van een KM veroorzakende bacterie te voorspellen: de ontwikkelde beslisboom had een nauwkeurigheid van 90,6% wat wil zeggen dat 90,6% van alle KM gevallen een juiste voorspelling krijgt van de Gramstatus van bacterie die de KM veroorzaakt. Van alle 96 kwartiermelkingen in de training set werd voor 74% daarvan door de beslisboom een hoge kans voorspeld om geïnfecteerd te zijn met een Grampositieve of een Gramnegatieve bacterie. Wanneer alleen die gevallen geselecteerd werden, vertoonde de beslisboom een nauwkeurigheid van 97,2%. De beslisboom was helaas geen robuust model; de nauwkeurigheid van de beslisboom

daalde naar 54,4% wanneer de beslisboom getest werd op de data in de test set. Met de sensorinformatie bleek het niet mogelijk te zijn om de KM veroorzakende bacterie zelf te voorspellen; deze beslisboom had een nauwkeurigheid van 52% op basis van de data in de training set. De belangrijkste conclusie van de studie beschreven in *Hoofdstuk 5* was dat er een sterke indicatie is dat sensor data van AMS gebruikt kunnen worden om een onderscheid te maken tussen Grampositieve en Gramnegatieve KM veroorzakende bacteriën.

Het toevoegen van andere sensorinformatie aan al bestaande detectiemodellen werd als tweede mogelijkheid onderzocht om de detectie van KM te verbeteren. *Hoofdstuk 6* van dit proefschrift beschrijft de mogelijk toegevoegde waarde van het 'on-line' meten van het koemelk celgetal voor de detectie van KM. Data werden verzameld op een Nieuw-Zeelands proefbedrijf voor AMS (the Greenfield Project, Hamilton, Nieuw-Zeeland). Drie detectiemodellen werden ontwikkeld en onderling vergeleken: het eerste model maakte alleen gebruik van de elektrische geleidbaarheid als voorspeller voor uiergezondheid, het tweede model van het 'on-line' gemeten koemelk celgetal, en het derde model combineerde deze twee sensorinformatie bronnen met behulp van een 'fuzzy logic' model. De sensitiviteit van alle drie de modellen werd vastgezet op 80%. Het aantal vals positieve attenties per 1.000 koemelkingen was vergelijkbaar tussen de eerste twee modellen. Wanneer de sensorinformatie echter werd gecombineerd, was er een twee- tot drievoudige daling te zien in het aantal vals positieve attenties per 1.000 koemelkingen. De conclusie van deze studie was dat het detectievermogen verbeterde wanneer koemelk celgetal sensorinformatie toegevoegd werd aan een KM detectiemodel gebaseerd op sensorinformatie van de elektrische geleidbaarheid.

De derde onderzochte mogelijkheid voor het verbeteren van de detectie van KM was het toevoegen van niet-sensor koe-informatie. *Hoofdstuk 7* van dit proefschrift beschrijft een studie waarbij koe-informatie (niet gebaseerd op sensormetingen) wordt toegevoegd aan een aangepaste versie van de basis beslisboom die beschreven is in *Hoofdstuk 4*. Deze aangepaste beslisboom gebruikt sensorinformatie om een kans voor KM op koemelkings niveau te geven. Deze kans op KM is vervolgens aangepast door niet-sensor koe-informatie toe te voegen (pariteit, maand van lactatie, seizoen, celgetal geschiedenis en KM geschiedenis). In tegenstelling tot wat verwacht werd, bleek het detectievermogen niet te verbeteren wanneer niet-sensor koe informatie werd toegevoegd aan een detectiemodel gebaseerd op sensorgegevens. Het detectiemodel gebaseerd op sensorinformatie en niet-sensor koe-informatie had zelfs een slechter detectievermogen van KM dan het detectiemodel dat alleen sensorinformatie

gebruikte. De conclusie van deze studie was dat niet-sensor koe-informatie geen toegevoegde waarde heeft voor de detectie van KM bij AMS.

