

# Individualised automated lameness detection in dairy cows and the impact of historical window length on algorithm performance

D. Piette<sup>1</sup> , T. Norton<sup>1†</sup>, V. Exadaktylos<sup>2</sup> and D. Berckmans<sup>1,2</sup>

<sup>1</sup>M3-BIORES, Division Animal and Human Health Engineering, Department of Biosystems, KU Leuven, Kasteelpark Arenberg 30, 3001 Heverlee, Vlaams-Brabant, Belgium; <sup>2</sup>BioRICS nv, Technologielaan 3, 3001 Leuven, Vlaams-Brabant, Belgium

(Received 13 November 2018; Accepted 18 June 2019)

*Lameness is an important economic problem in the dairy sector, resulting in production loss and reduced welfare of dairy cows. Given the modern-day expansion of dairy herds, a tool to automatically detect lameness in real-time can therefore create added value for the farmer. The challenge in developing camera-based tools is that one system has to work for all the animals on the farm despite each animal having its own individual lameness response. Individualising these systems based on animal-level historical data is a way to achieve accurate monitoring on farm scale. The goal of this study is to optimise a lameness monitoring algorithm based on back posture values derived from a camera for individual cows by tuning the deviation thresholds and the quantity of the historical data being used. Back posture values from a sample of 209 Holstein Friesian cows in a large herd of over 2000 cows were collected during 15 months on a commercial dairy farm in Sweden. A historical data set of back posture values was generated for each cow to calculate an individual healthy reference per cow. For a gold standard reference, manual scoring of lameness based on the Sprecher scale was carried out weekly by a single skilled observer during the final 6 weeks of data collection. Using an individual threshold, deviations from the healthy reference were identified with a specificity of 82.3%, a sensitivity of 79%, an accuracy of 82%, and a precision of 36.1% when the length of the healthy reference window was not limited. When the length of the healthy reference window was varied between 30 and 250 days, it was observed that algorithm performance was maximised with a reference window of 200 days. This paper presents a high-performing lameness detection system and demonstrates the importance of the historical window length for healthy reference calculation in order to ensure the use of meaningful historical data in deviation detection algorithms.*

**Keywords:** healthy, reference, back, posture, data

## Implications

The challenge in developing automated lameness detection tools is that lameness evolves differently for each individual cow. Hence, accurate lameness monitoring on farm scale can only be achieved via individualised lameness detection. In this study, a healthy reference is calculated for each cow based on historical data and deviations from that reference are used to automatically classify a cow as lame or healthy. The effect of the historical window on the lameness detection performance is evaluated to demonstrate the importance of tuning the deviation thresholds and the quantity of the historical data being used in deviation detection algorithms.

## Introduction

Lameness corresponds to abnormal gait resulting from injury or weakness in the back, feet or legs; and in dairy cows, it is a major cause of compromised health and welfare, production loss and reduced fertility (Baggott and Russell, 1981; Alban *et al.*, 1996; Barnes *et al.*, 2011). The prevalence of lameness on dairy farms is underestimated by modern farmers (Archer *et al.*, 2010; Leach *et al.*, 2013; Van Nuffel *et al.*, 2015b). Due to the increasing farm sizes and the continued consolidation of the dairy industry, farmers now have less time to look after each individual cow (Guarino *et al.*, 2017). Additionally, inadequate knowledge and inconsistencies in terminology between farmers and technology providers or advisors to identify cows as sound or lame often result in underestimation of the severity of lameness and the urgency for treatment (Horseman *et al.*, 2014; Sadiq *et al.*, 2017). Finally, the adaptive behaviour cows

<sup>†</sup> E-mail: tomas.norton@kuleuven.be

develop to hide their lameness in the presence of a human observer makes detection of early lameness more challenging as well (Sadiq *et al.*, 2017). Consequently, lame cows in the herd are often detected (if at all) when they are already in an advanced state. It is not surprising that cost-effective, automated and continuous monitoring systems for lameness in dairy cows are now being sought after to enable farmers to improve their daily herd management and detect lameness in an earlier stage (Van Nuffel *et al.*, 2015b).

Remote sensing solutions such as 2D or 3D video cameras have excellent potential as lameness monitoring systems. However, there are challenges when developing algorithms for such devices as one algorithm has to work for multiple animals despite the fact that individual cows have their own specific way of walking and that lameness is expressed in an individual way (Berckmans, 2006). To meet this challenge, real-time lameness detection systems must account for the normal healthy behaviour of the cow, so that abnormalities can then be picked up quickly. Such an approach requires maximising the usage of historical and real-time data. Individualised monitoring systems using animal-level historical data have been shown to achieve better detection accuracies when compared with population-based monitoring systems (Tambuyzer, 2018). Historical data can be used to tune a model for expected behaviour, that is, a healthy reference in the case of lameness monitoring. Then as new data are continuously collected, deviations from the healthy reference can be used to determine the lameness status of a cow (Dórea *et al.*, 2013).

