Computers and Electronics in Agriculture 220 (2024) 108877

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

A predictive model for hypocalcaemia in dairy cows utilizing behavioural sensor data combined with deep learning

Meike van Leerdam [a,*], Peter R. Hut [a], Arno Liseune [b], Elena Slavco [a], Jan Hulsen [c], Miel Hostens [a,d]

[a] Faculty of Veterinary Medicine, Utrecht University, Yalelaan 7, 3584CL, Utrecht, The Netherlands
[b] Faculty of Economics and Business Administration, Ghent University, Tweekerkenstraat 2, B-9000 Ghent, Belgium
[c] Vetvice, Moerstraatsebaan 115, 4614 PC, Bergen op Zoom, The Netherlands
[d] College of Agriculture and Life Sciences, Cornell University, 273 Morrison Hall, NY 14853 Ithaca, United States of America

ARTICLE INFO

Dataset link: https://github.com/Bovi-analytics/van-leerdam-et-al, sensor_data.zip (Original data)

Keywords:
Dairy cattle
Hypocalcaemia
Sensors
Deep learning
Prediction
Transition period

ABSTRACT

(Sub)clinical hypocalcaemia occurs frequently in the dairy industry, and is one of the earliest symptoms of an impaired transition period. Calcium deficiency is accompanied by changes in cows' daily behavioural variables, which can be measured by sensors. The goal of this study was to construct a predictive model to identify cows at risk of hypocalcaemia in dairy cows using behavioural sensor data. For this study 133 primiparous and 476 multiparous cows from 8 commercial Dutch dairy farms were equipped with neck and leg sensors measuring daily behavioural parameters, including eating, ruminating, standing, lying, and walking behaviour of the 21 days before calving. From each cow, a blood sample was taken within 48 h after calving to measure their blood calcium concentration. Cows with a blood calcium concentration ≤2.0 mmol/L were defined as hypocalcemic. In order to create a more context based cut-off, a second way of dividing the calcium concentrations into two categories was proposed, using a linear mixed-effects model with a k-Means clustering. Three possible binary predictive models were tested; a logistic regression model, a XgBoost model and a LSTM deep learning model. The models were expanded by adding the following static features as input variables; parity (1, 2 or 3+), calving season (summer, autumn, winter, spring), day of calcium sampling relative to calving (0, 1 or 2), body condition score and locomotion score. Of the three models, the deep learning model performed best with an area under the receiver operating characteristic curve (AUC) of 0.71 and an average precision of 0.47. This final model was constructed with the addition of the static features, since they improved the model's tuning AUC with 0.11. The calcium label based on the cut-off categorization method proved to be easier to predict for the models compared to the categorization method with the k-means clustering. This study provides a novel approach for the prediction of hypocalcaemia, and an ameliorated version of the deep learning model proposed in this study could serve as a tool to help monitor herd calcium status and to identify animals at risk for associated transition diseases.

1. Introduction

The most challenging time in the lifespan of a cow is around parturition, more commonly known as the transition period (Grummer, 1995). The cow has to adapt homeorhetic from a pregnant state to a non-pregnant, and more importantly, lactating state (Bauman and Currie, 1980). In this period, most infectious diseases and metabolic disorders occur or originate, ranging from ketosis and retained fetal membranes to displaced abomasum and mastitis (Drackley, 1999). One of the arising problems is hypocalcaemia, more commonly known as milk fever. Once a cow starts lactating, she looses more calcium in her milk, urine, and faeces than she can replenish through intestinal reuptake. The mechanisms to rebalance calcium take a while to initiate, resulting in a calcium dip right after calving (Horst et al., 1994). Recently, Horst et al. (2021) suggested that hypocalcaemia could also be explained as a result of an inflammatory reaction seen around calving. There are two forms of hypocalcaemia; clinical hypocalcaemia (CH), with visible clinical signs like increased heart rate, cold ears and recumbency, and subclinical hypocalcaemia (SCH) which has no recognizable symptoms but is associated with impaired postpartum health and performance (Serrenho et al., 2021).

