University of Twente

Master thesis - CS

192199508

Detection of subclinical ketosis in dairy cows using behaviour sensor data

Author
R. Monshouwer

Committee
dr. M. Poel
dr.ing. G. Englebienne

Supervisors
dr. M. Poel
A. Harbers (Nedap)
R. Aly (Nedap)

November 12, 2020

UNIVERSITY OF TWENTE.

nedap

Acknowledgement

On a personal note, I would like to thank my supervisor Mannes Poel. He has guided me through this large project and gave good feedback when I needed it.

Also, I would like to thank the company Nedap, and especially Arnold Harbers and Robin Aly, for giving me the chance to do something so interesting. The weekly meetings were always helpful, productive and encouraging. They made me feel part of the company, also when it was required to work from home.

At last, I would like to thank all my family and friends for supporting me along the way in this long project.

Abstract

As precision dairy farming technologies become more common, new possibilities arise in the management of health in dairy cows. Ketosis is a transition cow disease, caused by the start of milk production after calving. Even at the subclinical level, it has an impact on milk yield, reproduction, and other cow diseases causing significant costs on a farm level. Currently, detecting subclinical ketosis by sampling blood is the golden standard, but this is infeasible on a large scale. With behaviour data captured with cow-mounted sensors, large-scale detection is possible. Initial studies to produce a detection model from behaviour data with machine learning have been performed, but a focus on all aspects of the machine learning methodology is lacking. By presenting a complete methodology and comparing time windows, normalisation, features and machine learning models, this study provides an exploratory view on using cow behaviour data to detect subclinical ketosis with machine learning. Using BHBA measurements from two on-farm experiments as targets and behaviour data as input, the detection models are compared. Additionally, a regression variant is proposed to produce a better estimation of subclinical ketosis. The behaviour data was windowed by taking all data relative to either calving-date or measurement-date and the measurement date based window was not significantly better. To evaluate the effect of farm-specific behaviour, herd normalisation of behaviour was compared to other normalisation techniques. The non-normalised values were significantly better with a mean AUC of 0.65. With static risk factors such as BCS, parity and dry period length as the baseline feature set, several feature sets derived from the behaviour data were compared. However, the behaviour features were not able to beat the static features. Comparing several standard machine learning models, Random Forest and Gradient Boosting performed best on the task. In the direct prediction of BHBA levels using regression, high levels of BHBA were not predicted accurately. Overall, the subclincal ketosis detection model using behaviour data of this study are not reliable enough, but a solid methodology is presented for future studies using machine learning to detect cow diseases using behaviour data.

Contents

1 Introduction

Modern dairy cows are adapted to produce high amounts of milk for to supply the world demand for dairy products. One of the problems modern dairy cows are facing today is the disease called ketosis. This disease is caused by the sudden change in energy demand, generally happening after parturition. Cows start giving milk right after calving and this creates a negative energy balance. The severeness of the negative energy balance cannot be countered and the cow suddenly shows less movement, eating, ruminating and less milk yield. At the subclinical level, there is a prevalence of 8% to 40 % on a farm during parturition on a global level [13] and the costs caused by ketosis are substantial [46]. While the detection of subclinical ketosis is proven effective using blood samples, this is found to be expensive and time-consuming on a large scale. Moreover, blood sampling is invasive to the cow. New methods are needed to detect subclinical ketosis to prevent further costs, treatment and improve well-being of the cows. The rise of Precision Dairy Farming technologies provides new ways to monitor cows on the individual level, especially on large-scale systems. Behaviour sensors mounted on the individual cow monitor the behaviour of cows in near real-time. This allows systems to be developed that are able to correlate the behaviour of the cow to certain dairy cow diseases, such as ketosis.

1.1 Problem statement

Early behavioural studies showed a correlation of decreased eating and rumination time around calving when cows were suffering from ketosis [31, 37, 68, 39]. Other behaviour parameters such as lying time and activity have also been related to ketosis[37, 68]. To implement a system that can detect subclinical ketosis, there needs to be reference values to those behaviour parameters that can distinguish healthy cows from ketotic cows. As it turns out, cow behaviour has a high variation between cows and this makes developing a system of reference values hard. Therefore, we can utilise machine learning techniques to learn those reference values by the system itself. Machine learning can find subtle differences and are able to generalise to build a robust prediction system. Already, some initial studies in detecting ketosis have developed methods using machine learning with behaviour data. However, these studies are either limited by data set size or lack in the application of important parts of the machine learning methodology. To date, these studies have not used normalisation techniques, experimented with different window sizes, or compared different features. Moreover, since the threshold for subclinical ketosis is disputed[22], predicting BHBA directly with regression can deliver more precise predictions for cows. Therefore, there is a need for more exploratory research in subclinical ketosis detection using behaviour data, focused on every step in the machine learning methodology.

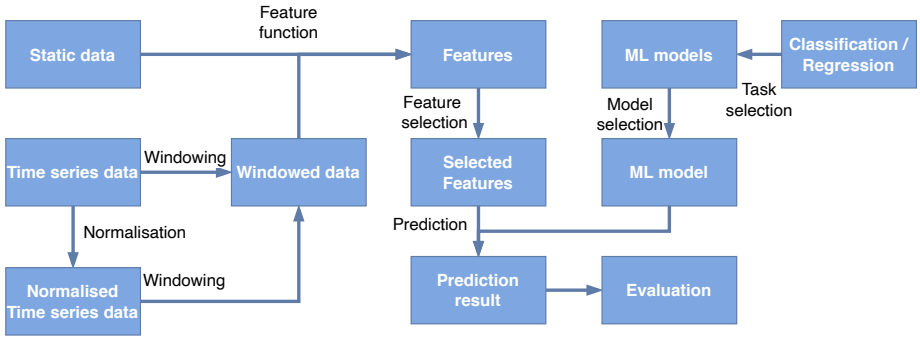


Figure 1.1: Different steps in (subclinical ketosis) classification with time series and machine learning. Each block represents data and each arrow represent a transformation.

1.2 Research question

That leads us to formulate to the main question for this research.

> To what extent can subclinical ketosis be detected in dairy cattle using peripartum data in a machine learning approach?

This study aims to contribute in different areas of the steps involved in ketosis detection using machine learning. These steps are defined in Figure 1.1. They are largely based on the fundamental steps of supervised machine learning. The key areas of this research are around windowing, normalisation, feature functions, machine learning models, and classification compared to regression. These areas are investigated to explore different approaches for SCK detection.

How do a time-window based on the day of measurement of BHBA and a time-window based on the day of calving compare with respect to the quality of ketosis detection? is the first sub question for this research, as the state-of-the-art have never used time windows relative to the day of calving. They have exclusively used time-windows based on the day of measurement of BHBA. However, the literature suggests that ketosis related behavioural changes also happen at an earlier moment in time and ketosis has a strong correlation to the period around calving. Therefore, a calving-date based time window is introduced and compared to the measurement-date based time window.

How does normalisation on herd level compare to other normalisation methods and non-normalised data? is the second question of this research, as farm management varies and this has an impact on the behaviour of the entire herd of cows. The behaviour of a cow is different every day and different for every cow and this is caused by a lot of different factors. To be able to distinguish normal variation in behaviour from abnormal variation caused by SCK, a method to minimise normal variation is needed. Normalisation is a standard practice within machine learning to achieve this. As this study uses data from several different farms, one way to apply normalisation is to normalise to the entire herd. This eliminates changes that caused the entire herd to behave differently, for example when there is a visit from a veterinarian. This herd-normalisation is compared to other individual normalisation methods and non-normalised data.

What is the effect of different feature sets on subclinical ketosis detection? is the third question for this research. State-of-the-art have used mostly raw data from behaviour sensors as features for their ketosis detection models. This raw data is either steps per day or the total time (eat-

ing/ruminating/lying,...) per day. However, raw values are susceptible to daily variation and result in overfitting. Moreover, the literature suggests that SCK is sustained over multiple days. Therefore, features can capture multiple days have the potential to improve the detection performance. In this study, the raw behaviour features are compared to statistical features, trend features. In addition, a new promising method of quantifying regularity in the prepartum period is added as features to compare with raw values.

What is the performance of different machine learning models in detecting subclinical ketosis is the fourth question. Given a set of machine learning models, we want to find the best model for ketosis detection. In the previous attempts at ketosis detection found in the literature, usually a set of basic supervised learning models were used and compared. As some of these models work better under different circumstances, this study also evaluates and compares different models. Like related work, the set of models is restricted to a few different basic machine learning models, as this is an exploratory research.

How precisely can BHBA values be predicted using a regression model and how does this model compare to the classification model? is the last sub question for this study. The classification in this study is performed by transforming the measured BHBA levels in healthy and sick cows by applying the standard SCK threshold value (1.2 mmol/L). However, this threshold value for SCK is disputable[22] and might mark healthy cows incorrectly. The reaction to certain levels of BHBA in blood differs per cow. Instead, if an estimation can be made on the value of BHBA it can provide a finer grained decision on how to treat a cow. Regression, as opposed to classification, is able to predict continuous values, so BHBA values can be directly predicted using regression. Moreover, the predicted BHBA values can be transformed into a classification result. By utilising the standard SCK threshold, a direct comparison between regression and classification is possible. Therefore, a set of regression models is trained on the same data as the classification models and evaluated on regression-specific metrics and evaluated on classification metrics.

The five questions introduced above are summarised below. Their place in Figure 1.1 is highlighted in bold.

1. How do a measurement-date based time-window and a calving-date based time-window compare with respect to the quality of ketosis detection? **(Windowing)**
2. How does herd-normalisation compare to other normalisation methods and non-normalised data? **(Normalisation)**
3. What is the effect of different feature sets on subclinical ketosis detection? **(Feature function)**
4. Which machine learning model has the best at detecting subclinical ketosis? **(Model prediction)**
5. How precisely can BHBA values be predicted using a regression model and how does this model compare to the classification model? **(Regression)**

2 Background

In this chapter, information is given about the dairy cow disease called ketosis, the risk factors for ketosis and the current state of precision dairy farming. Furthermore, fundamental background information is given on time series, normalisation, machine learning and evaluation as they form the building blocks of this study.

2.1 Ketosis

Ketosis or hyperketonemia is a disease mainly seen in the transition period of dairy cattle. This transition period is defined as the three weeks before and after calving [21]. During the transition period, cows experience a negative energy balance, weight loss, hypocalcemia and reduced immune function [40]. The transition period is a vulnerable and critical period for dairy cows [38]. Around 30 to 50% of dairy cows develop metabolic or infectious diseases around calving [40]. The negative energy balance occurs when the cow milk production is increasing, but its feed intake does not increase accordingly [1]. This early lactation stage has the highest disease incidence of the lactation-gestation cycle [35]. When energy and nutrient consumption of the cow is lacking, the probability increases of developing clinical or subclinical diseases including ketosis, hypocalcemia and metritis [57]. These diseases relate to the process of metabolising body fat [53], categorised as metabolic disorders [66]. Ketosis, defined as an excessive amount of ketone bodies in the blood, can present itself as clinical ketosis (CK) by a visible decrease in appetite, uncoordinated movement, acetonic breath, weight loss and a decrease in milk production. It can also present itself as subclinical ketosis (SCK), defined as having excessive amounts of ketone bodies in blood without the visible (clinical) signs mentioned. In its subclinical and clinical state, ketosis has been related to decreased milk yield and increased chances for other fresh cow disorders [20, 33, 22, 43, 54, 6]. For example, cows diagnosed with SCK had 6.1 more chances of displaced abomasum [43], a disease in which treatment is more costly.

Prevalence The prevalence of SCK differs in literature. It ranges from 1.8% up to 55% [23, 45]. The study of McArt et al. [43] found 43% of the cows were diagnosed with SCK, with the peak prevalence (70%) at 5 days in milk. Duffield et al. [23] reported a prevalence of SCK between 8.9% to 34%. More studies found its peak prevalence of SCK within the first month of lactation [52, 23]. On a global level, SCK prevalence ranged from 8.3% to 40.1% [13]. Differences in prevalence are attributed to study methodology, regional differences and differences in parity distribution per study [77]. These studies show that SCK occurs on many dairy farms, most prominent in the first weeks of lactation of a cow.

Costs Costs of ketosis depend on incidence rates, treatment costs and milk price. The majority of the costs are caused by reduced reproduction rates and loss of milk yield [22, 46]. Reproductive efficiency is the key to a profitable dairy farm [63]. McArt et al. [46] estimated the cost of SCK at $289 per case, accounting for associated diseases such as metritis and displaced abomasum. Others claimed that costs involving the disease could be accounted for twice and they provided a recalculated estimation of the costs [57], at a total cost of €257 per case.

Treatment cost consist of diagnostic costs, therapeutic costs, labour costs, and these costs are estimated at $6 (2015, [46]) and €22 (2015, [57]). The large difference could be explained as the latter study considered CK treatment costs. However, both did not factor in extra labour costs. Therefore, the direct costs of SCK are low compared to the indirect costs of SCK.

The impact of SCK is also visible on an ecological level, as agriculture is responsible for a significant amount of greenhouse gasses (9.8%, [15]). The effect of impaired reproductive performance and reduced milk yield also contributed to the greenhouse gas emissions [49]. This study showed that SCK and related diseases contributed to a 2.3% increase in greenhouse gasses.

Treatment Treatment can be applied in a proactive or reactive manner. Proactive treatment is meant to prevent the development of ketosis and consist of specific feed intake during the dry period. However, the success of preventive treatment is variable [23]. Reactive treatment is applied after diagnosis of (subclinical) ketosis. Studies show that after (veterinary) treatment of SCK, most cows recover in 5 days time [45]. Treatment also reduces the risk of clinical ketosis and increases milk yield [45]. Therefore, detection of SCK in an early state can minimise the costs involved [32].

Diagnosis Clinical ketosis in cows is often noticed by the farmer, shown by reduced eating time, weight loss, uncoordinated movement, acetonic breath and decreased milk yield [2]. Near-zero feed intakes have been reported on the day of diagnosis [32]. The clinical diagnosis is usually between ten days to three weeks after calving [30].

In SCK, invasive tests have to be used to diagnose ketosis reliably. By definition, there are no visible signs of illness and farmers cannot notice cows with SCK. Presence of ketone bodies in blood indicates a negative energy balance, signalling ketosis at high levels [30, 1]. From the three ketone bodies present, Beta-hydroxybutyrate (BHBA) has become the standard detection method of ketosis [47]. BHBA can be measured in urine, milk and blood [71]. Blood is reported to be the most accurate [71] and it is considered the gold standard method [47]. This gold standard is measured in the laboratory, but more convenient cow-side tests are available [47]. The definition of SCK based on BHBA concentrations is based on a certain threshold value. The cutoff threshold commonly used is 1.2 mmol/L, based on distribution skewness of this value [52]. However, this value has been claimed as arbitrary [22]. Care is needed when measuring BHBA, as variations throughout the day have been reported [41]. Blood sampling is considered invasive and time-consuming [76]. Therefore, in practice blood sampling is considered unfeasible on large scale [81] and new types of tests have the potential to detect SCK on a large scale.

2.1.1 Detection or Prediction

In identifying cows with ketosis, there is a distinction between detecting ketosis and predicting ketosis based on the relative time of decision making. Currently, the detection of ketosis is defined

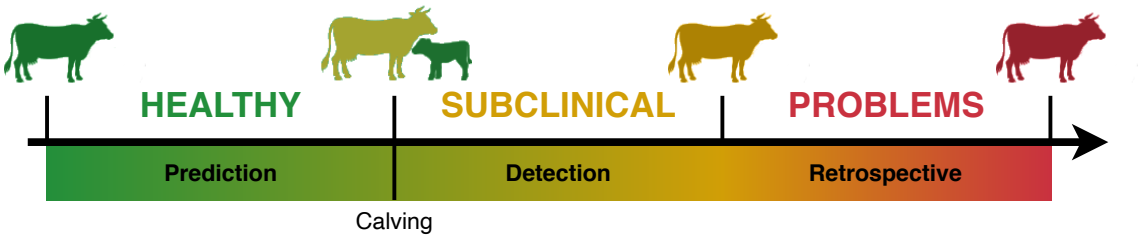


Figure 2.1: Timeline of the health status progression in the transition period of a cow with ketosis. On the time axis, three types of analysis are related to stages of ketosis. This figure shows that the progression of SCK is a gradual process and thus prediction and detection have a dependency on time.

as finding high BHBA concentrations in blood samples of the cow. Prediction of SCK is when there is a reliable forecast about the future state of SCK of the cow, which can only be verified afterwards. This prediction can be used as an early warning signal and stimulate preventive treatment. In the retrospective scenario, other diseases caused by ketosis have been diagnosed and costs have been made. In terms of treatment costs, the retrospective scenario is considered the most expensive, followed by detection. Prediction is considered the least expensive as an early response may benefit treatment [32], which proves the added value of prediction over detection. The separation over time of detection, prediction and retrospective analysis in SCK is illustrated in Figure 2.1.

2.2 Non-invasive measurable risk factors of SCK

Several studies have shown that other cow measurable parameters are related to SCK [31, 37, 16, 62, 68, 39, 67, 65, 38, 26, 51]. These can be categorised into behaviours like eating or ruminating and other parameters like milk yield. This distinction is based on the type of sensor involved, where the behaviour can be measured using a similar sensor.

2.2.1 Behavioural risk factors

Behaviour of an animal is considered as a good indicator of its physiological state [29]. Rumination behaviour, eating behaviour and the amount of activity are seen as signs closely related to health and productivity [7].

Eating Eating behaviour consists of feed intake, duration, and eating repetitions [19]. Postpartum diseases are frequently related to eating behaviour during the transition period [31]. Metabolic disorders such as ketosis have been reported to have a great impact on eating behaviour [19]. The study of Goldhawk et al. [31] and Itle et al. [37] reported up to 28% reduction in eating time compared to healthy animals in the postpartum period leading to diagnosis. In the prepartum period, eating time is also reported to be reduced [16]. The type of feed has been identified as a risk factor, with less quality feed 1.5 higher odds of SCK compared to high quality feed [6]. Specific eating behaviour is also found in ketotic cows. Sahar et al. found ketotic cows eating less in the

first 90 minutes after fresh feed delivery [62].

Rumination Cows need to ruminate to acquire nutrients. This is a vital process for a cow, so differences in rumination behaviour can suggest a disorder. Like eating, rumination behaviour is summarised as total rumination time, rumination spouts(repetitions) and rumination duration. Most of the studies investigating the association between rumination and ketosis used total rumination time as a behavioural indicator. Stangaferro et al. [68] reported a decrease in rumination time in the period of 5 days before diagnosis up to the diagnosis day. Decrease in rumination time was also reported in the study of King et al. [39]. Other studies reported a rumination time decrease in ketotic cows in the week prepartum [67], especially for multiparous cows [65, 38]. Soriani et al. [67] even negatively correlated rumination time with BHBA levels, suggesting that rumination time has a direct relation with SCK.

Lying Due to the role of lying in rumination [19], it is considered a vital part of dairy cow behaviour. Lying amounts to a large part of a cows daily activity and cows prefer lying over eating and social behaviour [50]. Deviation in prepartum lying behaviour have been associated with ketosis, Itle et al. [37] showed that ketotic cows show reduced lying time in the week prepartum.

Activity The activity of a cow is usually measured as the daily step count of a cow. Activity has also been associated with ketosis. A study of Stangaferro et al. noted decreased activity up to five days before diagnosis [68]. Similar reductions in activity have been seen in other studies [42, 26, 39, 51]. However, questions have been raised on whether decreased activity is a consequence or a cause of ketosis [51].

2.2.2 Other risk factors

Other factors have been identified to associate with SCK. Body Condition Score (BCS) is a visual assessment of the ratio of body fat of a dairy cow. A high BCS was identified as a risk factor to develop ketosis [77, 58, 44]. Locomotion scores, visual assessments of the mobility problems of a cow, are associated with SCK [14]. Especially higher lameness scores show higher levels of BHBA. Parity is also considered an important factor for SCK. Compared to primiparous, multiparous showed higher prevalence of SCK [77, 6]. Time of year due to different feeds has also been related to ketosis [77]. Milk parameters such as colostrum yield, previous lactation length and dry period length were considered increased risk for SCK and CK [77]. Milk yield was also reported to be reduced before diagnosis of subclinical ketosis [39].

2.3 Precision dairy farming

Precision dairy farming is defined as the technology to support or replace the farmers' observation of their livestock [61]. Its purpose is to monitor the individual behaviour of cows in close to real-time [5]. This is especially important in high-value livestock, such as dairy cows [78]. Using various individual cow sensors, data is gathered and processed by cow management systems. These systems provide farmers with detailed information of each individual cow. Farming industries' interest in precision dairy farming has increased over the last years [7]. This individual monitoring

has become increasingly important, since cows per dairy farm are increasing, but employees per farm are not increasing accordingly [51, 7]. Reducing workload and increasing efficiency using precision dairy farming are the largest factors of interest for farmers [7]. Its potential accuracy of livestock production management is unprecedented and allows fine-grained management of the farm [78]. However, care must be taken with the monitored parameters, as they are biological and inherently unpredictable [67].

Sensor types In the last years, there has been a growth in the implementation of multiple devices to monitor behaviour and physiological parameters [61]. Pedometers were the first sensors to be applied to dairy farming [24]. Electrical conductivity and milk colour sensors in milk systems are used in mastitis detection. Several types of boluses are also available: measuring rumen pH, temperature or detecting oestrus [61, 24]. Special rumination sensors are available, for detection of rumination via a microphone sensor [64]. These sensors belong to the wearable sensor category. Other wearable sensors are tri-axial accelerometers mounted in the neck, leg, ear or tail of the cow. Whereas the neck, leg or ear sensors capture general behaviour, there are tail sensors specifically designed for calving prediction. Wearable sensors which capture the general behaviour of a cow (eating, ruminating, lying, or walking) are widespread and validated in several studies [75, 12].

These sensors are often integrated into cow management software. This system uses the data from the sensors and produces alerts whenever an event is detected (oestrus) or some behaviour is irregular (reduced eating time). The behavioural alerts are based on threshold differences from the cows past behaviour. The event detection mostly uses binary targets [82], for example, oestrus or no-oestrus, without providing insight on the certainty of this event. This leads to many false alerts.

Usage The usage of precision dairy farming was categorised by Rutten et al. [61]. It consists of 1) (sensor) measurements, 2) classification of measurements, 3) integration with other sources, 4) decision making. On Dutch farms, the adoption rate of activity sensors for oestrus detection is around 20% [69]. Other sensor technologies are less adopted [60]. New development in precision technology presents a unique problem in case of subclinical detection. The farmer must act on alert without any clinical signs [68]. Studies show that the output of sensor data is already intensively used [58]. Observation of cows behaviour has seen increased importance in the early identification of a disease [19].

Usefulness Surveys show that farmers want affordable technology with a clear cost-to-benefit ratio [10]. Many systems provide farmers raw data or indicators that are not clearly related to disease or treatment [61]. The value of unprocessed data is limited [29]. Alerts based on behaviour are not sensitive enough, a lot of alerts are false positives. They must be improved and produce an actionable change for farmers [25]. Although sensitivity is lacking for alerts and other new functions, it can still be useful when no other investments have to been made [25].

