Prioritizing alarms from sensor-based detection models in livestock production - A review on model performance and alarm reducing methods

Final preprint (accepted version) of article published in Computers and Electronics in Agriculture. Please cite as:


DOI: 10.1016/j.compag.2016.12.008

This research was funded by the Danish Council for Strategic Research (The PigiIT project, Grant number 11-116191)
Prioritizing Alarms from Sensor-based Detection Models in Livestock Production -
a review on model performance and alarm reducing methods

K.N. Dominiak¹, A. R. Kristensen²

¹HERD - Centre for Herd-oriented Education, Research and Development, Department of Large Animal Sciences, University of Copenhagen, Grønnegårdsvej 2, 1870 Frederiksberg C, Denmark
²Department of Animal Science, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark

Abstract

The objective of this review was to present, evaluate and discuss methods for reducing false alarms in sensor-based detection models developed for livestock production as described in the scientific literature. Papers included in this review are all peer-reviewed and present sensor-based detection models developed for modern livestock production with the purpose of optimizing animal health or managerial routines. The papers must present a performance for the model, but no criteria were specified for animal species or the condition sought to be detected. 34 papers published during the last 20 years (1995 - 2015) are presented in three groups according to their level of prioritization: “Sheer detection models” based on single-standing methods with or without inclusion of non-sensor-based information (19 papers), “Improved detection models” where the performance of the described models are sought to be improved through the combination of different methods (12 papers) and “Prioritizing models” where the models include a method of ranking or prioritizing alerts in order to reduce the number of false alarms (3 papers). Of the three methods that rank or prioritize alerts; Fuzzy Logic, Naive Bayesian Network (NBN) and Hidden phase-type Markov model, the NBN shows the greatest potential for future reduction of alerts from sensor-based detection models in livestock production. The included detection models are evaluated on three criteria; performance, time-window and similarity to determine whether they are suitable for implementation in modern livestock production herds. No model fulfills all three criteria and only three models meet the performance criterion. Reasons for this could be that both sensor technology and methods for developing the detection models have evolved over time. However, model performance is almost exclusively presented by the binary epidemiological terms Sensitivity (Se) and Specificity (Sp). It is suggested that future research focus on alternative approaches for the output of detection models, such as the prior probability or the risk of a condition occurring. Automatic monitoring and early warning systems offer an opportunity to observe certain aspects of animal health, welfare, and productivity more closely than traditionally accomplished through human observation, and the opportunities for improving animal welfare should continue to be a driving force throughout the field of precision livestock farming.

Keywords: sensor, early warning system, performance, sensitivity, specificity, automatic monitoring, livestock production

1. Introduction

Livestock production has moved from extensive production to intensive production over the last few decades (Sorensen et al., 2010). Society’s demand for high-quality animal products is continuously increasing while the number of farmers producing the products is decreasing (Kashiha et al., 2013; Berckmans, 2014). The natural consequence of this situation is a centralization of the production units with increasing numbers of animals at each site (Sorensen et al., 2010; Kashiha et al., 2013).

This centralization, together with the 2008 financial crisis, has changed the conditions of the whole managerial situation leaving the farmer with fewer personnel and less time for each of the daily management tasks creating an increasing market for technical solutions. Technology in livestock production includes automatic monitoring and management information systems (MIS), which gathers available information, and decision support systems (DSS), which analyses the available information, in order to detect and control the health and welfare status of the animals at any given time, by providing early warnings of potential problems (Sorensen et al., 2010; Kashiha et al., 2013; Berckmans, 2014).

Giving the right alarm at the right time is a crucial property of an early warning system, and too many false alarms represent a recurring challenge throughout the field of building models for early warning systems. The false alarms are time-consuming and diminish the trust in the system which in livestock production might lead to the consequences of farmer or personnel either ignoring the alarms from time to time or making personal prioritization of the alarms based on experience, time expenditure, gut feeling and work enthusiasm. In such cases, both animal welfare and gross margin are at risk of being compromised and in order to optimize the benefit of an early warning system for the farmer, a prioritization of alarms must be made ensuring communication of only the relevant alarms to the farmer.

Prioritization of alarms can be done at two levels; either by a reduction in the number of false alarms produced by the early warning system, or by a prioritization of alarms. A reduction in the number of alarms can be done through a satisfying level of performance of the early warning system, while a prioritization of alarms seek to rank true positive (TP) and false positive (FP) alarms. Ranking can be done according to severity of the condition in focus, for example lameness, from those that need immediate attention to those that can be attended within a given period of time. The ranking can be made according to different overall motivations such as animal welfare, costs or production efficiency.

The aim of this review is to evaluate methods for prioritizing...
sensor-based alarms in livestock production in order to reduce the number of false alarms. The evaluation will be done through a presentation of the different methods described in the scientific literature. Then, the advantages and disadvantages of the methods, for their realistic implementation in commercial livestock production, are discussed.

The studies included in this review are of such a variety in terms of study-designs, conditions in focus, and definitions of case (a condition, which should be detected by the model) vs non-case (a condition, which should not be detected by the model), that a true comparison of methods and results are not possible. Therefore, this review does not focus on one species, one condition, or on one type of sensor. Instead, it strives to elucidate the general development of sensor-based detection models with a focus on the prioritizing methods. The challenging task of expressing biological variation through statistical methods at an implementable level of accuracy is hereby sought illustrated.

2. Conceptual framework

2.1. Sensor-based detection systems

The idea of a sensor-based detection system is to automatically detect a condition based on observations from one or more sensors installed in the pen or the barn. Examples of conditions include oestrus, parturition, diseases or impaired productivity. In most cases, the outcome is binary in the sense that the condition is either present or not present at a certain time. Therefore, this review does not focus on one species, one condition, or on one type of sensor. Instead, it strives to elucidate the general development of sensor-based detection models with a focus on the prioritizing methods. The challenging task of expressing biological variation through statistical methods at an implementable level of accuracy is hereby sought illustrated.

The basic principles behind a detection system can be described as follows: Assume that a sensor system observes the value of a variable $x_t$ at time $t = 1, \ldots, T$. The variable can either be univariate (i.e. a scalar) or multivariate (i.e. a vector). We shall denote as $D_t$ the set of all observations until time $t$, i.e. $D_t = \{x_1, \ldots, x_t\}$.

The detection system will typically provide some kind of summary statistic $s_t = f(D_t)$ based on the available information until now. The function $f$ can be very simple, for example $f(D_t) = x_t$ (returning the most recent observation) or $f(D_t) = (x_{t-n+1} + \ldots + x_t)/n$ (returning the average value of the $n$ most recent observations). However, the $f$ function can also be derived through more sophisticated advanced methods like Kalman filtering, neural networks or other computer intensive methods.

The detection is (either literally or conceptually) based on the comparison of the summary statistic $s_t$ to a predefined threshold $\tau$. An alert is given if the summary statistic $s_t$ exceeds the threshold. Thus, at time $t$, we will either have the event $A^+_t(\tau) : s_t > \tau$ “Alert at time $t$” or the event $A^-_t(\tau) : s_t \leq \tau$ “No alert at time $t$”.

As a very simple example of this framework, assume that we wish to detect a certain disease in an animal. The disease is known to cause fever, so a temperature sensor is attached to the animal. The temperature is logged every hour and transmitted to a computer. In this case $x_1, \ldots, x_t$ are simply hourly temperature measurements. A simple summary statistic would be the current temperature implying that $s_t = f(D_t) = x_t$, but also the average over the last few hours might be relevant.

In order to finish the detection system we need to define a threshold, $\tau$. Assuming that the normal temperature of the animal in question is $\tau_0$, it would be natural to choose a higher threshold $\tau = \tau_0 + \delta$ where $\delta > 0$. It is not straight forward to choose the threshold. It is obvious that if $\delta$ is small, many alerts will be given. It has the advantage, that most of the disease cases will be found (true positives), but on the other hand, we will also have cases where the temperature is above the threshold for other reasons (oestrus, measurement errors or other conditions). In other words, a low threshold will lead to many false positive cases. If, on the other hand, a high threshold is chosen, the number of false positive cases will decrease but on the cost of sometimes not detecting true cases (for instance if they are less severe). Thus, we are at risk of having many false negative cases.

This illustrates the general problem in detection methods, namely that there is a built-in conflict between few false positive and few false negative cases. Methods for measuring the performance of detection systems are therefore needed. The traditional approach has been to characterize a detection method by two conditional probabilities known as the sensitivity and the specificity. For given threshold, $\tau$, the sensitivity, $se_\tau$, and the specificity, $sp_\tau$, are defined as follows

\[
se_\tau = P(A^+_t(\tau)|E^+_t) \quad (1)
\]
\[
sp_\tau = P(A^-_t(\tau)|E^-_t), \quad (2)
\]

where $E^+_t$ and $E^-_t$ are the true presence and absence, respectively, of the condition we try to detect.

It should be noticed that all performance indicators introduced so far are specific for the chosen threshold. Since, in many cases, the threshold can be chosen so that the sensitivity becomes 1 (or close to one) it will be at the cost of a lower specificity. It is, therefore, necessary always to look at both primary performance indicators simultaneously.

In order to estimate an over-all performance indicator (independently of a threshold), the Receiver Operating Characteristic Curve, roc, is often used. The curve is defined by the following parametrization:

\[
roc = \{(fpr(\tau), se(\tau)) : \tau \in R\}, \quad (3)
\]

where $fpr(\tau) = 1 - sp(\tau)$. The over-all performance indicator is the Area Under Curve, auc, determined as

\[
auc = \int_{-\infty}^{\infty} se(\tau)fpr'(\tau)d\tau. \quad (4)
\]

A perfect system will have an auc = 1 so, in general, values close to 1 are preferred.

A study by Aparna et al. (2014) has chosen a completely different approach, where the summary statistic is defined as the expected time to next condition. Thus, if the random variable $\Theta$ is the time to next condition, then

\[
s_t = f(D_t) = E(\Theta|D_t). \quad (5)
\]

Hence, there is no comparison with a chosen threshold. This seems to be a natural approach in cases where the condition will eventually happen (e.g. oestrus or parturition) or will happen with high probability.

An overview of the symbols, concepts and definitions is given in Table 1.
Table 1: Conceptual framework and performance assessment of sensor based detection systems