Previous studies on lameness monitoring have incorporated historical data into their detection algorithms (Pastell and Madsen, 2008; Alsaad *et al.*, 2012). Pastell and Madsen (2008) developed a four-balance system to measure the distribution of weight carried by the cow's legs in a milking robot. Using cumulative sum charts, they compared this healthy reference value (determined over 10 to 40 visits) to automatically detect lameness. Alsaad *et al.* (2012) instead derived activity and lying behaviour variables from pedometers to detect lameness. For their algorithm, a 14-day period free of lameness for each cow was selected to compute the mean and standard deviation of the behaviour variables. These were then used to obtain abnormal behaviour (via deviations) on the individual level from the real-time data. While both of these studies define a short and relatively fixed window to determine a healthy reference per cow and limited work was done to optimise this window, studies in other research fields suggest that short time windows may not guarantee the best result (Lafrance and Miller, 2010; Vial and Berezowski, 2015). Lafrance and Miller (2010) studied serum creatinine levels in humans during baseline and hospitalisation to develop a classifier for acute kidney disease. They observed that the sensitivity of abnormality detection for kidney disease was increased when the healthy reference window was increased by 3 to 12 months. Furthermore, the authors noted the potential for the healthy reference to change over time. This makes sense, considering that the physiology of living organisms changes over time. This

was confirmed by Vial and Berezowski (2015) who found that for syndromic surveillance in livestock in general, a long history for the healthy reference is required in order to fit deviation detection algorithms more easily (Vial and Berezowski, 2015). Therefore, it can be inferred that the healthy reference of disease processes would benefit from longer time windows and regular updating of this window.

In this paper, we focused on defining a methodology for selecting (1) individualised thresholds for lameness detection and (2) the length of historical data needed to calculate a healthy reference. The study is part of an overall research effort to design an early warning system for lameness using 3D camera technology (Van Hertem *et al.*, 2014). Back posture values that have been automatically extracted from top view 3D images of the cows' back are used as a way of measuring the degree of lameness (Poursaberi *et al.*, 2010; Pluk *et al.*, 2012; Van Hertem *et al.*, 2014). In the paper, we first present a novel methodology to calculate a healthy reference for back posture values based on the historical data for each individual cow. Then, we present a deviation detection algorithm that was developed to automatically identify lameness. Finally, the effect of historical window on algorithm performance is presented.

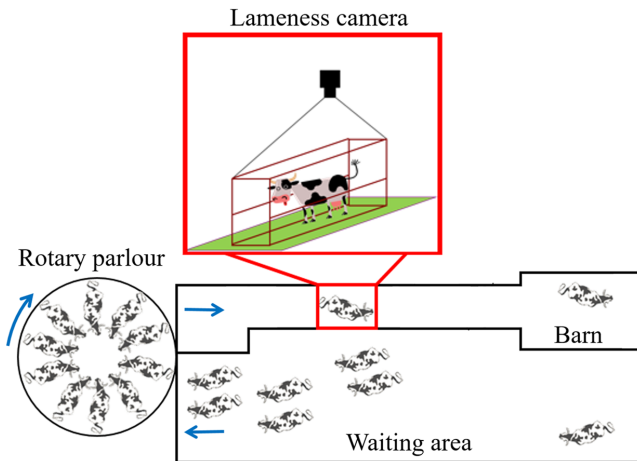
## Material and methods

### *Subjects and infrastructure*

Data were collected in a commercial Swedish dairy farm with more than 2000 Holstein Friesian cows. Cows were housed in a free stall barn with concrete floor and straw bedded cubicles, were fed total mixed rations and had ad libitum access to feed and water. Cows were milked twice a day, produced 29.6 kg of milk per day on average and had an average parity of 1.9. During milking time, cows walked from a waiting area to a rotary milking parlour (Figure 1). After leaving the milking parlour, cows walked through a corridor on the way back to the barn. In this corridor, the cows passed under a 3D camera connected to a computer that calculated the back posture value of each cow based on the recorded images. The back posture value was calculated based on the curvature of the cow's spine. The higher the back posture value, the more curved the cow's spine, and the more severe the lameness problem. We refer to the study published by Van Hertem *et al.* (2014) for detailed information on how this back posture value is calculated. Back posture data collection for each cow started between 2 and 5 days after calving and ended the day before dry off.

### *Gold standard*

To evaluate the performance of the automated system for lameness detection, the outcome of the system must be tested against the gold standard for lameness, that is, manual score. For manual lameness scoring, the Sprecher scale was used (Sprecher *et al.*, 1997) by a live observer who gave the cows a score between 1 and 5 depending on their gait. A manual score of 1 was given when the cow



**Figure 1** (colour online) Ground plan with set-up of the dairy cow lameness monitor in the farm.

had a sound gait. A manual score of 5 was given to a cow with a severe deviation in gait, that is, a severely lame cow. For further data analysis, cows with a manual score of 1 and 2 were considered healthy. In order to minimise the intra-observer variability, a highly experienced observer was used for this process. For this study, manual scoring was performed on a weekly basis during a period of 6 weeks between 18 September 2015 and 23 October 2015. The scoring was performed at the same place where the 3D images were recorded and was carried out simultaneously to the 3D recordings.

Given that it is not possible to manually score all cows in 1 day, some cows lacked manual scores on 1 or more days. The goal of the manual scoring was to create a representative and reliable data set of cows that were (1) not lame, (2) continuously lame or (3) developing a lameness problem (newly lame) during the 6-week scoring period. For that purpose, data from cows with only three or fewer manual scores out of six were rejected from the data set because it was not possible to accurately determine their lameness status or evolution throughout the scoring period. Next, the evolution of the manual scores of each cow in the 6-week scoring period was evaluated. When the manual scores of a cow were varying on a daily basis between healthy scores (1 to 2) and lame scores (3 to 5), this cow was not considered for further analysis. The first objective of eliminating these cows from the data set was to eliminate erroneous manual scores as a result of cows concealing their lameness on 1 day but not necessarily the others. A second aim was to eliminate cows experiencing ephemeral lameness that can be the result, for instance, of stepping on a sharp stone.

#### Back posture values

In order to capture sufficient historical data on each cow, back posture values were collected from August 2014 until November 2015. Given that cows were milked twice a day, two back posture values were available per cow per day if video recording and video processing were successful for both milking sessions (Viazzi *et al.*, 2013; Viazzi *et al.*, 2014). When two back posture values were available for a

cow on the same day, the average back posture value for that day was computed.