2.4 Time series

Time series is a type of data in which data points are bound together by time. The value at time 2 is dependant on the value of time 1. This type of data is commonly seen in climate data (temperature over years), financial data (stock market values over time), and sensor data. Early machine learning

on time series concentrates on time series forecasting: the prediction of future values using value from the past. This means forecasting the temperature for the next ten years or predicting the stock market value of tomorrow.

Classification of time series instead of forecasting has seen recent interest as wearable sensors are integrated in modern society. The classification of tasks based on time series is not concerned about what values are next, but rather what patterns can be extracted to distinguish two or more types of classes[3]. For example, a sensor may be placed on a human subject and the sensor data over a day is captured. This data can be used to classify the different types of movement this person did during the day (walking, sitting, or lying) by learning specific patterns in the time series data. The prediction of movement types is called time series classification and common methods to classify time series are to compare whole segments of time series to each other, find small unique patterns for each class, or extract higher-order features from the time series[3]. Among the popular algorithms, Dynamic Time Warping has achieved success in time series classification and it is often used as a baseline in time series classification tasks[3]. Recent advancements in deep learning have also enabled to improve on time series classification, with the availability of enough data[36].

This study aims the detection of SCK using behaviour data and this is a time series classification task, because behaviour data is time bound and behaviour data from different cows has to be classified as either healthy or SCK. For example, Figure 2.2 shows the hourly values of the eating time of cows A and B. This figure of a three-day period shows the time correlation of eating: the eating time of one hour depends on whether the cow has eaten before. The SCK detection model would use this behaviour data and predict whether or not cows A or B are healthy or sick.

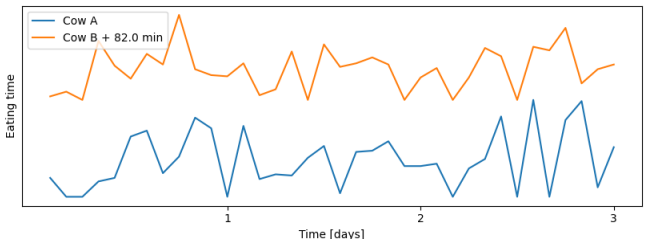


Figure 2.2: Eating time of two cows in a three day period. Both cows show lower eating time around midnight, indicating a day-night pattern.

2.5 Windowing

Time series are often large, windowing is a way of limiting time series data into smaller parts. By taking a certain reference point in the time series with a certain window size, it creates a different subset of the time series data. Often, a single value in a time series is not informative enough. Therefore, by taking multiple values in the neighbourhood, there is more information to be found. For example, in time series forecasting, a single stock value contains only small information about the stock value in the future. However, taking the last 30 days into account, we can calculate a trend from the past 30 days to make a better prediction.

Taking account of the last 30 days can be done at any point in the time series. If this procedure is performed on a data set containing a full year, we can calculate the stock price for every next day by “sliding” our operation over the time series. This procedure in this case is called a sliding

window procedure. The sliding window is a common method when there is a need for a continuous output at every point in time. The sliding window enables to look back time steps in the past to make a more accurate prediction. Often this is enough to make a reliable prediction without needing the entire time series.

In time series classification, the sliding window approach is also used when there is a label for each point in time. When a prediction is only needed at a single point in the time series, windowing can be used to limit the amount of data. Limiting the data might be necessary to reduce the amount of dimensions, reduce noise or when samples in a data sets have different lengths in their time series. In this study, we do not have label for each point in time, so we use the single point windowing to limit the amount of data per cow.

2.6 Normalisation

Normalisation is the technique of unifying the numerical scale of different features to reduce the impact of large-scale features and reduce the effect of external factors. For example, in loan application data, the age of a person is much smaller than the net income of this person. The numerical value of the net income is much larger. To avoid this problem, we can apply normalisation. This scales all data points to the same order of magnitude. The two most common techniques in normalisation are *min-max scaling* and *z-score normalisation*.

Min-Max scaling maps all data points to the range of [0,1] by utilising the maximum and minimum of each attribute. More specifically, let min_i and max_i be the minimum and maximum of attribute i, then the min-max scaling for the jth data point x_i^j is defined as follows

$$y_i^j = \frac{x_i^j - min_i}{max_i - min_i} \tag{2.1}$$

The problem with this approach is that outliers have a lot of impact on the scaling.

Z-Score normalisation does not have this problem. It takes the mean μ_i and standard deviation σ_i of attribute i and transforms the jth data point x_i^j as follows

$$y_i^j = \frac{x_i^j - \mu_i}{\sigma_i} \tag{2.2}$$

2.7 Time series features

To derive features from time series, the naive method is to use all values within the window of the time series. Depending on the window size, this might be small enough to be used as features itself, or it might be too much data. "Raw" values can be too specific, because they are bound to a specific time.

An alternative is to aggregate the time series data to calculate other features. Descriptive statistics like mean, variance, minimum and maximum are statistics that can be calculated on a time series.

We can also fit a polynomial function to the time series, like a linear fit. Such a linear fit can return a slope which tells us whether the values are increasing, stabilising or decreasing over time.

2.7.1 Time representation

In a time series, the time/date is often also a feature. However, the standard representation for time does not account for the cyclic nature of time itself. For example, on a number line, the time 00:00 is very far from 23:00, while it only is an hour apart. This makes it difficult for machine learning models to notice their close relation. Therefore, we introduce another representation for time by extracting it into two features with a sinus and cosine transformation (see Equations 2.3 and 2.4). In these formulas, x is the time variable in a zero-based interval $0 \leq x \leq X$. This transformation works for every cyclic (time) variable (weekdays, month, quarters, hours, seconds).

$$\text{sine}(x) = \sin\left(\frac{2\pi \times x}{X+1}\right) \tag{2.3}$$

$$\text{cosine}(x) = \cos\left(\frac{2\pi \times x}{X+1}\right) \tag{2.4}$$

The results of this transformation are points in a 2-dimensional space. An example of such a transformation is given in Figure 2.3. This figure shows the transformation of the months of the year. On an ordinary number line, January is at the far left while December is at the far right. After the transformation, January is close to December in the 2-d space as shown in the figure.

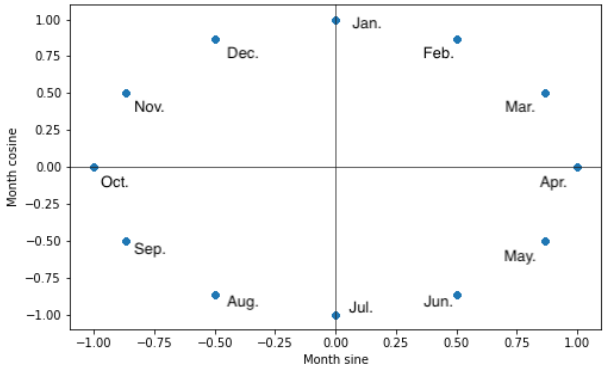


Figure 2.3: Month represented by sine and cosine transformation

2.7.2 Nonperiodicity and Fast Fourier Transform

The study of van Dixhoorn et al. [74] introduced a new statistic for cow behaviour named non-periodicity. It resembles a measure of regularity in the daily pattern of the behaviour of a cow. Their assumption was that cows with more regular behaviour have better capacity to undergo the transition period around parturition. Nonperiodicity is defined as the mean squared error between

the autocorrelation of the time series and a sinusoid with a 24 hour cycle and an amplitude of 0.25. The 24 hour cycle was chosen to represent the diurnal cycle of the behaviour of cows. The amplitude of 0.25 seems arbitrary and follow-up questions to the original authors revealed that this value was chosen because it had the best fit to their test subjects.

$$\text{Nonperiodicity}(x) = \frac{1}{T}\sum_{t=1}^{T}\left(\text{autocorr}_1(x_t) - 0.25\sin\left(\frac{2\pi t}{u}\right)\right)^2 \tag{2.5}$$

where `autocorr` is the autocorrelation function with one-step lag. Autocorrelation is visualised for a cow in Figure 2.4 together with the sinoid with the parameters mentioned.

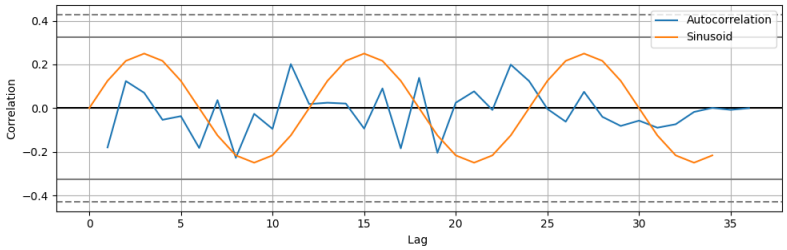


Figure 2.4: Autocorrelation plot of eating time of a cow and a sinusoid with a period of 24 hours and an amplitude of 0.25. The nonperiodicity (mean squared error between these two time series is high (0.17) van Dixhoorn et al. [74]

The nonperiodicity is closely related to other techniques to analyse frequencies in a signal. When there are periodic signals in a time series, we can apply Fast Fourier Transform (FFT) to analyse the frequencies of these signals. These frequencies or the strength of the frequencies can then be used as features. An example of a FFT of the eating behaviour of a cow is shown in Figure 2.5. To limit the number of frequencies, a threshold line is calculated for each individual FFT by the following formula

$$\text{Peak threshold} = q75 + 3 * (q75 - q25) \tag{2.6}$$

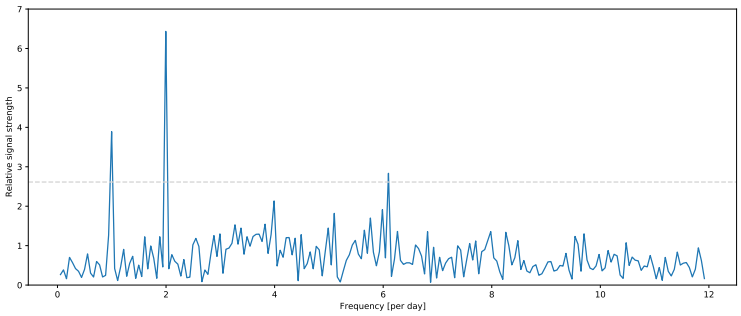


Figure 2.5: Fast Fourier Transform of eating behaviour of a cow over a 18 day period. The dotted line is the peak threshold line. Peaks occur at 1 time, 2 times and 6 times a day.

This formula is derived from the standard outlier definition which is a box plot ($median \pm (q75 - q25)$ As the variation within the FFT is high, the threshold is set at a high level. From this FFT, we can calculate the following features.

- Strength at frequency 1, 2, ..., 6 times per day. Peaks are rounded if slightly higher or lower than the integer value. This value is 0 if below the peak threshold
- Mean and variance of all frequencies
- Highest strength value
- Frequency with the highest strength
- Total strength
- Total number of frequencies with a peak above the threshold
- Average strength of all nonzero frequencies

2.8 Machine learning models

In this study, four supervised machine learning models are used: Random Forest, Gradient Boosting, Multilayer Perceptron and Random Forest. They are basic classical machine learning models, which suits the exploratory aspect of this study.

The Random Forest (RF) [72] classifier consists of an ensemble of decision trees. Multiple trees are created by taking a random subset of the training data and fitting a decision tree, a technique called bagging. These trees produce a vote in which the majority wins. The splitting criteria in these trees are based on the seen training data. In this study the training data is similar to testing data, the random forest classifier has good initial performance. Common hyperparameters for Random Forest to be tuned are the number of trees, maximum number of features per tree, maximum depth of each tree, the minimal samples per split and the minimum samples per leaf.

Gradient Boosting (GB) [27] is an ensemble, like Random Forest, of decision trees. The difference is the building process of trees. Instead of taking a subset of data for each tree, an initial tree is fitted to the data set. Then, using the gradient of a predefined loss function, a new tree is trained on these gradients. The Gradient boosting hyperparameters to be tuned have overlap with the Random Forest, as both use decision trees as the base learner. However, the hyperparameter values cannot be shared as the learning process is different.

The Naïve Bayes (NB) [28] classifier is based on the conditional probability of features explaining the target under the assumption of mutual feature independence. For a few number of features, this assumption might hold and if the conditional probability is not uniform, the Naive Bayes classifier also provides good performance without any configuration. This classifier fits the static feature set well, as features are scarce and mostly independent. For continuous data, a Gaussian probability model is needed to predict unseen values. The Naive Bayes does not have any hyperparameters to be tuned.

The Multilayer Perceptron (MLP) model is a feed-forward neural network [8]. Using a nonlinear activation function, MLP is able to learn complex functions, which makes it a good classifier algorithm. The optimiser used for this study was the Adam optimiser. Common hyperparameters to be tuned are learning rate, hidden layer size, and activation function.

Hyperparameters Hyperparameter optimisation is the process of finding the best set of hyperparameters for a machine learning model. In contrast to the normal parameters of the model, hyperparameters cannot be learned during training. Some machine learning models have several hyperparameters (e.g., the number of trees and the subset of features per tree) or none (e.g. k-Nearest Neighbour). The goal of the hyperparameter optimisation is to find the set of parameters which minimises a predefined performance metric in cross-validation. The optimal way is to perform a grid search over a manually defined hyperparameter space. When this space is large, it takes a lot of computing time to find the optimal set of hyperparameters. Another approach is to randomly sample from this hyperparameter space, limiting the number of samples to be taken. Both approaches are easily executed in parallel, because the hyperparameters values are independent of each other. Other approaches exists, such as Bayesian optimisation, gradient-based optimisation and evolutionary optimisation. These approaches are deemed too complex to fit into this research. Random hyperparameter search is the chosen approach for hyperparameter optimisation of machine learning models.

2.9 Evaluation metrics

2.9.1 Classification metrics

Classification can be evaluated by a number of statistics. The basic evaluation process is by comparing the test values and the predicted values. From these, we derive a confusion matrix containing the number of true positives, false positives, true negatives and false negatives (in the case of binary classification). Precision is the percentage of true positive values in the set of all predicted positive values. Recall is the percentage of true positive values in the set of all positive values. This is also called sensitivity. Specificity is the percentage of true negative values in the set of all negative values. Accuracy is the percentage of all correctly predicted values in the set of all values.

Most supervised machine learning models can also produce a probability value instead of an actual binary prediction. Then these probabilities are transformed into the positive or negative class using a threshold value, with 0.5 as the default threshold. If we would change this threshold value, the confusion matrix would change. If this threshold is 0.8, the number of predicted positive values are lower and the number of predicted negative values are higher. We can vary the threshold in such a way that we can get 100% Sensitivity or 100% Specificity and everything in between. The curve that follows is called the ROC-curve. This curve contains information of what Sensitivity score we can get at a certain Specificity score and vice versa. This allows for fine-grained control over the the trade-off between the number of predicted positives and negatives, especially in an imbalanced data set. From this curve, the area underneath can be calculated and this number is another metric called the Area under the ROC-curve(AUC).

In the same fashion, we can create a Precision-Recall curve by adjusting the positive threshold. Then the result is a function of Precision at a certain recall, or vice versa.

2.9.2 Regression metrics

Common regression metrics include mean squared error, mean absolute error and R^2. The mean squared error(MSE) is the squared difference of the predicted value and the actual value averaged over all predictions. The mean absolute error(MAE) is the absolute difference of the predicted value

and the actual value averaged over all predictions. The R^2 score is the *coefficient of determination* which is representation of the proportion of explained variance in the predictions. The best value is 1 and a value of 0 means the model always predicts the expected value over all values (mean).

In a data set where outliers are important, the MSE and MAE are not able to capture the performance of those outliers, as they are both the average scores. Rather, by analysing the distribution of all errors, we can derive a useful quantity to quantify the performance on outliers. For instance, we can take the 90% percentile of the errors to see what the error is at the higher errors. If a model is performing well on the well-represented numbers, but worse on the outliers, than this percentile includes those errors.

3 Related work

Combining the sensors from precision dairy farming and the known behavioural deviations of cows related to SCK presents new possibilities in the detection of SCK. In fact, a number of studies already have used precision dairy farming technologies to detect SCK and also in other related problems. This related work is split into two categories. The first category consists of all research conducted on the detection or prediction of ketosis. The second category consists of other research deemed similar enough to provide insight into the data, methods and outcomes.

3.0.1 Search procedure

On February 26 2020, a literature search on the Scopus database has been performed using the following query:

> (Prediction OR Detection OR Identification) AND (ketosis or hyperketonemia) AND (activity OR behaviour OR behavior) AND dairy AND (cattle OR cows)

This initial search resulted in 28 papers. A second search step was executed to include relevant research on ketosis and precision dairy farming. This process consisted a forward citation and backward reference search and selecting the relevant papers based on the presence of terms in the title or abstract. These terms were based on similar disease(ketosis), data(behaviour) or objective(detect or predict ketosis). Some of these papers were unavailable, non-English or had an region specific source. Finally, 11 papers have been identified as related work. An overview of these papers is presented in Table 3.1.

3.1 State-of-the-art on ketosis detection

As technological advancement have been predicted to include automatic monitoring of metabolic health [55, 19], initial studies into ketosis detection have been performed. Background on the methods and models used in these studies can be found in Chapter 2.

A study by Eckelkamp [25] used activity and behaviour to improve the performance of disease alerts using machine learning models, under the assumption that alerts signalled a disease such as ketosis. Using data from a validated commercial behaviour sensor, several machine learning model have been tested in a large field study of 1374 cows (4 farms), of which 213 cows were diagnosed

with SCK. This study used 30% decrease from a 10-day moving average in eating, lying or steps as features. As it only considers daily totals, the features do not capture any information within daily pattern of a cow. The best performance this study produced reached 83% sensitivity and 83% specificity with a Random Forest model in a time window of 5 days before to the day of diagnosis.

The study of Stangaferro et al. [68] used a proprietary score to predict ketosis. The method of scoring was not published and this research is only reproducible with proprietary sensors. They also used a validated commercial behaviour sensor that produced a Health Index Score (HIS). This HIS was used to signal ketosis and other metabolic disorders. The time window of HIS consisted of daily score in a period of 5 days before to two days after diagnosis. The study was performed on 1080 cows, of which 54 cows were diagnosed with SCK. This study reached 91% sensitivity, but did not report a specificity score.

Ushikubo et al. [73] have used several machine learning models to detect subclinical ketosis in an early stage. Data such as feed intake, several milk parameters and activity were as features. Using feature selection techniques, only variance of activity was used as input to the machine learning models. Their ketosis target was defined as 0.1 mmol/L BHBA measured in milk, which has been showed to be an unreliable method of testing. This study compared daily data from two days and three days before diagnosis and therefore this study produced a predicting model. However, the analyses was still performed retrospectively, which means that in production use could give a lot more false positives, and thus real-world application is limited. Using several different feature selection methods and machine learning models, their best performance (using SVM) was 93% sensitivity using a Support Vector Machine model, but they did not report specificity.

Bonfatti et al. [9] used infrared spectroscopy to predict BHBA levels. This new technology have potential to be integrated into the milk systems to quickly detect ketotic cows. They used blood samples as reference, but only reached 76% sensitivity and 82% specificity, which was labelled as too low to be valuable.

Paudyal et al. [56] compared rumination on the day of diagnosis to the (herd) average rumination of 3 to 5 days before diagnosis and transformed these values into a cow index system. If one of these values passed a certain threshold, the cow was marked as sick. Cows were grouped based on days in milk. These threshold values were determined graphically using a Receiver Operating Curve (ROC) on the entire data set. No separation between training and test sets was made. Sensitivity and specificity reached both 80%.

Xu et al. [81] presented a metabolic status indicator based on a k-Means clustering and predicted those statuses using a combination of milk data, Dry Period Length (DPL), Body Weight (BW) and Parity (PAR). This milk data was sampled in a weekly frequency and at the end of the week, the data was used as features. The metabolic status of poor was related to ketosis, but predictions on poor status did not reach high performance and were omitted from the original paper. In the supplemental paper, the performance on all statuses reached an error rate of 24% with a Random Forest Model.

Steensels et al. [70] used rumination, activity and milk data together with logistic regression to detect ketosis. The data was sampled on a daily frequency in the period of 4 days before to 2 days before diagnosis. Their targets were based on urine samples. Their model reached a performance of 70% sensitivity and specificity. All related work used short time windows, up to 10 days, but literature suggests changes related to SCK also happen earlier. No study used combination of all the different behavioural parameters.

Early detection of ketosis may benefit treatment response and stop the progression into other

diseases [68, 62, 80]. This is illustrated in Figure 2.1. An improvementment to the current studies is to combine real-time measurements into the model [80]. The impact of different farms can also be significant, with other validation farms showing a decreased detection accuracy [70].

3.2 Other related work

Detection in dairy science using machine learning have been widely applied in oestrus and mastitis detection [11]. Fuzzy logic has been applied to oestrus and mastitis detection by de Mol et al. [18], combining milk yield, milk temperature and activity of a cow. Detection of estrus, general illness or calving are already widespread in dairy farming [11, 68] An improvement of 55% to 80% has been reported in the case of oestrus detection [69]. Borchers et al. [11] used behaviour date to predict the calving day and 2h period. Benaissa et al. [4] combined localisation data with accelerometer data to predict calving and oestrus. Both have improved sensitivity compared to accelerometer data alone. Purpose built sensors have also had their success in calving prediction, with a tail sensor achieving high performance [48].

Study	Objective	Data source	Time window	Model(s)	Evaluation	Score	Cows (% SCK)
[25]	Predict disease (e.g. ketosis)	Feeding, activity	-5d to d0, daily	**RF**, LDA, PC-ANN	Se/Sp/Acc	83%/82%/83%	1,374 (16%)
[68]	Identify metabolic disorders	Rumination, activity	-5d to 2d, daily	HIS (proprietary)	Se	91%	1080 (5%)
[73]	Early prediction of ketosis	Feed intake, Milk data, Activity	-3d, daily	kNN, **SVM**, Logit, RF, XGB	Se	93%	75 (?)
[56]	Detect health disorders (CK)	Rumination	-5d to d0, daily	Index value cutoff	Se/Sp/AUC	80%/80%/0.8	198 (23%)
[81]	Predict metabolic status	Milk data, DPL, BW, PAR	d0, weekly	DT, NB, BN, SVM, ANN, BA, **RF**, kNN	Error rate	24%	295 (?)
[70]	Detect ketosis	Rumination, Activity, Milk data	-4d to -2d, daily	Logistic regression	Se/Sp	70%/70%	706 (29%)
[9]	Predict BHBA and ketosis	Milk infra red spectra	n/a	PLS discriminant	Se/Sp	76%/82%	542 ($\approx$ 56%)
[11]	Predict calving	Behaviour data		RF, LDA, NN	Se/Sp	100%/87%(d) 83%/80%(8h)	53
[17]	Detect oestrus	Milk data and Activity		Fuzzy logic	Se/Sp	67%/99%	179
[32]	Test possible indicators of diseases	Feeding behaviour	-7d to d0, daily	Standard deviation from 7d rolling mean	Pr	91%	$\approx$100
[59]	Predict calving	Activity, Rumination, Feeding, Temperature		Logistic regression	AUC	0.929	400

Table 3.1: Overview of relevant papers. See the glossary for abbreviations. Time window is defined as the period in which data is used and the time steps size of these data points. Models marked in bold scored best on ketosis detection or prediction.