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Formula/Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_t$</td>
<td>Observation at time $t = 1, \ldots, T$</td>
<td>From sensors</td>
</tr>
<tr>
<td>$D_t$</td>
<td>Set of all observations until now</td>
<td>$D_t = {x_1, \ldots, x_t}$</td>
</tr>
<tr>
<td>$s_t$</td>
<td>Summary statistic at time $t$</td>
<td>$s_t = f(D_t)$</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Threshold at time $t$</td>
<td>Decided</td>
</tr>
<tr>
<td>$E_t^+$</td>
<td>Condition (true) at time $t$</td>
<td>Gold standard</td>
</tr>
<tr>
<td>$E_t^-$</td>
<td>No condition at time $t$</td>
<td>Gold standard</td>
</tr>
<tr>
<td>$A_t^+(\tau)$</td>
<td>Alert at time $t$ with threshold $\tau$</td>
<td>$s_t &gt; \tau$</td>
</tr>
<tr>
<td>$A_t^-(\tau)$</td>
<td>No alert at time $t$ with threshold $\tau$</td>
<td>$s_t \leq \tau$</td>
</tr>
<tr>
<td>$se(\tau)$</td>
<td>True sensitivity with threshold $\tau$</td>
<td>$se(\tau) = P(A_t^+(\tau) \mid \tau)$</td>
</tr>
<tr>
<td>$sp(\tau)$</td>
<td>True specificity with threshold $\tau$</td>
<td>$sp(\tau) = P(A_t^-(\tau) \mid \tau)$</td>
</tr>
<tr>
<td>$er(\tau)$</td>
<td>True error rate with threshold $\tau$</td>
<td>$er(\tau) = P(E_t^- \mid A_t^+(\tau))$</td>
</tr>
<tr>
<td>$fpr(\tau)$</td>
<td>False positive rate with threshold $\tau$</td>
<td>$fpr(\tau) = 1 - sp(\tau)$</td>
</tr>
<tr>
<td>roc</td>
<td>Receiver Operating Curve</td>
<td>$roc = \int_{\tau} \frac{se(\tau)}{sp(\tau)} d\tau$</td>
</tr>
<tr>
<td>auc</td>
<td>Area Under Curve</td>
<td>$auc = \sum_{\tau} \frac{se(\tau)}{sp(\tau)}$</td>
</tr>
<tr>
<td>TP</td>
<td>Number of true positive cases</td>
<td>$TP = \sum_i h(A_i^+(\tau) \cap E_i^+)$</td>
</tr>
<tr>
<td>FP</td>
<td>Number of false positive cases</td>
<td>$FP = \sum_i h(A_i^-(\tau) \cap E_i^-)$</td>
</tr>
<tr>
<td>TN</td>
<td>Number of true negative cases</td>
<td>$TN = \sum_i h(A_i^+(\tau) \cap E_i^-)$</td>
</tr>
<tr>
<td>TN</td>
<td>Number of false negative cases</td>
<td>$FN = \sum_i h(A_i^-(\tau) \cap E_i^+)$</td>
</tr>
<tr>
<td>$Se_\tau$</td>
<td>Estimated sensitivity</td>
<td>$Se_\tau = \frac{TP}{TP + FN}$</td>
</tr>
<tr>
<td>$Sp_\tau$</td>
<td>Estimated specificity</td>
<td>$Sp_\tau = \frac{TN}{TN + FP}$</td>
</tr>
<tr>
<td>$SR_\tau$</td>
<td>Estimated success rate</td>
<td>$SR_\tau = \frac{TP}{TP + FN + FP}$</td>
</tr>
<tr>
<td>$ER_\tau$</td>
<td>Estimated error rate</td>
<td>$ER_\tau = 1 - SR_\tau$</td>
</tr>
<tr>
<td>$FPR_\tau$</td>
<td>Estimated false positive rate</td>
<td>$FPR_\tau = 1 - Sp_\tau$</td>
</tr>
<tr>
<td>$FAR_\tau$</td>
<td>Estimated false alarm rate</td>
<td>$FAR_\tau = FP_\tau / T$</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Curve</td>
<td>$ROC = {fpr(\tau), se(\tau) : \tau \in S}$</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under Curve</td>
<td>Numerical integration</td>
</tr>
</tbody>
</table>

1 $I(\cdot)$ is the indicator function.
2 $S$ is a set of tested values of $\tau$.

2.2. Estimating the performance of detection methods

Even though Eqs. (1) and (2) define the most common performance indicators it is, in most cases, not possible to calculate them analytically. Instead, they must be estimated from data. A necessary condition is that a gold standard allowing us to know the true state of the system (whether the condition is present or absent) is available. The gold standard is seen as a perfect reference standard.

Given a detection system as described in the previous section, a time series of observations $x_1, \ldots, x_T$ and a gold standard, the detection system can be run with a given threshold, $\tau$, for $t = 1, \ldots, T$. This will result in a time series of events drawn from the following four different combinations of detection result ($A_t^+(\tau)$ or $A_t^-(\tau)$) and true state ($E_t^+$ or $E_t^-$):

- **True positive**: $A_t^+(\tau) \cap E_t^+$
- **False positive**: $A_t^+(\tau) \cap E_t^-$
- **True negative**: $A_t^-(\tau) \cap E_t^-$
- **False negative**: $A_t^-(\tau) \cap E_t^+$

The next step in measuring the performance is to count the number of occurrences of each of the four event combinations. Denoting the numbers as $TP_\tau$, $FP_\tau$, $TN_\tau$ and $FN_\tau$, respectively, the estimated sensitivity, $Se_\tau$, and specificity, $Sp_\tau$, are calculated as:

- True negative: $Se_\tau = \frac{TP_\tau}{TP_\tau + FN_\tau}$
- False positive: $Sp_\tau = \frac{TN_\tau}{TN_\tau + FP_\tau}$

Other similar performance indicators like success rate (SR), error rate (ER), false positive rate (FPR) and false alert rate (FAR) are also occasionally estimated (see Table 1 for an overview).

The estimated ROC curve is constructed by choosing a large set $S = \{\tau_1, \tau_2, \ldots, \tau_n\}$ of possible threshold values (where $\tau_1 < \tau_2 < \ldots < \tau_n$). For each $\tau_i \in S$, the sensitivity, $Se_\tau$, and false positive rate $FPR_\tau$, are estimated and plotted as ($FPR_\tau, Se_\tau$) for $i \in \{1, 2, \ldots, N\}$. Finally, the AUC is determined by numerical integration.

2.3. The curse of false positives

In traditional diagnostic tests, focus is often on the level of sensitivity, because the test usually is carried out only once. With only one test result available it is therefore very important that as many true disease cases as possible are detected. In sensor-based detection systems, on the contrary, tests are carried out continuously (or at least regularly). Accordingly, there will be many opportunities to detect a condition so the demands on the sensitivity can be relaxed. Therefore, the true vulnerable point of sensor-based detection systems is the specificity.

Monitoring sensor data from several different data sources has a built-in risk of generating too many false alarms. This can also be the case when only one time series is monitored. The number of false positives may be a problem, even in systems where the specificity of the detection method is very high. This was for instance a problem with an automatic heat detection method (Ostersen et al., 2010) for sows returning to oestrus that had a specificity around 99%. Nevertheless, the error rate (as defined in Table 1) the ratio of false positive out of all alarms) exceeded 95%. This is a natural consequence of sows returning to oestrus being a relatively rare condition.

The phenomenon is easily illustrated using the notation of Table 1. Assume that the condition being detected occurs with probability $p$ at an arbitrary time $t$ (i.e. the prevalence is $p$). Then, $P(E_t^+ | E_t^-) = p$. The error rate is the conditional probability $P(E_t^- | A_t^+(\tau))$. According to Bayes’ Theorem, we have:

- True positive: $A_t^+(\tau) \cap E_t^+$
- False positive: $A_t^+(\tau) \cap E_t^-$

\[
\begin{align*}
\text{error}(\tau) & \quad = \quad P(E_t^- | A_t^+(\tau)) \\
& \quad = \quad \frac{P(A_t^+(\tau) | E_t^-) P(E_t^-)}{P(A_t^+(\tau) | E_t^-) P(E_t^-) + P(A_t^-(\tau) | E_t^-) P(E_t^-)} \\
& \quad \quad = \quad \frac{(1 - sp(\tau))(1 - p) + se(\tau) p}{(1 - sp(\tau))(1 - p) + se(\tau) p}.
\end{align*}
\]

Figure 1 illustrates the error rate as a function of prevalence for three values of specificity and with a sensitivity of 0.8. As is the case in the figure, even a specificity of 0.99 and a prevalence of 0.01 leads to an error rate above 0.5. In other words, more than half of the alarms will be false positive. If only a specificity of 0.9 is assumed, the error rate will be more than 0.9 with a prevalence of 0.01. When time series from different data sources are...
monitored simultaneously there is an even higher risk of false alarms.

Only some of the raw alarms will, therefore, require intervention, and it is therefore important to have methods for prioritizing alarms in order to reduce the number of false alarms.

3. Criteria for implementation

According to Hogeveen et al. (2010) three criteria must be fulfilled for a detection model to be implemented in commercial livestock production. These are A) a high performance in terms of sensitivity (Se) and specificity (Sp), B) a time window corresponding to the necessary response time for the specific condition, and C) a high degree of similarity between the study design and the real everyday conditions in commercial farms. The level of value created by the warning system, relative to the investments needed by the farmer for sensors or equipment, could be added as fourth criteria - but first and foremost models fulfilling the three basic criteria must be developed. Throughout this review, the performance criteria is generally given the highest influence when considering the implementability of a model. If the performance is too poor, neither time window nor similarity will be considered further. Should the performance level fulfill the minimum demands (as described in subsections “Performance considerations” and “Performance - minimum requirements”), the lengths of time windows and the criteria of similarity will be considered according to relevance in the given article.

3.1. Performance considerations

The nature of the condition to be detected must be taken into consideration when defining the level of satisfying performance. So must the costs and consequences of false alarms in monetary, welfare and production efficiency terms. The performance needed for detecting conditions like oestrus or clinical mastitis (CM), which both need immediate response, is fairly high (Rasmussen, 2002; Ostersen et al., 2010) whereas the demands for detecting conditions like lameness, or impaired daily gain, are considered to be lower, hence reflecting a less urgent condition highly influences the requirements to the performance. A high Se is desirable when identifying a condition with high prevalence, while a high Sp is necessary when a condition with low prevalence - like CM or oestrus - is sought to be detected (Rasmussen, 2002).

Although the epidemiological terms of Se and Sp are traditionally used for expressing the performance of a detection model, Friggens et al. (2007) state that Se and Sp are of limited value when it comes to monitoring continuous conditions, time series, and progressive scales of conditions. These limitations and the risk-based alternatives to Se and Sp will be discussed in Section 5.3.

Some authors (Firk et al., 2002; Sherlock et al., 2008; Claycomb et al., 2009) have preferred to describe the performance of detection models by SR, ER (Firk et al., 2002), FAR (Sherlock et al., 2008) or FPR (Viazziet al., 2013). SR is defined as the proportion of true alarms out of all alarms (cf. Table 1) and provides as such an easily interpretable expression of how often the model is right when giving an alarm. Likewise, the ER is the proportion of false positive out of all given alarms. Thus, both SR and ER relate to the number of given alarms, but do not give any information on whether the detection model identifies all cases or has an acceptable level of false negative observations.

FAR, on the other hand, is defined as the proportion of false positive out of the total number of observations. This indicator is used by Hogeveen et al. (2010) and Viazziet al. (2013). Sherlock et al. (2008) suggest that FAR is expressed as the proportion of false positive out of a given, predefined number of observations - for instance 1000 milkings when detecting CM. Communicating to the end user of the alarm system, how many times out of 1000 milkings one must expect a false positive alarm, is easily done, and this interpretation is used by Kamphuis et al. (2008b); Claycomb et al. (2009).

3.2. The missing gold standard

Throughout the literature, the definitions of case vs non-case are individually set for each study dependent on the study design. In defining a case of CM such different definitions as “presence of clinical signs like clots in the milk or swollen quarters” (de Mol et al., 1997, 1999), “Somatic Cell Count (SCC) above 100,000 cells/ml or treatment performed” (Cavero et al., 2007) and “one or more alerts given in a defined period around the recorded date of an observed case” (de Mol and Ouweltjes, 2001) illustrate that there is no reference to a generally accepted definition for automated detection of CM, since none currently exists. Mein and Rasmussen (2008) suggest a less stringent definition of a TP case (CM detection) than the one defined in the International Standard (ISO 20966, 2007 in Mein and Rasmussen, 2008, Annex C). This is done as an attempt to agree on a general definition that both maintains the robustness of the gold standard, is practically assessable, and is strengthening the statistics for calculation of the performance of a detection model. The suggestion has not led to a consensus on the matter.

Visual scoring the degree of lameness on a lameness score scale (LS) is widely used as a detection tool. These scorings are often considered the gold standard, although it is a highly subjective method, where the reliability of the scoring result is positively correlated with the experience of the observer (Tello et al., 2011). More than 20 different types of lameness score scales, both discrete and continuous, exist (Tello et al., 2011).
Often scorings on a four- or five-point scale are reduced to a three-point trait (Garcia et al., 2014) or even to a binary (Alsaaed et al., 2012), which illustrates the difficulties of ranking lameness in detailed degrees using this method.

Since the terms FP, FN, TP and TN are based on the ability of the detection models to recognize a case or a non-case, it can be argued that with no consensus in the case/non-case definitions, a direct comparison of performance measures is like comparing apples to oranges. This review, however, illustrates the difficulties in obtaining implementable results regardless of the choices of species, conditions and underlying definitions. In Tables 3, 4 and 5, all inputs are reported with the same terminology as is used in the respective publication when listing methods, variables and performances.