#### Healthy reference calculation

The objective of the lameness detection system is to monitor the evolution of back posture values from low values (healthy) to high values (lame). Given that each cow has an individually different morphology, a certain back posture value could indicate soundness for one cow but lameness for another. Therefore, an individual threshold per cow, rather than a group-level threshold for the herd, is desirable for the classification of back posture values as lame or not lame.

In order to define this individual threshold, the healthy reference back posture value must be established per cow. In this study, this reference was defined as the mean of the 5% lowest back posture values observed for a cow in its entire (available) history. This threshold of 5% was carefully chosen by the researchers after visual inspection of the back posture values of every individual cow. By taking 5% of the lowest values, the effect of outliers is minimised and the definition of healthy reference becomes more robust than when the minimal back posture value is chosen. This definition of the healthy reference requires that a minimal data set of historical back posture values exists for every cow. In this study, the minimal length of this historical data set was arbitrarily set to 30 'monitored' days, which means that a cow needed to have back posture data of at least 30 days to be able to calculate a healthy reference. Once a historical window of 30 days was achieved, it was increased by 1 day every day, ever increasing the size of the historical window of each cow for as long as it stayed in the farm. Note that this method assumes that in its first 30 monitored days in the herd, a new cow entering the herd is sound for at least 5% of the time in order to establish an accurate healthy reference. Although this is mostly true, it is possible that a new cow is constantly lame in this period. In that case, the healthy reference for that cow in the first 30 monitored days will not be correct. Assuming the cow will not be chronically lame and will become sound at a certain moment, its healthy reference will be corrected automatically since the historical window for lameness detection increases with time.

#### Lameness detection and performance evaluation

In this study, a threshold is defined as the amount by which the back posture value can deviate from the healthy reference before a cow is classified as lame by the monitoring system. More specifically an individual threshold is used in this paper, as defined by equation (1).

$$\text{Threshold} = \text{factor} \cdot \text{healthy reference} \quad (1)$$

with 'factor' representing a constant that was optimised on group level and 'healthy reference' representing the healthy back posture reference value of an individual cow.

By implementing equation (1) when a back posture value of a cow exceeded the threshold, a cow was classified as lame. If the individual threshold was not exceeded, the cow was classified as healthy.

In order to determine the optimal factor and to evaluate the performance of lameness detection, threefold cross validation was performed for different factor values. For this purpose, the data set was randomly split in three parts. One third of the data was used to build the algorithm and the remaining two third of the data was used for validation. This was repeated two more times. The factor value varied between 1.05 and 1.5 in steps of 0.05, assuming that deviations in back posture value between 5% and 50% of the healthy reference would be indicative of lameness. For every factor value, the individual threshold was calculated and cows were classified as lame or healthy by the algorithm. Then validation was performed on a data point basis. Each time a back posture value and a manual score were available for a certain data point, it was labelled as true positive (TP), true negative (TN), false positive (FP) or false negative (FN). For this purpose, the 5-point manual scoring system was translated to a binary classification of healthy (manual scores 1 to 2) versus lame (manual scores 3 to 5), analogously to Van Hertem *et al.* (2014). The sensitivity, specificity, accuracy and precision of the lameness detection were calculated for each factor value using equations (2) to (5). A receiver–operating curve (ROC) was constructed and its operating point was identified and the corresponding factor value and lameness detection performance were noted.

$$\text{Sensitivity} = \frac{\sum TP}{\sum TP + \sum FN} \quad (2)$$

$$\text{Specificity} = \frac{\sum TN}{\sum TN + \sum FP} \quad (3)$$

$$\text{Accuracy} = \frac{\sum TP + \sum TN}{\sum TP + \sum FN + \sum TN + \sum FP} \quad (4)$$

$$\text{Precision} = \frac{\sum TP}{\sum TP + \sum FP} \quad (5)$$

#### Historical window

We hypothesised that as the length of the historical window increases, the definition of the healthy reference becomes more accurate and reliable. The rationale behind this was that the more the data available for one cow, the more accurate the healthy reference will be and by extension the more reliable the lameness classification will be. To test this hypothesis, the historical window to calculate the healthy reference was varied between 30 days and 250 days in steps of 20 days. Then the cows were classified as lame or healthy using the individual threshold and the performance of the classification was evaluated using threefold cross validation. Note that for this analysis, a subset of the data set with 209 cows was taken. The reason for this is that not all cows in the 209 cow data set had 250 days of historical data. In order to perform this particular analysis, only cows that had 250 or

more days of historical data were withheld. This resulted in a new data set of 172 cows for this analysis.