4 Materials and methods

This chapter is divided into six sections. Section 4.1 contains a detailed description of the behaviour, blood measurement and other data used in this study. Section 4.2 shows the steps taken in this study to prepare this data for ketosis detection. Section 4.3 introduces two different time windows as proposed in the first sub research question. Section 4.4 introduces the different feature sets used in this study. It also defines the normalisation techniques: herd-normalisation, prepartum normalisation and within-window normalisation. Section 4.5 presents the machine learning models used in this study and Section 4.6 defines the evaluation process and metrics to answer the different research questions.

4.1 Data description

For this study, data from two experiments are presented: SenseOfSensors and EFRO. They recorded blood sample data, behaviour data and other relevant data of dairy cows from farms in the Netherlands. Extra data from the cow management system is also available. This section gives a description of the data collection methods and the types of data collected.

4.1.1 Data sources

In the period between 2016 and 2018 several on-farm experiments have been performed. These experiments recorded blood serum measurements and additional data such as calving dates, Body Condition Score (BCS), locomotion, parity and more. Cows from these studies were equipped with behaviour sensors during the experiments. The blood measurements of interest for this study are around the first week and second week postpartum, falling in the 'Detection' zone in Figure 2.1. Other related information such as parity and calving dates on cows have also been recorded.

SenseOfSensors SenseOfSensors (SoS) was the largest experiment, consisting of eight farms in the Netherlands. The experiment took place from November 2016 to May 2018. During this period, cows were monitored by veterinarians around their calving period. For each cow, several things were noted. Most importantly, cows were scored on their BCS and locomotion at different times during parturition, calcium and BHBA concentrations were measured from blood serum sampled in the first week postpartum and BHBA concentrations were measured in the second week postpartum. For this study, the prepartum scoring data and the first week BHBA data were important, so the other data was removed. As the data from this study contained missing data, some filtering has

been performed. Cows without identification number (to match with behaviour data) and calving date (an important reference date) have been removed from the data. After that, cows without a first measurement of BHBA were removed from the data.

EFRO EFRO was a smaller experiment, consisting of four farms in the Netherlands. The experiment took place between March 2017 and August 2018. Blood serum have been sampled in the two weeks prepartum, in the first week postpartum and five weeks postpartum. The blood serum samples were analysed for concentrations of calcium, albumin, BHBA, Hatpoglobin, NEFA and Ureum. For this study, only the BHBA concentrations of the first week prepartum were important and therefore the rest of this data was removed. Like in the previous data set, filtering of the data was needed. The filtering process was adjusted, because there was no calving date in the EFRO data and only the second measurement of the EFRO experiment is important to this study.

The initial amount of data points and amount of data points after filtering is shown in Table 4.1. The filtering on calving date and ID was performed first and then all data points without first week BHBA measurement were removed.

Filter	**SoS**	**EFRO**	**Total**
No filter	1740	181	1921
Calving date & ID	1551	181	1732
BHBA measurement	1403	178	1581

Table 4.1: The amount of calvings per study with filtering on 1) calving date and cow identification number and 2) BHBA measurements. The latter was performed after the filter of calving date. For SoS, the BHBA filter was based on either the existence first measurement value. For EFRO, the BHBA filter was based existence on the second measurement value.

Cow management system Extra data about the cows from the experiments was available through data from the cow management systems from all farms involved. This data consisted of behaviour data, animal data data and calendar data. The animal data was used for additions and corrections on the experiment data, explained in Section 4.2.2

4.1.2 Blood measurements

BHBA levels have been measured in blood serum for the experiments. As stated in Section 2.1, measurement of BHBA is the standard detection method of SCK and the value of 1.2 mmol/L is used as threshold for SCK.

Histograms of BHBA values in the SoS experiments show an almost equal distribution between the first and second measurement, both peaking at around 1 mmol/L (Figure A.4). Also, the change in measurements between the first week and second week shows a normal distribution (Figure A.5).

With the cutoff value determined at 1.2 mmol/L BHBA, the amount of cows with subclinical ketosis and clinical ketosis is presented in Table 4.2.

Measurement	SoS	EFRO	Total
2 weeks prepartum	n/a	0	0
1st week postpartum	297	8	305
2nd week postpartum	353	n/a	353
5th week postpartum	n/a	21	21

Table 4.2: Cows with BHBA at 1.2 mmol/L or higher at different measurements in the two studies. SoS did not have a third measurement. Moments of measurements are different for the experiments. SoS measured in the first week postpartum and in the second week postpartum. EFRO measured in two weeks prepartum, first week postpartum and five weeks postpartum.

4.1.3 Sensor data

For both experiments, the behaviour data was captured using Nedap neck and leg mounted sensors. The neck sensor is mounted with a collar around the neck of the cow and the leg sensor is strapped to one of the front legs of the cow. These mounting points enable capturing and categorising of different cow behaviours. The neck sensor captures behavioural activity every minute and outputs this as subsequent blocks. Behaviours is categorised as inactive, ruminating, eating and other. The neck sensor also captures a head movement count, which resembles the cows neck activity. The leg sensor captures a summary of behavioural activity in the last 15 minutes. Behaviour is categorised as walking, standing and lying. Each 15 minutes, a summary of the minutes walking, standing and lying is outputted from the sensor. In addition to that, the leg sensor also captures the number of the cows standups and the number of steps. For each sensor, the behaviour data is mutually exclusive. For example, when a cow is ruminating, it cannot be inactive or eating. Likewise, when a cow is lying, it cannot be standing or walking. Therefore, a high correlation between the different behavioural activities exist. As the sensors work independently, different combinations of neck and leg behaviour data exists. For example, cows can be ruminating while lying, standing or walking. A summary of the sensor properties is presented in Table 4.3. The sensor data is stored on the sensor for 24 hours. In the barn, there is a receiver which transports the sensors data to either an on-site database or to Nedap online storage.

Sensor	Nedap SmartTag Neck	Nedap SmartTag Leg
Mounting position	Neck	Front leg
Behaviour data (time)	Eating, Ruminating, Inactive and Other	Lying, Standing, Walking
Behaviour output data	Subsequent periods	15 minutes
Activity data (amount)	Neck movements	Steps
Additional data	n/a	Number of standups

Table 4.3: Sensor properties

An example of daily eating time derived from sensor data is given in Figure 4.1.

Both sensors are validated in a study of Van Erp-van der Kooij et al [75]. Compared with visual observation of behaviour, eating, rumination, inactive, lying, standing show high Pearson correlation scores ($r > 0.8$), only walking time showed a low Pearson correlation score ($r = 0.25$).

As both studies concerned of multiple farms, some few external factors are not present in the

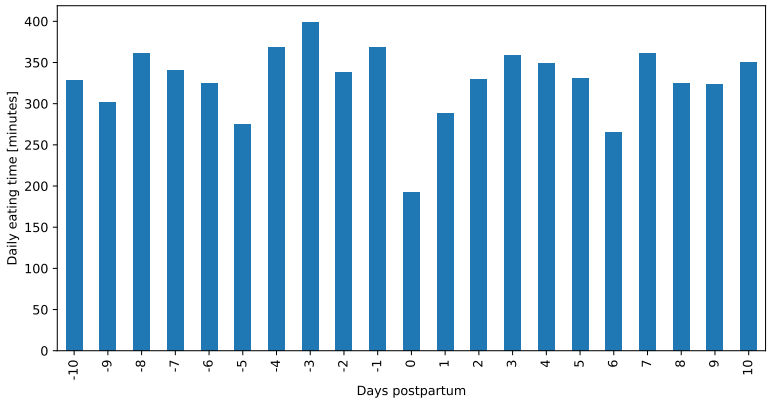


Figure 4.1: Daily eating time of a random cow from the experiments around parturition.

behaviour data. Firstly, as cows enter the dry period, the length of this period depends on management practices. Often, farms have a special area for dry cows and cows on different farms experience different environments including beddings, irregular food or different surrounding cows. Also, external effects such as the season and temperature can have effect on cows behaviour.

Behaviour data is merged from different sources and this resulted in data loss. Also the amount of cows with behaviour data in the prepartum period increases up until the measurement period (see Figure A.11; Appendix A). This affects the choice of the time window, because the machine learning models benefit from more training samples and testing is more reliable with more samples.

4.1.4 Other related data

Additionally, other cow-specific data have been logged such as parity, calving date, BCS and locomotion score. This includes parity and the calving date. Parity has been shown to have effect on the SCK prevalence.

As SCK is a fresh-cow disease, so having the calving date is very important. The negative energy balance occurs during the first part of lactation, so having SCK prepartum is unlikely.

Calvings Farms from both studies differ in the amount of calvings in the SoS and EFRO experiments (shown in Figure A.1. For example, the largest farm 2.5 times as much calvings registered compared to the smallest in the SoS experiment. As the amount of data is an important factor in machine learning, the skewed distribution introduces a bias towards the larger farms.

The SoS study span over multiple years, so it is possible that a cow can have multiple calvings in the span of the study. Indeed, about 20% of the cows had multiple calvings (shown in Figure A.3). The EFRO study did not have multiple calvings per animal recorded.

Parity As mentioned in the previous chapter, parity is a risk factor for developing SCK. The distribution of parity in the samples is shown in Figure A.2. Especially the distinction between

primiparous and multiparous cows is important, as multiparous cows have higher risk for SCK. Almost three quarters of cows are multiparous in both studies.

BCS & Locomotion The SoS experiment recorded BCS and locomotion scores from a veterinary assessment. This scoring was at the end of the dry-period of each cow according to the experiment protocol. Figure A.7 show the median scoring is ten days before calving. Higher BCS and locomotion scores have been identified as risk factors for SCK in Section 2.2.2. Figure A.8a and A.9a show the distribution of BCS and locomotion scores respectively. Higher BCS was related to SCK in literature and this is visible in Figure A.8b. Higher locomotion score was also associated with SCK, Figure A.9b also shows this pattern, with locomotion score of four showing elevated levels of BHBA.

4.1.5 Training-test split

As good practice, the data set was split up into a training set and testing set. The training set is used as input for the machine learning model. For training and testing, a (80%/20%) stratified split of the data is applied. As the data is is imbalanced, stratification is applied to have an relative equal amount of positives in training and test set. With 1581 BHBA measurements in the first week, this results into 1265 training and 316 testing samples.

4.2 Data mining & Preprocessing

From a large database of cow behaviour, certain data has to be filtered in order to be relevant for this study. We need to deal with missing data and outliers, merge the blood data with the behaviour data, define the unit of time steps, define a time window and deal with gaps in behaviour data. All data mining and preprocessing is performed using the Python[1] programming language and Pandas[2] data processing package.

4.2.1 Outliers

Values that differ significantly from the other values are marked as outliers. Outliers occur in any sampled probability distribution by means of chance, but they can also be caused by sensor misreadings or human errors. Outliers have been identified in measurement dates, as shown in the histograms in Figure A.6. The criterium for outliers was that measurement dates should between calving and 14 days after calving. The outliers are likely to be typing errors in calving dates or measurement dates. Removal of the outliers show that the remaining measurements are within the time frame set by the experiment (4-10 days after calving). Outliers are corrected by hand where possible, i.e. if typing errors are off by a year or a month. They are also corrected with system data. The amount of outliers unresolved are reported in the results.

[1] `https://www.python.org`
[2] `https://pandas.pydata.org`

4.2.2 Data correction

The experiment data contained errors and gaps and to reduce erroneous data, correction were made to this data set. Some life identifications were wrongly formatted, so the data is only partly matched. The format of these life ids is 'NL xxxxxxxxx', where the x's form a 9 digit number. Errors include the absence of the 'NL' part, the absence of the space and missing a number in the digits. Also some cows had duplicate life ids or missing life ids, which are likely to be input errors.

To recover as much as possible, the errors were corrected in the following order:

1. Absence of the 'NL' or the space was checked with a regular expression and corrected.
2. Missing life ids were filled by matching the combination of the farm number and animal number from the experiment data with the cow management system. Animals numbers can be reused on a farm, so the calving date from the experiment data validated the right match.
3. Missing parity was also filled if there was inserted if available on the cow management system.
4. Missing measurement dates were computed
5. All life ids were checked against the cow management system and the misspelling of some life ids were corrected.
6. Calving dates, blood measurement dates and parity were also checked against the cow management system and likely type errors were corrected by hand.

4.2.3 Merging data

Behaviour data comes from Nedap sensors and blood data comes from different studies. Both data sources contained life ids for the cows, which were corrected in the previous section. To be able to link the behaviour data to the experiment data, each data point from the behaviour data was marked with the correct life id. This enables us to match it with the experiment data and select the relevant behaviour data in a time window.

4.2.4 Transformation & Resampling

The two sensors have different outputs for their behaviour. In order to use both sensor data together, we have to transform the data into compatible shared format. The neck sensor outputs subsequent periods of behaviour and the leg sensor outputs summaries of a period of 15 minutes (See Figure 4.2). The transformation is only possible in one direction, as the summary does not contain the order of the behaviour.

The neck behaviour data was transformed to a summary similar to the leg sensor by calculating in which time frame(s) a certain behaviour period belongs and dividing this time according to their contribution of the total time in this time frame. This is illustrated by the transformation of the neck behaviour in Figure 4.2 to the summarised format.

Moreover, this study also resamples the data to limit the amount of behaviour data points. To illustrate, the sampling rate of the behaviour from the leg sensor is 15 minutes. This results in

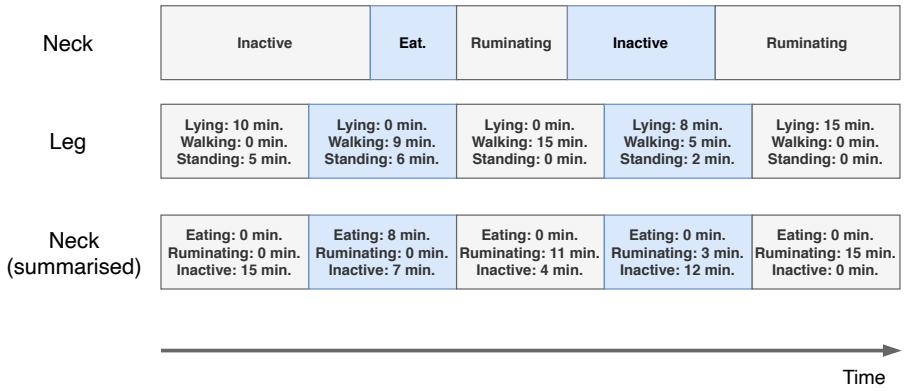


Figure 4.2: Output of behaviour data from the neck and leg sensors. The neck sensor outputs behaviour in subsequent periods. The leg sensor outputs behaviour in a summary of 15 minute blocks. The summarised neck behaviour is dividing the periods to the fixed time blocks.

480 data points over a period of 5 days for a single behaviour. Resampling means summarising multiple data points into one by taking the sum, mean or some other statistic. This will sacrifice some detail, but it will allow the model training to be faster and decrease the amount of outliers in the behaviour data.

For this study, we chose to resample the sampling rate of the data into periods of 24 hours by taking the sum of the values. This period was defined as 24h starting at 0:00 in the local timezone. This yields into total eating/ruminating/etc.. time of each day. The daily time scale is preferred, because cows have a diurnal pattern and therefore daily values are directly comparable.

4.2.5 Gaps in behavioural data

For several reasons, there are gaps in the behavioural data. The sensor itself has a limited memory and occasionally some data is lost. An overview of the gaps in behaviour data per sensor is found in Figure A.10 in Appendix A. To reduce the amount of data loss, a distinction is made between different amounts of data missing:

1. Missing values within a day. The rest of the values within this day are scaled linearly to a total of 24 hours. The correction function defined in Equation 4.1 is the function to correct within a day, where x_i is the daily behaviour minutes for behaviour i and J is the set of all behaviour types.
2. Missing values of a day or more. These days are be marked as missing.

$$\text{corrected}(x_i) = x_i \times \frac{24 * 60}{\sum_{j \in J} x_j} \tag{4.1}$$

When a time window selects a missing day for a given subject, the subject will be dropped from the data set.

4.2.6 Imputation

The experiment data is not complete and missing values have been imputed by a 3-Nearest-Neighbours approach to minimise data loss. The training and test set from Section 4.1.5 contain 459 entries and 121 entries respectively which have one or more missing values (See Appendix A.4 for more statistics on missing values). Imputation is the method of filling in missing data by using information contained in the entire data set. The naive approach for imputation is to fill the missing values by the mean of non-missing values on the entire data set. The 3-Nearest-Neighbour approach differs by utilising information contained in row to produce a better estimate for the missing value. For example, consider a data set D with entries containing values for X and Y. Some entry A is missing value X in this data set. Imputation with the mean fills the value of X_A in A with the mean value of X in the entire data set. Imputation with 3-Nearest-Neighbours fills the value of X_A in a with the mean value X of the 3 nearest entries based on the value in value Y. The nearest value is based on Euclidean distance. As data values are often correlated, the 3-Nearest-Neighbour approach provides a better estimation compared to a mean value approach. The imputation was fitted on the training set and then applied to training and test set.

4.3 Time windows

To limit the amount of input data, a time window has to be defined. As stated in Section 1.2, the first research question is to experiment with a calving based time window. The relation of the transition period of a cow in relation to the data of this study highlights two important dates: measurement day and calving day. Firstly, the measurement day is the moment where the cow is sampled for her blood. Secondly, as seen in Figure 5.1 and 5.2, the day of calving marks a change in the behaviour of cows. These two events are used to define two time windows. Since SCK occurs mostly in the two weeks postpartum, we use this fact to define our time windows.

While the assumption that some behavioural patterns are visible in cows with SCK is supported by literature, it is still unknown whether these patterns are long or short, occur once or multiple times. In Section 3, multiple studies used measurement-date based time windows. As stated in Section 2.2.1, this approach is sub optimal, since different behavioural patterns related to SCK occur at different moments relative to calving. Therefore we define two time windows for behaviour data and evaluate which is better suited to SCK detection.

1. Fix the size of the time window to include x days before and up until measurement (measurement-date based);
2. Fix the time window to be x days before and y days after calving (calving-date based).

In a cross validation on the training set, results showed that a 5 day window was best on this data set. Likewise, the start of calving-date based was set at 2 days after calving.

The options are graphically represented in Figure 4.3. This shows that the measurement date is not fixed relative to calving, therefore a measurement-date based (MB) window selects different days relative to calving for every cow. It also shows that the end of this window is equal to the day of measurement, so the state of BHBA concentration in the cow at last day in the window is always correct.

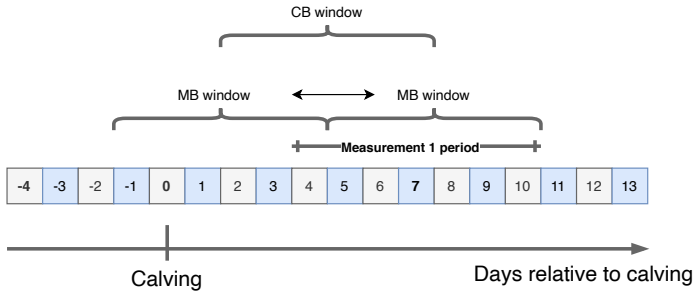


Figure 4.3: Graphical overview of all time windows. The calving date is fixed, but the measurement date is different for each cow. Therefore, the measurement-date based window has lower and upper bound.

This in contrast to the calving-date based (CB) window. There are cows which are tested at 10 days after calving, but most of cows are tested before. This results in a time window where the BHBA value measured is somewhere in this time window. Therefore, the complete time-window does not represent the current state of an animal.

The calving allows for a better recognition of pattern of subclinical ketosis might cause, but it has a weaker relation to BHBA value directly compared to measurement-date based time window. In the context of predicting or detecting SCK, the usage of the calving-date based makes the classification in some cases a prediction and in other cases a detection, as the calving-date based overlaps some measurement dates.

4.4 Features

Different feature sets are derived from the data available. We distinguish three categories in the feature sets: the static feature set, behaviour feature sets and combined feature sets. The static feature set contains features derived from information about the cow which is known beforehand and is considered a risk factor for subclinical ketosis. The behaviour feature sets contain feature sets which have a unique feature definition. The combined feature sets contain combinations of the feature sets introduced earlier to evaluate their combined potential.

The amount of samples per feature set and time window varies. We set the amount of samples to all animals with complete behaviour data in the given time window. This causes the Fast Fourier Transform features and individual normalisation features to contain significantly less samples compared to the other feature sets. Table 4.4 provides an overview of the amount of samples per feature set and time window. Features derived from the daily totals (with herd-normalisation or without normalisation) do not change the amount of samples.

4.4.1 Static features

The static feature set contains the dry period length, the previous lactation length, the gestation length, the prepartum body condition score, the prepartum locomotion score and the month of the calving date. Dry period length was calculated as the difference in days between the dry-off

Feature set	MB	CB
Static features*	961 / 231	982 / 233
Daily behaviour totals	961 / 231	982 / 233
Herd-normalised totals	953 / 228	982 / 233
Prepartum normalisation	902 / 210	907 / 212
FFT features**	n/a	845 / 194
Combined features	953 / 228	982 / 236

Table 4.4: The amount of samples per time window and feature set. *The static feature set is limited to all samples found in the daily behaviour time windows for fair comparison. **FFT feature set is based on prepartum data, which is calving based.

date and calving date. Previous lactation length is calculated as the difference in days between the previous calving date and the dry-off date. The gestation length is calculated as the difference in days between the last insemination date and the calving date. The dry-off date, previous calving date and last insemination date are all gathered from the cow management system and suffer from administration issues. The prepartum body condition score and locomotion score were gathered from experiment data. The month of the calving date was represented as a combination of a sine and cosine transformation of the value using the method defined in Section 2.7.1.

To summarise the above, the following features are considered in the static feature set.

1. Body Condition Score
2. Locomotion score
3. Parity
4. Dry period length
5. Gestation length
6. Previous lactation length
7. Sine transformation of the calving month
8. Cosine transformation of the calving month

The amount of samples in static data is larger than in any of the behaviour feature sets using any time window, because not all cows were fitted with a behaviour sensor, behaviour data was incomplete or the data contained gaps. Therefore, this feature set was evaluated twice using different sets of training and test data. It was evaluated on the training and test samples present in the measurement-date based time window and once it was evaluated on the training and test samples present in the calving-date based time window. This allows to have a fair comparison between behaviour features and static features.

4.4.2 Behaviour features

Daily behaviour totals

Daily behaviour totals contain the sum of each behaviour type of each animal per day as described in the resampling process (Section 4.2.4). Through time windows, daily totals of the cows are the same relative to the reference point (calving or measurement date). Literature suggests that some behaviour time decreases (eating, activity) and other behaviour time increases (inactivity, lying), so daily totals are a good starting point as a feature set. The amount of features is equal to the amount of behaviour parameters multiplied by the amount of days specified by the time window. For the measurement-date based time window, the amount features is equal to 6(days) $\times$ 10(behaviours) = 60 features, for calving-date based time window this is equal to 8(days) $\times$ 10(behaviours) = 80 features. As suggested by related work, cows suffering from ketosis show abnormal behaviour.