3.3. Performance - minimum requirements

For detection of CM, two minimum requirements for sensitivity are defined in the literature whereas there is only one defined minimum requirement for specificity. In the International Standard (ISO 20966, 2007 in Mein and Rasmussen, 2008, Annex C) a target value for sensitivity is suggested to be 70%, and the target specificity to be above 99%, before a cow is registered on a mastitis attention list. Rasmussen (2002), on the other hand, defines the minimum requirements for sensitivity as 80% and 99% for the specificity, as it is done in Annex C. Since the main reason for building sensor-based detection models is to provide a foundation for better decision support than what human experts can give (Quimby et al., 2001; Kristensen et al., 2010), and since the highest obtained accuracy by human observation is found to be 80% (Rasmussen, 2005), the higher of the two minimum demands to sensitivity is well supported. There is no consensus in the choice of minimum requirement to a threshold though, and both the definitions by Rasmussen (2002) (Hoegevan et al., 2010; Kamphuis et al., 2010b; Huybrechts et al., 2014) and those of Annex C (Kramer et al., 2009; Steeneveld et al., 2010a; Miekley et al., 2012) are used in publications on CM detection.

No standard requirements for performance in lameness detection - or detection of the onset of farrowing - are found in the literature. The performance requirements defined for CM detection will therefor be generally applied when discussing these models.

Some studies define diseases in disease blocks defined as uninterrupted sequence of “days in disease” in association with a detection of the condition (Miekley et al., 2013a). The performance is then expressed in block specificity and block sensitivity (Kramer et al., 2009; Cavero et al., 2007; Miekley et al., 2012, 2013a). Disease blocks can be defined similar to time windows (Kramer et al., 2009), which will be discussed in the next subsection. A reason for the use of disease blocks can be to focus on early detection (Kramer et al., 2009; Miekley et al., 2013a) but it is important to notice that by calculating block sensitivity instead of sensitivity at day level or even at case level, information on the number of successive alerts is neglected (de Mol, 2013). This can cause the sensitivity of the models to be higher since the number of observations registered by the model is reduced, and the chance of TP is increased.

3.4. Time window

Time windows define time frames for conditions in livestock production. The gold standard is the true clinical status, and the time window can be defined as the minimum expected length of the true clinical status (Sherlock et al., 2008). The length of a time window can be based on direct or on indirect indicators. Direct indicators are for example SCC (Hojsgaard and Friggens, 2010), laboratory analyzed hormone levels (de Mol et al., 1997) or visual observations by the farmer for CM in cows (de Mol et al., 1997; Kamphuis et al., 2010b; Miekley et al., 2013a). Indirect indicators can be back pressure test for oestrus in sows (Cornou et al., 2008), changes in feeding behavior for lameness in cows (Quimby et al., 2001) or animal activity as in Cornou et al. (2008) and Cornou and Lundbye-Christensen (2011). Time windows can overlap if the condition is occurring multiple times with short intervals. In such cases, the time windows can be merged into disease blocks (Cavero et al., 2006; Kramer et al., 2009; Miekley et al., 2012, 2013a), and Se and Sp of disease blocks can be reported as block sensitivity and block specificity (Cavero et al., 2007; Kramer et al., 2009) which might have both advantages and disadvantages as mentioned in the previous subsection.

If a condition is present and an alarm occurs within the time window, the alarm is TP. If an alarm occurs before or after the defined time window, it is considered FP. On the other hand, if a condition is present but no alarm occurs within the defined time window, the situation is FN whereas it is TN if no alarm occurs in the absence of the condition of interest (Figure 2).

The length of the time window has great influence on the performance of the detection model. A very long time window like 17 days for CM (de Mol et al., 1997) heightens the chance of TP and increases the performance of the alarm system because all alarms occurring within the time window are classified as TP. It can be argued that a system that generates an alarm anywhere between 1 and 10 days before to a week after an intervention is needed, might be of little practical use to the farmer. A detection model with a short time window, like 6 hours for oestrus detection (Ostersen et al., 2010), or at the very milking where CM occurs for the first time, an alarm is of great use as a managerial tool (Kamphuis et al., 2010a). The downside of this is that such short time windows increase the demands to the accuracy of the model.

3.5. Similarity between models and real life

Similarity between the study population and commercial livestock production populations is of utmost importance if the
developed detection models should have a chance of performing well under field conditions and later be implemented. Three reasons for dissimilarities between the study population and commercial livestock populations are: a narrow data set that does not depict the variety of commercial farms, indistinct definitions of case (TP) and non-case (TN) in the study design, and the capability of the model to handle missing data (Hogeveen et al., 2010). If these criteria are not fulfilled, the risk of a disappointing level of detection performance in a commercial herd is high (Hogeveen et al., 2010).

A detection model must be validated externally to prove its accuracy under conditions other than the ones it is created under. A high degree of accuracy is reflected in high sensitivity, specificity and reproducibility (Liu et al., 2009).

When evaluating models with promising performances, it is highly relevant to include the validation method to get a more fulfilling picture of the potential for implementation. Financial limitations or different types of deadlines can be reasons for designing the study validation in a way that does not meet the similarity criteria - and it bight be of greater importance to build a model first, and then validate it under conditions less challenging than in herds representative for the average production form in the given area of interest.

The strongest validation is on data, which is completely independent from the data set used for training and learning the model, as for instance data coming from another herd. If it is not possible to obtain suitable data from an independent herd, and if the data set is large enough, validation can be done by dividing the original data set into test data, learning data and validation data (Witten and Frank, 2005). Often the data set is too small for such a division, and other methods must be considered. A commonly accepted validation method is a 10-fold cross validation as used by Viazzi et al. (2013). With this method, the data set is randomly divided into ten subgroups, one subgroup is then retained as validation data, and the model is trained on the remaining nine subgroups. The validation is strongest with this method, when the process is repeated ten times, each time with a new subgroup used for validation data, although this is not always done (Witten and Frank, 2005).

Another validation method, which is used by Liu et al. (2009), is “leave-one-out” cross validation. This method is to some extend similar to the 10-fold cross validation, only it is n-fold, where n is the number of observations/animals in the data set. The validation is performed n times, with each observation left out in turn, and the rest of the data set used as training data (Witten and Frank, 2005). Both cross-validation methods mentioned above are relatively narrow in an implementation aspect due to the high degree of dependency between training and validation data.

Basing a model on data from a few animals (Cornou and Lundbye-Christensen, 2010, 2011; Aparna et al., 2014), animals from a single herd (Bressers et al., 1995; Ostersen et al., 2010; Viazzi et al., 2013; van Hertem et al., 2013, 2014; Garcia et al., 2014) or from herds where the managerial status differs from the average commercial herd, as might be the case in a research herd (de Mol et al., 1997, 2013; de Mol and Ouweltjes, 2001; Caver et al., 2006, 2007; Pastell and Kujala, 2007; Kamphuis et al., 2008a; Kramer et al., 2009; Steeneveld et al., 2010a; Maertens et al., 2011; Miekley et al., 2012, 2013a,b; Kashiha et al., 2014) can have a high impact on the similarity between the study population and commercial herds. This is either because the biological variety of the whole herd is poorly represented in the small study population, or because the routines are adjusted according to demands of the study design. In the case of research herds, the available resources might differ from what is possible in commercial herds.

The similarity of a model is also highly affected in studies where data is collected in herds with extraordinary high/low prevalence of the condition of interest compared to prevalence in average commercial herds (Miekley et al., 2013a; van Hertem et al., 2013, 2014). And the same is true for studies where animals from the same herd are divided into subgroups in order to define learning data and test data (Kramer et al., 2009; Cornou and Lundbye-Christensen, 2010) since this approach ignores any herd-specific correlation, such as genetics or managerial factors.

An obvious reason for not validating the model under field conditions, even though it strengthens the model, is that it can be very time consuming (de Mol et al., 2001; Nielsen et al., 2005). This is often the reason for cross validating on a subgroup of the study population (van Hertem et al., 2014; Viazzi et al., 2013) or using the same data for training and validating the model (Aparna et al., 2014; de Mol and Woldt, 2001; Liu et al., 2009).

The definition of case/non-case can - as previously mentioned - be very individual in some studies (de Mol et al., 1997, 1999; Caver et al., 2006; Miekley et al., 2012; Kamphuis et al., 2013; Garcia et al., 2014) whereas other studies use definitions and routines that are already used by the personnel in the farm where the data is collected (Maatje et al., 1997; de Mol and Ouweltjes, 2001; Kamphuis et al., 2010a,b; Miekley et al., 2013b; van Hertem et al., 2013, 2014; Huybrechts et al., 2014).

Since it is common in commercial production herds that data are missing at a more or less influential level, a detection model must be able to handle missing data as well. In some studies, data sets with missing data are left out during the model developing process for different reasons (Pastell and Kujala, 2007; Steeneveld et al., 2010a; Maertens et al., 2011; de Mol et al., 2013; van Hertem et al., 2013, 2014; Garcia et al., 2014) whereas other models are based on incomplete - but more realistic - data sets from commercial farms (Bressers et al., 1995; Liu et al., 2009; Kamphuis et al., 2010a,b; Miekley et al., 2013b; Huybrechts et al., 2014), hence showing a higher level of similarity.

4. Criteria for inclusion in this review

4.1. Primary criteria

Papers included in this review are all peer-reviewed and present sensor-based detection models developed for modern livestock production with the purpose of optimizing animal health or managerial routines. Papers on models that are based on parameters analysed in laboratories (Barkema et al., 1998; Nielsen et al., 2005; Chagunda et al., 2006; Frijggen et al., 2007; Steeneveld et al., 2008; Hofsgaard and Friggens, 2010), parameters assessed by humans (Barkema et al., 1998; Steeneveld et al., 2010b), or where the condition in focus is artificially applied to the animal as a part of the study design (Mithner et al., 1996; Abell et al., 2014) are therefore not included.

Papers included must furthermore present results from a performance analysis. Papers where methods for detecting, monitoring or assessing parameters for early warning systems are developed, tested or evaluated, but the results are presented as the method having a future potential, are therefore not included. This criterion leaves out several studies (Bressers et al., 1994; Moshou et al., 2001; White et al., 2004; Madsen and Kristensen, 2005; Madsen et al., 2005b; Oliviero et al., 2008; Xi-
4.2. Conditions detected

A variety of conditions are sought to be detected in papers included in this review. Some papers present models detecting more than one condition (de Mol et al., 1997, 1999) or several methods for detecting the same condition (Cavero et al., 2007; van Hertem et al., 2014). Some papers combine two methods in order to improve the overall performance (Kramer et al., 2009; Kamphuis et al., 2010a; Cornou and Kristensen, 2014b; Porto et al., 2014; Mehdizadeh et al., 2015; Dutta et al., 2015).

4.3. Sensor types

Multiple sensor types are included in this review representing the technological evolution through the last two decades (from 1995 to 2015). Data from automatic milking systems (AMS) form the basis for the vast majority of sensor-based detection models, but a variety of other sensor types are included as well. Sensors for monitoring movement include video cameras, different 2D movement sensors (pedometers and neck transponders), and 3D movement sensors (accelerometers and pressure sensitive sensors like force plates and load cells). While other sensor types (flow meters, feeding troughs with sensors, weight scales and climate computers) also provide valuable information in several studies.

4.4. Methods - presentation

The included papers are presented in three groups according to their level of prioritization:

**Group 1 (Table 3):** Sheer detection models based on single-standing univariate or multivariate methods with or without the inclusion of non-sensor-based information.

**Group 2 (Table 4):** Improved detection models where the performance of the described models are sought to be improved through the combination of different methods.

**Group 3 (Table 5):** Prioritizing models where the model includes a method of ranking or prioritizing alerts in order to reduce the number of false alarms.