## Results

### Gold standard

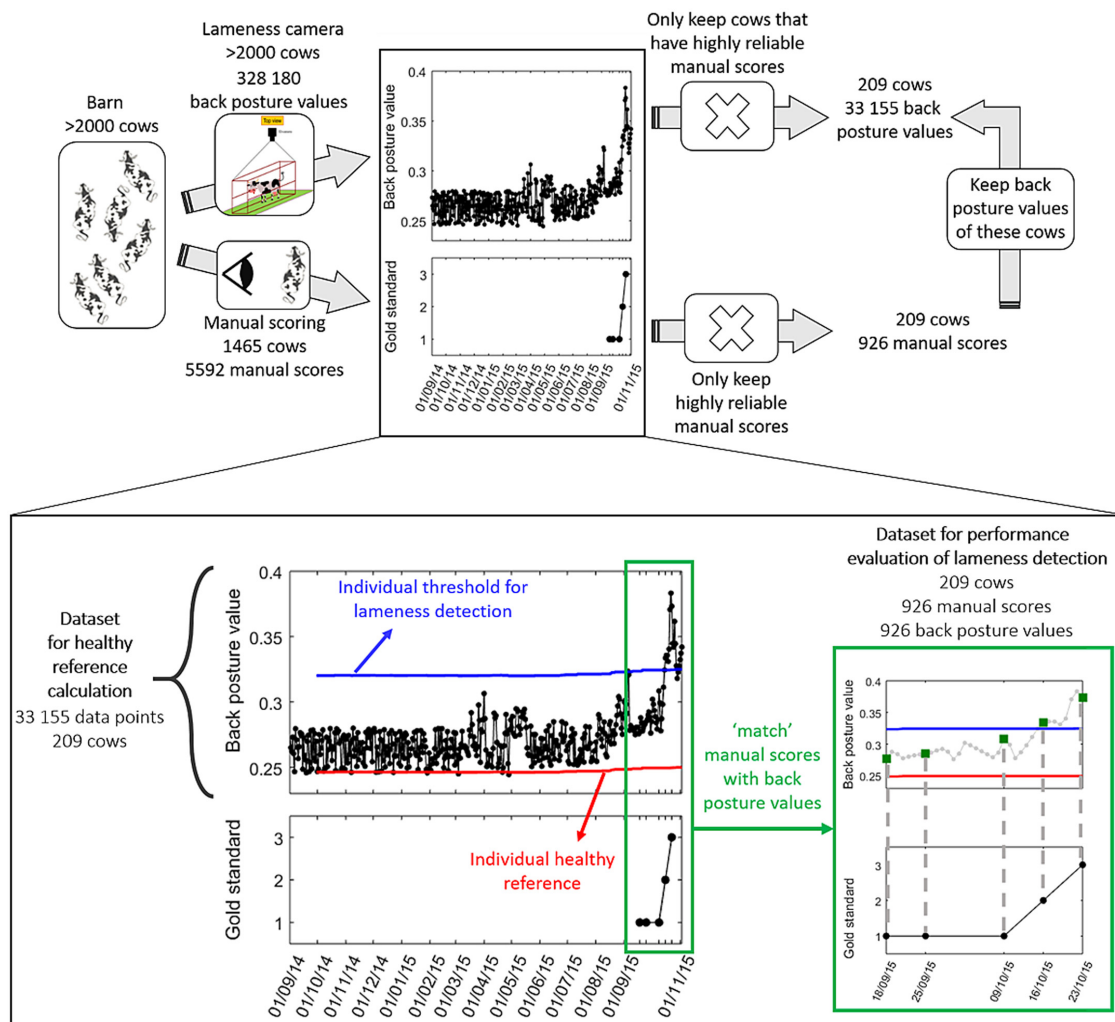
In total, 5592 manual scores belonging to 1465 cows were collected during the 6-week scoring period. After rejecting data from cows with only three or fewer manual scores out of six from the full data set of the 1465 scored cows, a sub data set of 886 cows with a total of 4325 manual scores was retained. Next the evolution of the manual scores of each cow in the 6-week scoring period was evaluated visually. The goal of the manual scoring was to create a time series data set of cows that were (1) consistently healthy, (2) consistently lame or (3) developing a lasting lameness problem during the 6-week scoring period. Cows with ephemeral lameness and cows that recovered from a lameness problem during the 6-week scoring period were thus removed from the data set, resulting in a final data set of 209 cows with 926 manual scoring data points in total. During the 6-week scoring period, 35% of the cows in this farm were either developing a lameness problem or suffering from a longer term lameness problem.

### Back posture values

From August 2014 until November 2015, more than  $3.28 \cdot 10^5$  back posture values were collected for 2165 different cows. Back posture values were averaged per day if more than one score was available (60% of the time) and only the data of the 1465 cows that received manual scores were retained, after which a back posture value data set with more than  $2.06 \cdot 10^5$  historical data points was obtained. Removing the cows from the data set of which the manual scores were considered ‘unreliable’ resulted in a first distinct data set for healthy reference calculation with over  $3.3 \cdot 10^4$  historical data points from August 2014 until November 2015 belonging to 209 cows. Finally, a second distinct data set for algorithm performance evaluation was generated by only keeping the back posture scores on days where a gold standard manual score was given in the 6 weeks period between 18 September 2015 and 23 October 2015. This data set contained 926 data points belonging to 209 cows. Figure 2 gives a comprehensive overview of the data collection and illustrates the difference between the data set for healthy reference calculation and the one for performance evaluation.

### Healthy reference

Figure 3 shows the manual scores of a cow during the 6-week manual scoring period. As can be seen, the manual score of this cow was 1 for the first three scorings and increased to a score of 3 at the end of the 6-week scoring period. This means that this cow developed a lameness problem during the manual scoring period. Figure 3 also shows the back posture



**Figure 2** (colour online) Comprehensive scheme of the data collection and database generation for the development of a dairy cow lameness monitor.

values of this cow during the same period. It is clear that, as the manual score of the cow increased, the back posture value of the cow also rose, as is indicated by the arrows in the figure.

For this example it can be said that back posture values above 0.32 indicate that the cow was lame and that a back posture value under 0.32 indicates that the cow was healthy. However, this cut-off value differs between cows. This is illustrated in Figure 4, which shows the variation of more than  $3.3 \cdot 10^4$  historical back posture values belonging to the 209 studied cows, for the different manual scores. Note that none of the cows received a manual score of 5. It is clear that for the same manual score, there is a large variability in back posture values between different cows.