The daily totals are also used as input for the derived features below.

Behaviour statistics

The behaviour statistics features set contain the mean, variance, minimum and maximum of all behaviour types aggregated of the daily behaviour totals within the specified time window. It also contains the ratio between the means of different behaviour types. The ratio between different behaviour types removes the absolute amount of behaviour time. For instance, some cows eat more than others, but if their rumination time is relatively the same, then the ratios are equal. As ketosis concerns energy, the ratio between ruminating and activity gives an approximation to amount of energy gained and spend. Compared to the daily behaviour totals, the statistics removes the time component from the features as natural variation between days occurs. Removing the time component reduces the effect of natural variance of behaviour cows experience between days. For instance, a cow with ketosis can have their activity on two days on 70% of normal activity while another cow with ketosis can have activity on 50% and 90% for two days. In this case their is a large difference between daily behaviour totals, however their mean is equal. The statistics feature set contains 4(statistics) $\times$ 10(behaviours) + (3 $\times$ 3)(ration) = 49 features regardless of the time window.

Behaviour trend

The behaviour trend feature set contains a slope and offset for each behaviour type derived from a linear fit on the daily behaviour totals within the window. Behaviour is not necessarily a linear function, it also have a shape of some other function. However, if the general trend is downwards, the linear fit slope will show this trend. The linear fit is defined as a linear function $f(x) = ax + b$ where a(slope) and b(offset) are coefficients which minimise the mean squared error. Like the statistics feature set, the trend feature set removes the time component within the feature. The trend line represents the development of behaviour over time. The statistics feature set contains 2(statistics) $\times$ 10(behaviours) = 20 features regardless of the time window.

Prepartum FFT features

The Fast Fourier Transform (FFT) feature set contains features derived from a FFT from a two week prepartum time window based on hourly data to find regularity in the behaviour of a cow. This deviates from all the other feature sets, as they use daily data instead. However, to find regularity within a day, a smaller time scale was needed. The FFT is an extension to the non-periodicity feature introduced by van Dixhoorn et al. [74]. Where the non-periodicity is the mean squared difference between the autocorrelation and a 24h sine wave, the FFT feature set applies a FFT on the hourly data and extracts the frequency of each behaviour per day.

From the FFT, a threshold line is calculated and values above this threshold line are defined as peaks. For each frequency, the strength of the peak is saved if it passed the threshold. Also the maximum strength, the frequency with maximum strength and the number of peaks are calculated as well as some other statistics. The entire feature definition is found in Section 2.7.2.

4.4.3 Normalised behaviour features

There is a lot of variation in behaviour among cows. This variation has the effect that normal behaviour for one cow can be abnormal for another. In order to mitigate these some of these normal abnormalities and also to normalise the magnitude of the features, we apply different normalisation techniques and evaluate their effect on ketosis detection.

The basis for normalisation in this study is the procedure of z-score normalisation: values are subtracted by some mean $\bar{x}$ and divided by some standard deviation s. The sample set k for the mean $\bar{x}_k$ and standard deviation s_k is varied throughout different normalisation methods. By changing this set k, we can normalise on various aspects of cow behaviour.

$$x_{\text{norm}} = \frac{x - \bar{x}_k}{s_k} \tag{4.2}$$

Herd normalisation

As cows are situated in barns with other cows, there is social behaviour, specific eating times and inter-barn movement. This changes the behaviour of cows, but its not disease related. For example, the entire herd can moved to another barn. This is difficult to detect with individual cow data and these changes may look like outliers or noise when compared to cows from another farm. However, they could be explained by the behaviour of the herd.

Moreover, research has indicated that social behaviour of a cow changes when it suffers from SCK. For example, cows with SCK showed less feed intake in the first 90 minutes of fresh feed delivery compared to healthy herd [62]. Without extra data, it would be hard to detect such behaviours.

Therefore, the normalisation based on the herd is applied. The herd is defined as all cows in the same group: lactating or dry, because cows can be moved into a dry-off pen. As the cow management system registers dry-off events, we cannot assume that dry cows belong to same herd as the lactating cows. Based on the individual cow, we calculate the herd mean and herd standard deviation of either all dry cows or all lactating cows. This includes cows that did not enter the experiments, but were fitted a sensor and their behaviour was logged in the cow management system. Then, each

daily total of a cow is subtract with the mean of its herd, divided by the standard deviation of this herd, producing an aforementioned z-score normalisation.

The behaviour statistic feature set and behaviour trend feature set are also applied to the herd-normalised values.

Prepartum normalisation

The prepartum normalisation feature set contains the daily behaviour totals which have been z-score normalised by the mean and standard deviation of the first week before calving (see Figure 4.3. Individual cows have different behaviour and a normalisation on individual level could show abnormal behaviour. The amount of this data for this feature set is smaller than the others, because the number of animals having prepartum data is smaller than the number of animals in both time windows.

Within-window normalisation

The within-window normalisation feature set contains the daily behaviour totals which have been individually z-score normalised. The daily behaviour totals are subtracted by the mean within time window and divided by the standard deviation within the time window.

4.4.4 Combined feature sets

The combined power of the behaviour features is also evaluated in this study, with the exception of the FFT features and prepartum normalised features, because these feature sets the amount of data significantly (see Table 4.4. Finally, a different set using the combined behaviour features and the calendar data is evaluated.

In conclusion, the following feature sets are tested (Table 4.5).

Feature set name	Normalisation	Feature functions
Static features	None	Static features
Totals	None	Daily totals
Statistics	None	Statistics
Trend	None	Trend
FFT features	None	FFT features
HN-Totals	Herd	Daily totals
HN-Statistics	Herd	Statistics
HN-Trend	Herd	Trend
PN-Totals	Prepartum	Daily totals
WWN-Totals	Within-window	Daily totals
Combined behaviour features	None + Herd	Daily Totals + Statistics + Trend
All features	None + Herd	Static features + Totals + Statistics + Trend

Table 4.5: Summary of all feature sets tested in this study. HN is herd-normalised values, PN is prepartum normalised and WWN is within-window normalised values.

4.4.5 Feature selection

Feature selection was applied to each feature set to limit the amount of features and reduce the amount of noise in the feature set. The feature selection method used in this study is based on the frequently used wrapper method[34]. The wrapper was based on a initial fit with a Random Forest classifier/regressor which scored the features according to their importance in building the trees in this model. Importance is calculated as the normalised reduction of Gini impurity of the training set. Features which scored higher than the mean importance were kept and features which scored lower than the mean were dropped. On the calendar feature set, feature selection was not applied, because the amount of features was already low.

4.5 Machine learning models

For this study a set of classical machine learning algorithms is used. Machine learning has not been applied often in SCK detection and the size of the data set is not large enough to apply deep learning. This study uses learning models based on different principals: Bayesian probability, decision trees and neural networks. The models used in this study are Random Forest (RF), Naïve Bayes (NB), Multilayer Perceptron (MLP) and Gradient Boosting (GB). For an introduction of the models, see Section 2.8. These models have proven to be successful on smaller data sets and among the models that are used in exploration machine learning research.

Except for Naive Bayes, all machine learning models introduced above have a regression variant. The linear regression model is added as replacement for Naive Bayes as it also has a very simple definition.

The hyperparameter optimisation for the given models are randomly searched of a set parameter space. Each experiment first computed their hyperparameters on the training set, so hyperparameters differ per experiment. This benefits either both small and large number of data points per sample, as the models can adapt their hyperparameters which makes the chosen model more flexible. See Appendix D for all parameters searched in the hyperparameter optimisation for each model. For hyperparameter optimisation, the training set is used, as we do not want to optimise our model on the test set.

4.6 Evaluation

The evaluation procedure of this study concerns hyperparameter optimisation using cross validation, predictions on the test set with given performance metrics and performance comparison by pair-wise significance testing or mean ranking scores. As said earlier, for training and testing, the split is a (80%/20%) stratified split of the data is applied. As the data is imbalanced, stratification is applied to have an relative equal number of positives inthe training and test set. For all approaches, 5-fold cross-validation is used to optimise the hyperparameters of each model using the training data set.

4.6.1 Task definitions

Classification The main task of this study is to detect SCK in dairy cows. Like Section 4.1.2, we use the cutoff point of 1.2 mmol/L to distinguish healthy and SCK cows and we can define this task as a binary classification task. Given a data set with certain features, produce a target label (SCK or healthy).

Let X be all behavioural data from calvings, $f(x)$ be the function to transform the raw values into features for any $x \in X$ and $\mathcal{L} = \{\text{SCK}, \text{Healthy}\}$ is the set of labels. Then our classification model C defines a mapping $C : f(X) \to \mathcal{L}$. Note that feature function $f(x)$ is the identity function $f(x) = x$ for the daily behaviour totals feature set.

Regression

One of the sub questions of this research proposes to investigate a regression model to predict BHBA directly, instead of applying a threshold like the classification task. Thus the task is given a set of features, predicting a BHBA value. Then the predicted BHBA values can be compared to the original BHBA values, or both can be transformed to binary values by applying the threshold. The binary values can then be evaluated against all non-probabilistic evaluation metrics.

The regression task is formally defined as follows. $\mathcal{X}$ and $f(x)$ are the same as classification, but the targets are now drawn from the continuous domain $\mathbb{R}+$. Then our regression model R defines a mapping $R : f(\mathcal{X}) \to \mathbb{R}+$ Then these predicted BHBA values can also be transformed into classification labels using 1.2 mmol/L threshold. The adjustment is then a mapping $R_c; f(\mathcal{X}) \to \mathbb{R}+ \to \mathcal{L}$

4.6.2 Metrics

All related studies use the evaluation methods of sensitivity and specificity. To be able to compare with related work, this study also produces sensitivity and specificity scores for the classification task and regression task. Care must be taken when using sensitivity and specificity. When the balance between positive and negative cases is skewed, the sensitivity and specificity can be both high, while the practical usage of such a model is limited. Therefore, we introduce additional metrics that give more insight into the performance of our model. For detailed information on all evaluation metrics used, refer to Section 2.9.

Classification metrics

For the classification models, the test set is evaluated on Area under the Receiver Operating Curve (AUC), Sensitivity (Recall), Specificity, Precision, Accuracy. In addition to these metrics, the Sensitivity at 95% Specificity (Se@95Sp) and Precision at 95% Recall (Pr@95Re) are calculated. In subclinical ketosis detection, the Sensitivity at 95% Specificity indicates sensitivity score at a false negative rate which is acceptable in real-life settings, because the literature suggests a high amount of false negatives reduces the amount of trust in the system. The Precision at 95% Recall is a useful metric to see if the classifier would retrieve all positive cases, what percentage of all classified positives is actually positive. In a ketosis detection system, this metric shows how many

healthy cows would be treated if all almost all SCK cows detected. The 95 percentage is chosen, because the threshold of subclinical ketosis is not valid for all cows. By taking 95% Specificity or Recall, we can eliminate the cows that are healthy but still test positive for subclinical ketosis and vice versa.

Regression metrics

Metrics for the regression models are the (root) mean squared error, the mean absolute error and the R^2 score. In addition to these metrics, the error in the 95% percentile is also calculated. As shown in the distribution plots of the BHBA value in week 1, most measurements are around 1. However, this study is interested in values of 1.2 and up. In the distribution of all prediction errors of BHBA values, the 95% percentile error shows the larger errors, which presumably are in the higher BHBA values. The 95% percentile error is preferred over the maximum error, as the BHBA value can be high while the mean is low.

As shown earlier, using the SCK threshold value, the regression values can be transformed to binary classification values. Then, the non-probabilistic metrics of classification can be used as well. This allows us to compare the regression model to the classification model.

4.6.3 Significance testing

This experiment produced 168 different results, through a combination of different time windows, machine learning models and tasks. These different combinations can be compared to each other to answer the four subquestions presented in Section 1.2. This regards the comparison between a measurement-date based based time window and a calving-date based time window. As Table 4.6 shows, the assumption that the metric values are normally distributed does not hold every time under a significance level of $\alpha = 0.05$. Therefore, this study uses the nonparametric Wilcoxon signed rank significance test[79] for all metrics. This rank test is used one-sided to test if the mean of one distribution is greater than the other with an alpha value of 0.05. In some cases, we use the same experiments to search for multiple significant differences. In this case, we apply Bonferroni correction for multiple testing by dividing the alpha value by the number of tests.

Metric	p-value
AUC	0.09
Se@95Sp	0.07
Pr@95Re	6.8e-07
Accuracy	5.5e-14

Table 4.6: Shapiro-Wilk test for normality. The null hypothesis is that the metric values are normally distributed. This hypothesis is rejected for Pr@95Re and Accuracy.

4.6.4 Ranking

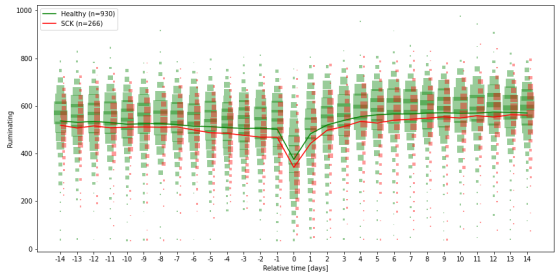
In the case of feature set comparisons, there are not enough test results to apply significance testing. Therefore, a ranking method is applied. This works by grouping all test results of the same

parameters and computing the ranking for each of the groups. The mean rank is then calculated for each attributed that needs to be compared. Let X, Y, Z be the parameters of an experiment and S all scores of the experiment. To compute the mean rank of all values in X, all experiments are grouped by their values of Y and Z and for each group, their scores in S are used to create a ranking for the value X. Then all rankings are averaged on the values in X to produce the mean ranking of values in X.

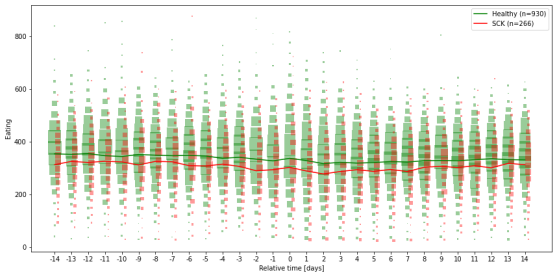
5 Results

5.1 Data visualisations

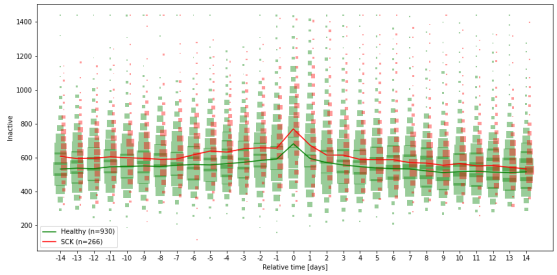
Before applying machine learning to the problem, the behaviour data is first visualised over time. The behaviour data was split into a healthy cow set and a SCK cow set, based on the cutoff value mentioned in Section 4.1.2. The behaviour data was limited to two weeks prepartum up until to weeks postpartum for each animal and the data was sampled into daily totals. Figure 5.1, 5.2 and 5.3 show the histograms for each day for each set of cows. The median of each set is calculated and visualised over time on top of the histogram.



(a) Rumination time



(b) Eating time



(c) Inactive time

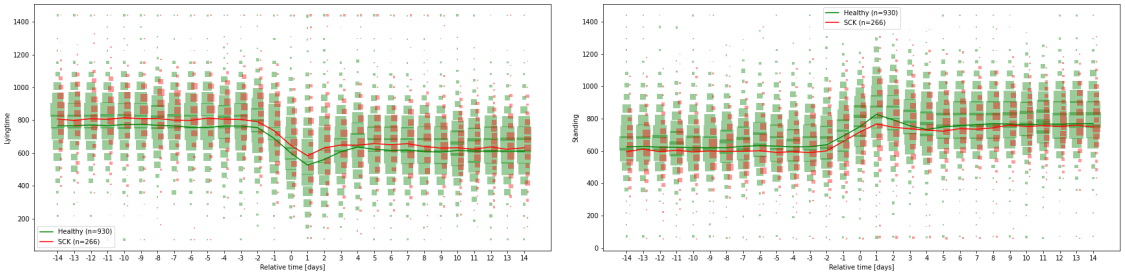
Figure 5.1: Median neck behaviour (line) and histogram of each day compared between SCK and healthy cows. n is the amount of samples for each class. Visible differences are present at eating time, ruminating and inactive time in the median line. The spread of each day is large.

In general, the overlap in the histograms of both groups is large. No significance tests are performed as this data is only a initial insight into behavioural differences.

Behaviour derived from neck sensor Rumination time in both data sets sees a large drop off directly postpartum, as seen in Figure 5.1a. From there on, the rumination time increase to a higher point than prepartum. In contrast to rumination time, eating time in Figure 5.1b sees no notable change in around calving. Inactivity time has a large increase around calving, as seen in Figure 5.1c. This is expected as the neck behaviour is mutually exclusive.

As mentioned in literature, the rumination time for SCK cases is lower compared to healthy animals. Just like in rumination time, the SCK cases show lower eating time. The median inactivity is visibly higher for ketotic animals compared to healthy animals, but the spread of inactivity is also larger.

The daily median neck activity can be found in Figure 5.3a. Like in inactivity time, activity (measured in number of head movements) sees a notable increase postpartum. Differences in the groups are only marginal postpartum, prepartum no difference is visible.



(a) Lying time

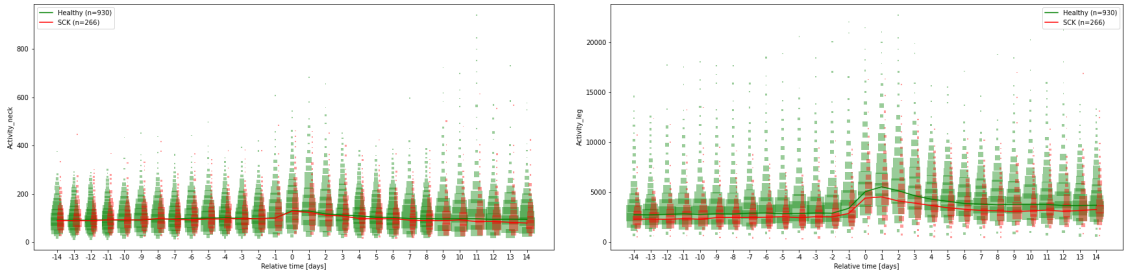
(b) Standing time

(c) Walking time

Figure 5.2: Median leg behaviour (line) and the histogram of each day compared between SCK and healthy cows. n is the amount of samples for each class. Walking, lying and standing show small visible differences between the classes, the spread of values is large.

Behaviour derived from leg sensor Lying time sees a significant drop around calving as seen in Figure 5.2a, with small differences between the two groups. As lying is decreased around, standing time is increased as seen in Figure 5.2b. Standing time also show small differences between the two groups. Walking time is also increase around calving, as is the amount of standups a cow makes. These can be seen in Figure 5.2c respectively. The activity sensor of the leg (Figure 5.3b shows the same curve as walking time.

The distribution of BHBA is visualised in Figure 5.4. This figure shows farms have different distributions of BHBA levels.



(a) Neck activity **(b)** Leg activity

Figure 5.3: Activity (line) and histograms of each day compared between SCK and healthy cows. n is the amount of samples for each class. Leg activity median line. Spread across each day is large.

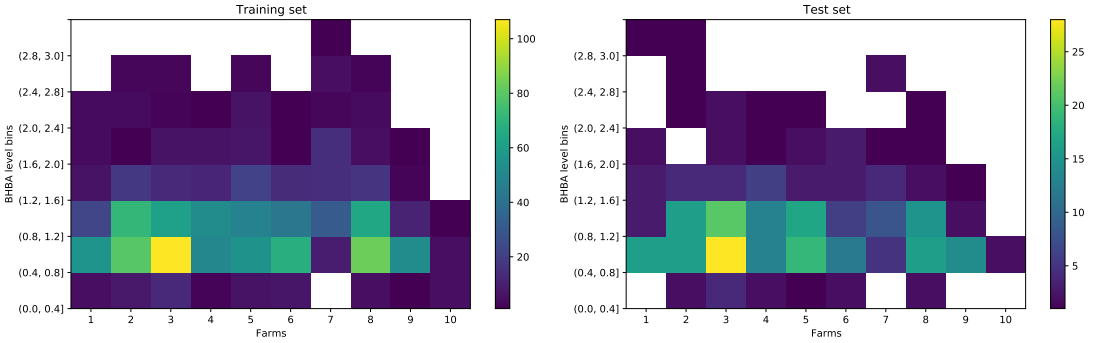


Figure 5.4: BHBA values spread around farms for training and test set. The BHBA measurements are put into bins. The BHBA levels are not spread out the same on each farm.

5.2 Windowing

In this scenario we compare test results of calving-date based (CB) window and measurement-date based (MB) based wind by comparing pair-wise using the same feature-set and model. We compare these two on four different metrics: AUC, Se@95Sp, Pr@95Re and Accuracy. Therefore by the Bonferroni correction $\alpha = \frac{0.05}{4} = 0.0125$ The mean Area under the ROC-curve(AUC) score for a measurement-date based time window is 0.65($\pm$ 0.01) compared to the mean AUC score of 0.63($\pm$ 0.01) for a calving-date based time window. With a value of $\alpha = 0.0125$, the p-value of 0.417 shows we cannot conclude that the measurement-date based time window is better than the calving-date based window. As seen in Table 5.1, also Se@95Sp and Accuracy is higher, but not significantly. Pr@95Re is higher for calving-date based windows and is even significantly better with an inverse p-value of $1 - 0.997 = 0.003$.

Metric	MB	CB	Samples	p-value
AUC	0.648 ($\pm$0.0086)	0.632 ($\pm$0.0084)	44	0.417
Se@95Sp	14.3 ($\pm$1.2)	13 ($\pm$1.2)	44	0.35
Pr@95Re	27.8 ($\pm$0.23)	28.7 ($\pm$0.25)	44	0.997
Acc	71.4 ($\pm$1.2)	70 ($\pm$1.3)	44	0.124

Table 5.1: Comparison of different time windows. Given are the mean ($\pm$SE) metric values. (measurement-date based (MB) is a time window of 5 days before up to and including the day of measurement. calving-date based (CB) is a time window of 2 days after until 10 days after calving.)

When we analyse per machine learning model (Table 5.2), in general RF and GB have more variation between the windows. RF and GB perform better in measurement-date based windows on AUC, Se@95Sp and better in calving-date based window on Pr@95Re. NB and MLP are almost equal between the windows on all metrics. As there are a lot of tests here, so no significance testing is performed.

5.3 Normalisation

In Section 4.4.3, we presented three normalisation methods: herd-normalisation, individual normalisation and z-score normalisation. Each of these used daily values for each behaviour as features. From these three, initial tests showed most promise for herd-normalisation. Therefore, unlike the other normalisation methods, we also created the same derived feature sets: statistics and trend. This resulted in 24 different paired test results, with the herd normalisation being the difference. With $\alpha = \frac{0.05}{4} = 0.0125$ (by Bonferroni correction), we found that the AUC score of raw values was significantly better with a p value of $1 - 0.999 = 0.001$. In the other metrics, raw values were not significantly better, see Table 5.3.