In de algemene discussie (*Hoofdstuk* 8) worden de specifieke eisen aan een automatisch KM detectiemodel (eisen aan het detectievermogen, het gebruiken van beperkte tijdsintervallen, het gebruiken van velddata, en het geven van informatie van de bacterie die de KM veroorzaakt) als ook de drie onderzochte manieren (gebruik van andere modellen, toevoegen van nieuwe sensorinformatie, en toevoegen van niet-sensor informatie) om de detectie van KM te verbeteren bij AMS bij elkaar gevoegd en bediscussieerd. *Hoofdstuk 8* bediscussieert daarnaast ook dat het voor de toekomst van het ontwikkelen van een KM detectiemodel essentieel is om meer inzicht te krijgen in wat melkveehouders van een detectiemodel verwachten en waarvoor ze precies een attentie willen krijgen. Daarnaast beschrijft het hoofdstuk het belang van internationale afspraken over de gouden standaard, over de toegepaste tijdsintervallen, en over eisen aan het detectievermogen om in de toekomst KM detectiemodellen beter onderling te kunnen vergelijken.

Dankwoord (Acknowledgement)

In eerste instantie vond ik het maar een raar idee…het dankwoord in het Nederlands terwijl de rest van het proefschrift in het Engels is geschreven. Een dappere poging volgde om deze vreemde traditie te doorbreken. Helaas… deze pogingen belandden allemaal in de digitale prullenbak. Uiteindelijk bleek het toch alleen mogelijk om mijn dank naar verschillende mensen uit te spreken (schrijven?) wanneer ik dit in het Nederlands deed. En zelfs dat bleek nog een ware beproeving omdat er zoveel mensen zijn geweest die een (bewuste dan wel onbewuste) bijdrage hebben geleverd om mijn promotietraject te laten slagen. Daarom, in het Nederlands en in willekeurige volgorde:

DANKJEWEL…

Henk Hogeveen, Diederik Pietersma en Erwin Mollenhorst voor jullie dagelijkse begeleiding. Henk, bedankt voor je vertrouwen in mij. Je stimulerende en motiverende woorden tijdens de 2-wekelijkse besprekingen hebben mij vaak verder geholpen. Je schriftelijke kanttekeningen waren, zeker wanneer ik die eindelijk had ontcijferd, altijd nuttig. Ook vele malen dank voor je steun en begrip als ik aan je deur klopte met de onheilspellende woorden "Heb je even tijd voor me…?". Dank dat je mij deze kans hebt aangeboden en dat je mij kennis hebt laten maken met de wondere wereld van mastitis en wetenschap. Diederik, je begeleiding in het eerste jaar kan niet anders dan inspirerend en motiverend omschreven worden. Je enthousiasme en positivisme zijn een groot voorbeeld voor mij geweest, en ze maakten het leven in het bezemhok een stuk gezelliger. Helaas heb je na één jaar afscheid moeten nemen van jouw project. Hopelijk heb ik je niet teleurgesteld in de afronding daarvan. Erwin, waardig opvolger van Diederik. Ook jij bent een voorbeeld voor mij geweest. Ik heb het zeer gewaardeerd dat je altijd tijd vrijmaakte wanneer ik je kantoor binnenliep om weer eens vragen te stellen over SAS, om beslisbomen te bespreken, of om allerlei stukken door te lezen. Het gepuzzel om waarnemingen aan melkingen te koppelen waren niet alleen leerzame momenten, maar ook goede team-building momenten. Ik hoop dat ik iets van je rust en je geduld over mag nemen.

Mijn paranimfen Wilma en Kirsten. Wilma, mijn partner-in-crime in dit promotietraject, dank voor je steun, voor je heldere en duidelijke inzichten, en voor die overvloed aan geduld wanneer ik weer eens daarin tekort schoot. Dank ook voor het beschikbaar stellen van je slaapkamer, voor het lekker in de rondte kletsen, voor het geduldig aanhoren van mijn zwangerschapsperikelen, en voor een geweldig avontuur in Nieuw-Zeeland. Kortom, dank omdat je het 4 jaar met mij volgehouden hebt. Kirsten, bedankt voor de vele lachsalvo's en de heerlijke nachten in Zeewolde (eindelijk bijslapen!). Wat heb ik een bewondering voor je talent om met deadlines te werken, hoe meer hoe beter lijkt wel jouw motto daarin.

Bedankt voor de vele motiverende gesprekken. Je hebt me vaak het gevoel gegeven dat ik meer kan dan ik in eerste instantie zelf denk. Ik ben blij zulke paranimfen achter mij te hebben staan.