The reported prevalence of hypocalcaemia differs between studies, but lies between 14%–40% of the cows after parturition overall and

* Corresponding author.
E-mail address: m.b.vanleerdam@students.uu.nl (M. van Leerdam).

https://doi.org/10.1016/j.compag.2024.108877
Received 2 September 2022; Received in revised form 15 March 2024; Accepted 21 March 2024
Available online 3 April 2024

for higher parity cows up to 69% (Serrenho et al., 2021). Since the disease is very common, many preventive measures are being applied in an attempt to reduce milk fever occurrences. For instance; feeding a pre-calving diet low in calcium, feeding a diet with a negative dietary cation-anion difference and oral calcium drenching around calving, are regularly used (Thilsing-Hansen et al., 2002; DeGaris and Lean, 2008). Not all risk factors of hypocalcaemia are related to nutrition, (DeGaris and Lean, 2008) and none of the measures have succeeded to fully prevent hypocalcaemia (Venjakob et al., 2017; Ribeiro et al., 2013).

Devices that measure physiological or behavioural parameters, are increasingly used in the dairy industry. In a survey held among Dutch dairy farmers in 2015, 39% of the farmers reported using at least one sensor system (Steeneveld and Hogeveen, 2015), a number that probably has been growing ever since. Sensors can be used for oestrus detection (Firk et al., 2002), lameness detection (Chapinal et al., 2010), mastitis detection (Cavero et al., 2008; Jensen et al., 2016), and numerous other applications. Nowadays, 129 different sensor systems are commercially available (Stygar et al., 2021).

Overton et al. (2017) has shown that there are differences in prepartum behaviour between healthy cows and those affected by metabolic disease postpartum. The authors thereby suggested that this difference could be used for disease prediction. Other researchers, such as Soriani et al. (2012), Liboreiro et al. (2015), Hendriks et al. (2020) and Gusterer et al. (2020), have likewise explored the connection between cow behaviour and disease using sensor-based techniques. These studies also reported behavioural differences between disordered and healthy cows, suggesting a possible predictive value of behaviour for disease. Stangaferro et al. (2016) has shown that these sensor-based systems can be used for the detection of diseases, and de Mol et al. (2015) has demonstrated a system to identify deviations of behaviour which are associated with metabolic disease. However, at the present time, for as far as we know, models for specifically hypocalcaemia prediction with behavioural sensor data do not yet exist (Garcia et al., 2020), despite the suggestion of Overton et al. (2017) and the potential of precision livestock farming to perform such tasks (Garcia et al., 2020; Wathes et al., 2008).

In this study, we tried to build the first prediction model for a specific disease using activity data. Given the high prevalence, impact, and focus on the prevention of hypocalcaemia in practice (LeBlanc et al., 2006), a model was built to predict hypocalcaemia.

The descriptive models using animal behaviour data described above all used traditional machine learning models, but in this paper we chose to approach this using a deep learning model. Deep learning is better scalable and often outperforms traditional data analysis methodologies when handling large and complex data (Janiesch et al., 2021). We hypothesized that given the complexity of behavioural data, it therefore will result in a more accurate model. A LSTM deep learning model was chosen, since this network can analyse sequential data and has the ability to recognize temporal patterns, as can be seen in a series of behavioural data, using a long term memory (Hochreiter and Schmidhuber, 1997). The downside of a neural network is, next to high computational cost, the black-box principle, making it hard to distinguish which behaviours are important for a models' prediction. This problem is partly solved in this study with a cross-validated permutation feature importance, although this method is labour-intensive and cumbersome.

The goal of the study was to create a predictive model able to differentiate at parturition which cow is at risk for hypocalcaemia after parturition. This could then serve as a tool for the prevention and control of metabolic disease.

2. Materials and methods

2.1. Data collection

This study was a part of the Sense Of Sensors project. In this project, cows of eight Dutch dairy farms were equipped with leg and neck sensors measuring animal behaviour. The Smarttag sensors were provided by Nedap (Nedap, Groenlo, the Netherlands) and measured five different features throughout the day. The leg sensors measured the total minutes per day spent standing, lying and walking, while the neck sensors measured the minutes spent eating and ruminating. For this study, the sensor values recorded during the 21 days before calving were used, a period generally accepted as the close up period and a typical time for a cow to prepare for calving. So for each cow, there were 5 different values each day for 21 days, resulting in 105 values per cow in total. In a paper by Hut et al. (2021), the farms and sensors used are described elaborately. The data was collected between November 2016 and May 2018 and included both data of dairy cows in the transition period and pre-fresh heifers.