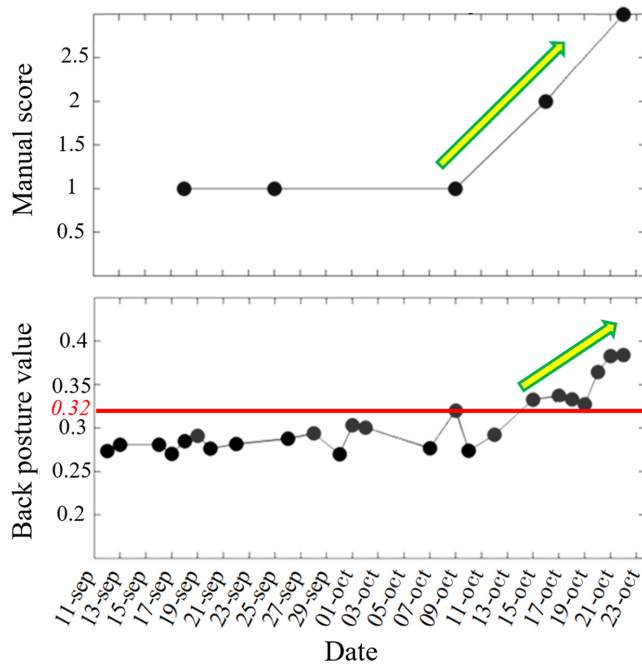
#### Lameness detection performance

The threefold cross validation using the 926 data points to define the individual threshold using different factor values resulted in the ROC as shown in Figure 5. This ROC shows that the optimal performance of the lameness detection is obtained when 'factor' is equal to 1.3, resulting in a

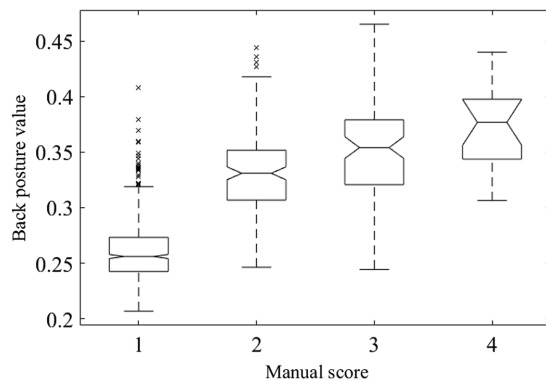
sensitivity of 79%, a specificity of 82.3%, an accuracy of 82% and a precision of 36.1%.

#### Effect of historical window on lameness detection performance

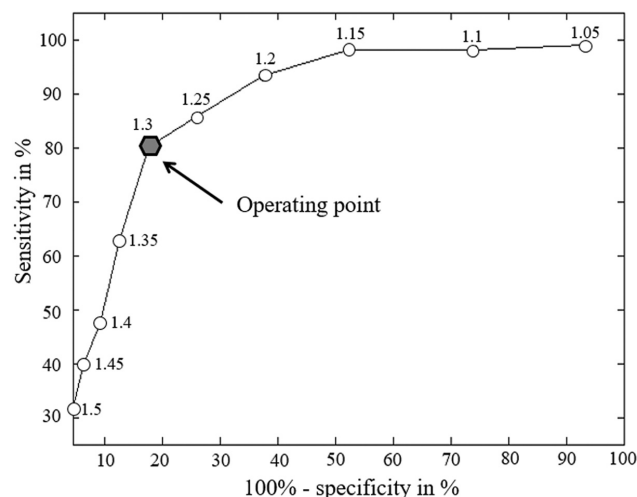
Figure 6 shows the ROC for the different historical windows. It can be seen from the figure that as the historical window increased, every decrease in specificity goes along with an even bigger increase in sensitivity. On average, for every unit decrement in specificity, there was a 5-unit increment in sensitivity. Thus, it can be said that lameness detection performance improves as the length of the historical data set to calculate the healthy reference increases. This improvement was more important for shorter historical windows (30 to 150 days) than for longer historical windows (170 to 250 days), as seen in Figure 6. Optimal performance was reached for a window of 200 days. Note that in this analysis the lameness detection algorithm reached a slightly lower sensitivity (76%) due to a difference in data set (172 out of 209 cows). The accuracy and precision corresponding to the points in Figure 6 are given in Table 1. The accuracy is very similar



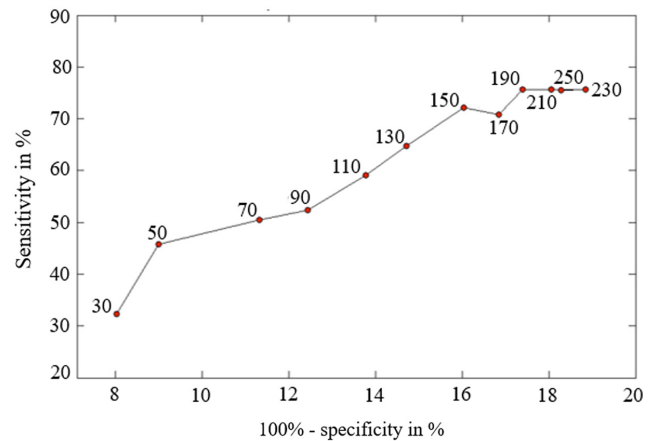
**Figure 3** (colour online) Illustration of increasing manual score and back posture value in a newly lame cow.



**Figure 4** Between-cow variation in back posture values for different manual scores, data from 209 cows.



**Figure 5** Receiver-operating curve (ROC) to determine factor. Data from 209 cows.



**Figure 6** (colour online) Average impact of historical window for healthy reference calculation on algorithm performance. The historical window is varied between 30 and 250 days, in steps of 20 days. Data from 172 cows.

for all window lengths, though it is slightly higher for a window of 50 days. The precision is maximal for a window of 250 days.

## Discussion

It was observed that 35% of the cows in this farm were either developing a lameness problem or suffering from a longer term lameness problem. Knowing that the worldwide average lameness prevalence is approximately equal to 35% (Schlageter-Tello *et al.*, 2015), it can be said that the data set used in this paper was a representative data set for lameness research.