Some methods (for example, regression models and control charts), are represented in more than one group (Tables 3, 4, and 5), and appear in the specific group based on whether the method is used alone or in combination with other methods.

In many studies, performance indicators are reported several times due to different thresholds or different subgroups of animals. Therefore an approach has been taken in order to compare the highest level of performance obtained by any method under any circumstances given in the relevant study. In the tables, the notations HSe and HSp are used. HSe is the highest sensitivity achieved in the study, and the specificity in brackets is the corresponding specificity. Equivalently, HSp is the highest specificity achieved in the study, and the corresponding sensitivity is shown in brackets.

The notations “HSe $\tau_m (Sp\ y)$” and “HSp $\tau_m (Se\ y)$” are mathematically defined as

\[
(x_m, y) = \left(Se_y, Sp_y, \tau_y \right), \quad \tau_y = \arg\max_{\tau} \{ Sp_y | r \in S \} \quad (10)
\]

\[
(x, y_m) = \left(Se_y, Sp_y, \tau_y \right), \quad \tau_y = \arg\max_{\tau} \{ Sp_y | r \in S \}, \quad (11)
\]

$S$ is the set of thresholds tested in the study.

4.5. Literature search strategy

For the initial search the following keywords were used: automatic monitoring, livestock production, sensors, ranking, prioritizing, alarms and detecting. These keywords were then combined with words like mastitis, lameness, estrus, gain, cow, sow, broiler. From those basic searches, backward searches were done through references and bibliographies of relevant authors. The databases used for the searches were Ovid (CAB Abstracts, Web of Science, Agricola) and Sciedirect in the period from November 2014 to June 2015.

5. Method description

In this section, the methods used for building detection models in the reviewed papers are described according to their level of prioritization. In both groups 1 Sheer detection models (Table 3) and 2 Improved detection models, (Table 4) some papers present a technique where the level of one or both performance parameters are fixed (Kamphuis et al., 2010b, 2013), or defined with a minimum level (Cavero et al., 2006; Kramer et al., 2009; Miekley et al., 2012, 2013a) when doing performance analyses. With a fixed parameter, it is possible to calculate the corresponding threshold for detecting a condition under given circumstances, and hereby reduce the number of false alarms generated by the detection model. However relevant, according to the alarm-reducing characteristics, this technique is not a part of the construction of the detection model and will not be described further.

5.1. Sheer detection models

An overview of the sheer detection models (Group 1) identified for this review is shown in Table 3.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Animal category</th>
<th>Number of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Mastitis (CM)</td>
<td>Cow</td>
<td>17</td>
</tr>
<tr>
<td>Lameness</td>
<td>Cow and sow</td>
<td>14</td>
</tr>
<tr>
<td>Oestrus</td>
<td>Cow and sow</td>
<td>9</td>
</tr>
<tr>
<td>Other diseases</td>
<td>Cow and sow</td>
<td>5</td>
</tr>
<tr>
<td>Parturition</td>
<td>Sow</td>
<td>2</td>
</tr>
<tr>
<td>Activity types</td>
<td>Sow</td>
<td>1</td>
</tr>
<tr>
<td>Weight estimation</td>
<td>Weaned pigs</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: Distribution of conditions covered by presented detection models. Detection of CM and of lameness have had the highest focus overall.

4.5. Literature search strategy

For the initial search the following keywords were used: automatic monitoring, livestock production, sensors, ranking, prioritizing, alarms and detecting. These keywords were then combined with words like mastitis, lameness, estrus, gain, cow, sow, broiler. From those basic searches, backward searches were done through references and bibliographies of relevant authors. The databases used for the searches were Ovid (CAB Abstracts, Web of Science, Agricola) and Sciedirect in the period from November 2014 to June 2015.
et al., 2006; Kamphuis et al., 2008b; Kramer et al., 2009). de Mol et al. (1997, 1999) present a multivariate cow-dependent approach and an AutoRegressive Integrated Moving Average (ARIMA) for analyzing time series with Kalman filter. Later de Mol and Ouweltjes (2001) use an unspecified time series model, where milking intervals and milking frequencies are included as variables. The specificity for CM detection in de Mol et al. (1997) is based on milk sampled with a two month interval, and cows with no CM pathogens or elevated SCC counts in any samples during the study period were defined as TN. This means that a TN cow with one or more alarms was considered FP. This case definition ignores any CM cases that begin and end between two samples, and creates optimal - but unrealistic - conditions for the detection model. The multivariate methods presented by de Mol et al. (1997, 1999) is, however, a novel approach through the incorporation of the animal history and traits, and it is widely implemented in later publications (de Mol and Ouweltjes, 2001; Cavero et al., 2007; Claycomb et al., 2009; Kramer et al., 2009; Kamphuis et al., 2010b; Steeneveld et al., 2010a; Garcia et al., 2014; Huybrechts et al., 2014). General practice at the time of several of these studies was milking in milking parlours, and the use of AMS was in its modest beginning (Kamphuis et al., 2008a; Rutten et al., 2013) which made the inclusion of sensor-based variables limited compared to later studies.

Performances presented in early studies by de Mol et al. (1997, 1999); de Mol and Ouweltjes (2001); de Mol et al. (2001), are fairly high, with either sensitivities or specificities fulfilling the minimum requirements by Rasmussen (2002). The requirements are not met at the same time for both performance measures, though. Not even extremely long time windows, a variety of case definitions, or different techniques for performance analysis in de Mol et al. (1997) led to both parameters meeting the requirements at the same time. Authors agree that the performance of the presented models is too poor for practical implementation and suggest either improvement of both sensors and alert rules (de Mol et al., 1997, 1999) or addition of temperature sensors that have proven informative in detecting CM (de Mol and Ouweltjes, 2001). Simple control charts and local regression were tested and showed to have poor performance in Cavero et al. (2007), and these methods are only used in combination with other methods in later research (Cornou et al., 2008; Lukas et al., 2009; Cornou and Lundbye-Christensen, 2011; Miekley et al., 2012, 2013a; Huybrechts et al., 2014). Even though a model for milk yield based on time series was suggested already by Deluyker et al. (1990), it was on a general cow level, and de Mol et al. (1999) seem to be the first to model cow-specific “normal” behaviour through time series based on sensor data.

A binary classification is bound to misclassify some “grey zone” cows. The use of a lower SCC threshold of 100,000 cells/ml in defining healthy/sick cows as used by Miekley et al. (2012); Cavero et al. (2006) raises another concern of employing an artificially high sensitivity (Claycomb et al., 2009) due to too many healthy cows being classified as sick (false positive). Even though the chosen threshold of 100,000 cells/ml is following the definitions from “Deutsche Veterinärmedizinische Gesellschaft e.V.” for mastitis, it appears to be too low since a number of papers have reported average bulk tank SCC’s from 151,000 cells/ml to >800,000 cells/ml (Maatje et al., 1997; Cavero et al., 2006; Kamphuis et al., 2008b; Claycomb et al., 2009; Kramer et al., 2009; Miekley et al., 2012). Mein and Rasmussen (2008) even suggest that cows could be classified as “true negatives” if the SCC is <200,000 cells/ml and all foremilk samples are without clinical signs.

Fuzzy logic is a method where variables that can obtain multiple linguistic values are determined relative to the connection in which they appear. The linguistic variables can be “many, few, almost all, several”, and they are given a numeric value (degree of membership) between 0 and 1 before they are included in for instance statistical models (Klir, George J. and Folger, Tina A, “Fuzzy Sets, Uncertainty and Information” 1988). When this method is used in models for CM detection (Cavero et al., 2006; Kamphuis et al., 2008b; Kramer et al., 2009), it is applied through three steps of a fuzzy logic system called fuzzification, fuzzy inference and defuzzification:

**Fuzzification** transforms the sensor-measured input variable to a fuzzy value that is a combination of linguistic interpretation and grade of membership (Kamphuis et al., 2009).

**Fuzzy inference** applies a set of IF THEN rules generated on expert knowledge for each trait described by fuzzy values and combines them like IF (all X is Z) AND (no Y is X) THEN (no Y is Z) (Klir and Folger, 1988).

**Defuzzification** transforms the fuzzy values into one numeric value that is compared with a threshold to determine for instance if a cow has got CM or is healthy (Kamphuis et al., 2008b).

The Fuzzy Logic method was first applied by Cavero et al. (2006) who used it on AMS sensor variables. The thresholds for case definitions were very low, which resulted in high performance (in terms of Se and Sp) and large error rates. Fuzzy logic has been used later for detecting CM with both in-line and on-line SCC (Kamphuis et al., 2008b), and for detecting both CM and lameness (Kramer et al., 2009), but no results suitable for implementation in commercial herds were achieved.

The method is good at representing the form of uncertainty that is naturally imbedded in modelling traits with biological variation. By using the so-called Fuzzy Expert System, crisp values can be fuzzified (Klir, George J. and Folger, Tina A, “Fuzzy Sets, Uncertainty and Information” 1988) before applying rules and defuzzification. It would be interesting to see Fuzzy logic applied to categorization of lameness degrees in cows since lameness is a trait with a high degree of biological variation.

Cavero et al. (2006); Kamphuis et al. (2008b); Kramer et al. (2009) all use numeric sensor measurements as input variables though, and the numeric values are first fuzzified to linguistic values then defuzzified back to numeric values. This process does not seem intuitively as the most obvious method. The Fuzzy logic method is used for combining sensor-based alerts with subjective human judging of CM in the study by de Mol and Woldt (2001), and this will be discussed further in Section 5.3.

5.1.2. Sheer detection models based on behaviour and movement sensors

A variety of behaviour and movement sensors are used in detecting changes in the behaviour or movement pattern of an animal. The changes detected are either due to lameness, or the onset of a condition associated with well known behavioural changes like oestrous or farrowing. Numerous studies employ a variety of techniques for assessing activities. These include pressure platforms measuring weight distribution (Pastell and Kujala, 2007; Oliviero et al., 2008; Pastell and Madsen, 2008; Pastell et al., 2008a,b; Pluym et al., 2013; Mohling et al., 2014), pressure sensitive mats monitoring irregularities in gait patterns
(Maertens et al., 2011; Pluk et al., 2012; Nuffel et al., 2013), and accelerometers measuring types of activity in two or three dimensions (Cornou and Lundbye-Christensen, 2010; Cornou et al., 2011; van Hertem et al., 2013; Cornou and Kristensen, 2014b). Activity sensors fastened to the animal (Alsaad et al., 2012; Kamphuis et al., 2013; Miekley et al., 2013b; Dutta et al., 2015) or infrared sensors fastened on inventory (Frenson et al., 1998; Aparna et al., 2014) are also used in multiple studies. Although several types of statistical methods have been used for building sheer detection models based on behaviour or movement sensors, the performance in general follows the same trend as the sheer detection models based on AMS sensors with either a high sensitivity or a high specificity, and with consensus in the finding that multivariate models outperform univariate (Maatje et al., 1997; Kamphuis et al., 2013; van Hertem et al., 2013).

A study by Miekley et al. (2013b) found missing values are causing up to 30% information loss for some cows when using principal component analysis (PCA), whereas (Pastell and Kujala, 2007) found that other methods, like probabilistic neural network (PNN), handle incomplete data sets better. The use of infrared sensors in detecting onset of oestrus in sows is tested and found inadequate for implementation by Freson et al. (1998) since TN and FN could not be distinguished.

Maertens et al. (2011) present an impressive highest accuracy (HSe 90, HSp 100) in detecting lameness amongst dairy cows using a spatiotemporal approach. This accuracy is however only on identification of severely lame cows whereas the overall performance of the model is presented as a success rate above 80% without specification of Se, Sp or FP. The spatiotemporal approach is new in lameness detection of livestock animals though, and this is investigated in further research (Pluk et al., 2012; Nuffel et al., 2013; Meijer et al., 2014).