Blood samples were taken from the cows by a veterinarian on the day of calving (0), the day after calving (1) or two days after calving (2), in order to measure the blood calcium concentration. The samples were taken from the coccygeal vein using a vacutainer and collected into a heparinized blood collection tube. At the same day, the collected samples were centrifuged for ten minutes at 4500 rpm (Centrifuge 5804 R; Eppendorf Germany) and afterwards manually pipetted into Eppendorf cups. The samples were stored at −20 degrees Celsius, awaiting quantitative analysis of total calcium serum concentration using the Calcium Arsenazo method (Leary et al., 1992). This method was executed by the Olympus AU680 with a limit of quantitation of 1 mmol/L and an end point determination of 660 nm. One cow was removed from the dataset because of an extraordinary high blood calcium value of above 3, 4 mmol/L due to the administration of a calcium infusion just before sampling. Since the research was conducted over a longer period of time, 21 cows participated multiple times, but with a different parity. However, each unique animal calving date combination was seen as a different test subject. This selection process resulted in 609 unique dairy cow calving date combinations deemed appropriate for this research.

For 416 cows, the body condition score (BCS) was determined by a trained veterinarian at the end of the dry period. The scores were described on a scale between 1 and 5, with 0.25 increments, as defined by Ferguson et al. (1994). At the same observation for 414 cows, the locomotion score was determined on a scale between 1 and 5 based on posture and gate, but with the use of integers only, according to Sprecher et al. (1997).

Calving seasons were extracted from the recorded calving dates and were defined as 3-month periods according to Sanders et al. (2009). Summer, for instance, was defined as the months of July, August, and September.

2.2. Label preprocessing

Cows were divided into two categories; hypocalcaemic cows with a blood calcium concentration equal to or lesser than 2.0 mmol/L and normocalcaemic cows with a blood calcium concentration above 2.0 mmol/L. This threshold was chosen according to Reinhardt et al. (2011). There is, however, increasing discussion whether the cut-off value of 2.0 mmol/L is a valid number to define SCH, or is in fact chosen arbitrarily in the past and therefore not evidence based (Serrenho et al., 2021). Therefore, a second way of splitting the two categories was proposed using a linear mixed-effects model, in combination with k-means clustering. This method corrects for parity and day of measurement and results in a more fluent context based cut-off, as recommended by Serrenho et al. (2021). The calcium concentration was used as the response variable for the linear mixed effect model and the day of blood sampling relative to calving (0, 1 or 2), parity (1, 2 or 3+) and farm were the predictor variables. As measurements of cows from the same farm are correlated, grouping of the data must be taken into account. Therefore, a linear mixed effect model was chosen, which is a hierarchical multilevel model, and allows for different regression coefficients for each predictor per farm and thereby includes both

the variation within a farm and between farms. Then the k-means clustering method was used based upon the residuals of the calcium predictions and the absolute values of the calcium concentrations. This resulted in two clusters; one group with relatively low calcium concentrations and a group with relatively high calcium concentrations.

From the 609 cows, 365 cows were randomly attributed to the train set, which equals approximately 60%. Of the remaining 244 cows, 50% were attributed to the validation set and the remaining 122 cows were attributed to the test set.

As neural networks cannot function with incomplete data (Ennett et al., 2001), missing BCS and locomotion values were imputed using the sklearn SimpleImputer (Python), which replaced null values by the score that was most frequent in the train set.

2.3. Feature preprocessing

Min–Max scaling was used to normalize the sensor data. The normalization was fitted on the values of the train set per behaviour. The resulting normalization was also used for the test and validation set without resetting the minimal and maximal value.

For numerous reasons, including sensor malfunctioning and administrative errors, data points were missing from the dataset. In fact, approximately 14.6% of the sensor values were misrecorded. As each cow had 105 different data points, omitting all cows with one or more missing values would have resulted in a too great reduction in cows. In order to therefore replace the missing values, a SimpleImputer (Python) was used based upon the train set, which transformed the missing values into the mean per feature of the according day before calving. The distribution of missing values is visualized with a heatmap in the appendix, Fig. 6.