Hans Heesterbeek voor de fijne samenwerking. De voortgangsgesprekken waren altijd bijzonder ontspannend, en de samenwerking om het vierde artikel tot stand te brengen heb ik als zeer prettig ervaren. Je opmerkingen, suggesties en humor zijn zaken die ik zeer gewaardeerd heb. Ook bedankt voor al je hulp met mijn laatste loodjes om het proefschrift af te ronden.

Rik van der Tol, Martin Wiedemann en Ad Feelders voor jullie bijdrage aan het eerste, dan wel het tweede artikel. Rik, ook bedankt voor de altijd weer gezellige, maar soms frustrerende, bezoekjes aan de veehouders om de logcomputers aan te sluiten en draaiende te houden. De vergadering die we hebben gehad waren altijd inspirerend en gezellig.

Jenny Jago, Rob Sherlock, Graeme Mein, and Kevin and Barbara Bright for making the NZ adventure one of my highlights of the past 4.5 years. Jenny, many thanks for letting me feel at home during my stay. Your enthusiasm and seemingly endless amount of energy are two things I can only dream about. Many thanks for offering me the opportunity to start a whole new adventure in Hamilton. I'm truly honored. Graeme and Rob, thanks for the collaboration during the CellSense project. I enjoyed the discussions we had in NZ and afterwards. Your comments on the endless versions of the manuscript (wooly, meaty, fuzzy?) are greatly appreciated. It was great to have a chat during the IDF in Christchurch and the Italian dinner. Many thanks also to Kevin and Barbara Bright. I had so much fun with the two of you during my stay in NZ in 2007, and again in 2010 in the Abel Tasman Reserve. I consider myself a fortunate girl that had the chance to meet you. Hope this friendship will continue for a long time.

Mede (mastitis) AiO's en collega's van Landbouwhuisdieren. Collega's dragen zoveel bij aan een fijne werkdag en jullie hebben dat super gedaan. Marian en Wietske, voor jullie nog een extra bedankje. Marian voor de vele zen- en zelfreflectie momenten, hopelijk volgen er nog vele. Ik ben je er zeer dankbaar voor onze bijzondere vriendschap. Wietske, jammer dat we niet langer collega's zullen zijn; had mij prima kunnen passen. Bedankt voor een super gezellige tijd op kamer 1.05.

Alle deelnemende veehouders voor hun inzet om de data verzameling rond te krijgen. Dank voor de kopjes koffie, de lekkere tomatensoep, de suggesties die gedaan zijn tijdens de bezoekjes, en voor de tomeloze inzet alle voor ons belangrijke zaken te noteren. Zonder jullie geen data en dus geen promotietraject.

NS, voor 4 jaar lang (bijna) probleemloos treinen.

Jos en Agnes Morsink voor het leggen van de basis, en voor het bieden van een tweede 'thuis'. Als ik niet naar de deel was gelopen om Jos te zoeken, had ik hier

nu niet gestaan (maar toeval bestaat niet, toch?). Agnes, bedankt voor je steun, altijd en overal, op wat voor vlak dan ook. Ik zal onze gesprekken missen.

De Wageningen-Meiden-Club voor alle heerlijke gierende en oorverdovende kippenhok momenten. Wat heb ik gelachen, buikpijn! Onze afspraakjes hebben me gestimuleerd vol te houden, en het waren perfecte gelegenheden om mijn ei kwijt te raken. Dappere, sterke en lieve Leonie, een speciaal bedankje voor jou, omdat je me hebt laten zien dat je dromen wel kunt waarmaken.

Alle vriendjes en vriendinnetjes, bedankt dat ik bij jullie mijn werk los kon laten. Lekkere wijntjes, gezellige thee en koffie momenten, lunches, en gewoon eventjes over iets anders kletsen dan mastitis en koeien. Agnes, inderdaad, je moet durven dromen en die dromen ook achterna gaan. Suus, Girl-Power-Girl, bedankt voor het laten zien van zoveel kracht. Jacq, bedankt voor altijd een luisterend oor en voor het gezamenlijk in het zweet fietsen. Mayo, ik heb alle dagen gelachen, bedankt voor dit advies. En Chris, bedankt voor je grandioze raad…helaas had Mr. Shell geen tijd voor mij.