Some authors discuss improvements by inclusion of walking speed (Meijer et al., 2014) or longer pressure mats to measure more gait cycles within one measurement (Nuffel et al., 2013). The use of sensor mats for lameness detection is still a relatively new area in research, and Pluk et al. (2012) naturally focus more on improving the techniques and choosing the most informative variables and methods instead of on implementation in commercial farms.

The study by Cornou and Lundbye-Christensen (2010) on classifying activity levels of sows prior to farrowing, does not reach the performance defined in Rasmussen (2002), but still the results are remarkable since the corresponding Se and Sp are both 96% as contrary to most other studies that reach either high Se or high Sp. The performance is on identifying a sow in activity (walking, feeding, rooting merged) correctly as opposite to lying down either laterally or sternally. A reliable detection of activity category is valuable in predicting conditions that follow a change in activity level - like oestrus or parturition.

For models built on data from video cameras, infrared cameras or 3D cameras, methods like decision trees or different types of regressions have been used in detecting different conditions. None of the presented models detect with a performance high enough for implementation in commercial farms, and the results by van Hertem et al. (2014) detecting lameness in cows reach neither sensitivities nor specificities matching the definitions in Rasmussen (2002).

Although Viazzi et al. (2013) have simplified the lameness score scale from a 5-point to a 3-point, the performance presented as TruePositiveRate and FalsePositiveRate is too low for implementation. Bressers et al. (1995) only present the success rate and a notice of presence of many false positive in detecting oestrus by monitoring sows’ visits to a boar, hereby indicating a high sensitivity and a low specificity. A similar study was later conducted by Ostersen et al. (2010) with more complex methods that will be presented in Table 5.

5.2. Improved detection models

The models in Table 4 all have in common that methods are combined to create an improvement in model performance. The improvements added are different types of control charts (Cornou et al., 2008; Cornou and Lundbye-Christensen, 2011; Miekley et al., 2012, 2013a; Huybrechts et al., 2014), further development of decision trees (Kamphuis et al., 2010a,b), or of various regression methods (Liu et al., 2009; Kashiha et al., 2014). Combinations of DLM with other methods (Ostersen et al., 2010; de Mol et al., 2013) and partial least squares discriminant analysis fitted by linear regression and improved by reducing the number of variables through backward variable selection (Garcia et al., 2014) are also included.

Combining different types of control charts with wavelet filtering, autoregressive methods, time series, or either univariate or generalized DLMs does not result in a performance high enough for implementation (Cornou et al., 2008; Cornou and Lundbye-Christensen, 2011; Miekley et al., 2012, 2013a; Huybrechts et al., 2014). Using CUSUM in detecting the onset of parturition does however result in both a sensitivity and a specificity of 100% for a subgroup of nine sows based on activity level, and a sensitivity of 100% combined with a specificity of 95% when including all 19 sows in the study (Cornou and Lundbye-Christensen, 2011).

Although the performance obtained by Cornou and Lundbye-Christensen (2011) is impressive, there is an overlap between the individual parameters used in both methods (DGLM and CUSUM) and animal specific reference days which may have increased the performance. This is mentioned by the author as a subject for future change if a large-scale study should be conducted. Few alarms appear at time zero (that is at the actual onset of the farrowing) but the majority of the alarms based on the CUSUM method occur between 12 and 24 hours before onset of farrowing (mean 4.7 - 14.8 hours, SD 4.9 - 9.1 hours), while the DGLM method produces alarms in average 15 hours before farrowing (SD 4.3 - 7.5 hours). An alarm long time before farrowing with a relatively large standard deviation is suboptimal if the purpose is to be present during farrowing. If the purpose, on the other hand, is to prepare the sow or the farrowing crate in order to reduce piglet mortality as in the later discussed study by Aparna et al. (2014), an alarm long time before would in most cases be sufficient.

Different methods for improving DLMs are presented by Ostersen et al. (2010); de Mol et al. (2013). de Mol et al. (2013) use quadratic trend models fitted with DLM to detect lameness in cows. The presented performance of the model is not suitable for implementation, but the authors mention that both discount factors and threshold for the Bayes factor in the DLM can be adjusted. Adjustments can prioritize a higher or lower Se according to the needs of the end user, which means that the threshold for alarms can be adjusted - or prioritized - according to individual needs.

Ostersen et al. (2010) detect oestrus via both duration of a sow’s visit to a boar, the frequency of the visit, and a combination of the two parameters. Ostersen et al. (2010) combine a multiprocess DLM with Markov probabilities of the DLM components in the duration model and develop a DGLM for
the frequency model. The detection model combining both duration and frequency is based on Bayes Theorem and calculates a combined probability of the sow being in oestrus.

The multivariate model surprisingly enough performs worse than the univariate duration model. An explanation for this finding could be that the duration model includes the time distance between the visits which is closely related to the frequency. The results reported in the paper are remarkable due to the extremely high Sp of 99.4%, but remarkably enough the corresponding ER is as high as 93%. This illustrates the almost impossible task of achieving an overall satisfying performance of a detection model when using solely sensor-based data for detection of conditions with very low prevalence.

Kamphuis et al. (2010a,b) present decision trees with different data mining techniques or cost matrices added as improvements for detecting CM. Even though the inclusion of cost matrices in a model designed for decision support is highly relevant, it does not improve the performance enough for implementation.

5.3. Prioritizing methods

As seen in the descriptions of sheer and improved detection models, there is a general problem with fulfilling the described criteria for implementation. Scientific literature describes three overall alternative approaches to this problem; a higher extent of added knowledge, in the form of non-sensor information, to the original detection model (Figure 3 A), an acceptance of the original performance level plus a postprocessing step of prioritization or ranking of the alarms into TP or FP (Figure 3 B), or a presentation of the model output as a time gradient or a risk of case vs. non-case (Figure 3 C) disregarding the source of model input variables. To some extent, a customization of thresholds according to the risk attitude of the farmer can be regarded as a prioritizing measure, but this approach implies that the model is adjusted to the specific herd at time of implementation and possibly multiple times hereafter as the health or managerial status is dynamic and will change.

Fuzzy logic is used by de Mol and Woldt (2001) to combine sensor-based output from earlier developed detection models with additional information about the cow (Figure 3 A) in order to formalize the manager’s reasoning when manually judging alert lists for CM and oestrus. Hereby, they reduce the number of FP on the CM alert list from 1265 to 64 and the number of oestrus alarms by 32%.

The CM model combines the AMS alerts from de Mol and Ouweltjes (2001) with average and variance of sensor measurements, while the oestrus model combines alerts from de Mol et al. (1997) with both qualitative and quantitative non-sensor-based cow information. By combining qualitative and quantitative parameters, de Mol and Woldt (2001) are fulfilling the basic demands of the Fuzzy logic method, but as remarkable as their results are, they must be interpreted with some care, since the Se and Sp respectively are calculated on different divisions of the data.

Naive Bayesian Networks (NBN) as a tool for discriminating between TP and FP alerts from AMS (Figure 3 B) is demonstrated in a study by Steeneveld et al. (2010a) where the number of FP alerts are reduced by 35%. Unfortunately, the model misses 10% of the TP alerts meaning that the specificity is too low for implementation. A satisfying performance level cannot be expected in this study though since the initial performance of the AMS providing the alert list has an Se of 70% and an Sp of 97.8%. The results do show a potential for NBN as a prioritizing method and more research should be done using this method.

A completely different approach for detecting or predicting a condition is used by Aparna et al. (2014). The paper focuses on predicting the exact onset of farrowing, in order to reduce piglet mortality caused by hypothermia. The underlying model is based on Hidden Phase-type Markov methodology, where the time spent in each defined phase of a given condition is modeled. For this study the well-defined behavioural phases preceding a farrowing is used.

The study is based on sows already inserted into farrowing section which makes the probability of the sow actually farrowing very high - almost definitely known to happen - unlike any other condition included in this review. Well-defined behavioural phases are known for a few conditions like parturition and to some extent oestrus in both sows and cows.

This phase-based method is, however, difficult to apply on conditions like CM or tail biting, where the chronological success ision of phases is unknown, and different phase-patterns can lead to the same condition. Also a crucial difference between predicting the onset of farrowing and predicting events of CM and tail biting, is not known beforehand, whether the condition will occur at all or not.

Interestingly enough, Aparna et al. (2014) do not operate with the traditional performance parameters (sensitivity, specificity and error rate) but produce estimates of time to occurrence of farrowing (Figure 3 C) hereby providing decision support to the farmer in choosing which sow to attend to first. By combining water and activity sensors, the model produced 97% true positive predictions with a mean of 11.5 hours and an SD of 4.6 hours. These results fulfill the aim of the paper to provide sufficient warning time for preparing the crate and sow for farrowing, but the SD is too long to provide accurate alarms for the exact onset of the farrowing with the purpose of providing timely aid to complications.
6. Method discussion

The previous description of sheer, improved and prioritizing detection models illustrates a trend in the development of detection models over the last two decades (1995-2015). This trend is not depicting a straight-forward evolution of detection methods but rather a correlated evolution in both model complexity and general evolution of sensor technology.

6.1. Evolutionary trends of methods and sensors

The evolution of sensor-based detection models is facilitated both by the technological evolution causing lower market prices and smaller, more precise devices in general, and by the joint scientific experiences made through peer-reviewed studies and research. In that sense, the evolution of sensor-based detection models has generally moved from univariate models on general species level (Deluyker et al., 1990) or comparing data with a simple threshold (Bressers et al., 1995) through improving detection accuracy by including non-sensor-based animal-specific information like “day of treatment” (Cavero et al., 2007), “calving dates”, or “days in lactation” (de Mol et al., 2001). Parallel to including non-sensor-based information, more multivariate models were developed (Cavero et al., 2006; Kamphuis et al., 2008b; Kramer et al., 2009).

With performance still not reaching a satisfying level, researchers have continued to develop models focusing on prioritizing the generated alarms through the use of Fuzzy logic (de Mol and Woldt, 2001), Naïve Bayesian Network (Steeneveld et al., 2010a) or variations of DLM (Cornou et al., 2008; Ostersen et al., 2010; de Mol et al., 2013). During the same period in time (1995-2015), the technical evolution of sensors has made it possible for the precision in CM detection to move from udder level to quarter level. Also data is available much faster, going from on-line monthly or weekly pooled reference data like SCC, to in-line sensors (Kamphuis et al., 2008b) providing the possibility of detecting a CM case during the actual milking.

A similar evolution of both sensor types and method complexity is also found in models detecting conditions like lameness in cows and oestrus in both sows and cows, but since the history of automatic detection is shorter for these conditions compared to CM, the evolutionary changes are not as profound.

Rajkondawar et al. (2002) developed a fully automatic detection model using limb-specific kinetic measures, and later several studies were based on partly automatic measures, using manual gait score as gold standard (Pastell and Kujala, 2007; Maertens et al., 2011). It has, however, been the development of force load cells (Liu et al., 2009) and pressure sensitive mats (Maertens et al., 2011), which has made a huge difference in lameness detection for cows. The former of the two sensor types, has even been used in the development of a commercially implemented product for lameness detection, which Liu et al. (2009) sought to make more accurate in their study.

6.2. The perfect performance - does it exist?

Despite the technological evolution and the increased complexity of methods in sensor-based detection models, the accuracies of these models are generally at a level that does not fulfill the criteria of implementation (Hogeveen et al., 2010). A great variation in model performance throughout the different studies is revealed when the performance is visualised. The performance of HSp and the corresponding Se is shown in Figure 4 for the papers that present both Se and Sp. Nine studies reach a Se above 80% and ten reach a HSp above 99% but only three papers (Liu et al., 2009; Maertens et al., 2011; Cornou and Lundbye-Christensen, 2011) present models that fulfill both performance criteria when including subgroups of the data sets.