The sensor values were placed into a three-dimensional matrix of 21 days by 5 behaviours by the number of cows. The calcium categories and static features were also extracted per cow and put into separate lists with an index matching the sensor matrix. The feature preprocessing is visualized in the appendix Fig. 5.

There were approximately 2.7 times more cows in the normocalcaemia category, causing class-imbalance. However, most prediction models give the best results with a balanced dataset (Johnson and Khoshgoftaar, 2019). Therefore, upsampling was performed whereby the cows with hypocalcaemia were extracted from the train set and randomly sampled with replacement until there were as many cows with hypocalcaemia as cows with normocalcaemia. This upsampling was only performed on the train set.

2.4. Model building

In order to predict the probability to fall within a specific category of calcium concentrations, three models were built: a logistic regression model, a XgBoost model and a LSTM deep learning model.

The logistic regression model was chosen as the representative for a traditional approach to prediction models, in order to be able to make a comparison with the novel methods. For this model, the 3D sensor value array was flattened to an array with the shape (number of cows, 21 days · 5 features). This array served as the input feature of the model, while the calcium group served as the output label. For the model, the liblinear solver algorithm was used. Due to the limited amount of hypocalcaemic cows, it was hypothetized that the model could focus too much on the healthy cows, therefore the model was trained both with and without random upsampling of the train set. A second way to deal with class imbalance, namely adding class weights, was performed, introducing cost-sensitive learning (Johnson and Khoshgoftaar, 2019). The accompanying cost matrix was defined by a grid search of a range of possible weights.

The second model used was a XgBoost model; a relatively recent developed, but already widely used, machine learning model using tree boosting (Chen and Guestrin, 2016). Research conducted by Gertz et al. (2020) has already demonstrated that XgBoost is a proficient and user-friendly approach for classifying motion-sensor data in cattle. The input features and output labels used for this model were the same as for the logistic regression model. The validation set was used for early stopping and hyperparameter tuning, which was performed automatically using random search. The hyperparameters tuned are noted in Table 1. One parameter to point out is a positive class weight as a possible solution to class imbalance, set to be the total number of cows with normocalcaemia divided by the total number of cows with hypocalcaemia. This method was used since it is the default solution for class imbalance in a XgBoost model.

The third model was a LSTM deep learning model. The choice of an LSTM (Long Short-Term Memory) model is fitting due to its capability to analyse sequential data and its proficiency in recognizing temporal patterns within the data (Hochreiter and Schmidhuber, 1997). A LSTM model consist of cells, that next to providing an output, also provide a cell-state which functions as the memory for the next cell. The matrices of the 5 behavioural features were used as model input, where for each time step a LSTM cell was formed. The cells were aligned in a chronological order, connected by the cell state. Each cell has three gates; a forget gate, to forget unnecessary information passed on by the previous cell, an input gate layer, to process the new input and add it to the cell state, and an output gate, which provides a filtered version of the acquired cell-state as output. In the end, this resulted in a one dimensional vector containing a summary of the information the LSTM layers filtered from the sensors. This vector was passed on to a classic multilayer perceptron (MLP) layer, which converted the vector to a value between 0 and 1 using a sigmoid function. This value was used as the probability that a cow will be in the low calcium category post-partum. The architecture of the LSTM deep learning model is visualized in Fig. 1.

2.5. LSTM model training

The training process of the LSTM model started with random initiation of all the weights of the equations in the deep learning model, thereby forming the untrained version of the model. The calcium category acquired with these random weights was compared with the desired result. This difference is expressed as the loss function. The LSTM model was then trained using the backpropagation algorithm (Rumelhart et al., 1986) in combination with the Adam gradient-based optimization algorithm (Kingma and Ba, 2015), adjusting the weights and biases in order to minimize the loss function, using the training data. The model was programmed to stop training when the loss of the validation set did not decrease for three consecutive rounds, a process which is called early stopping. This was used to prevent overfitting on the training data.