In Table 5.4, we present the highest possible AUC scores for each normalisation method with a daily values feature set. Raw daily values outperformed all normalisation method proposed on AUC scores.

As we have also created two feature sets from herd-normalised data and compared it to the raw variant. Table 5.5 present the highest possible AUC scores of these feature sets. The statistics feature set performed better than the trend feature set, both for raw values and herd-normalised

Metric	Model	MB	CB	samples	
AUC	RF	0.683 (±0.014)	0.642 (±0.02)	11	0.143
AUC	NB	0.617 (±0.011)	0.632 (±0.013)	11	0.857
AUC	MLP	0.621 (±0.014)	0.621 (±0.013)	11	0.857
AUC	GB	0.671 (±0.02)	0.632 (±0.021)	11	0.212
Se@95Sp	RF	17.8 (±2.1)	13.5 (±2.3)	11	0.124
Se@95Sp	NB	11.1 (±1.9)	11.9 (±2.7)	11	0.465
Se@95Sp	MLP	10.4 (±1.7)	11 (±1.9)	11	0.571
Se@95Sp	GB	17.9 (±2.8)	15.8 (±2.7)	11	0.605
Pr@95Re	RF	27.2 (±0.48)	29.2 (±0.46)	11	0.996
Pr@95Re	NB	28 (±0.39)	28.3 (±0.46)	11	0.762
Pr@95Re	MLP	27.9 (±0.43)	28.3 (±0.54)	11	0.703
Pr@95Re	GB	28.1 (±0.53)	29.1 (±0.57)	11	0.923
Acc	RF	74.9 (±0.63)	73.3 (±0.77)	11	0.124
Acc	NB	64.6 (±3.3)	60.1 (±3.8)	11	0.212
Acc	MLP	71.5 (±2.2)	72.8 (±0.73)	11	0.465
Acc	GB	74.6 (±0.6)	74 (±0.79)	11	0.5

Table 5.2: Time window comparison per machine learning model. Given are the mean (±SE) metric values. (measurement-date based (MB) is a time window of 5 days before up to and including the day of measurement. calving-date based (CB) is a time window of 2 days after until 10 days after calving.)

values. Moreover, the highest possible scores of the herd-normalised statistics feature set is on par with the best feature set with raw values.

5.4 Features

For the feature set comparison, we apply a different comparison method to the previous sections. As the amount of experiments for each feature set was limited, we apply a ranking method the feature sets. As seen in Table 5.6, the Static Feature set has the highest mean rank in AUC, Se@95Sp and Pr@95Re. Only in Accuracy, the All Features set performed better. The Fast Fourier Transform feature set performed worst on all metrics, except for Pr@95Re. In this metric, it was third best.

5.5 Machine learning models

In the model comparison tests, Random Forests performed best compared to the other machine learning models with a mean AUC score of 0.664 compared to Multilayer Perceptron (0.619), Naive Bayes (0.626) and Gradient Boosting (0.651), see Table 5.7. In Sensitivity at 95% Specificity, Gradient Boosting was better than Random Forest. In other metrics, they performed on par, but MLP and Naive Bayes are always worse.

Metric	Raw	Herd-normalised	Samples	p-value
AUC	0.651 (±0.0063)	0.628 (±0.0072)	24	0.999
Se@95Sp	12.8 (±1)	11.5 (±1.3)	24	0.747
Pr@95Re	27.8 (±0.23)	27.5 (±0.22)	24	0.964
Acc	71 (±1.3)	69.4 (±2.3)	24	0.718

Table 5.3: Comparison of raw values and herd-normalised values under different feature sets (daily values, statistics and trend), different windows and different models. Given are the mean (±SE) metric values. (Raw values are daily summations of each behaviour under a time window with optional feature extraction. Herd normalised values are individual deviations from the herd on a daily basis.)

Metric	Daily values	Herd-normalisation	Prepartum normalisation	Within-window normalisation
AUC	0.70	0.65	0.62	0.61
Se@95Sp	18	15	22	5
Pr@95Re	28	26	31	27
Accuracy	75.32	73.68	72.38	68.83

Table 5.4: Daily values (with normalisation) under measurement-date based time window and a Random Forest model

5.6 Regression

The regression model performed best on the Static Feature set (see Table 5.8, with a root mean squared error of 0.42, a Q95E of 0.84 and a R^2 score of 0.16. The Q95E score means that in the worst case, a BHBA value is predicted 0.84 higher or lower than reality. Since BHBA values lower than 0.4 are few, this error will always transform a healthy BHBA value to a value above the threshold, and vice versa. The R^2 value shows that there is small linear correlation between the predicted value and the actual BHBA value.

The mean accuracy of of the classifiers was 74.0(± 0.275), compared to the mean accuracy of the regressors 70.5(± 1.36). With a significance threshold of $\alpha = 0.05$, we can conclude that classification task perform significantly better with a p-value of 0.024 under 44 paired samples.

Metric	Statistics	Herd-normalised statistics	Trend	Herd-normalised trend
AUC	0.70	0.71	0.65	0.63
Se@95Sp	18	15	17	19
Pr@95Re	28	27	26	26
Accuracy	75.32	73.25	72.29	73.68

Table 5.5: Aggregated feature sets (with normalisation) under measurement-date based time window and a Random Forest model. The normalisation is based on on difference from the herd (see Section 4.4.3) The statistics features contain the mean, variance, minimum and maximum of daily (normalised) values within the time window. The trend features contain the slope and offset of a linear fit on the daily (normalised) values within the time window.

	AUC	Se@95Sp	Pr@95Re	Accuracy
Static features	**1.12** (±0.12)	**1.88** (±0.52)	**2.50** (±0.85)	3.88 (±0.90)
All Features	3.88 (±1.09)	3.12 (±0.72)	6.50 (±1.41)	**2.88** (±0.69)
Trend	4.38 (±0.84)	7.12 (±0.72)	7.25 (±0.92)	5.88 (±1.11)
All behaviour features	5.62 (±1.12)	4.38 (±1.27)	7.88 (±1.25)	4.00 (±1.13)
Daily values	5.62 (±0.65)	4.50 (±0.73)	5.88 (±1.23)	5.75 (±1.11)
Statistics	5.75 (±0.56)	8.62 (±0.94)	7.38 (±1.16)	7.00 (±1.24)
Daily values herd normalised	6.12 (±1.01)	6.00 (±0.98)	7.38 (±0.71)	6.12 (±1.01)
Statistics herd normalised	7.12 (±0.77)	7.88 (±0.77)	7.88 (±1.17)	8.75 (±1.03)
Daily values within window normalised	8.38 (±0.94)	8.38 (±1.31)	6.50 (±1.02)	7.00 (±0.80)
Trend herd normalised	8.62 (±0.91)	8.00 (±0.80)	8.50 (±0.60)	5.88 (±0.83)
Daily values prepartum normalised	10.00 (±0.85)	7.38 (±1.12)	2.88 (±0.67)	9.38 (±0.89)
FFT features	10.75 (±0.25)	9.50 (±1.19)	3.00 (±0.58)	11.00 (±0.71)

Table 5.6: Feature set mean rank (±SE) sorted on AUC mean rank. The highest average rank is highlighted in bold. The rank was calculated by comparing all feature sets under the same conditions (equal window and model)

Metric	GB	MLP	NB	RF
AUC	0.651 (±0.014)	0.619 (±0.0093)	0.626 (±0.0082)	0.664 (±0.012)
Se@95Sp	17 (±1.9)	10.5 (±1.2)	11.5 (±1.5)	16.2 (±1.6)
Pr@95Re	28.6 (±0.38)	28.1 (±0.32)	28.2 (±0.29)	28.2 (±0.38)
Accuracy	74.3 (±0.47)	71.9 (±1.1)	62.4 (±2.4)	74.2 (±0.49)

Table 5.7: Mean metrics scores (±SE) of models on all experiments. GB is Gradient Boosting, MLP is Multilayer Perceptron, NB is Naive Bayes and RF is Random Forest. On all metrics, Gradient Boosting performed best on a small margin, in AUC Random Forest performed best.

Metric	Static Features	Behaviour Features	All Features
RMSE	0.42	0.43	0.43
Q95E	0.84	0.84	0.86
R^2	0.16	0.11	0.13
Accuracy	77.06	75.44	75.44

Table 5.8: Regression statistics with a measurement based window and a Random Forest Regression model

6 Discussion

This chapter discussed the quality of this research in three sections. Section 6.1 discusses the results in relation to the state-of-the-art. Section 6.2 discusses the quality of the data used in this study, Section 6.3 discusses the quality of the results and

6.1 Quality of the state-of-the-art

Compared with other related studies, this study was one of the largest in the data samples. By combining the two experiments, we gathered 1581 samples of which 1192 (or 1215 depending on the time window) were usable. The largest study in the related work consisted of 1374 cows[25], followed by 1080[68] and 706[70]. However, these studies were limited by their methodology.

Stangaferro et al. used a proprietary model, which makes it impossible to reproduce. Furthermore, their evaluation metric was sensitivity, which can be altered as high as possible using a prediction probability and threshold (see Section 2.9. In the study of Steensels et al., the specificity score of 70% is considered very low, because the prevalence of ketosis in that study was 29%. This means there are 150 false positives, which will generate a lot of attentions for the farmer, in turn reducing the trust in the system.

During the study, we also executed a reproduction of the study of Eckelkamp et al.[25]. This study reported higher sensitivity scores compared to this study. The experiments in this study also measured BHBA values using the same measurement device. When this value was higher than 1.2 mmolL, it was marked as ketosis. The study also used daily data, however no time window was used. Instead, each day was a separate sample and days without measurement were considered negative for ketosis. Furthermore, data from days before calving was also added and marked as negative. As seen in Section 2.1, the risk for ketosis is low during this period, so the relevancy of this data is questionable. Therefore, this study has improved on the methodology of the state-of-the-art by contributing to many more aspects of the ketosis detection process.

Although the initial number of BHBA values in this study was large compared to the state-of-the-art, the resulting data set size is still considered small in machine learning. In addition to a relatively small total size of the data set, the small number of cows per farm and the restriction of a single measurement after calving were also limiting factors. The total number of samples is due to a limited experiment setup with only ten farms and the removal of certain animals because of missing data. Because the experiments occurred in the Netherlands, where the typical size of a farm is about 100-200 cows, the number of cows per farm is low. Single measurements can be misleading in the case of ketosis. The sample taken is an approximation of the cow well-being.

Studies show that the time of day has influence on the level of BHBA within the blood, due to daily variations such as rest and feeding times. Furthermore, the test is inherently susceptible to errors, because the testing machinery is not 100 percent accurate. Therefore, a set of measurements over a few days period is more representative of the state of ketosis within a cow and should be a better target for ketosis detection systems.

6.2 Quality of the data

The experiment data contained a lot of errors and missing values, especially in the SenseOfSensors experiment. We were able to correct some of these errors and missing values using our procedure, but still imputation was needed to fill the gaps in the experiment data. This increases the dependence on the training data and thereby increases the amount of overfitting to the data set. As seen in Figure A.10, the amount of data for each cow was not equal.

The behaviour data contained gaps in their behaviour data, it lacked data in the days before calving and some cows did not have any behaviour data. This affected the number of usable samples in training and testing and thereby decreases the performance of the machine learning models. The number of positive cases of subclinical ketosis is imbalanced compared to the negative cases; 305 of 1581 experiment samples in the first week. Therefore, missing data has more impact on the generalisation of positive cases compared to the negative cases. Moreover, 59 (measurement-date based) to 75 (calving-date based) samples lacked prepartum data, which affected the performance comparison between either prepartum normalisation or prepartum features as the number of samples were unequal.

The behaviour data was represented as minutes spent at a certain behaviour per day. However, the measurement of time of a certain behaviour does not fully represent the dynamics of this behaviour. For instance, the behaviour data contains eating time, but this is not the same as feed intake. The number of minutes eaten does not tell the speed of feed intake, the amount per bite or the amount of swallows. Therefore while the general behaviour is captured in the data, the dynamics of the behaviour are hidden. Moreover, this behaviour is not a raw sensor measurement. Instead, the sensors have an integrated a classification algorithm based on human-created features from the accelerometer data. This means that the behaviour is processed from raw sensor data. Even though its correlation is high compared to human observation[75], it still introduces extra noise in the data. As the validation study is performed on a single farm, it is unknown whether the correlation is high in a general setting. This affects the reliability of the behaviour data

6.3 Quality of the results

In this study, we used four evaluation metrics in the results. In the pairwise comparison of the different research questions, we used a Bonferroni correction to account for the multiple testing problem. This is viewed as a conservative measure in the multiple testing world. However, this did not impact any results in this study; the p-values were either above the $\alpha = 0.05$ threshold or below the Bonferroni corrected $\alpha = \frac{0.05}{4} = 0.0125$ threshold.

As a general remark, we note that the results presented in Chapter 5 are all based on the same experiments. The difference is the grouping of the results to compare time windows, feature sets or models. Important to note is that the set of compared variables in one aspect also affects the

others. For example, the machine learning models includes the MLP model, but in the results, this model seems to perform randomly, irrespective of time window or feature set. In the within-window normalisation feature set with measurement-date based window, the MLP model scores 0.64 AUC, much higher than the other models (see Table B.5. However, in the daily values feature set with measurement-date based window, the MLP scores 0.58 AUC, but other models perform much better (see Table B.6). The comparison between feature sets is affected here by the inclusion of the MLP model. This lead to insignificant results in the time window comparison and normalisation comparison and in the average feature ranking.

The results presented aimed to answer the five sub-research questions proposed in Section 1.2. Before making any conclusions, there are some remarks regarding these questions. The results on the first research question, *How do a measurement-date based time-window and a calving-date based time-window compare with respect to the quality of ketosis detection?*, were affected by the moment of measurement in the experiments. As seen in Figure 4.3, the calving window could fail to overlap with the moment of measurement, which in this case could render the target incorrect, as it is unknown whether the cow was suffering from ketosis somewhere in the calving window. This could be prevented by enlarging the window size of the calving-date based window. However, the size of the time window was set at five days for each window to equalise the amount of input features between the MB window and CB window and limit the amount of samples to be discarded, as many samples had incomplete behaviour data.

Moreover, the samples each window contained were not equal. This also affected the comparison, as both time windows had other training and test samples. When analysing the the different samples in the training set, the mean BHBA of samples not in the calving-date window was 1.36 compared to the mean BHBA of 0.83 of the samples not in the measurement-date window. This means that the calving-date based window had fewer positive SCK samples to train on.

In the results of the second sub research question, *How does herd-normalisation compare to other normalisation methods and non-normalised data?*, we note that the size of some of the participating farms is very small. This has an impact on the herd-normalisation, which relies on the mean and standard deviation of the herd. Then individual cows have a relatively large influence on the mean.

When performing the experiments on single farms, the results are mixed. For instance, farm 4 performs worse (0.60 AUC; Table C.3 on daily values compared to the entire data set (0.70 AUC; 5.4), whereas farm 8 performs much better (0.81; Table C.4. The BHBA level spread in these farms is almost equal (Figure 5.4, so there are other factor factors which influence the prediction. These results explain why herd-normalisation did not improve ketosis detection. When individual farm classifiers would all perform better, there must be specific farm features lost in the entire data set, which should become visible after normalising for each farm.

Regarding the results of the third sub-research question, *What is the effect of different feature sets on subclinical ketosis detection?* , remarks on the results have to be made. The performance of raw daily values under the four metrics (see Table 5.6) can be limited to this test set. As the training and test data set is a random split on the same data set, there is a large chance that the performance does not translate to other farms in different countries.

As a result, static risk factors perform best in the feature ranking. The problem is that this feature set only has 8 features, while the behaviour sets have a minimum of $10 \times x$ where x is the number of features derived from the time windowed behaviour . For example, daily raw values have 50 features. While feature selection was applied, the number of features is still multiple times bigger. This impact is seen in the difference between the "All feature" set result and the static feature set

result. All feature set contains more than 200 features including risk factors. However, the result is performed worse on all metrics.

The results of the fourth sub-research question, *Which machine learning model has the best at detecting subclinical ketosis?*, Naive Bayes was unable to deal with the high number of features. This was signified by the result of Naive Bayes in the Static feature set (0.71 AUC) compared to the All feature set(0.60 AUC).

The MLP had the issue of random results, as mentioned earlier. This is probably due to overfitting to the training set caused by the high number of iterations (10000) used to train the MLP. These were necessary to converge the model in a stable state, however it hurt the model by overfitting.

The last sub-research question, *How precisely can BHBA values be predicted using a regression model and how does this model compare to the classification model?*, for this research concerned the application of regression models. The assumption was that the SCK threshold was not a good fit for all cows and that some cows marked with SCK were actually healthy and vice versa. However, the regression approach did not yield better results. Especially the high BHBA levels were difficult to predict, illustrated by the difference between the Q95 error and the root mean squared error, where Q95E is twice the RMSE for the best regressor (Table 5.8). When looking at the predicted-actual plots (Figure B.3; Appendix B), it shows that the regressors predict around the mean of the BHBA values. There is a very low correlation between the predicted values and actual values in general, seen in these plots and with the R^2 value of 0.16 of the best regressor. Thus, the current regression models do not fit the problem very well.

7 Conclusion

The aim of this research was to find out to what extent subclinical ketosis can be detected in dairy cows using behaviour data focusing on all steps in the machine learning process. Behaviour data derived from neck and leg sensors were combined with BHBA measurements. The machine learning classifiers were scored on AUC, Sensitivity@95%Specificity, Precision@95%Recall and Accuracy. Using both raw and normalised data, different feature sets were formed and compared.

How do a measurement-date based time-window and a calving-date based time-window compare with respect to the quality of ketosis detection? Two windowing methods were created to limit the amount of behaviour data for each cow. Measurement-date based windows were already used in the initial experiments for ketosis detection, but these were not significantly better than calving-date based windows in any metrics with p-values of 0.417, 0.35, 0.997 and 0.124.

How does herd-normalisation compare to other normalisation methods and non-normalised data? The variation in behaviour by dairy cows is known to be significantly different between cows. In the data visualisations from Section 5.1 we saw a large spread of behaviour for each day. Therefore, normalisation should be answer to equalise these differences and really focus on the oddities. However, as shown in the results, normalisation did not improve classification. Specific normalisation techniques to factor out farm-specific behaviour were applied, but the AUC of daily values were significantly better with 0.651 mean AUC compared to 0.628 mean AUC for herd normalised features.

What is the effect of different feature sets on subclinical ketosis detection? Static prepartum features were added as a baseline. In a mean ranking order, the static features outranked all behaviour feature sets, including the set with static features included with an average rank of 1.12, 1.88, 2.50 and 2.88 on AUC, Se@95Sp, Pr@95Re and Accuracy respectively.

Which machine learning model has the best at detecting subclinical ketosis? Four different classical machine learning models were trained on the feature sets and compared to each other. The ensemble models of Random Forest and Gradient Boosting outperformed on each metric with average AUC scores of 0.664 and 0.651, respectively.

How precisely can BHBA values be predicted using a regression model and how does this model compare to the classification model? At last, the continuous values of BHBA were used directly to perform a regression learning to provide a more nuanced approach to subclinical ketosis detection. However, the regression model failed to predict higher BHBA values with a 95 percentile error of 0.84 and a R^2 score of 0.16 with Static features and a Random Forest regressor. Moreover, compared to the classification model, the predicted BHBA values were significantly worse at detecting SCK with the standard threshold value. The average accuracy score of the regression model was 70.5 compared to the average accuracy of 74.0.

To what extent can subclinical ketosis be detected in dairy cattle using peripartum data in a machine learning approach? The best model in this study consisted of static features with a Random Forest model with an AUC score of 0.76. Overall, the ketosis detection model is not reliable enough for commercial usage. However, this study presented a thorough methodology for ketosis detection with behaviour data and machine learning. Moreover, with small adjustments in the time window and target, this methodology is also applicable for other (transition) cow diseases.

7.1 Recommendations

To develop a more robust SCK detection, the focus should lie on the gathering of more cows with more samples on larger farms. The data set of this study contained 1581 cows with BHBA measurement. Depending on the methodology used, this should be increased by a factor 10 for the current methods, or by a factor 1000 for more complex methods such as deep learning.

In this study, data was lost because of unavailable prepartum behaviour data and incomplete experiment data. From the 1581 BHBA measurements, only 65% (1037) could be used in the prepartum features. Furthermore if multiple BHBA measurements were taken, progression of the disease can be used to create new features which are better at distinguishing healthy cows and cows suffering from SCK.

Herd normalisation was applied unsuccessfully in this study. One of the reasons discussed in Section 6.3 was the size of the farms. This normalisation technique should perform better if larger farms are considered. Typical Dutch farms contain about 150 cows, however in other countries farms with more than 1000 cows exist. Even when herd-normalisation would not work on those larger farms, the data is still useful to develop other methods to normalise farm-specific behaviour.

Addition of other types of automated measurable data of cows is worth considering. Milk data is missing in this study, but related work showed that this data is valuable in relation to ketosis detection. Moreover ketosis has been related to lower amounts of socialising within the herd, so features representing that can prove to be valuable.

Deep learning neural networks have shown to outperform all other methods on a variety of tasks given enough data. Furthermore, in time series classification, deep learning has seen a recent popularity. It has beaten the widely used DTW algorithm. However, the data in this study was limited and since the DTW experiments showed very little promise, we decided to limit this study to classical learners such as Random Forest and Naive Bayes. Given more data samples and more tests, the proven capabilities of deep learning can provide improvements to subclinical ketosis detection.

Bibliography

[1] A.A. Adewuyi, E. Gruys, and F.J.C.M. van Eerdenburg. Non esterified fatty acids (NEFA) in dairy cattle. A review. *Veterinary Quarterly*, 27(3):117–126, September 2005.

[2] Lennart Andersson. Subclinical Ketosis in Dairy Cows. *Veterinary Clinics of North America: Food Animal Practice*, 4(2):233–251, July 1988.

[3] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, May 2017.

[4] S. Benaissa, F.A.M. Tuyttens, D. Plets, J. Trogh, L. Martens, L. Vandaele, W. Joseph, and B. Sonck. Calving and estrus detection in dairy cattle using a combination of indoor localization and accelerometer sensors. *Computers and Electronics in Agriculture*, 168, 2020.

[5] Daniel Berckmanns and European Conference on Precision Livestock Farming. *Precision livestock farming '13 papers presented at the 6th European Conference on Precision Livestock Farming, Leuven, Belgium, 10-12 September '13*. Univ., Leuven, 2013. OCLC: 885442316.

[6] Anna C. Berge and Geert Vertenten. A field study to determine the prevalence, dairy herd management systems, and fresh cow clinical conditions associated with ketosis in western European dairy herds. *Journal of Dairy Science*, 97(4):2145–2154, April 2014.