Liu et al. (2009) detect lameness in cows and use logistic regression in combination with B-spline transformation to obtain Se of 100 % and Sp of 100% when detecting lameness on forelimbs, and Se of 99.5% and Sp of 100% when detecting lameness at cow-level. The authors convert a five-point lameness score to a binary (sound-lame) and furthermore validate the model by the leave-one-out cross validation method on a data set consisting of 261 cows. It is reasonable to assume that leaving out the information provided by only one cow for validation, using the remaining 260 cows to train the model a total of 261 times, is close to learning and testing the model on the same data which will result in a high level of performance. Therefore, the study does not fulfill the similarity criteria. Maertens et al. (2011) obtain Se of 90 % and Sp of 100% when detecting severe lame cows (gait score 3 on a three-point lameness score) using linear regression on kinematic variables from pressure sensitive mats, but the aim of lameness detection is primarily to point the farmer towards the cow that needs extra focus rather than those who need acute treatment (Pastell and Kujala, 2007), and with this model not fulfilling the performance-criteria for sound or mild-lame cows, it seems to be of little use in the production. Cornou and Lundbye-Christensen (2011) use CUSUM to detect the onset of farrowing based on the sow’s activity pattern and obtain Se of 100% and Sp of 100% for a subgroup of 9 sows with the sow’s individual variance. The level of performance when including all sows (n = 19) using individual variance is Se of 100% and Sp of 95% whereas the performance for all sows using group variance is Se of 95% and Sp of 89% thus not fulfilling the performance criteria. In the discussion, the authors mention that using individual variance might be over optimizing the model since the reference days of each sow were known beforehand. They recommend the study to be repeated in a large scale experiment where this bias is avoided and suggest inclusion of more animals and different setup of time windows.

The potential of sheer and improved sensor-based detection models is well exploited and they generally do not detect at an implementable level of accuracy. This calls for alternative approaches with a higher degree of customization and adaptability to individual needs at herd-, farmer-, or animal-levels.

Figure 4: Performance (Highest Specificity (HSp), corresponding specificity (Se)) for the 28 papers that present model performance with sensitivity and corresponding specificity. Lines indicate performance criteria (sensitivity 80% and specificity 99%).
6.3. Customization and prioritizing strategies

Throughout the literature, three strategies of prioritizing methods and few concrete suggestions for customizing models are described. Figure 3 (A, B, C) illustrate three different strategies for improving the performance of a model or for ranking or prioritizing the output of the detection models.

6.3.1. Customization

Customization of detection models based on DLM is suggested by Cornou et al. (2008); Ostersen et al. (2010); de Mol et al. (2013) who all present different variations of DLM in their detection models and discuss further adaptation for implementation.

The DLM is not a prioritizing method as the previously discussed fuzzy logic and NBN in the manner of ranking alerts according to a given preference or classifying alerts as true or false. The DLM as a statistical method can predict - or produce a forecast for - the state of the condition of interest one step forward and compare the prediction with the following observation. Nevertheless, the method as presented in these three papers is capable of adjusting to individual circumstances through described strategies for changes in the discount factor of the DLM which alter the adaptability of the model.

The herd-specific adjustments could be on the prevalence of the condition in focus (Hogeveen et al., 2010), the level of management (Huijs et al., 2010) and the farmer’s risk attitude. The latter might differ in terms of both economic consequences (Rutten et al., 2014) and workload associated with accepting a lower, or a higher, level of false alarms (Mollenhorst et al., 2012). In addition to this strategy de Mol et al. (2013) describe how changing the threshold for the Bayes factor of the DLM influences the Se and Sp of the model so it can be adjusted to the risk attitude or level of management at the individual herd.

6.3.2. Prioritizing strategies

Prioritizing strategy (A) combines sensor data with additional non-sensor information at animal-, section- or herd-specific level in a detection model in order to increase the level of performance. This strategy is followed to some extent by Maatje et al. (1997); Ostersen et al. (2010) who mention the potential of combining sensor and non-sensor data and by Garcia et al. (2014) where the parity of the cow is used as classification parameter when defining groups in the data set. Different methods can be used for combining sensor data and non-sensor data, and NBN has been used with interesting results in studies by Steeneveld et al. (2009, 2010b) who use cow-specific information to provide probability distributions for pathogens and for prioritizing alerts from AMS alert lists.

NBN is also used by Jensen et al. (2016) for combining sensor data and cow-specific information in a CM detection model. Using NBN for combining data from different sources is not common in livestock production but has been done previously by Steeneveld et al. (2009, 2010b) and also recently in the world of computer security where Benferhat et al. (2013); Bouzar-Benlabiod et al. (2015) combine sensor alerts and expert knowledge to improve performance of computer security models.

Even though animal-specific information has great impact on the performance of a detection model, the use of cow-specific information alone is not always enough as proven by Steeneveld et al. (2008, 2010a). Animal-specific biological markers, as used by Chagunda et al. (2006) in a dynamic deterministic biological model, can however show that detailed cow-specific information in combination with laboratory analysis of the enzyme L-lactate dehydrogenase (LDH) can present an impressive performance level with Se 82% and Sp 99% - including no other AMS information. This type of model is yet not implementable due to technological demands.

Prioritizing strategy (B) describes a different approach where the sub-optimal performance obtained by a detection model, whether based either solely on sensors or on combined information types, is initially accepted and the generated alarms are prioritized or ranged by combining them with additional non-sensor data in a following postprocessing step as it has been done by de Mol and Woldt (2001); Steeneveld et al. (2010a) (see Fig Flow chart B 1,2). Two different methods using strategy (B) are described in the literature; fuzzy logic and NBN.

By using fuzzy logic de Mol and Woldt (2001) reduce the number of false positive alerts from earlier developed statistical models detecting CM and/or oestrus (de Mol et al., 1997, CM and oestrus); (de Mol and Ouwertjes, 2001, CM). Two separate fuzzy logic models are created - one for each condition. In the CM model de Mol and Woldt (2001) reduce the number of false positive alarms from 1265 to 64 by combining the output of the statistical model with fuzzified additional information. The information is added on standardized deviation in electric conductivity of each quarter as well as measured conductivity at quarter level. This use of fuzzy logic raises the same issue as seen in Caverio et al. (2006); Kamphuis et al. (2008b); Kramer et al. (2009) where numerical values are first fuzzified to linguistic values, and then defuzzified to numerical again. Since fuzzy logic is a method meant for quantifying linguistic - or fuzzy - values, it seems more obvious to use the method on qualitative factors like reproductive status, level of activity or a description of lameness degree parallel to lameness scores in lameness detection.

Furthermore, the input to this fuzzy logic model is the output of a statistical model where the performance is obtained through long time windows and a high degree of selectivity in the choice of included data (de Mol and Ouwertjes, 2001). As apposed to the CM model, the oestrus model include qualitative parameters like reproductive status and information on activity level. The number of false positive alerts were reduced by 32%, and the false alarms were sought to be further reduced through manual and computational optimization of the model but without noteworthy improvements. In their discussion the authors discuss that the model might have been improved further by including the use of “expert knowledge” from the herdsman or personnel. Even though de Mol and Woldt (2001) reduce the number of false positive alarms, and present a method of prioritization, it is this author’s opinion that fuzzy logic should be used with care on data sets consisting of large amount of quantitative information like sensor-based data as the method is not well suited for this.

Steeneveld et al. (2010a) also follow strategy (B) and use Naive Bayesian Network (NBN) to classify which of the alerted cows on an AMS alert list need further investigation for CM. This is done by calculating the probability of an alert being TP or TN based on information from either one variable or combinations of variables. The variables originate either solely from AMS, solely from additional cow-specific information, or from combining this information. The AUC clearly shows that combining the two sources of information perform the best. NBN is well suited for expressing uncertainties, which will inevitable be a part of describing large individual variation. Even though NBN is the simplest version of Bayesian Classification models, assuming no correlation between the included variables, more
advanced Bayesian Networks have been tested on the same data sets without improving the results (Steeneveld et al., 2010a).

Interestingly enough when analyzing the impact of single variables, Steeneveld et al. (2010a) find that from the non-AMS cow information (parity, days in milk, season of year, SCC history, CM alert history) only days in milk were significantly different between FP and TP alerts. On the opposite, high levels of SCC found in the SCC history of the cow were evenly distributed amongst the cows generating FP and TP alerts. Because the level of SCC is considered a very important indicator of CM (Steeneveld et al., 2008), and the SCC level measured by Steeneveld et al. (2010a) was significantly different between TP and TN milkings, these findings indicate that the alert list is based solely on SCC. This indication makes the ranking of alarms based on multiple cow-specific parameters - not only on SCC - highly relevant. High levels of SCC provide valuable information, but just generate too many FP, perhaps even detecting both CM and subclinical CM when used as single variable (Rasmussen and Bjerring, 2005; Steeneveld et al., 2010a).

Steeneveld et al. (2010a) do not reach a satisfying accuracy when discriminating between TP and FP alerts, but the number of FP is reduced by 33%, and the use of NBN as a simple prioritizing tool in livestock production herds warrants further consideration. The capability of NBN to combine information through adding prior probabilities for any relevant information, sensor-based or not, enables the incorporation of managerial factors. These factors could be changes in feed composition, treatments, and herd-specific routines. Information on the herd-specific health status is also relevant for evaluating if the conditions of interest is of higher or lower prevalence than in average herds.

An important aspect in customizing an early warning system to a specific herd or risk-attitude of a farmer is the farmer’s preferences to the detection system. Mollenhorst et al. (2012) have asked farmers what preferences they have to a CM detection system, and the result is that a low number of false alerts and alerts given in good time with emphasis on the more severe cases is the most important feature. The adaptability to individual circumstances is also important for the farmers. The probabilities for any relevant information can be combined with herd-specific thresholds according to the risk attitude of the farmer (Steeneveld et al., 2010a). NBN shows a high degree of adaptation, which meets the demands for customization, characteristic for modern farmers with ambitions (Mollenhorst et al., 2012).

Prioritizing strategy (C) represents an alternative to performance presented by the epidemiological terms of Se and Sp. This alternative is to present the output of the detection model as a gradient or a risk of a condition occurring. Se and Sp are designed for binary outputs, which essentially does not conditions like CM, tail biting, or lameness, which are gradually evolving, and in nature more complex than binary (Friggens et al., 2007, 2010). Detection models in livestock production are, however, traditionally based on discrete measurements in time (Sherlock et al., 2008) which simplify the picture of a complex condition. Presenting the alarms in the form of a risk indicator (Nielsen et al., 2005; Friggens et al., 2007; Hojsgaard and Friggens, 2010) or as a time gradient leading up to the occurrence of a condition known to happen (Aparna et al., 2014) has been seen. In addition to these output types, the posterior probabilities for a condition to occur as calculated by NBN could be a future approach worth focusing on. Strategy (C) is well suited as a decision support tool because it provides detailed information on the individual animal and at the same time allows the farmer to evaluate the alarms personally and use both experience and knowledge of the herd in combination with well substantiated information from the detection model.

6.4. What is more important - priorities are dynamic

In this review, the overall perspective for evaluating the prioritizing detection models has been to reduce the number of false alarms communicated to the farmer. Traditionally the models have generated alarms indicating what animal to attend to, but other motivations for prioritizing can be mentioned. Decision support for which intervention to choose if multiple are possible, or which alarms to attend to first if more monitoring systems are installed at the same farm generating alarms at the same time are both relevant. The optimal prioritization is not a static solution. It might change on a weekly or even daily basis according to multiple factors, and different interests could generate different optimal prioritization outputs. Market prices or costs associated with an intervention (man-hours, equipment etc.) could be used as added information parameters in a prioritizing model. Such a cost minimizing approach would most likely generate a different output than using animal health parameters or welfare parameters. From the farmers perspective, it might be of high priority to optimize his or hers life quality by generating more free time to spend with the family or by increasing the social acceptance in society.