2.6. Static cow features

In order to improve the sensor based model performance, static cow features were added as input features. The following features were included: calving season, parity (1, 2 or 3+) and the day of blood sampling compared to calving. Since neural networks can only process numerical data as input, calving season and parity were converted to a binary variable using the sklearn OneHotEncoder. This process resulted in 8 static features to include in the models. The models were also tested with the addition of BCS and locomotion score measured at the end of the dry period to the static cow features, resulting in 10 static features included in the model. The preprocessing of the static features is visualized in appendix Fig. 5. For the logistic regression model and the XgBoost model, the static features were combined with the sensor values in one array. However, in order to combine static and sequential input for the deep learning model, a functional model was build where the output vector of the LSTM layer was combined with the static features in a concatenation layer, which was subsequently processed using a traditional MLP layer with a ReLu activation function and converted into a binary output using an MLP layer with a sigmoid activation function. This process is visualized in Fig. 1.

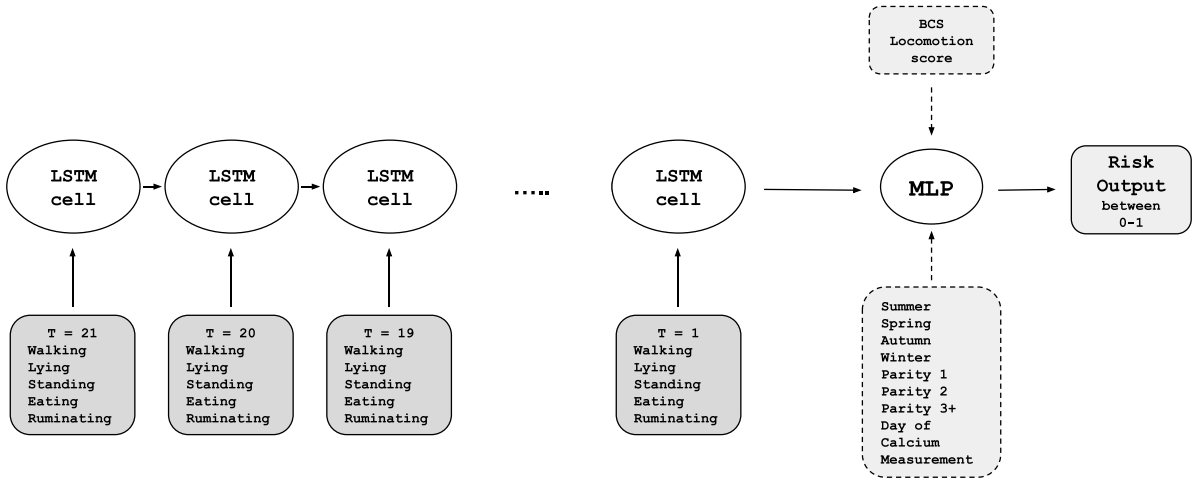


Fig. 1. Overview of the LSTM deep learning model architecture. The static features, delineated with dashed boundaries, were optional.

2.7. Model tuning

Since neural networks are prone to overfitting (Krogh, 2008), measures were taken in order to prevent overfitting. The number of nodes were kept low to limit the size of the models, and a dropout and batch normalization layer were added. Among others, these hyperparameters were tuned and can be found in Table 1. The tuning process of the models was based on random search, using the validation set to compare the different hyperparameter combinations. There were six different sets of input features; with only sensor values both upsampled or not, sensors with parity, day of measurement relative to calving and calving season upsampled or not and sensors with all the static features, including BCS and locomotion score, both upsampled or not. These combinations of the different sets of input features with the calcium category to predict, based on either the clusters or on the cut-off value, were seen as separate models, yielding 12 different models to tune per model category. These combinations are listed in Table 2. Each model was individually tuned and for each model 200 random combinations of hyperparameters were tried, selecting the hyperparameter settings with the best results. Only the logistic regression model was tuned using grid search, thereby testing all sixty possible combinations. This method was more appropriate since there were less potential combinations of hyperparameters.

2.8. Model evaluation and comparison

The best performing hyperparameter configuration was evaluated using bootstraps in order to quantify model consistency. A bootstrap is a random sample with replacement of cows from the validation set with the same size as the validation set. Fifty bootstraps were created, each consisting of a unique combination of re-sampled cows. First, the best hyperparameter configuration based on the area under the receiver operating characteristic curve (AUC) while predicting the calcium category of the validation set, was selected for each model. Then this model was tested using the bootstraps, rendering fifty AUC values each. Finally, the mean AUC and standard deviation (SD) of the AUC were calculated and used to compare the different models. The SD is a metric for consistent model performance, thus models with a lower SD are more precise. Models were compared initially using the mean AUC score and when this value was equivalent between two models, the model with the least input features was preferred. The three models with the best mean AUC score were evaluated on the test set. In Fig. 2 a schematic overview of the methodology for the model evaluation is given.