[7] J.P. Bikker, H. van Laar, P. Rump, J. Doorenbos, K. van Meurs, G.M. Griffioen, and J. Dijkstra. Technical note: Evaluation of an ear-attached movement sensor to record cow feeding behavior and activity. *Journal of Dairy Science*, 97(5):2974–2979, May 2014.

[8] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006.

[9] V. Bonfatti, S.-A. Turner, B. Kuhn-Sherlock, T.D.W. Luke, P.N. Ho, C.V.C. Phyn, and J.E. Pryce. Prediction of blood β-hydroxybutyrate content and occurrence of hyperketonemia in early-lactation, pasture-grazed dairy cows using milk infrared spectra. *Journal of Dairy Science*, 102(7):6466–6476, July 2019.

[10] M.R. Borchers and J.M. Bewley. An assessment of producer precision dairy farming technology use, prepurchase considerations, and usefulness. *Journal of Dairy Science*, 98(6):4198–4205, June 2015.

[11] M.R. Borchers, Y.M. Chang, K.L. Proudfoot, B.A. Wadsworth, A.E. Stone, and J.M. Bewley. Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. *Journal of Dairy Science*, 100(7):5664–5674, July 2017.

[12] M.R. Borchers, Y.M. Chang, I.C. Tsai, B.A. Wadsworth, and J.M. Bewley. A validation of technologies monitoring dairy cow feeding, ruminating, and lying behaviors. *Journal of Dairy Science*, 99(9):7458–7466, September 2016.

[13] Nikolaus Brunner, Stephan Groeger, Joao Canelas Raposo, Rupert M Bruckmaier, and Josef J Gross. Prevalence of subclinical ketosis and production diseases in dairy cows in Central and South America, Africa, Asia, Australia, New Zealand, and Eastern Europe1. *Translational Animal Science*, 3(1):84–92, January 2019.

[14] D.F. Calderon and N.B. Cook. The effect of lameness on the resting behavior and metabolic status of dairy cattle during the transition period in a freestall-housed dairy herd. *Journal of Dairy Science*, 94(6):2883–2894, June 2011.

[15] European Commission. Climate change - driving forces.

[16] Ruan R. Daros, Hanna K. Eriksson, Daniel M. Weary, and Marina A.G. von Keyserlingk. The relationship between transition period diseases and lameness, feeding time, and body condition during the dry period. *Journal of Dairy Science*, 103(1):649–665, January 2020.

[17] R. M. De Mol. *Automated detection of oestrus and mastitis in dairy cows*. PhD Thesis, S.n., 2000.

[18] R.M. de Mol and W.E. Woldt. Application of Fuzzy Logic in Automated Cow Status Monitoring. *Journal of Dairy Science*, 84(2):400–410, February 2001.

[19] I. Dittrich, M. Gertz, and J. Krieter. Alterations in sick dairy cows' daily behavioural patterns. *Heliyon*, 5(11):e02902, November 2019.

[20] I. R. Dohoo and S. W. Martin. Subclinical ketosis: prevalence and associations with production and disease. *Canadian Journal of Comparative Medicine: Revue Canadienne De Medecine Comparee*, 48(1):1–5, January 1984.

[21] J. K. Drackley. ADSA Foundation Scholar Award. Biology of dairy cows during the transition period: the final frontier? *Journal of Dairy Science*, 82(11):2259–2273, November 1999.

[22] T.F. Duffield, K.D. Lissemore, B.W. McBride, and K.E. Leslie. Impact of hyperketonemia in early lactation dairy cows on health and production. *Journal of Dairy Science*, 92(2):571–580, February 2009.

[23] T.F. Duffield, D. Sandals, K.E. Leslie, K. Lissemore, B.W. McBride, J.H. Lumsden, P. Dick, and R. Bagg. Efficacy of Monensin for the Prevention of Subclinical Ketosis in Lactating Dairy Cows. *Journal of Dairy Science*, 81(11):2866–2873, November 1998.

[24] E.A. Eckelkamp. Invited Review: Current state of wearable precision dairy technologies in disease detection. *Applied Animal Science*, 35(2):209–220, April 2019.

[25] Elizabeth A. Eckelkamp. *On-Farm Utilization of Precision Dairy Monitoring: Usefulness, Accuracy, and Affordability*. PhD thesis, University of Kentucky Libraries, 2018.

[26] J.L. Edwards and P.R. Tozer. Using Activity and Milk Yield as Predictors of Fresh Cow Disorders. *Journal of Dairy Science*, 87(2):524–531, February 2004.

[27] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, October 2001.

[28] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian Network Classifiers. *Machine Learning*, 29(2/3):131–163, 1997.

[29] A.R. Frost, C.P. Schofield, S.A. Beaulah, T.T. Mottram, J.A. Lines, and C.M. Wathes. A review of livestock monitoring and the need for integrated systems. *Computers and Electronics in Agriculture*, 17(2):139–159, May 1997.

[30] J.P. Goff and R.L. Horst. Physiological Changes at Parturition and Their Relationship to Metabolic Disorders. *Journal of Dairy Science*, 80(7):1260–1268, July 1997.

[31] C. Goldhawk, N. Chapinal, D.M. Veira, D.M. Weary, and M.A.G. von Keyserlingk. Prepartum feeding behavior is an early indicator of subclinical ketosis. *Journal of Dairy Science*, 92(10):4971–4977, October 2009.

[32] L.A. González, B.J. Tolkamp, M.P. Coffey, A. Ferret, and I. Kyriazakis. Changes in Feeding Behavior as Possible Indicators for the Automatic Monitoring of Health Disorders in Dairy Cows. *Journal of Dairy Science*, 91(3):1017–1028, March 2008.

[33] Thomas H. Herdt. Ruminant Adaptation to Negative Energy Balance. *Veterinary Clinics of North America: Food Animal Practice*, 16(2):215–230, July 2000.

[34] I. Iguyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[35] K.L Ingvartsen, R.J Dewhurst, and N.C Friggens. On the relationship between lactational performance and health: is it yield or metabolic imbalance that cause production diseases in dairy cattle? A position paper. *Livestock Production Science*, 83(2-3):277–308, October 2003.

[36] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, July 2019.

[37] A.J. Itle, J.M. Huzzey, D.M. Weary, and M.A.G. von Keyserlingk. Clinical ketosis and standing behavior in transition cows. *Journal of Dairy Science*, 98(1):128–134, January 2015.

[38] E.I. Kaufman, S.J. LeBlanc, B.W. McBride, T.F. Duffield, and T.J. DeVries. Association of rumination time with subclinical ketosis in transition dairy cows. *Journal of Dairy Science*, 99(7):5604–5618, July 2016.

[39] M.T.M. King, K.M. Dancy, S.J. LeBlanc, E.A. Pajor, and T.J. DeVries. Deviations in behavior and productivity data before diagnosis of health disorders in cows milked with an automated system. *Journal of Dairy Science*, 100(10):8358–8371, October 2017.

[40] Stephen Leblanc. Monitoring Metabolic Health of Dairy Cattle in the Transition Period. *Journal of Reproduction and Development*, 56(S):S29–S35, 2010.

[41] Amanda Lee. AN EVALUATION OF PHYSIOLOGICAL AND BEHAVIORAL INDICATORS OF POSTPARTUM DISEASES AND HEAT STRESS IN DAIRY COWS. *Theses and Dissertations–Animal and Food Sciences*, January 2018.

[42] Daniela N. Liboreiro, Karine S. Machado, Paula R.B. Silva, Milton M. Maturana, Thiago K. Nishimura, Alice P. Brandão, Márcia I. Endres, and Ricardo C. Chebel. Characterization of peripartum rumination and activity of cows diagnosed with metabolic and uterine diseases. *Journal of Dairy Science*, 98(10):6812–6827, October 2015.

[43] J.A.A. McArt, D.V. Nydam, and G.R. Oetzel. Epidemiology of subclinical ketosis in early lactation dairy cattle. *Journal of Dairy Science*, 95(9):5056–5066, September 2012.

[44] J.A.A. McArt, D.V. Nydam, and G.R. Oetzel. Dry period and parturient predictors of early lactation hyperketonemia in dairy cattle. *Journal of Dairy Science*, 96(1):198–209, January 2013.

[45] J.A.A. McArt, D.V. Nydam, P.A. Ospina, and G.R. Oetzel. A field trial on the effect of propylene glycol on milk yield and resolution of ketosis in fresh cows diagnosed with subclinical ketosis. *Journal of Dairy Science*, 94(12):6011–6020, December 2011.

[46] J.A.A. McArt, D.V. Nydam, and M.W. Overton. Hyperketonemia in early lactation dairy cattle: A deterministic estimate of component and total cost per case. *Journal of Dairy Science*, 98(3):2043–2054, March 2015.

[47] Jessica A.A. McArt, Daryl V. Nydam, Garrett R. Oetzel, Thomas R. Overton, and Paula A. Ospina. Elevated non-esterified fatty acids and β-hydroxybutyrate and their association with transition dairy cow performance. *The Veterinary Journal*, 198(3):560–570, December 2013.

[48] G. A. Miller, M. Mitchell, Z. E. Barker, K. Giebel, E. A. Codling, J. R. Amory, C. Michie, C. Davison, C. Tachtatzis, I. Andonovic, and C.-A. Duthie. Using animal-mounted sensor technology and machine learning to predict time-to-calving in beef and dairy cows. *animal*, pages 1–9, January 2020.

[49] P.F. Mostert, C.E. van Middelaar, E.A.M. Bokkers, and I.J.M. de Boer. The impact of subclinical ketosis in dairy cows on greenhouse gas emissions of milk production. *Journal of Cleaner Production*, 171:773–782, January 2018.

[50] Lene Munksgaard, Margit B. Jensen, Lene J. Pedersen, Steffen W. Hansen, and Lindsay Matthews. Quantifying behavioural priorities—effects of time constraints on behaviour of dairy cows, Bos taurus. *Applied Animal Behaviour Science*, 92(1-2):3–14, July 2005.

[51] Nour-Addeen Najm, Lisa Zimmermann, Oliver Dietrich, Anna Rieger, Rainer Martin, and Holm Zerbe. Associations between motion activity, ketosis risk and estrus behavior in dairy cattle. *Preventive Veterinary Medicine*, 175:104857, February 2020.

[52] M. Nielen, M. G. Aarts, A. G. Jonkers, T. Wensing, and Y. H. Schukken. Evaluation of two cowside tests for the detection of subclinical ketosis in dairy cows. *The Canadian Veterinary Journal = La Revue Veterinaire Canadienne*, 35(4):229–232, April 1994.

[53] P.A. Ospina, D.V. Nydam, T. Stokol, and T.R. Overton. Association between the proportion of sampled transition cows with increased nonesterified fatty acids and β-hydroxybutyrate and disease incidence, pregnancy rate, and milk production at the herd level. *Journal of Dairy Science*, 93(8):3595–3601, August 2010.

[54] P.A. Ospina, D.V. Nydam, T. Stokol, and T.R. Overton. Evaluation of nonesterified fatty acids and β-hydroxybutyrate in transition dairy cattle in the northeastern United States: Critical thresholds for prediction of clinical diseases. *Journal of Dairy Science*, 93(2):546–554, February 2010.

[55] T.R. Overton, J.A.A. McArt, and D.V. Nydam. A 100-Year Review: Metabolic health indicators and management of dairy cattle. *Journal of Dairy Science*, 100(12):10398–10417, December 2017.

[56] S. Paudyal, F. P. Maunsell, J. T. Richeson, C. A. Risco, D. A. Donovan, and P. J. Pinedo. Rumination time and monitoring of health disorders during early lactation. *animal*, 12(7):1484–1492, July 2018.

[57] D. Raboisson, M. Mounié, E. Khenifar, and E. Maigné. The economic impact of subclinical ketosis at the farm level: Tackling the challenge of over-estimation due to multiple interactions. *Preventive Veterinary Medicine*, 122(4):417–425, December 2015.

[58] J.R. Roche, N.C. Friggens, J.K. Kay, M.W. Fisher, K.J. Stafford, and D.P. Berry. Invited review: Body condition score and its association with dairy cow productivity, health, and welfare. *Journal of Dairy Science*, 92(12):5769–5801, December 2009.

[59] C.J. Rutten, C. Kamphuis, H. Hogeveen, K. Huijps, M. Nielen, and W. Steeneveld. Sensor data on cow activity, rumination, and ear temperature improve prediction of the start of calving in dairy cows. *Computers and Electronics in Agriculture*, 132:108–118, January 2017.

[60] C.J. Rutten, W. Steeneveld, A.G.J.M. Oude Lansink, and H. Hogeveen. Delaying investments in sensor technology: The rationality of dairy farmers' investment decisions illustrated within the framework of real options theory. *Journal of Dairy Science*, 101(8):7650–7660, August 2018.

[61] C.J. Rutten, A.G.J. Velthuis, W. Steeneveld, and H. Hogeveen. Invited review: Sensors to support health management on dairy farms. *Journal of Dairy Science*, 96(4):1928–1952, April 2013.

[62] Mohammad W. Sahar, Annabelle Beaver, Daniel M. Weary, and Marina A.G. von Keyserlingk. Feeding behavior and agonistic interactions at the feed bunk are associated with hyperketonemia and metritis diagnosis in dairy cattle. *Journal of Dairy Science*, 103(1):783–790, January 2020.

[63] M Saint-Dizier and S Chastant-Maillard. Towards an Automated Detection of Oestrus in Dairy Cattle: Automated Oestrus Detection in Dairy Cattle. *Reproduction in Domestic Animals*, 47(6):1056–1061, December 2012.

[64] K. Schirmann, M.A.G. von Keyserlingk, D.M. Weary, D.M. Veira, and W. Heuwieser. Technical note: Validation of a system for monitoring rumination in dairy cows. *Journal of Dairy Science*, 92(12):6052–6055, December 2009.

[65] K. Schirmann, D.M. Weary, W. Heuwieser, N. Chapinal, R.L.A. Cerri, and M.A.G. von Keyserlingk. Short communication: Rumination and feeding behaviors differ between healthy and sick dairy cows during the transition period. *Journal of Dairy Science*, 99(12):9917–9924, December 2016.

[66] Kirsten Schulz, Jana Frahm, Ulrich Meyer, Susanne Kersten, Dania Reiche, Jürgen Rehage, and Sven Dänicke. Effects of prepartal body condition score and peripartal energy supply of dairy cows on postpartal lipolysis, energy balance and ketogenesis: an animal model to investigate subclinical ketosis. *Journal of Dairy Research*, 81(3):257–266, August 2014.

[67] N. Soriani, E. Trevisi, and L. Calamari. Relationships between rumination time, metabolic conditions, and health status in dairy cows during the transition period1. *Journal of Animal Science*, 90(12):4544–4554, December 2012.

[68] M.L. Stangaferro, R. Wijma, L.S. Caixeta, M.A. Al-Abri, and J.O. Giordano. Use of rumination and activity monitoring for the identification of dairy cows with health disorders: Part I. Metabolic and digestive disorders. *Journal of Dairy Science*, 99(9):7395–7410, September 2016.

[69] W. Steeneveld and H. Hogeveen. Characterization of Dutch dairy farms using sensor systems for cow management. *Journal of Dairy Science*, 98(1):709–717, January 2015.

[70] Machteld Steensels, Ephraim Maltz, Claudia Bahr, Daniel Berckmans, Aharon Antler, and Ilan Halachmi. Towards practical application of sensors for monitoring animal health; design and validation of a model to detect ketosis. *Journal of Dairy Research*, 84(2):139–145, May 2017.

[71] Elise H. Tatone, Jessica L. Gordon, Jessie Hubbs, Stephen J. LeBlanc, Trevor J. DeVries, and Todd F. Duffield. A systematic review and meta-analysis of the diagnostic accuracy of point-of-care tests for the detection of hyperketonemia in dairy cows. *Preventive Veterinary Medicine*, 130:18–32, August 2016.

[72] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282, Montreal, Que., Canada, 1995. IEEE Comput. Soc. Press.

[73] Sho Ushikubo, Chikara Kubota, and Hayato Ohwada. The Early Detection of Subclinical Ketosis in Dairy Cows Using Machine Learning Methods. In *Proceedings of the 9th International Conference on Machine Learning and Computing - ICMLC 2017*, pages 38–42, Singapore, Singapore, 2017. ACM Press.

[74] I.D.E. van Dixhoorn, R.M. de Mol, J.T.N. van der Werf, S. van Mourik, and C.G. van Reenen. Indicators of resilience during the transition period in dairy cows: A case study. *Journal of Dairy Science*, 101(11):10271–10282, November 2018.

[75] E. Van Erp-Van der Kooij. Validation of Nedap Smarttag Leg and Neck to assess behavioural activity level in dairy cattle. 2016.

[76] R. J. van Hoeij, A. Kok, R. M. Bruckmaier, M. J. Haskell, B. Kemp, and A. T. M. van Knegsel. Relationship between metabolic status and behavior in dairy cows in week 4 of lactation. *animal*, 13(3):640–648, March 2019.

[77] T. Vanholder, J. Papen, R. Bemers, G. Vertenten, and A.C.B. Berge. Risk factors for subclinical and clinical ketosis and association with production parameters in dairy cows in the Netherlands. *Journal of Dairy Science*, 98(2):880–888, February 2015.

[78] C.M. Wathes, H.H. Kristensen, J.-M. Aerts, and D. Berckmans. Is precision livestock farming an engineer's daydream or nightmare, an animal's friend or foe, and a farmer's panacea or pitfall? *Computers and Electronics in Agriculture*, 64(1):2–10, November 2008.

[79] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80, December 1945.

[80] L. Wisnieski, B. Norby, S.J. Pierce, T. Becker, J.C. Gandy, and L.M. Sordillo. Predictive models for early lactation diseases in transition dairy cattle at dry-off. *Preventive Veterinary Medicine*, 163:68–78, February 2019.

[81] Wei Xu, Ariette T.M. van Knegsel, Jacques J.M. Vervoort, Rupert M. Bruckmaier, Renny J. van Hoeij, Bas Kemp, and Edoardo Saccenti. Prediction of metabolic status of dairy cows in early lactation with on-farm cow data and machine learning algorithms. *Journal of Dairy Science*, 102(11):10186–10201, November 2019.

[82] Nils Zehner, Joël J. Niederhauser, Matthias Schick, and Christina Umstatter. Development and validation of a predictive model for calving time based on sensor measurements of ingestive behavior in dairy cows. *Computers and Electronics in Agriculture*, 161:62–71, June 2019.

Glossary

beta-hydroxybutyrate A keton body. 7, 60

calving The act of a cow delivering a calf. 6

clinical Disease showing distinct physical signs. 6

hypocalcemia A postpartum disease in which a cow has a shortage of calcium caused by start of lactation. 6

lactation Period in which a cow produces milk. 6

metritis Infection of the uterus. 6

negative energy balance When energy intake (food) is lower than energy consumption. 6

postpartum After giving birth (after calving). 8, 22–24, 29, 39, 40

prepartum Before giving birth (before calving). 8, 9, 23–25, 39, 40

subclinical Disease hiding distinct physical signs. 6

Acronyms

Acc Accuracy. 21, 36

ANN Artificial Neural Network. 21

AUC Area under the Receiver Operating Curve. 21, 36

BA Bootstrap Aggregation. 21

BCS Body Condition Score. 9, 22, 26

BHBA Beta-hydroxybutyrate. 7–9, 19, 21, 23, 24, 26, 64, 65

BN Bayesian Network. 21

BW Body Weight. 19, 21

CB calving-date based. 4, 5, 29–32, 37, 42, 43, 48, 49

CK clinical ketosis. 6, 7, 9, 21, 23

DPL Dry Period Length. 19, 21

DT Decision Tree. 21

FFT Fast Fourier Transform. 14, 15, 30, 31, 33, 34, 43

GB Gradient Boosting. 15, 35

HIS Health Index Score. 19

kNN k-Nearest Neighbours. 21

LDA Linear Discriminant Analysis. 21

Logit Logistic Regression. 21

MB measurement-date based. 4, 5, 29–32, 37, 42–45, 48, 49

MLP Multilayer Perceptron. 15, 35

NB Naïve Bayes. 15, 21, 35

NN Neural Network. 21

PAR Parity. 19, 21

PC-ANN Principal Component Artificial Neural Network. 21

PLS Partial Least Squares. 21

Pr Precision. 36

Pr@95Re Precision at 95% Recall. 36

RF Random Forest. 15, 21, 35

ROC Receiver Operating Curve. 19

SCK subclinical ketosis. 4–9, 11, 18, 19, 23, 25, 26, 29, 30, 33, 35–37, 39, 40, 48–52

Se Sensitivity. 21, 36

Se@95Sp Sensitivity at 95% Specificity. 36

SoS SenseOfSensors. 22, 23, 25, 26, 48

Sp Specificity. 21, 36

SVM Support Vector Machine. 19, 21

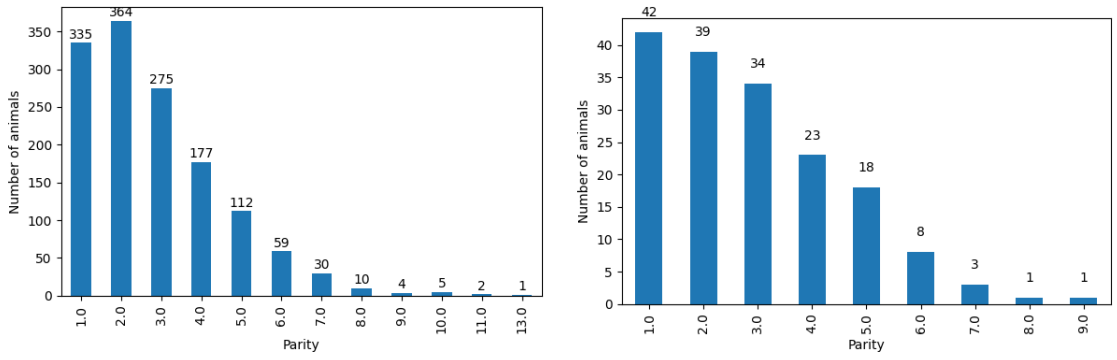
XGB XGBoost. 21

A Data exploration

A.1 Calvings and parity

(a) SenseOfSensors **(b)** EFRO

Figure A.1: Pie chart of calvings per farm. The SenseOfSensors study has a balanced distribution between farms, while the EFRO study is unbalanced.



(a) SenseOfSensors **(b)** EFRO

Figure A.2: Distribution of parity in both studies. The majority of cows has a parity ≤ 5. Three quarters of the cows are multiparous.