The box plots in Figure 4 clearly illustrated the large variability in back posture values for the same manual lameness score. Additionally, the large overlap between the whiskers of the box plots showed that one back posture value can indicate lameness for one cow but soundness in gait for another. This individual cow variation has already been pointed out in previous research and confirms that back posture values should be analysed and interpreted at the individual cow level and that the healthy reference should be calculated for every cow separately (Maertens *et al.*, 2011; Cook *et al.*, 2012; Abuelo *et al.*, 2016). Then an individual threshold was defined to detect deviations in the back posture data of a cow. By defining the threshold as the multiplication of a fixed factor and the healthy reference, a threshold for lameness detection was obtained that was individual for every cow, which is known to improve the discriminatory power of monitoring systems (Tambuyzer, 2018). It is important to note that the optimised fixed factor that was found in this study is most likely farm specific and will have to be calibrated for every new farm.

Based on the individually defined threshold, a cow was classified as lame or not lame, after which the algorithm results were compared with the manual scores, resulting in a system performance of 82.3% sensitivity, 79% specificity, 82% accuracy and 36.1% precision. This performance is

**Table 1** Algorithm accuracy and precision for different lengths of the historical window for lameness detection in dairy cows

	Window length in days											
	30	50	70	90	110	130	150	170	190	210	230	250
Accuracy (%)	86.8	87.3	86.4	85.5	85.5	85.6	85.5	84.7	84.4	84.4	84.0	85.0
Precision (%)	38.0	44.0	42.1	39.8	41.7	42.9	43.4	41.6	41.8	42.8	42.9	46.1

very similar to that of Van Hertem *et al.* (2014) who obtained an accuracy of 81.2% with a binary classifier based on 3D video recordings of the back of 186 cows (51 lame and 135 healthy). The difference, however, lies in the sensitivity and specificity of both algorithms. Van Hertem *et al.* (2014) achieved a high specificity of 94.1%, which means that their algorithm would generate very little false alarms, a very desirable trait in lameness detection systems. Still they obtained a significantly lower sensitivity of 47.1%. In that regard, the algorithm presented in this paper performs significantly better and provides a better compromise between the number of false alarms and the number of undetected lameness problems. The only other published research where 3D video recordings are used to detect lameness in dairy cows is a study by Jabbar *et al.* (2017) who used the recordings to detect height variations in hip joints during walking in 23 dairy cows (20 lame and 3 healthy), achieving a sensitivity of 100%, a specificity of 75% and an accuracy of 97.7%. Although this performance seems high, Jabbar *et al.* (2017) used a limited data set of 23 cows with a lameness prevalence of 87% which is higher than the prevalence of a typical European farm. The data set presented in this study comprised 209 cows, where lameness prevalence was equal to 35%, respectively, which is more representative of a modern commercial European dairy farm (Schlageter-Tello *et al.*, 2015).

The algorithm developed in this study performs well when comparing its performance to that of lameness detection systems that use different sensor technologies, which report at least two performance measures that include sensitivity, specificity, accuracy and/or precision. Van Nuffel *et al.* (2015) reported that sensitivity, specificity and accuracy of lameness detection systems using load cells, pressure sensors, gait measuring devices or a combination of sensor data vary between 51.9% and 89%, 57.5% and 93.8%, and 76% and 96.2%, respectively. The highest specificity (93.8%) is achieved by the Stepmetrix developed by Liu *et al.* (2009), which is also characterised by the lowest sensitivity (51.9%). The highest sensitivity and accuracy are obtained by Pastell and Kujala's (2007) weight distribution system (100% and 96.2%), which on its turn has a very low specificity (57.5%). The main drawback of these systems is that there is a poor compromise between sensitivity and specificity. This is not the case for the Gaitwise by Maertens *et al.* (2011), which has an average sensitivity of 83.7% and an average specificity of 92%. The disadvantage of this system in comparison to the 3D-camera system is the larger space

that is needed for installation. Another lameness detection system that shows a good trade-off between sensitivity and specificity is the combination of different sensor data including milk yield, neck activity and rumination time, which can perform with a sensitivity of 89%, a specificity of 85% and an accuracy of 86% (Van Hertem *et al.*, 2013). Note that the comparison of the performance mentioned in the aforementioned studies should be interpreted with care, given that different cut-offs for lameness were used in the different studies. Van Hertem *et al.* (2014) and Pastell and Kujala (2007) considered cows with a score of 1 or 2 as healthy, and cows with a score of 3 to 5 as lame. Jabbar *et al.* (2017) and Liu *et al.* (2009) only considered cows to be healthy if they had a score of 1. Maertens *et al.* (2011) did not apply binary classification (healthy *v.* lame). Instead, they aimed to distinguish between cows with a score of 1, 2 and 3. Finally, Van Hertem *et al.* (2013) did not use manual scoring as a reference but relied on the assessment of a veterinarian instead to label cows as sound or lame. The obtained algorithm precision in this paper could not be compared to the precision of the previous studies, since it was never reported. Summarised, it can be said that the lameness detection system presented in this paper provides a good trade-off between specificity and sensitivity and has a satisfactory overall performance while being amongst the most feasible in terms of installation requirements (e.g. space needed). A shortcoming of the developed lameness detection system is that it cannot distinguish between altered back posture due to pain coming from lameness or other causes such as lactation stage or age.

As regards to the algorithm development, in a first step, the healthy reference of each cow was calculated based on the entire historical data set of each cow, which can be up to 15 months long. This means that for certain cows, the healthy reference is based on data from different lactation numbers and/or lactation stages. Research has shown that lameness prevalence remains rather constant across lactation stages, being only slightly lower in the first 10 weeks of lactation (Weber *et al.*, 2013). Lactation number, however, has a significant impact on lameness prevalence (Weber *et al.*, 2013). For that reason, historical windows that are longer than a lactation stage might not be optimal for healthy reference calculation.