Performance of the final models was evaluated on the test set using the AUC and the average precision (AP); the area under the precision–recall curve, using the predicted and true calcium categories of the test set. These metrics were chosen since they are threshold independent and therefore more suitable to compare between models. The accuracy, sensitivity (true positive rate) and specificity (true negative rate), of the best performing models were calculated with a threshold value of 0.5.

2.9. Feature importance

In order to calculate feature importance of the best performing model, a method called cross-validated permutation feature importance was applied (Kaneko, 2022). In this method, one feature of the validation set is randomly shuffled, thereby breaking the relationship with the associated blood calcium concentration, while the other features remain the same. This is repeated fifty times per feature. The importance of the feature is then measured by the mean decrease in accuracy of the models' prediction after permuting the feature. This method was only applied for the deep learning model since for the XgBoost and the logistic regression model the input array was flattened resulting in 105 input features instead of five, making it difficult to individually assess the many different features. In order to evaluate whether there is a difference in feature importance when static features are included, both the best model with only the behavioural features and the best performing model overall were tested.

2.10. Programming framework and data

Data processing and analysis was performed using the programming language Python (Python Software Foundation, version 3.8.10, http://www.python.org) with the add-on packages Pandas (The Pandas Development Team, 2020; McKinney, 2010), NumPy (Harris et al., 2020), scikit-learn (Pedregosa et al., 2011), XgBoost (Chen and Guestrin, 2016), Ray Tune (Liaw et al., 2018), TensorFlow (Abadi et al., 2016), Keras (Chollet et al., 2015) and Mat plotlib (Hunter, 2007). For the linear mixed-effects model, the programming language R was used (R. Core Team, 2013) version 4.1.1, with the following packages: 'lme4' (Bates et al., 2015), 'dplyr' (Wickham et al., 2015), and 'ggplot2' (Wickham, 2016). For both R and Python the Apache Spark (Zaharia et al., 2016) cluster-computing framework was used. To contribute to open science as defined by the UNESCO recommendation on Open Science, UNESCO (2021) the data used in this research is made publicly available. The data is published in the form of an ontology. This was developed in order to make the structure and concepts of the data more comprehensible and to make it easier to extend the dataset with external data, thereby facilitating future research. The populated

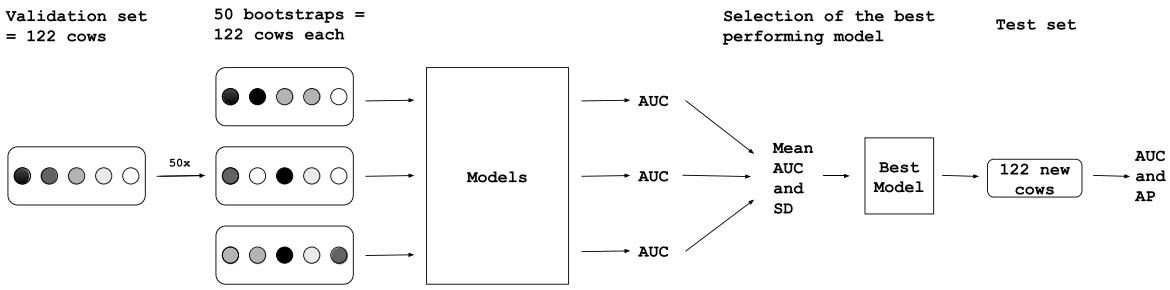


Fig. 2. Overview of model evaluation.

Table 1
For each of the three models, hyperparameters were tuned in order to select the hyperparameter configuration with the best results. The different hyperparameters are listed below, next to their possible settings, wherefrom 12 times 200 random combination were picked and tested. For the logistic regression model, all the combinations of settings were tested.