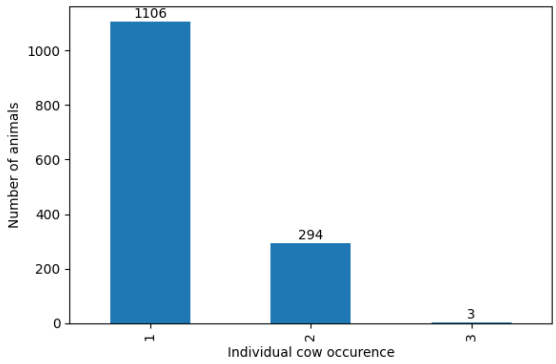
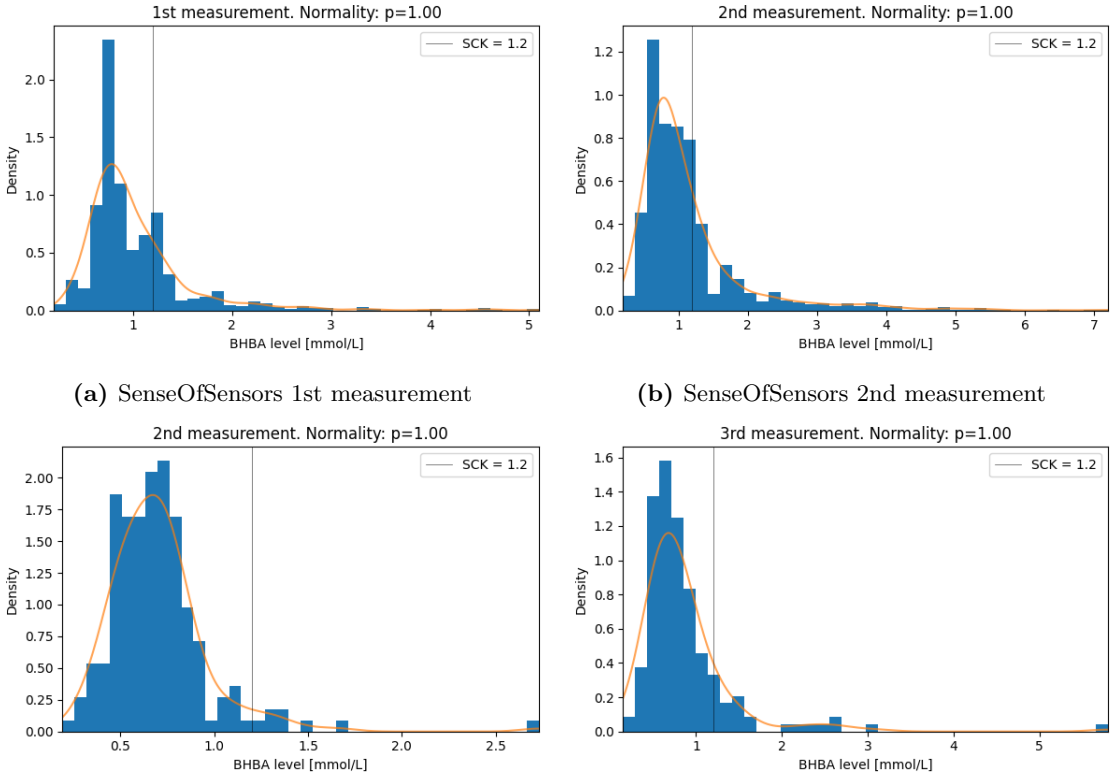


Figure A.3: Occurrence of an individual cow in the study for SenseOfSensors. About 20% of the cows had multiple calvings in this study. The EFRO study is omitted, because it lacks multiple calvings per cows

A.2 Blood measurements



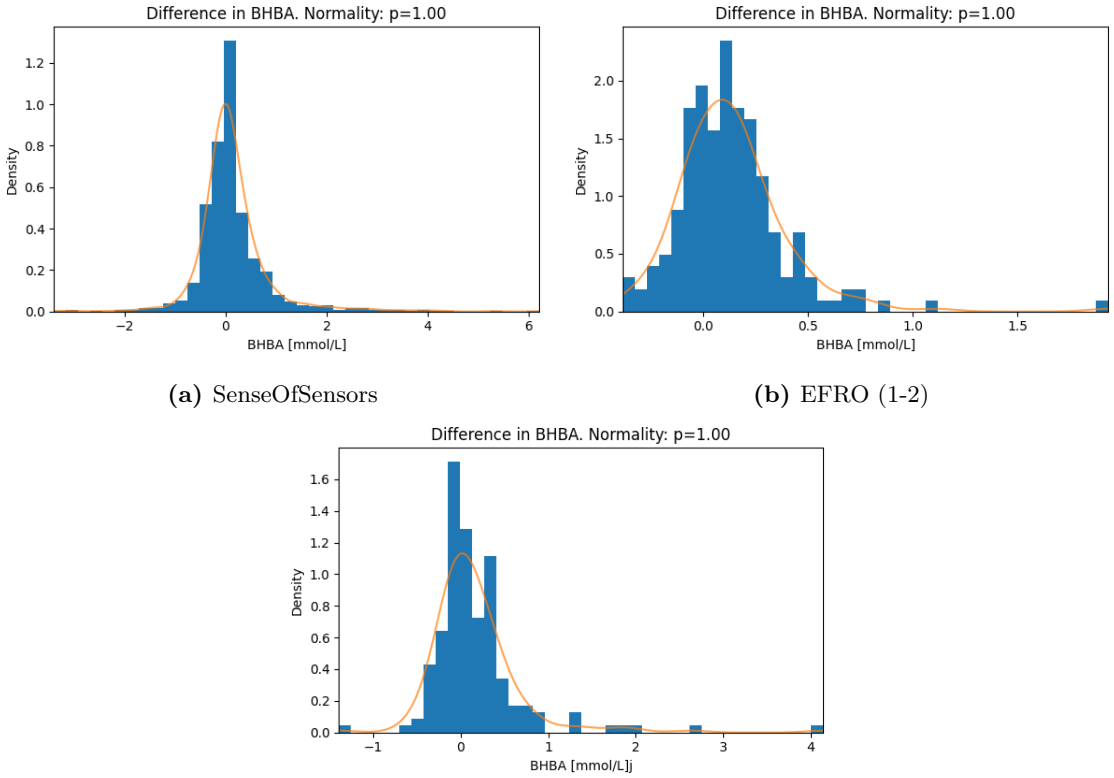
(a) SenseOfSensors 1st measurement

(b) SenseOfSensors 2nd measurement

(c) EFRO 2nd measurement

(d) EFRO 3rd measurement

Figure A.4: Histograms of BHBA measurements with a Kernel Density Estimator line and the SCK cut-off line. These histograms and lines show distributions skewed to the right. Therefore, they do not pass the normality test. Few measurements pass the SCK line.

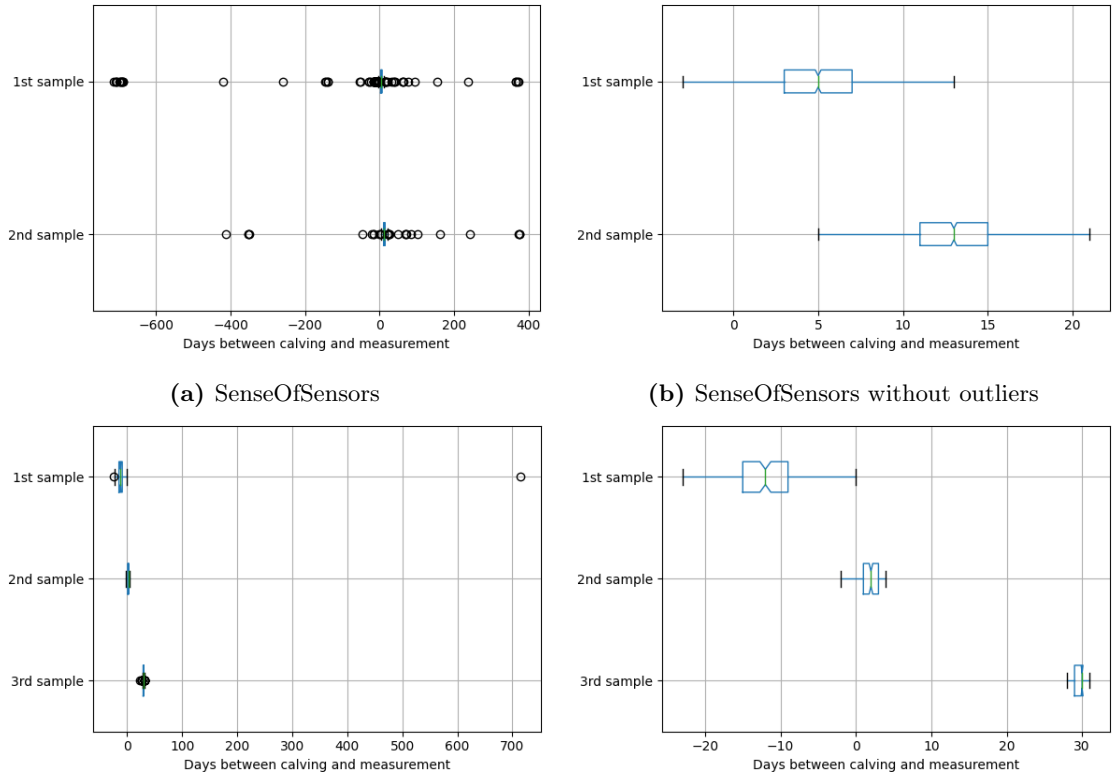


(a) SenseOfSensors

(b) EFRO (1-2)

(c) EFRO (2-3)

Figure A.5: Histograms of change between BHBA measurements. The change follows a normal distribution, with EFRO showed a slight translation of the mean to the right.



(a) SenseOfSensors

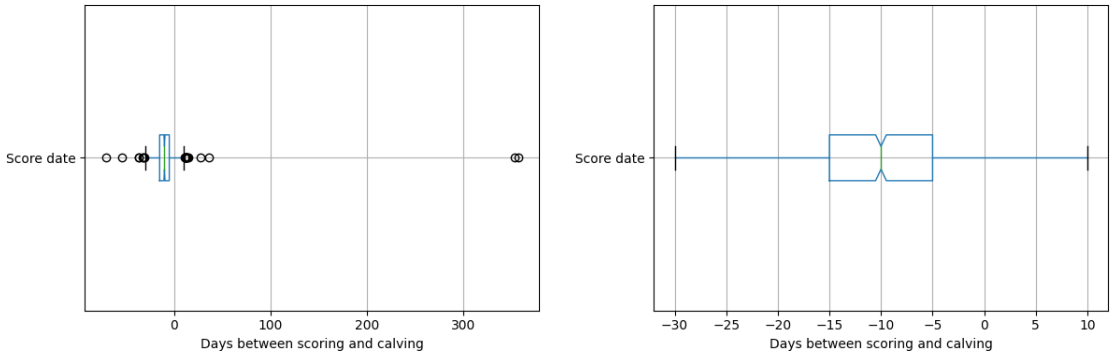
(b) SenseOfSensors without outliers

(c) EFRO

(d) EFRO without outliers

Figure A.6: Boxplots of relative measurement dates (difference between calving and measurement). SenseOfSensors contains a lot of outliers. Without outliers, the box plot confirm the study's specified measurement dates.

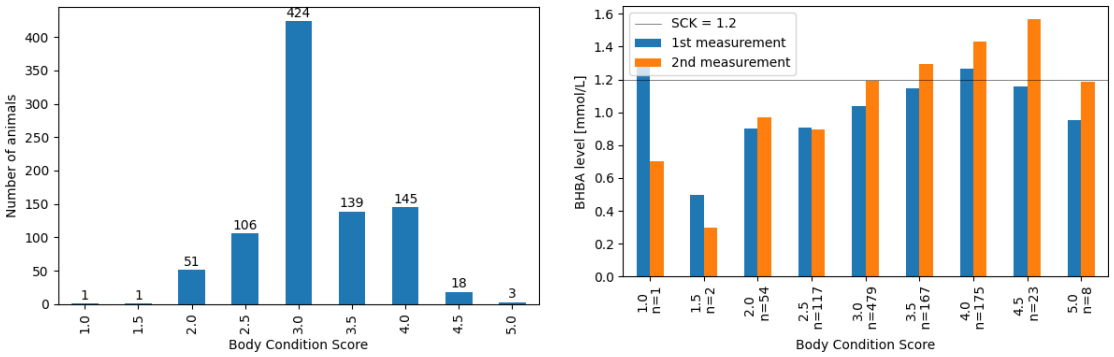
A.3 BCS & Locomotion



(a) SenseOfSensors

(b) SenseOfSensors without outliers

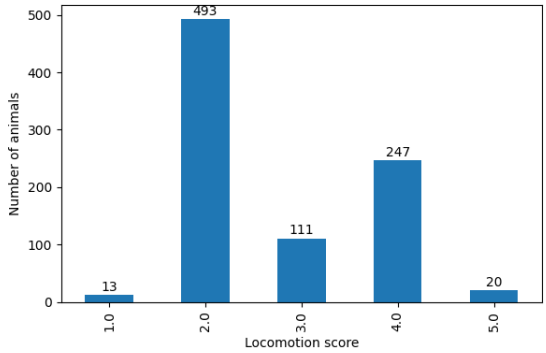
Figure A.7: Boxplots of relative scoring (BCS and locomotion) dates (difference between scoring and calving). Outliers are attributed to typing errors. On average, the scoring of BCS and locomotion is at 10 days prepartum.



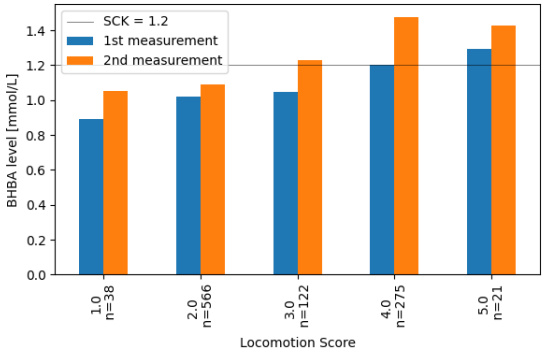
(a) Distribution of BCS

(b) Two measurements of BHBA grouped by BCS. The grey line indicated the SCK cutoff value.

Figure A.8: Body Condition Scores(BCS) in SenseOfSensors. The cow was scored at the end of the dry period. Low BCS means a shortage of body fat, high BCS means a superplus of body fat. Most cows show a BCS value of 3, which means a healthy amount of body fat. Higher BCS show higher levels of BHBA.



(a) Distribution of locomotion scores



(b) Two measurements of BHBA grouped by locomotion score. The grey line indicated the SCK cutoff value.

Figure A.9: Locomotion scores in SenseOfSensors. The cow was scored at the end of the dry period. Higher locomotion scores mean a higher impaired mobility. At higher scores, cows show a higher average BHBA measurement.

A.4 Missing values in experiment data

Feature	Training set	Test set
Parity	2	0
Lactation length	156	55
Dry period length	175	53
Gestation length	163	38
BCS	280	67
Locomotion	284	67

Table A.1: Missing values per feature based on a measurement-date index

Values missing	Training set	Test set
0	502	110
1	115	27
2	209	55
3	39	16
4	71	20
5	24	3
6	1	0

Table A.2: Amount of rows missing a number of values.

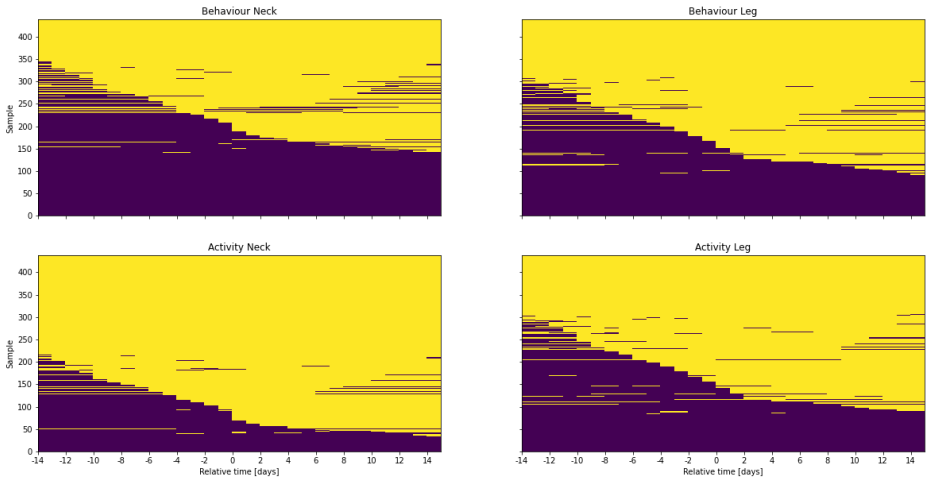


Figure A.10: Missing samples per data type.

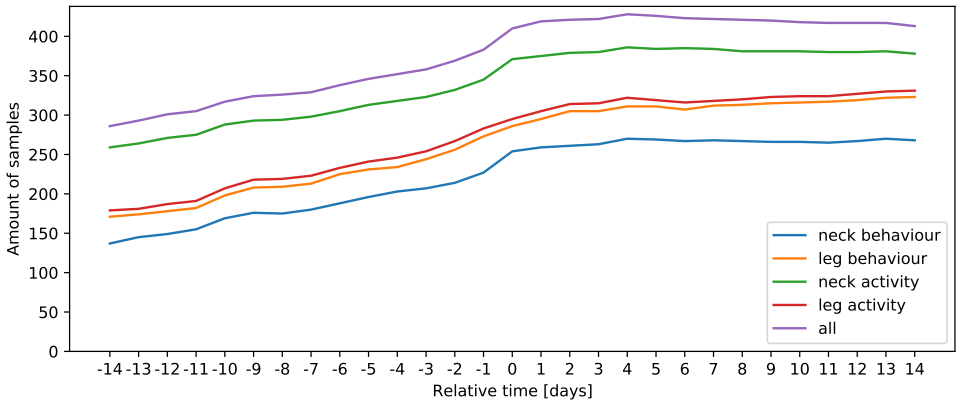


Figure A.11: Total amount of samples over time.

B Results

B.1 Classification

B.1.1 Measurement-date based window

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.74	0.22	0.27	8.47	97.63	55.56	74.56
NB	0.60	0.10	0.27	30.51	79.29	33.96	66.67
MLP	0.64	0.10	0.28	8.47	97.04	50.00	74.12
GB	0.75	0.22	0.27	15.25	97.63	69.23	76.32

Table B.1: All Features set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.72	0.22	0.27	16.95	95.27	55.56	75.00
NB	0.60	0.12	0.27	52.54	66.27	35.23	62.72
MLP	0.53	0.03	0.26	0.00	100.00	0.00	74.12
GB	0.70	0.17	0.27	15.25	98.82	81.82	77.19

Table B.2: All behaviour feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.65	0.15	0.26	15.25	94.08	47.37	73.68
NB	0.63	0.07	0.26	57.63	65.68	36.96	63.60
MLP	0.61	0.12	0.27	0.00	100.00	0.00	74.12
GB	0.64	0.19	0.27	16.95	96.45	62.50	75.88

Table B.3: Daily herd normalisation feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.62	0.22	0.31	5.17	98.03	50.00	72.38
NB	0.57	0.05	0.30	8.62	91.45	27.78	68.57
MLP	0.63	0.12	0.31	0.00	100.00	0	72.38
GB	0.63	0.10	0.29	10.34	94.08	40.00	70.95

Table B.4: Daily prepartum normalisation feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.61	0.05	0.27	18.33	86.55	32.35	68.83
NB	0.62	0.15	0.29	31.67	73.68	29.69	62.77
MLP	0.64	0.13	0.27	8.33	96.49	45.45	73.59
GB	0.59	0.05	0.29	1.67	96.49	14.29	71.86

Table B.5: Daily within-window normalisation feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.70	0.18	0.28	18.33	95.32	57.89	75.32
NB	0.63	0.15	0.28	46.67	73.68	38.36	66.67
MLP	0.58	0.12	0.28	71.67	42.11	30.28	49.78
GB	0.72	0.05	0.31	5.00	94.15	23.08	71.00

Table B.6: Daily values feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.70	0.13	0.27	16.67	94.15	50.00	74.03
NB	0.62	0.03	0.26	3.33	93.57	15.38	70.13
MLP	0.62	0.10	0.29	18.33	94.15	52.38	74.46
GB	0.69	0.17	0.26	6.67	97.66	50.00	74.03

Table B.7: Statistics feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.71	0.15	0.27	27.12	89.35	47.06	73.25
NB	0.60	0.10	0.27	81.36	17.16	25.53	33.77
MLP	0.64	0.03	0.28	8.47	93.49	31.25	71.49
GB	0.66	0.14	0.28	16.95	94.08	50.00	74.12

Table B.8: Statistics herd normalisation feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.65	0.17	0.26	33.33	85.96	45.45	72.29
NB	0.63	0.10	0.29	18.33	87.72	34.38	69.70
MLP	0.65	0.12	0.27	6.67	98.25	57.14	74.46
GB	0.64	0.08	0.27	8.33	95.32	38.46	72.73

Table B.9: Trend feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.63	0.19	0.26	35.59	86.98	48.84	73.68
NB	0.59	0.08	0.28	16.95	85.80	29.41	67.98
MLP	0.59	0.03	0.27	1.69	98.82	33.33	73.68
GB	0.61	0.10	0.27	11.86	92.31	35.00	71.49

Table B.10: Trend herd normalisation feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.75	0.20	0.31	38.33	88.30	53.49	75.32
NB	0.71	0.27	0.30	28.33	94.74	65.38	77.49
MLP	0.70	0.23	0.30	0.00	100.00	0.00	74.03
GB	0.71	0.20	0.28	13.33	96.49	57.14	74.89

Table B.11: Static feature set

B.1.2 Calving-date based window

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.74	0.23	0.35	14.75	97.14	64.29	75.85
NB	0.63	0.15	0.27	22.95	90.86	46.67	73.31
MLP	0.63	0.00	0.28	88.52	31.43	31.03	46.19
GB	0.74	0.16	0.32	13.11	96.00	53.33	74.58

Table B.12: All Features set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.69	0.16	0.31	6.56	97.71	50.00	74.15
NB	0.63	0.11	0.26	31.15	86.29	44.19	72.03
MLP	0.61	0.13	0.28	18.03	92.00	44.00	72.88
GB	0.67	0.08	0.27	6.56	96.00	36.36	72.88

Table B.13: All behaviour feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.55	0.04	0.31	18.18	84.17	31.25	65.46
NB	0.59	0.09	0.30	80.00	31.65	31.65	45.36
MLP	0.58	0.09	0.28	3.64	98.56	50.00	71.65
GB	0.56	0.07	0.30	9.09	92.81	33.33	69.07

Table B.14: FFT feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.62	0.16	0.27	14.75	96.57	60.00	75.42
NB	0.62	0.08	0.29	22.95	82.86	31.82	67.37
MLP	0.60	0.05	0.27	0.00	100.00	0.00	74.15
GB	0.62	0.07	0.26	6.56	96.00	36.36	72.88

Table B.15: Daily herd normalisation feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.52	0.08	0.30	1.67	99.34	50.00	71.70
NB	0.57	0.07	0.32	35.00	76.97	37.50	65.09
MLP	0.54	0.02	0.31	1.67	93.42	9.09	67.45
GB	0.50	0.05	0.31	3.33	96.71	28.57	70.28

Table B.16: Daily prepartum normalisation feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.59	0.10	0.27	6.56	98.86	66.67	75.00
NB	0.53	0.07	0.26	3.28	97.71	33.33	73.31
MLP	0.47	0.00	0.27	18.03	74.86	20.00	60.17
GB	0.63	0.07	0.27	1.64	96.57	14.29	72.03