The effect of the historical window on the performance of the lameness detection was investigated for different historical windows varying from 30 to 250 days in step of 20 days. It was observed that the overall lameness detection

performance improved with the length of the historical window, which might be linked to the fact that there are more sound than lame days. More specifically, sensitivity increased and specificity decreased to a lesser extent. The increase in sensitivity confirms previous findings from other research fields (Lafrance and Miller, 2010). The decrease in specificity would mean that increasing the historical window leads to a higher number of false positives. Similar observations were made in the kidney disease detection study by Lafrance and Miller (2010) who noted that extending the historical window resulted in more detection of subjects with a milder kidney disease, which could be thought of as a type of false positive. The most probable reason for this observation is that when too old data are used as a reference, it is no longer representative of that process.

This statement is strengthened by the observation that the increase in algorithm performance with increasing historical window is more important for shorter than for longer historical length. This shows that the historical window to calculate the healthy reference should have an upper limit. This makes sense if we take into account the fact that as a cow ages, its locomotion score and, therefore, its back posture value will naturally increase (Van Nuffel *et al.*, 2015). Consequently, taking too much history of a cow into account to calculate the healthy reference can possibly result in a biased and erroneous healthy reference calculation. This is in line with Vial and Berezowski (2015) who stated that the healthy reference may vary over time and should be updated regularly.

In this paper, a window of 200 days was identified as optimal. This means that lameness detection for an individual cow performs at its best once the cow has been in the herd for at least half a year. Taking into account that dairy cows stay in the herd for 3 to 4 years, this means that the lameness detection system will 'underperform' only about 15% of the cow's time in the barn. If the cow is reared in the same barn where it will be used for milk production, this period can even be reduced to zero by simply starting to monitor cows with the lameness detection system half a year before they start producing milk. Future research should investigate whether the historical window can be individualised per animal and whether this can further improve the performance of lameness detection systems. It is worthwhile to note that the findings of this paper can be applied in many other applications in the dairy sector as well as in various other fields of science where deviation detection algorithms are applied to monitor the health and/or well-being of individual living organisms.


## Conclusion

A first outcome of this paper is the presented methodology to select individualised thresholds for lameness detection in dairy cows, resulting in a high-performing lameness detection system with a satisfactory trade-off between sensitivity and specificity. A second outcome of this paper is the evidence that historical window can have a significant impact on the performance of a lameness detection system for dairy cows. The authors of this

paper suggest that the length of the historical window to calculate a healthy reference should be limited in order to ensure the use of meaningful historical data.

## Acknowledgements

This research was supported by DeLaval International, Sweden. We thank the farmer for making his infrastructure and data available for our study.

 D. Piette 0000-0002-2940-5031

## Declaration of interest

The authors declare that there is no conflict of interest.

## Ethics statement

The animals included in this study were managed in conformity with the Code of practice for the care and handling of dairy cattle (National Farm Animal Care Council, NFACC, 2018). Data were collected on a commercial farm of which the owner agreed to be involved in the project.

## Software and data repository resources

The data from this study are the property of the owner of the commercial farm and are commercially sensitive. For that reason, the data cannot be made available.

## References

- Abuelo A, Gandy JC, Neuder L, Brester J and Sordillo LM 2016. Short communication: markers of oxidant status and inflammation relative to the development of claw lesions associated with lameness in early lactation cows. *Journal of Dairy Science* 99, 5640–5648.
- Alban L, Agger J and Lawson LG 1996. Lameness in tied Danish dairy cattle: the possible influence of housing systems, management, milk yield, and prior incidents of lameness. *Preventive Veterinary Medicine* 29, 135–149.
- Alsaad M, Römer C, Kleinmanns J, Hendriksen K, Rose-Meierhöfer S, Plümer L and Büscher W 2012. Electronic detection of lameness in dairy cows through measuring pedometric activity and lying behavior. *Applied Animal Behaviour Science* 142, 134–141.
- Archer S, Green M and Huxley J 2010. Association between milk yield and serial locomotion score assessments in UK dairy cows. *Journal of Dairy Science* 93, 4045–4053.
- Baggott D and Russell A 1981. Lameness in cattle. *British Veterinary Journal* 137, 113–132.
- Barnes A, Rutherford K, Langford F and Haskell M 2011. The effect of lameness prevalence on technical efficiency at the dairy farm level: an adjusted data development analysis approach. *Journal of Dairy Science* 94, 5449–5457.
- Berckmans D 2006. Automatic on-line monitoring of animals by Precision Livestock Farming. In *Livestock production and society* (ed. R Geers and F Madec), pp. 287–294. Wageningen Academic Publishers, Wageningen, the Netherlands.
- Cook NB, Rieman J, Gomez A and Burgi K 2012. Observations on the design and use of footbaths for the control of infectious hoof disease in dairy cattle. *The Veterinary Journal* 193, 669–673.
- Dórea F, Revie C, McEwen B, McNab W, Kelton D and Sanchez J 2013. Retrospective time series analysis of veterinary laboratory data: preparing a historical baseline for cluster detection in syndromic surveillance. *Preventive Veterinary Medicine* 109, 219–227.
- Guarino M, Norton T, Berckmans D, Vranken E and Berckmans D 2017. A blueprint for developing and applying precision livestock farming tools: a key output of the EU-PLF project. *Animal Frontiers* 7, 12–17.