6.5. Research perspectives for early warning systems

The field of automatic monitoring and modeling is still relatively young, and concurrently with the technological evolution, future perspectives for developing decision-supporting tools for ambitious livestock producers continue to be an extremely interesting field of research and development. This review only includes papers that present a concrete performance, but many studies are exploring a range of topics, including lameness detection (Rajkondawar et al., 2002, 2006; Pastell et al., 2008a,b; Pastell and Madsen, 2008; XiangYu et al., 2008; Chapinal et al., 2009; Nielsen et al., 2010; Tanida et al., 2011; Pluk et al., 2012; Hoffmann et al., 2013; Nuffel et al., 2013; Pluym et al., 2013; Abell et al., 2014; Hothersall et al., 2014; Mohling et al., 2014; Wood et al., 2015), vision-based monitoring (White et al., 2004; Porto et al., 2014; Leroy et al., 2008; Cangar et al., 2008; XiangYu et al., 2008; Meh dizadeh et al., 2015; Kristensen and Cornou, 2011; Kashiha et al., 2013), methods for reducing animal mortality (Beltrán-Alcrudo et al., 2009; Bono et al., 2012, 2013, 2014), modeling of behavioral traits as welfare indicators (Bressers et al., 1994; Turner et al., 2000; Moshou et al., 2001; Madsen et al., 2005a; Madsen and Kristensen, 2005; Oliviero et al., 2008; Ferrari et al., 2010; Junge et al., 2012; Cornou and Kristensen, 2014a; Dutta et al., 2015), as well as the continuing focus on detecting CM in dairy cows (Kamphuis et al., 2008a; Claycomb et al., 2009; Lukas et al., 2009).

Skeptics might argue that further research in the development of early warning systems is of little use since the criteria for implementation are so difficult to fulfill. But looking at the broader perspectives, automatic monitoring and early warning systems offer an opportunity to observe the animals 24 hours a day 7 days week 365 days a year, which is far more than what is human possible in traditional livestock production. Early warning systems will always be a decision support tool for the farmer, and not a bullet-proof management manual. The farmer accepting a certain amount of false alarms, or relating to a given risk indicator for a condition occurring, is a realistic scenario after implementation of sensor-based early warning systems. The perspectives for improving animal welfare
through precision livestock farming are distinct, although more research is needed before warning systems with high enough accuracy are ready for implementation.

7. Conclusion

Three methods have been used for prioritizing sensor-based alarms in livestock production. Two of these methods, Fuzzy logic and Naive Bayesian Network, combine sensor data with non-sensor data whereas the third method, Hidden phase-type Markov model, generates a time gradient to the onset of farrowing - a condition known to happen. The use of Fuzzy logic reduces the number of alarms considerably but the method is not well suited for data consisting of large amounts of numerical values like sensor-based data.

Naive Bayesian Network reduces the number of alarms by 57%, and this method shows potential for further research in prioritizing true and false alarms. Hidden phase-type Markov model generates a continuous output which is an interesting alternative to the binary Se and Sp although the Hidden phase-type Markov model might not be the right choice for modelling conditions with no - or diffusely defined - phases or with varying probabilities of occurrence.

For 20 years, no sensor-based detection model has fulfilled the performance demands needed to generate a satisfyingly low level of false positive alarms, and these demands seem close to unreachable with the few models actually obtaining high performances being associated with high error rates. Instead of focusing on fulfilling unreachable demands based on binary performance parameters for more complex conditions, future research could seek alternative approaches for the output of detection models as for instance the prior probability - or the risk of a condition occurring or not. Alarms from detection models can be prioritized in order to optimize production efficiency, production costs, work load and animal health, and a future with automatic monitoring in livestock production looks promising considering both the life quality of the farmer and the welfare of the animals.


R. M. de Mol, A. Keen, G. H. Kroee, and J. M. F. H. Achten. Description of a detection model for oestrus and diseases in dairy cattle based on time


<table>
<thead>
<tr>
<th>Paper</th>
<th>Focus ($E_T$)</th>
<th>Method ($f(D_t)$)</th>
<th>Sensor</th>
<th>Variables ($x_t$)</th>
<th>Performance</th>
<th>Performance method</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>de Mol et al. (1997)</td>
<td>CM</td>
<td>Oestrus</td>
<td>Other diseases</td>
<td>Time series</td>
<td>Kalman filter</td>
<td>AMS$^2$ -like Activity tags Feeding troughs with sensors</td>
<td>Milk yield Milk temperature Activity EC Left over concentrates Cow status (for oestrus)</td>
</tr>
<tr>
<td>Maatje et al. (1997)</td>
<td>CM</td>
<td>Oestrus</td>
<td>Other diseases</td>
<td>Time series</td>
<td>Kalman filter</td>
<td>In-line QMC$^5$ In-line temperature Pedometer</td>
<td>QMC Milk yield Milk temperature Activity level General cow information</td>
</tr>
<tr>
<td>de Mol et al. (1999)</td>
<td>CM</td>
<td>Oestrus</td>
<td>Other diseases</td>
<td>Time series</td>
<td>Kalman filter</td>
<td>AMS-like sensors Step counter Feeding troughs with sensors Cow status (oestrus)</td>
<td>Milk yield Temperature EC Activity Concentrate intake and ration</td>
</tr>
<tr>
<td>de Mol and Ouweltjes (2001)</td>
<td>CM</td>
<td></td>
<td></td>
<td>Time Series</td>
<td></td>
<td></td>
<td>EC Milk yield</td>
</tr>
<tr>
<td>de Mol et al. (2001)</td>
<td>CM</td>
<td>Oestrus</td>
<td>Other diseases</td>
<td>Time series</td>
<td>Kalman filter</td>
<td>AMS-like sensors but in milking parlour</td>
<td>Milk yield Temperature EC Activity Concentrate intake</td>
</tr>
<tr>
<td>Caverio et al. (2006)</td>
<td>CM</td>
<td>Fuzzy logic</td>
<td>AMS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caverio et al. (2007)</td>
<td>CM</td>
<td>Moving Average EWMA$^7$ Locally weighted regression</td>
<td>AMS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pastell and Kujala (2007)</td>
<td>Lameness</td>
<td>Probabilistic neural network (PNN)</td>
<td>Balance platform</td>
<td></td>
<td></td>
<td>Weight distribution pr leg</td>
<td>HSe$^1$ 100 (Sp$^1$ 57.5)</td>
</tr>
<tr>
<td>Kamphuis et al. (2008b)</td>
<td>CM</td>
<td>Fuzzy logic</td>
<td>AMS Inline SCC (ISCC)</td>
<td></td>
<td></td>
<td>ISCC EC FSCC$^{12}$</td>
<td>HSR$^{13}$ 32% FAR$^{14}$ 1.2% Lowest FAR 1.2% (SR$^{15}$ 32%)</td>
</tr>
<tr>
<td>Kramer et al. (2009)</td>
<td>CM</td>
<td>Lameness</td>
<td>Fuzzy logic</td>
<td>Feeding and water troughs AMS neck transponders</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3: Sheer detection models**
<table>
<thead>
<tr>
<th>Paper</th>
<th>Focus ($E_i^*$)</th>
<th>Method ($f(D_i)$)</th>
<th>Sensor</th>
<th>Variables ($x_i$)</th>
<th>Performance</th>
<th>Performance method</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maertens et al. (2011)</td>
<td>Lameness</td>
<td>Unspecified. Described as: Specially-developed dedicated kinematic variable-analysis software</td>
<td>Pressure sensitive mat</td>
<td>20 kinematic gait variables in space-time Duration of each hoof imprint Location of each hoof imprint</td>
<td>HSe 90 (Sp 100) HSp 100 (Se 90)</td>
<td>Linear regression</td>
<td>Performance reported for gait score 3 Overall 84% of all cows were correctly classified in a three-scale classification</td>
</tr>
<tr>
<td>van Hertem et al. (2013)</td>
<td>Lameness</td>
<td>Correlation between binary output and input variables Logistic regression</td>
<td>Accelerometer</td>
<td>Milk yield Neck activity Ruminant time</td>
<td>HSe 89 (Sp 85) HSp 85 (Se 89)</td>
<td>Logistic regression</td>
<td>Milk yield higher correlated than neck activity or ruminant time Only strictly healthy cows included in data set</td>
</tr>
<tr>
<td>Kamphuis et al. (2013)</td>
<td>Lameness</td>
<td>Multivariate additive logistic regression</td>
<td>Weight scale Pedometer AMS</td>
<td>Liveweight Activity Milk-order - yield and duration</td>
<td>HSe 56.8 (Sp 80) HSp 90 (Se 41)</td>
<td>Multivariate additive logistic regression with AUC minimum 6</td>
<td>5 point lameness score scale reduced to binary Se calculated at fixed Sp levels of 80 and 90</td>
</tr>
<tr>
<td>Miekley et al. (2013b)</td>
<td>Lameness</td>
<td>Principal Component Analysis (PCA)</td>
<td>Milk meter Pedometer Feeding trough with sensors</td>
<td>Milk yield EC Activity Feed intake Number of feeding visits Feeding time</td>
<td>HSe 87.8 (Sp 61.9) HSp 76.7 (Se 83.3)</td>
<td>HSe for lameness HSp for CM PCA</td>
<td>Disease blocks of day of treatment plus 3-7 days before High error rate (99 for CM, 87.8 for lameness) TP 27 and FP 23 cows/day presented</td>
</tr>
<tr>
<td>Viazzi et al. (2013)</td>
<td>Lameness</td>
<td>Decision tree</td>
<td>Video Camera</td>
<td>BMP&lt;sub&gt;18&lt;/sub&gt; Back posture</td>
<td>TPR&lt;sup&gt;19&lt;/sup&gt; 0.94 (FPR&lt;sup&gt;20&lt;/sup&gt; 0.24) FPR 0.04 (TPR 0.25)</td>
<td>TPR (not lame) FPR (lame)</td>
<td>5 point LS&lt;sup&gt;21&lt;/sup&gt; scale reduced to 3 points (FPR) or binary scale (TPR)</td>
</tr>
<tr>
<td>van Hertem et al. (2014)</td>
<td>Lameness</td>
<td>Ordinal multinominal logistic regression</td>
<td>3D-camera Photocell Back-Posture-Measurement (BPM) Locomotion score (LS)</td>
<td>HSe 54.9 (Sp 90.4) HSp 94.1 (Se 47.1)</td>
<td>HSe Linear regression HSp Ordinal multinominal logistic regression</td>
<td>5-point LS&lt;sup&gt;22&lt;/sup&gt; transformed to binary Camera-data collected during night-milking due to sensor-sensitivity to sunlight</td>
<td></td>
</tr>
<tr>
<td>Bressers et al. (1995)</td>
<td>Oestrus</td>
<td>Threshold</td>
<td>Camera</td>
<td>Duration of visits to a boar Frequency of visits to a boar Combination of the two First standing response</td>
<td>95% of sows in oestrus detected in time for service Many false positive</td>
<td>Frequency of visits to a boar</td>
<td>Few details described in paper Oestrus validated by farm checklist</td>
</tr>
<tr>
<td>Freson et al. (1998)</td>
<td>Oestrus</td>
<td>Canonical discriminant analysis Logistic regression</td>
<td>Infra-red sensor</td>
<td>Body movement Behaviour</td>
<td>Se 79 (Sp 68)</td>
<td>Sensor performance - not model performance</td>
<td>TP and FP distinguished with probability of 99.9% Not possible to discriminate between TN&lt;sup&gt;27&lt;/sup&gt; and FN&lt;sup&gt;23&lt;/sup&gt;</td>
</tr>
</tbody>
</table>
Table 3: Sheer detection models (continued)

<table>
<thead>
<tr>
<th>Paper</th>
<th>Focus ($E_7$)</th>
<th>Method ($f(D_7)$)</th>
<th>Sensor</th>
<th>Variables ($x_t$)</th>
<th>Performance</th>
<th>Performance method</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cornou and Lundbye-Christensen (2010)</td>
<td>Classify activity types and categories</td>
<td>Uni- and multivariate DLM</td>
<td>Accelerometer Camera</td>
<td>Feeding Rooting Walking Lying sternally Lying laterally Active or Passive</td>
<td>HSe 96 (Sp 96)</td>
<td>Multivariate DLM with free variance structure</td>
<td>Activity category</td>
</tr>
</tbody>
</table>