Papa, Mama, Zussezusse en Broer. Bedankt voor al jullie liefde en vele steun. Klinkt als een cliché, maar het is echt waar: zonder jullie had ik hier niet gestaan. Pap, ja met koeien kun je dus je brood verdienen. Mama, de bezige bij, jouw woorden 'Als je maar je best doet' hebben mij gebracht waar ik nu sta. Nathalie en Marcus, ik kan me geen betere zus en broer wensen, en ik hoop nog lang te mogen steunen op jullie wanneer ik dat nodig heb. Pfff, wat zal ik jullie 4-en missen de komende twee jaar.

Beatrix en Corry, bedankt dat jullie het nooit vervelend vonden als ik weer eens verlaat thuis kwam van mijn werk. Beatrix, super dat je altijd een dag (of twee……of drie) extra kon oppassen en voor je bijdrage aan de voorkant van dit proefschrift.

Nynke en Hugo, mijn kleine wondertjes. Jullie hebben mijn promotie traject er niet altijd gemakkelijker op gemaakt, maar wel oneindig veel mooier (en efficiënter).

Willem. Ze zeggen wel eens dat de mensen die het dichts bij je staan het meest moeten ontgelden. Een waarheid als een koe. Ondanks dat we geen uitgebreide gesprekken hebben gehad over wat ik 'nu eigenlijk aan het doen ben', heb je zeker bijgedragen aan het slagen ervan. Bedankt voor het simpele feit dat je er was, dat je mijn keuze om dit te starten altijd gesteund en geaccepteerd hebt, en dat je de zorg voor Nynke en Hugo uit handen nam als ik dat aangaf. Mijn enorme dank voor je reactie ('Wat doen we dan nog hier?') op het Nieuw-Zeeland verhaal. Jij staat duidelijk 200% achter dit avontuur en deze mogelijkheid, en jouw steun betekent ontzettend veel voor mij. Ik zie dan ook enorm uit naar dit avontuur overzee. Onze droom die uitkomt en die we met ons 4-en gaan beleven. Ik kan me geen betere reisgenoten wensen.

About the author

Curriculum vitae

Claudia Kamphuis was born on April the 17th 1979 in Oldenzaal. After attending the Thijcollege High school in Oldenzaal, she started her bachelor study "Diergezondheidszorg" in 1998 at Hogeschool Larenstein, Deventer. After finishing that study in 2002, she moved to Wageningen to study 'Animal Sciences – Preventive Animal Health and Welfare' at Wageningen University. Her minor thesis considered the estimation of heritabilities of claw disorders and their correlation with conformation traits at the department of Animal Breeding and Genetics at Wageningen University. For her major thesis she participated at a project developing a new management tool for dairy farmers to monitor and manage udder health at their farm. This was done as collaboration project between the department of Quantitative Veterinary Epidemiology, Wageningen University, and the NRS (now CRV) in Arnhem, the Netherlands. After graduating in 2004, she started as junior researcher at Wageningen University, division Human Nutrition and Epidemiology. Here, she worked at an international meta-analysis project of the World Cancer Research Fund. In June 2006, she started as a PhD-student at the department of Farm Animal Health of the Faculty of Veterinary Medicine of Utrecht University. That project involved the improvement of clinical mastitis detection using sensor data from automatic milking systems. As part of that PhD-project, she went to DairyNZ Ltd, Hamilton, New Zealand to collaborate on a project regarding the additional value of a new on-line sensor (measuring cow somatic cell count) in the detection of clinical mastitis. By the end of 2010, she will finish her PhD-project at Utrecht University, to start as scientist at DairyNZ Ltd.

Publication list

Scientific publications from this PhD-project

Kamphuis, C., H. Mollenhorst, A. Feelders, D. Pietersma, and H. Hogeveen. 2010. Decision-tree induction to detect clinical mastitis with automatic milking. Computers and Electronics in Agriculture. 70(1):60-68.

Kamphuis, C., H. Mollenhorst, J.A.P. Heesterbeek, and H. Hogeveen. 2010. Detection of clinical mastitis with sensor data from automatic milking systems is improved by using decision-tree induction. Journal of Dairy Science. 93(8):3616-3627.

Kamphuis, C., H. Mollenhorst, and H. Hogeveen. 2010. Sensor measurement patterns revealed; predicting the Gram-status of clinical mastitis causal pathogens. Submitted.