- Horseman SV, Roe EJ, Huxley JN, Bell NJ, Mason CS and Whay HR 2014. The use of in-depth interviews to understand the process of treating lame dairy cows from the farmer's perspective. *Animal Welfare* 23, 157–165.
- Jabbar KA, Hansen MF, Smith ML and Smith LN 2017. Early and non-intrusive lameness detection in dairy cows using 3-dimensional video. *Biosystems Engineering* 153, 63–69.
- Lafrance J and Miller D 2010. Defining acute kidney injury in database studies: the effects of varying the baseline kidney function assessment period and considering CKD status. *American Journal of Kidney Diseases* 56, 651–660.
- Leach K, Paul E, Whay H, Barker Z, Maggs C, Sedgwick A and Main D 2013. Reducing lameness in dairy herds - Overcoming some barriers. *Research in Veterinary Science* 94, 820–825.
- Liu J, Neerchal NK, Tasch U, Dyer RM and Rajkondawar PG 2009. Enhancing the prediction accuracy of bovine lameness models through transformations of limb movement variables. *Journal of Dairy Science* 92, 2539–2550.
- Maertens W, Vangeyte J, Baert J, Jantuan A, Mertens KC, De Campeneere S, Pluk A, Opsomer G, Van Weyenberg S and Van Nuffel A 2011. Development of a real time cow gait tracking and analysing tool to assess lameness using a pressure sensitive walkway: the GAITWISE system. *Biosystems Engineering* 110, 29–39.
- NFACC 2018. Code of practice for the care and handling of dairy cattle. Retrieved on 13 November 2018 from [http://www.nfacc.ca/pdfs/codes/dairy\\_code\\_of\\_practice.pdf](http://www.nfacc.ca/pdfs/codes/dairy_code_of_practice.pdf).
- Pastell M and Kujala M 2007. A probabilistic neural network model for lameness detection. *Journal of Dairy Science* 90, 2283–2292.
- Pastell M and Madsen H 2008. Application of CUSUM charts to detect lameness in a milking robot. *Expert Systems with Applications* 35, 2032–2040.
- Pluk A, Bahr C, Poursaberi A, Maertens W, van Nuffel A and Berckmans D 2012. Automatic measurement of touch and release angles of the fetlock joint for lameness detection in dairy cattle using vision techniques. *Journal of Dairy Science* 95, 1738–1748.
- Poursaberi A, Bahr C, Pluk A, Van Nuffel A and Berckmans D 2010. Real-time automatic lameness detection based on back posture extraction in dairy cattle: Shape analysis of cow with image processing techniques. *Computers and Electronics in Agriculture* 74, 110–119.
- Sadiq MB, Ramanoon SZ, Mossadeq WMS, Mansor R and Syed-Hussain SS 2017. Association between lameness and indicators of dairy cow welfare based on locomotion scoring, body and hock condition, leg hygiene and lying behavior. *Animals* 7, 79.
- Schlageter-Tello A, Bokkers EAM, Groot Koerkamp PWG, Van Hertem T, Viazzi S, Romanini CEB, Halachmi I, Bahr C, Berckmans D and Lokhorst K 2015. Comparison of locomotion scoring for dairy cows by experienced and inexperienced raters using live or video observation methods. *Animal Welfare* 24, 69–79.
- Sprecher D, Hostetler D and Kanneene J 1997. A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. *Theriogenology* 47, 1179–1187.
- Tambuyzer T, Baschun D and Aerts J-M 2018. Towards individualised model-based monitoring: from biology to technology. PhD thesis, KU Leuven University, Leuven, Belgium.
- Van Hertem T, Maltz E, Antler A, Romanini CEB, Viazzi S, Bahr C, Schlageter-Tello A, Lokhorst C, Berckmans D and Halachmi I 2013. Lameness detection based on multivariate continuous sensing of milk yield, rumination, and neck activity. *Journal of Dairy Science* 96, 4286–4298.
- Van Hertem T, Viazzi S, Steensels M, Maltz E, Antler A, Alchanatis V, Schlageter-Tello AA, Lokhorst K, Romanini ECB, Bahr C, Berckmans D and Halachmi I 2014. Automatic lameness detection based on consecutive 3D-video recordings. *Biosystems Engineering* 119, 108–116.
- Van Nuffel A, Zwervaeagher I, Pluym L, Van Weyenberg S, Thorup V, Pastell M, Sonck B and Saeys W 2015. Lameness detection in dairy cows: part 1. How to distinguish between non-lame and lame cows based on differences in locomotion and behavior. *Animals* 5, 838–860.
- Van Nuffel A, Zwervaeagher I, Van Weyenberg S, Pastell M, Thorup V, Bahr C, Sonck B and Saeys W 2015b. Lameness detection in dairy cows: part 2. Use of sensors to automatically register changes in locomotion or behavior. *Animals* 5, 861–885.
- Vial F and Berezowski J 2015. A practical approach to designing syndromic surveillance systems for livestock and poultry. *Preventive Veterinary Medicine* 120, 27–38.
- Viazzi S, Bahr C, Schlageter-Tello A, Van Hertem T, Romanini CEB, Pluk A, Halachmi I, Lokhorst C and Berckmans D 2013. Analysis of individual classification of lameness using automatic measurement of back posture in dairy cattle. *Journal of Dairy Science* 96, 257–266.
- Viazzi S, Bahr C, Van Hertem T, Schlageter-Tello A, Romanini CEB, Halachmi I, Lokhorst C and Berckmans D 2014. Comparison of a three-dimensional and two-dimensional camera system for automated measurement of back posture in dairy cows. *Computers and Electronics in Agriculture* 100, 139–147.
- Weber A, Stamer E, Junge W and Thaller G 2013. Genetic parameters for lameness and claw and leg diseases in dairy cows. *Journal of Dairy Science* 96, 3310–3318.