Table B.17: Daily within-window normalisation feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.69	0.10	0.31	16.39	93.14	45.45	73.31
NB	0.63	0.10	0.27	39.34	79.43	40.00	69.07
MLP	0.60	0.10	0.27	32.79	80.57	37.04	68.22
GB	0.64	0.10	0.29	9.84	95.43	42.86	73.31

Table B.18: Daily values feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.63	0.18	0.28	4.92	97.71	42.86	73.73
NB	0.65	0.15	0.26	6.56	97.71	50.00	74.15
MLP	0.60	0.18	0.28	0.00	100.00	0.00	74.15
GB	0.61	0.07	0.27	6.56	93.14	25.00	70.76

Table B.19: Statistics feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.67	0.15	0.30	1.64	100.00	100.00	74.58
NB	0.62	0.10	0.26	86.89	16.00	26.50	34.32
MLP	0.60	0.07	0.26	8.20	92.57	27.78	70.76
GB	0.64	0.08	0.28	6.56	96.00	36.36	72.88

Table B.20: Statistics herd normalisation feature set

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.62	0.18	0.27	18.03	94.19	52.38	74.25
NB	0.66	0.15	0.27	22.95	91.86	50.00	73.82
MLP	0.65	0.13	0.27	9.84	95.93	46.15	73.39
GB	0.64	0.03	0.27	3.28	95.35	20.00	71.24

Table B.21: Trend feature set

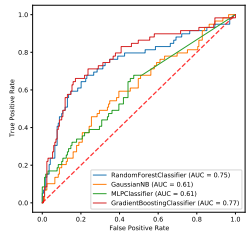
	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.60	0.11	0.26	8.20	99.43	83.33	75.85
NB	0.61	0.05	0.27	14.75	90.86	36.00	71.19
MLP	0.63	0.08	0.26	1.64	97.71	20.00	72.88
GB	0.62	0.10	0.27	9.84	95.43	42.86	73.31

Table B.22: Trend herd normalisation feature set

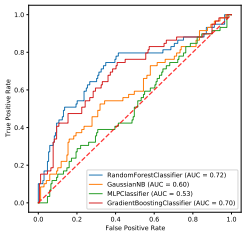
	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.76	0.28	0.30	37.70	87.43	51.11	74.58
NB	0.72	0.25	0.30	26.23	94.86	64.00	77.12
MLP	0.67	0.23	0.26	0.00	100.00	0.00	74.15
GB	0.72	0.25	0.31	8.20	97.14	50.00	74.15

Table B.23: Static feature set

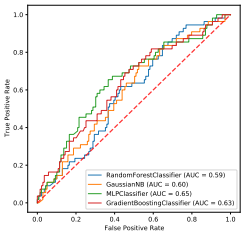
B.1.3 ROC plots



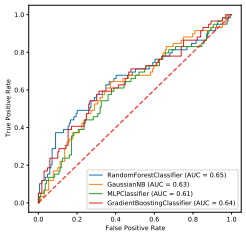
(a) All features



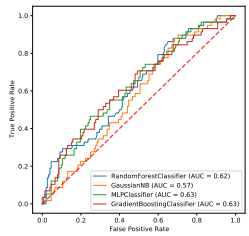
(b) All behaviour



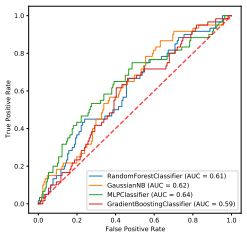
(c) FFT



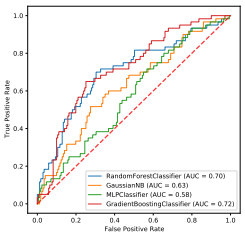
(d) Daily herd-norm.



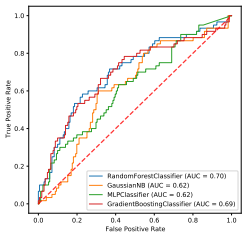
(e) Daily individual norm.



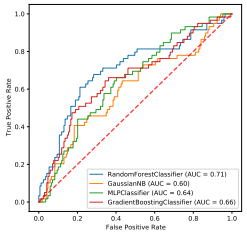
(f) Daily within-window norm.



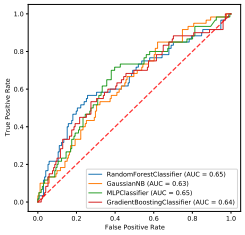
(g) Daily values



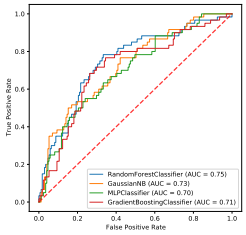
(h) Statistics



(i) Statistics herd



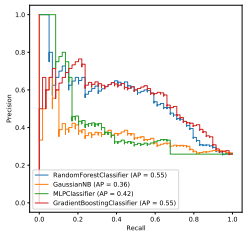
(j) Trend



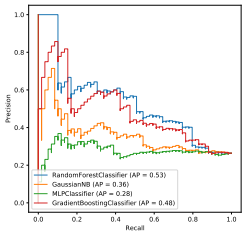
(k) Static

Figure B.1: ROC plots

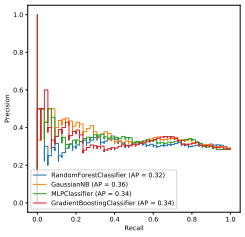
B.1.4 Precision-recall plots



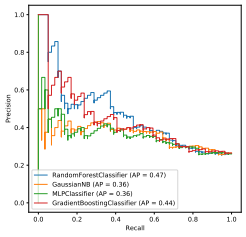
(a) All features



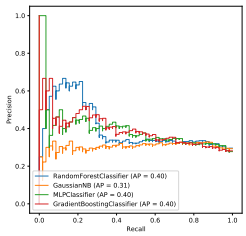
(b) All behaviour



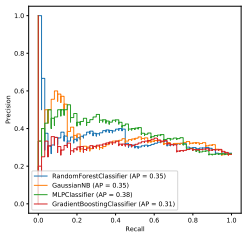
(c) FFT



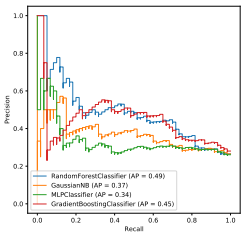
(d) Daily herd-norm.



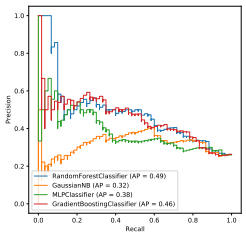
(e) Daily individual norm.



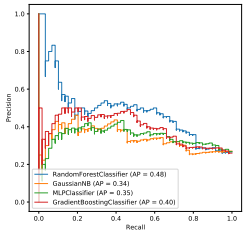
(f) Daily within-window norm.



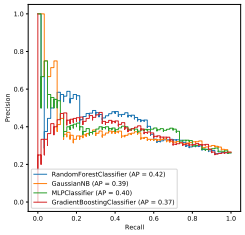
(g) Daily values



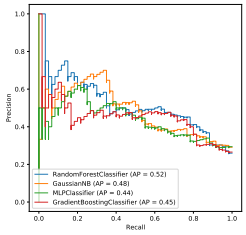
(h) Statistics



(i) Statistics herd



(j) Trend



(k) Static

Figure B.2: Precision-recall plots

B.2 Regression

B.2.1 Measurement-date based window

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.43	0.18	0.30	0.85	0.13	3.39	98.22	40.00	73.68
LR	0.43	0.19	0.31	0.82	0.11	30.51	92.90	60.00	76.75
MLP	0.46	0.21	0.34	0.94	0.00	0.00	100.00	0.00	74.12
GB	0.42	0.17	0.30	0.84	0.17	25.42	95.27	65.22	77.19

Table B.24: All Features set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.43	0.19	0.31	0.84	0.11	10.17	98.22	66.67	75.44
LR	0.46	0.21	0.32	0.85	0.02	18.64	92.31	45.83	73.25
MLP	0.46	0.21	0.33	0.95	-0.00	0.00	100.00	0	74.12
GB	0.44	0.20	0.31	0.85	0.06	15.25	94.67	50.00	74.12

Table B.25: All behaviour feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.44	0.20	0.32	0.93	0.07	6.78	98.82	66.67	75.00
LR	0.45	0.20	0.33	0.88	0.06	1.69	98.22	25.00	73.25
MLP	0.44	0.20	0.32	0.90	0.06	0.00	100.00	0	74.12
GB	0.44	0.19	0.32	0.91	0.08	15.25	96.45	60.00	75.44

Table B.26: Daily herd normalisation feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.45	0.20	0.33	0.95	0.08	3.45	96.71	28.57	70.95
LR	0.46	0.21	0.33	0.97	0.04	0.00	99.34	0.00	71.90
MLP	0.46	0.21	0.33	0.96	0.05	0.00	100.00	0	72.38
GB	0.46	0.22	0.33	0.99	0.03	12.07	94.08	43.75	71.43

Table B.27: Daily prepartum normalisation feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.45	0.21	0.33	0.94	0.04	6.67	97.66	50.00	74.03
LR	0.46	0.21	0.32	1.02	0.00	0.00	100.00	0	74.03
MLP	0.45	0.21	0.32	0.99	0.04	0.00	100.00	0	74.03
GB	0.46	0.21	0.32	0.97	-0.00	13.33	90.64	33.33	70.56

Table B.28: Daily within-window normalisation feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.44	0.19	0.31	0.91	0.09	8.33	97.08	50.00	74.03
LR	0.44	0.20	0.32	0.93	0.08	1.67	98.25	25.00	73.16
MLP	0.46	0.21	0.34	0.97	0.00	0.00	100.00	0	74.03
GB	0.45	0.20	0.31	0.96	0.05	11.67	94.15	41.18	72.73

Table B.29: Daily values feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.44	0.19	0.32	0.88	0.10	13.33	97.08	61.54	75.32
LR	0.46	0.21	0.33	0.92	0.01	5.00	98.83	60.00	74.46
MLP	0.46	0.21	0.34	0.98	0.00	0.00	100.00	0	74.03
GB	0.46	0.21	0.33	0.94	0.00	16.67	95.32	55.56	74.89

Table B.30: Statistics feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.44	0.20	0.32	0.96	0.07	8.47	96.45	45.45	73.68
LR	0.46	0.21	0.33	0.95	-0.00	5.08	95.86	30.00	72.37
MLP	0.46	0.21	0.33	0.89	0.01	1.69	98.82	33.33	73.68
GB	0.44	0.20	0.32	0.95	0.06	16.95	92.90	45.45	73.25

Table B.31: Statistics herd normalisation feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.44	0.19	0.31	0.90	0.09	13.33	98.83	80.00	76.62
LR	0.45	0.20	0.32	0.89	0.07	1.67	98.83	33.33	73.59
MLP	0.45	0.20	0.32	0.97	0.05	1.67	100.00	100.00	74.46
GB	0.45	0.20	0.32	0.92	0.05	11.67	98.25	70.00	75.76

Table B.32: Trend feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.45	0.20	0.33	0.86	0.04	10.17	97.63	60.00	75.00
LR	0.45	0.20	0.32	0.91	0.05	0.00	98.22	0.00	72.81
MLP	0.45	0.20	0.33	0.93	0.04	0.00	100.00	0.00	74.12
GB	0.46	0.21	0.34	0.92	-0.01	13.56	94.08	44.44	73.25

Table B.33: Trend herd normalisation feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.42	0.18	0.31	0.84	0.16	15.00	97.66	69.23	76.19
LR	0.44	0.19	0.32	0.92	0.10	3.33	98.25	40.00	73.59
MLP	0.44	0.19	0.33	0.88	0.10	0.00	100.00	0.00	74.03
GB	0.43	0.19	0.32	0.88	0.13	28.33	96.49	73.91	78.79

Table B.34: Static feature set

B.2.2 Calving-date based window

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.44	0.19	0.32	0.92	0.10	9.84	96.57	50.00	74.15
LR	0.45	0.21	0.32	0.91	0.04	18.03	93.71	50.00	74.15
MLP	0.46	0.22	0.34	1.05	-0.00	0.00	100.00	0.00	74.15
GB	0.44	0.19	0.32	0.88	0.09	22.95	93.71	56.00	75.42

Table B.35: All Features set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.44	0.20	0.32	0.92	0.08	8.20	96.00	41.67	73.31
LR	0.46	0.21	0.32	0.88	0.00	16.39	94.29	50.00	74.15
MLP	0.46	0.22	0.34	1.05	-0.00	0.00	100.00	0.00	74.15
GB	0.46	0.21	0.33	0.92	0.01	16.39	90.86	38.46	71.61

Table B.36: All behaviour feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.47	0.22	0.34	0.94	0.01	5.45	98.56	60.00	72.16
LR	0.47	0.22	0.33	0.93	0.01	0.00	97.84	0.00	70.10
MLP	0.47	0.22	0.34	0.92	0.01	1.82	98.56	33.33	71.13
GB	0.47	0.22	0.34	0.94	-0.01	12.73	94.24	46.67	71.13

Table B.37: FFT feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.46	0.21	0.33	0.93	0.01	6.56	97.71	50.00	74.15
LR	0.45	0.21	0.32	0.92	0.04	3.28	97.71	33.33	73.31
MLP	0.45	0.21	0.32	0.92	0.04	0.00	100.00	0.00	74.15
GB	0.49	0.24	0.35	0.94	-0.12	8.20	92.57	27.78	70.76

Table B.38: Daily herd normalisation feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.47	0.22	0.34	1.05	0.01	3.33	97.37	33.33	70.75
LR	0.48	0.23	0.35	1.01	-0.02	0.00	100.00	0.00	71.70
MLP	0.48	0.23	0.34	1.01	0.00	0.00	100.00	0.00	71.70
GB	0.49	0.24	0.36	1.08	-0.07	3.33	97.37	33.33	70.75

Table B.39: Daily prepartum normalisation feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.47	0.22	0.34	0.98	-0.02	0.00	98.29	0.00	72.88
LR	0.46	0.21	0.33	1.00	0.02	0.00	100.00	0.00	74.15
MLP	0.46	0.21	0.34	1.02	0.00	0.00	100.00	0.00	74.15
GB	0.49	0.24	0.36	1.02	-0.11	3.28	95.43	20.00	71.61

Table B.40: Daily within-window normalisation feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.47	0.22	0.34	0.91	-0.01	8.20	93.71	31.25	71.61
LR	0.46	0.21	0.33	0.98	0.03	6.56	96.00	36.36	72.88
MLP	0.46	0.22	0.34	1.04	-0.00	0.00	100.00	0.00	74.15
GB	0.51	0.26	0.37	1.05	-0.21	8.20	89.14	20.83	68.22

Table B.41: Daily values feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.45	0.20	0.32	0.87	0.06	3.28	97.71	33.33	73.31
LR	0.46	0.21	0.32	0.94	0.03	8.20	96.57	45.45	73.73
MLP	0.46	0.22	0.33	1.06	-0.00	0.00	100.00	0.00	74.15
GB	0.46	0.21	0.34	0.93	0.03	16.39	94.29	50.00	74.15

Table B.42: Statistics feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.44	0.19	0.32	0.83	0.11	6.56	98.86	66.67	75.00
LR	0.46	0.21	0.32	0.91	0.03	6.56	97.14	44.44	73.73
MLP	0.45	0.20	0.32	0.91	0.06	0.00	100.00	0.00	74.15
GB	0.44	0.19	0.31	0.96	0.12	14.75	95.43	52.94	74.58

Table B.43: Statistics herd normalisation feature set

	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.44	0.19	0.32	0.94	0.12	13.11	96.51	57.14	74.68
LR	0.43	0.19	0.31	0.90	0.13	8.20	97.67	55.56	74.25
MLP	0.46	0.21	0.32	1.00	0.02	0.00	100.00	0.00	73.82
GB	0.46	0.21	0.33	0.97	0.04	11.48	93.60	38.89	72.10

Table B.44: Trend feature set

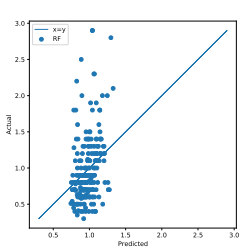
	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.45	0.21	0.33	0.89	0.04	4.92	97.14	37.50	73.31
LR	0.45	0.21	0.32	0.93	0.04	1.64	98.29	25.00	73.31
MLP	0.45	0.21	0.32	0.92	0.04	0.00	99.43	0.00	73.73
GB	0.48	0.23	0.34	0.95	-0.07	6.56	92.57	23.53	70.34

Table B.45: Trend herd-normalisation feature set

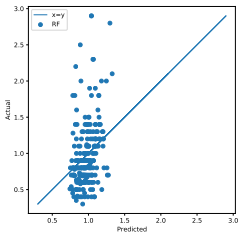
	RMSE	MSE	MAE	Q95E	R2	Se	Sp	Pr	Acc
RF	0.43	0.18	0.31	0.85	0.16	11.48	96.57	53.85	74.58
LR	0.44	0.19	0.31	0.93	0.09	3.28	98.29	40.00	73.73
MLP	0.46	0.21	0.34	1.01	0.02	0.00	100.00	0.00	74.15
GB	0.43	0.18	0.31	0.88	0.15	16.39	96.00	58.82	75.42

Table B.46: Static feature set

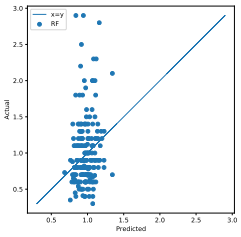
B.2.3 Predicted-Actual plots



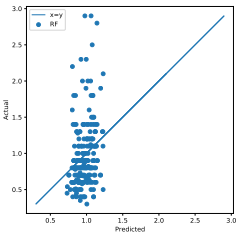
(a) All Features



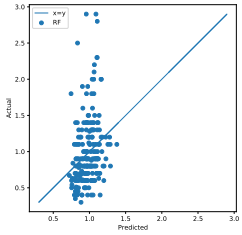
(b) All Behaviour features



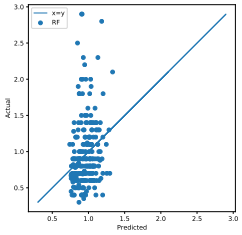
(c) FFT with CB window



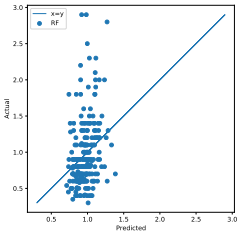
(d) Daily herd normalisation



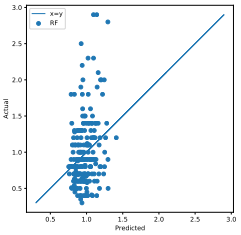
(e) Daily prepartum normalisation



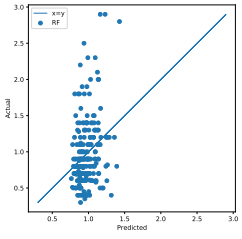
(f) Daily within-window normalisation



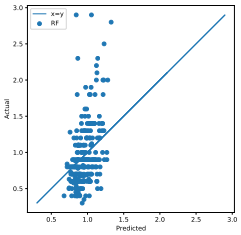
(g) Daily values



(h) Statistics



(i) Statistics herd-normalised



(j) Static feature set

Figure B.3: Predicted-Actual plots with a measurement-date based window and Random Forest regressor

B.3 Dynamic Time Warping Result

	DTW
AUC	0.50
Se@95Sp	0.00
Pr@95Re	0.26
Se	25.00
Sp	75.44
Pr	26.32
Acc	62.34

Table B.47: DTW classifier with daily values results on metrics

C Farm specific results

Measurement based windows have been trained and tested on the four biggest farms.

C.1 Classification

C.1.1 Measurement based

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.70	0.18	0.42	18.18	95.00	66.67	67.74
NB	0.56	0.09	0.35	54.55	60.00	42.86	58.06
MLP	0.53	0.00	0.35	27.27	80.00	42.86	61.29
GB	0.69	0.09	0.44	18.18	75.00	28.57	54.84

Table C.1: Daily values farm 2

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.67	0.20	0.19	0.00	100.00	0.00	81.48
NB	0.65	0.20	0.23	30.00	77.27	23.08	68.52
MLP	0.65	0.10	0.19	0.00	100.00	0.00	81.48
GB	0.58	0.10	0.19	20.00	93.18	40.00	79.63

Table C.2: Daily values farm 3

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.60	0.12	0.28	12.50	95.65	50.00	74.19
NB	0.58	0.12	0.27	50.00	47.83	25.00	48.39
MLP	0.58	0.12	0.26	25.00	69.57	22.22	58.06
GB	0.59	0.12	0.28	37.50	91.30	60.00	77.42

Table C.3: Daily values farm 4

	AUC	Se@95Sp	Pr@95Re	Se	Sp	Pr	Acc
RF	0.81	0.40	0.25	20.00	100.00	100.00	85.71
NB	0.57	0.20	0.20	20.00	86.96	25.00	75.00
MLP	0.53	0.20	0.20	60.00	39.13	17.65	42.86
GB	0.77	0.00	0.36	0.00	95.65	0.00	78.57

Table C.4: Daily values farm 8

D Hyperparameter search

D.1 Random Forest

The hyperparameter search for the Random Forest classifier (and regressor) was based on a 5-fold cross-validation in which the model with the best AUC score was selected. The parameters were searched using a randomised search over the parameter space. The randomised search had 30 iterations. These parameters with possible values are listed below. The parameters correspond to the `scikit-learn`[1] arguments for the Random Forest model.

```
n_estimators = [int(x) for x in np.linspace(start=100, stop=2000, num=20)]
max_features = ['auto', 'sqrt', None]
max_depth = [int(x) for x in np.linspace(10, 110, num=11)]
max_depth.append(None)
min_samples_split = [2, 5, 10]
min_samples_leaf = [1, 2, 4]
bootstrap = [True, False]
```

D.2 Gradient Boosting

The hyperparameter search for the Gradient Boosting classifier (and regressor) was based on a 5-fold cross-validation in which the model with the best AUC score was selected. The parameters were searched using a randomised search over the parameter space. The randomised search had 30 iterations. These parameters with possible values are listed below. The parameters correspond to the `scikit-learn` arguments for the Gradient Boosting model.

```
n_estimators = [int(x) for x in np.linspace(start=20, stop=200, num=10)]
max_features = ['auto', 'sqrt', None]
max_depth = [2**depth for depth in range(7)]
min_samples_split = [2, 5, 10]
min_samples_leaf = [1, 2, 4]
```

[1] `https://scikit-learn.org/stable/`

D.3 Multilayer Perceptron

The hyperparameter search for the Multilayer Perceptron classifier (and regressor) was based on a 5-fold cross-validation in which the model with the best AUC score was selected. The parameters were searched using a grid search over the parameter space. These parameters with possible values are listed below. The parameters correspond to the `scikit-learn` arguments for the Multilayer Perceptron model.

```
hidden_layer_sizes = [(x,) for x in range(2, 20, 2)] +
                     [(x, x) for x in range(2, 20, 2)]
activation=['relu', 'logistic'],
learning_rate_init=[0.001, 0.005, 0.01]
```