Hogeveen, H., C. Kamphuis, W. Steeneveld, and H. Mollenhorst. 2010. Sensors and clinical mastitis – the quest for the perfect alert. Sensors. 10:7991-8009.

Kamphuis, C., D. Pietersma, R. van der 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(2):169-181.

Kamphuis, C., R. Sherlock, J. Jago, G. Mein, and H. Hogeveen. 2008. Automatic detection of clinical mastitis is improved by in-line monitoring of somatic cell count. Journal of Dairy Science. 91(12):4560-4570.

Other scientific publications

Geelen, A., J.M. Schouten, C. Kamphuis, B.E. Stam, J. Burema, J.M.S. Renkema, E.J. Bakker, P. van 't Veer, and E. Kamphuis. 2007. Fish consumption, n-3 fatty acids, and colorectal cancer: a meta-analysis of prospective cohort studies. American Journal of Epidemiology. 166(10):1116-1125.

Van de Waaij, E.H., M. Holzhauer, E. Ellen, C. Kamphuis, and G. de Jong. 2005. Genetic parameters for claw disorders in Dutch dairy cattle and correlations with conformation traits. Journal of Dairy Science. 88(10):3672-3678.

Conference proceeding papers and seminars

Hogeveen, H., C. Kamphuis, W. Steeneveld, and H. Mollenhorst. 2010. Sensors and milk quality: The quest for the perfect alert. Pages 138-151 in Proceedings of the First North American Conference on Precision dairy management, Toronto, Canada.

Kamphuis, C., H. Mollenhorst, J.A.P. Heesterbeek, and H. Hogeveen. 2010. Data mining to detect clinical mastitis with automatic milking. Pages 568-572 in Proceedings of the 5th IDF International mastitis conference: Mastitis research into practice. J.E. Hillerton (ed.). Christchurch, New Zealand.

Mollenhorst, H., C. Kamphuis, W. Steeneveld, and H. Hogeveen. 2010. Improving automatic detection of abnormal milk. Page 152-153 in Proceedings of the first North American Conference on Precision Dairy Management, Toronto, Canada.

Kamphuis, C., H. Mollenhorst, A. Feelders, and H. Hogeveen. 2009. Decision tree induction as an automated detection tool for clinical mastitis using data from six Dutch dairy herds milking with an automatic milking system Pages 299-306 in Proceedings of the 4th European Conference on Precision Livestock Farming. C. Lokhorst, and P.W.G. Groot Koerkamp (eds.). Wageningen, the Netherlands.

Hogeveen, H., C. Kamphuis, R. Sherlock, J.Jago, and G. Mein. 2009. Inline SCC monitoring improves clinical mastitis detection in an automatic milking system. Pages 315-322 in Proceedings of the 4th European Conference on Precision Livestock Farming. C. Lokhorst, and P.W.G. Groot Koerkamp (eds.). Wageningen, the Netherlands.

Kamphuis, C., H. Mollenhorst, A. Feelders, and H. Hogeveen. 2008. Decision tree induction for detection of clinical mastitis using data from six Dutch dairy herds milking with an automatic milking system. Pages 267-274 in Mastitis Control – From science to practice. T.J.G.M. Lam (ed.). The Hague, the Netherlands.

Kamphuis, C., H. Mollenhorst, A. Feelders, and H. Hogeveen. 2008. Inline somatic cell counting improves clinical mastitis detection in an automatic milking system. Page 296 in Mastitis Control – From science to practice. T.J.G.M. Lam (ed.). The Hague, the Netherlands.

Kamphuis, C., D.Pietersma, and H. Hogeveen. 2007. Detection of abnormal milk and mastitis using sensor measurements of automatic milking machines. Pages

19-21 in Proceedings of the International Workshop Economic Decisions in Farm Animal Health. Foulum, Denmark.

Kamphuis, C., D. Pietersma, R. van der Tol, and H. Hogeveen. 2007. Analysis of sensor data patterns from an automatic milking system for abnormal milk detection. Pages 71-78 in Proceedings of the 3rd European Conference on Precision Livestock Farming. Cox, S. (ed.), Skiathos, Greece.

Book contributions

Contributor for the expert report 'Food, Nutrition, Physical Activity and the Prevention of Cancer: a Global Perspective' of the World Cancer Research Fund (2007).

Kloar!
‘t Book is oet.
Heanig an.