1 Clinical Mastitis  
2 Automatic Milking System  
3 Electric Conductivity  
4 Highest sensitivity reported in the article  
5 Specificity (Sp) is the specificity corresponding to HSe  
6 Highest specificity reported in the article  
7 Sensitivity (Se) is the sensitivity corresponding to HSp  
8 Quarter Milk Conductivity  
9 Exponentially Weighted Moving Average  
10 Somatic Cell Count  
11 Receiver Operating Curve  
12 Laboratory determined SCC  
13 Highest Success Rate  
14 False Alert Rate  
15 Success Rate  
16 True Positive  
17 False Positive  
18 Body Movement Pattern  
19 True Positive Rate  
20 False Positive Rate  
21 Lameness Score  
22 True Negative  
23 False Negative  
24 Dynamic Linear Model  
25 Multi Process Kalman Filter
<table>
<thead>
<tr>
<th>Paper</th>
<th>Focus ($E_7^1$)</th>
<th>Improvement</th>
<th>Method ($f(D_1)$)</th>
<th>Sensor</th>
<th>Variables ($x_t$)</th>
<th>Performance</th>
<th>Performance method</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al. (2009)</td>
<td>Lameness</td>
<td>Logistic regression</td>
<td>B-spline transformation of limb movement variables (LMV)</td>
<td>Force</td>
<td>All variables at both limb- and cow level: Peak ground reaction force Average ground reaction force Stance time of a limb Ground reaction force integral</td>
<td>HSe 100 (Sp 100)</td>
<td>HSe 15 knots</td>
<td>15 knots is highest degree of freedom LS² reduced from 5 point scale to binary No difference in accuracy related to three degrees of freedom Validated by “leave-one-out” 260 times</td>
</tr>
<tr>
<td>Kamphuis et al. (2010a)</td>
<td>CM²</td>
<td>Decision-tree</td>
<td>Pruning variated Cost matrix</td>
<td>AMS²</td>
<td>EC Color Visual scoring</td>
<td>HSe 56.7 (Sp 93.1)</td>
<td>HSe 5 nodes, low cost for false classification HP 26 nodes, high cost for false classification</td>
<td>Both sensor and non-sensor information used to identify positive cows</td>
</tr>
<tr>
<td>Kamphuis et al. (2010b)</td>
<td>CM</td>
<td>Decision-tree</td>
<td>Bagging</td>
<td>AMS</td>
<td>EC Color Milk yield Visual scoring</td>
<td>HSe 77.8 (Sp 97.9)</td>
<td>HP 99 (Se 71.4)</td>
<td>J48 combined with bagging</td>
</tr>
<tr>
<td>Miekley et al. (2012)</td>
<td>Mastitis</td>
<td>Wavelet filtering</td>
<td>CUSUM Selfstarting</td>
<td>AMS</td>
<td>EC CUSUM</td>
<td>HSe 83.6 (Sp 59.2)</td>
<td>HSp 85.5 (Se 63.5)</td>
<td>Block-sensitivity at minimum 70% Error rates up to 99.6% Diseases defined as disease blocks</td>
</tr>
<tr>
<td>Miekley et al. (2013a)</td>
<td>Mastitis</td>
<td>Wavelet filtering</td>
<td>Vector autoregressive model (VAR)</td>
<td>AMS</td>
<td>EC Milk yield Feeding pattern</td>
<td>HSe 78.9 (Sp 80.4)</td>
<td>HSp 81.0 (Se 74.2)</td>
<td>Block-sensitivity at minimum 70% Error rates up to 99.6% Diseases defined as disease blocks</td>
</tr>
<tr>
<td>de Mol et al. (2013)</td>
<td>Lameness</td>
<td>Quadratic trend models</td>
<td>Models fitted with DLM¹¹</td>
<td>AMS Concentrate feeder</td>
<td>10 activity variables Number of milkings Number of refusals Milk yield Concentrates left over</td>
<td>HSe 100 (Sp 95.8)</td>
<td>HSp 96.5 (Se 100)</td>
<td>Combined data sets Threshold for lame: 5 4 alerted variables required</td>
</tr>
<tr>
<td>Huybrechts et al. (2014)</td>
<td>CM</td>
<td>Time series</td>
<td>Asteregressive moving average</td>
<td>AMS</td>
<td>EC Milk yield CUSUM</td>
<td>HSe 63 (Sp -)</td>
<td>CUSUM</td>
<td>Specificity not reported</td>
</tr>
<tr>
<td>Garcia et al. (2014)</td>
<td>Lameness</td>
<td>Partial least squares discriminant analysis (PLS-DA)</td>
<td>Logistic regression Backward variable selection of originally 320 variables</td>
<td>AMS Activity tag</td>
<td>Parity 1: 17 variables Parity 2: 28 variables (variables not specified)</td>
<td>HSe 79 Sp 83</td>
<td>HSp 83 Se 79</td>
<td>Model for second parity 70% of observations (gait score 2 and parity 3) excludes Validated by “leave-one-out” 331 times</td>
</tr>
</tbody>
</table>
Table 4: Improved detection models (continued)

<table>
<thead>
<tr>
<th>Paper</th>
<th>Focus (E₇⁺)</th>
<th>Improvement</th>
<th>Method (f(Dₜ))</th>
<th>Sensor</th>
<th>Variables (xₜ)</th>
<th>Performance</th>
<th>Performance method</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cornou et al. (2008)</td>
<td>Oestrus</td>
<td>Lameness</td>
<td>Univariate DLM</td>
<td>CUSUM</td>
<td>Individual eating rank</td>
<td>HSe 75 (Sp 95.4)</td>
<td>(Oestrus) DLM and V-mask</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other health</td>
<td></td>
<td>ESF¹²</td>
<td></td>
<td>HSp 95.4 (HSe 75)</td>
<td></td>
<td>Study performed in three herds. Highest overall performance reported in this table. FP²³ 2–3 times higher with this method than in ESF alert list. Discusses customization through adjustments in the discount factor of the DLM.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>disorders</td>
<td></td>
<td>V-mask</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cornou and Lundbye-</td>
<td>Parturition</td>
<td></td>
<td></td>
<td>Eartag</td>
<td>Duration of visits to a boar</td>
<td>HSe 89.2 (Sp 96.9)</td>
<td>HSp 99.4 (Se 55.6)</td>
<td>Duration alone exceeds both combined model and frequency alone. Error rate for HSe is 97.1%. Error rate for HSp is 92%. Discusses customization through adjustments in the discount factor of the DLM.</td>
</tr>
<tr>
<td>Christensen (2011)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Frequency of visits to a boar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combination of the two</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes:
- CUSUM: Cumulative Sum of Squares
- ESF: Early Stages of Farrowing
- V-mask: Visual
- Eartag: Electronic identification tag
- MPKF: Multi-parameter Knowledge Fusion
- DGLM: Dynamic Generalized Linear Model
- RFID: Radio Frequency Identification
- HSe: High Sensitivity
- HSp: High Specificity
- Se: Sensitivity
- Sp: Specificity
- FP: False Positive

References:
Table 4: Improved detection models (continued)

<table>
<thead>
<tr>
<th>Paper</th>
<th>Focus ($E_t^+$)</th>
<th>Improvement</th>
<th>Method ($f(D_t)$)</th>
<th>Sensor</th>
<th>Variables ($x_t$)</th>
<th>Performance</th>
<th>Performance method</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kashiha et al.</td>
<td>Weight estimation</td>
<td>Linear regression</td>
<td>Transfer Function</td>
<td>Cameness</td>
<td>Top-view pig area</td>
<td>Highest $R^2$ 97.5%</td>
<td>Transfer Function</td>
<td>48 Transfer Function models calculated</td>
</tr>
<tr>
<td>(2014)</td>
<td></td>
<td>Mixed effects (non-linear)</td>
<td></td>
<td>Weight scale</td>
<td>Body weight</td>
<td></td>
<td></td>
<td>Only one presented in paper</td>
</tr>
</tbody>
</table>

1. Highest sensitivity reported in the article
2. Specificity (Sp) is the specificity corresponding to HSe
3. Highest specificity reported in the article
4. Sensitivity (Se) is the sensitivity corresponding to HSp
5. Lameness Score
6. Clinical Mastitis
7. Automatic Milking System
8. Electric Conductivity
9. Default AMS algorithm
10. Cumulative Sum
11. Dynamic Linear Model
12. Electronic Sow Feeding
13. False Positive
14. Dynamic Generalized Linear Model
15. Radio Frequency Identification
16. Multi Process Kalman Filter
17. Coefficient of determination $R^2$
<table>
<thead>
<tr>
<th>Paper</th>
<th>Focus ($E^*_F$)</th>
<th>Method ($f(D)$)</th>
<th>Input</th>
<th>Sensor</th>
<th>Variables ($x_i$)</th>
<th>Performance</th>
<th>Performance method</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dairy cows</td>
<td>Reduction of FP$^1$ in</td>
<td>Fuzzy logic. Default and manually</td>
<td>Results from statistical models presented in de Mol</td>
<td>AMS$^3$</td>
<td>Milk yield</td>
<td>HSe$^5$ 100 (Sp$^6$ 99.75) (CM)</td>
<td>CM: Fuzzy Logic</td>
<td>CM: Same data set used for learning and testing</td>
</tr>
<tr>
<td></td>
<td>detecting CM$^2$ and oestrus</td>
<td>optimized</td>
<td>and Ouweldes (2001) (CM)</td>
<td></td>
<td>Temperature</td>
<td>HSp$^7$ 99.75 (Se$^8$ 100) (CM)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>and de Mol et al. (1997) (oestrus)</td>
<td></td>
<td>Activity</td>
<td>HSe 79 (Sp 98.1) (oestrus)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>HSp 99.4 (Se 66) (oestrus)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The number of FP alerts reduced by 57% 10% of TP$^9$ alerts missing</td>
<td>NBN combining both sensor and non-sensor information</td>
<td></td>
</tr>
</tbody>
</table>

Steeneveld et al. (2010a) Identification of alerted cows that need further investigation for CM

Naive Bayesian Network (NBN) 

Logistic regression

Alert list from AMS

Non-AMS cow information

AMS-alert information

Combination of non-AMS and AMS information

AMS

EC

Color alert CM

Color alert Abnormal milk

Milk yield

Cow specific information

Default performance of AMS in the study does not meet minimum demands of satisfying detection performance
<table>
<thead>
<tr>
<th>Paper</th>
<th>Focus ( (E^*_f) )</th>
<th>Method ( f(D) )</th>
<th>Input</th>
<th>Sensor</th>
<th>Variables ( (x_t) )</th>
<th>Performance</th>
<th>Performance method</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aparna et al. (2014)</td>
<td>Time to farrowing</td>
<td>Hidden phase-type</td>
<td>Water valve</td>
<td>Water</td>
<td>Water consumption</td>
<td>Estimated time to farrowing</td>
<td>Water consumption and activity in combination gave best result for both time estimation and probability calculation</td>
<td>Condition is known to happen Predictions based on herd specific parameters</td>
</tr>
<tr>
<td></td>
<td>Probability of farrowing</td>
<td>Markov</td>
<td>Photo-cells</td>
<td>Activity</td>
<td>Behaviour</td>
<td>SD(^{10}) 4.6 h</td>
<td>97% True warnings</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Camera</td>
<td>Time of day</td>
<td>Time since mating</td>
<td>Probability of farrowing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SD(^{9}) 4.5 h</td>
<td>Threshold 12 h</td>
<td></td>
</tr>
</tbody>
</table>

1 False Positive  
2 Clinical Mastitis  
3 Automatic Milking System  
4 Electric Conductivity  
5 Highest sensitivity reported in the article  
6 Specificity \((Sp)\) is the specificity corresponding to HSe  
7 Highest specificity reported in the article  
8 Sensitivity \((Se)\) is the sensitivity corresponding to HSp  
9 True Positive  
10 Standard Deviation