Model category	Hyperparameter	Possible settings
Logistic regression	Upsampling	True or False
	Class weights	1:1, 1:2, 1:3, 1:4, 1:5
	Predicted variable	Based on Cut-off or Based on Cluster
	Use of static features	None, All, Without BCS and locomotion score
XgBoost	Learning rate	Log-Uniform between 0.0001 and 0.1
	Minimum loss reduction required for partition	0 or 1
	Maximum depth of a tree	2, 3, ..., 10
	Minimum sum of instance weight needed in a child	1, 2, 3, 4
	Class weights	1:1 or 1:2.7
	Upsampling	True or False
	Predicted variable	Based on Cut-off or Based on Cluster
	Use of static features	None, All, Without BCS and locomotion score
Deep learning model	Number of LSTM layers	0, 1, 2
	Size of hidden state	10, 20, ..., 100
	Use of ReLu activation LSTM	True or False
	Dropout Rate	0, 0.1, ..., 0.4
	Batch Normalization	True or False
	Batch Size	12, 22, 32, 42
	Use of static features	None, All, Without BCS and locomotion score
	Upsampling	True or False
	Predicted variable	Based on Cut-off or Based on Cluster
	Class weights	1:1, 1:2, 1:3, 1:4
	Size MLP layer for static features	10, 20, ..., 80

ontology, all the code written for this study, the code written for the ontology and a figure to visualize the structure of the ontology can be found on https://github.com/Bovi-analytics/van-leerdam-et-al.

3. Results

3.1. Calcium measurements

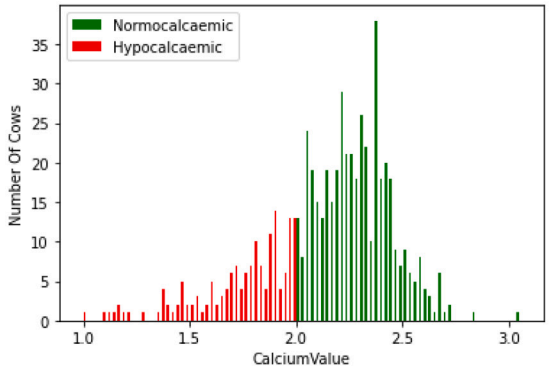
The mean calcium concentration was 2.15 mmol/L. Approximately 26.3% of the cows had a calcium concentration lower than or equal to 2.0 mmol/L. The results of the measurements on the blood samples taken from the cows within the 48 h after calving, as well as the division between the two categories, are presented in Fig. 3(a). The k-Means clustering of the measured calcium concentrations and residuals produced by the linear mixed-effects model resulted in two clustered categories, visualized in Fig. 4. Fig. 3(b) presents the distribution of calcium concentrations per clustered category. The percentage of cows attributed to the low calcium cluster was approximately 26.8%, thus the group sizes between the two methods were comparable.

3.2. Difference in behaviour between the two calcium categories

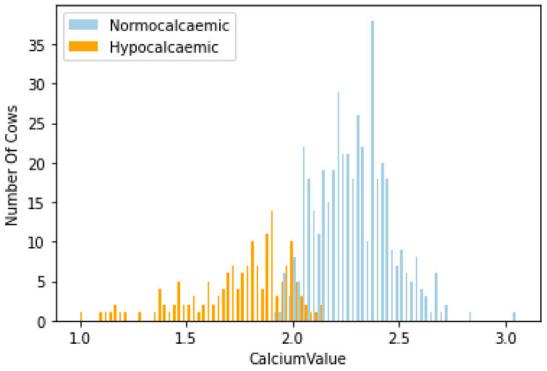
The changes in average daily minutes spent on a behaviour during the 21 days before calving per calcium category are visualized in appendix Fig. 7. When comparing the different calcium categories, the following observations were made, regardless of the method used of calcium categorization; normocalcaemic cows numerically spent more time walking and eating than hypocalcaemic cows during the entire 21-day period before calving. For eating behaviour, this difference increased closer to parturition. Low calcium cows spent fewer minutes lying and more minutes standing per day until day 5 before calving. Rumination had limited differences between normocalcaemic and hypocalcaemic cows, except for the 5 days before calving, when the hypocalcaemic cows spent more minutes ruminating.

3.3. Selecting the best model

All models predicted best using the cut-off categorization method. The results of bootstrapping for the different combinations of input features and labels are presented in Table 2. The static cow features improved the model performance by 0.11 for the LSTM deep learning model, 0.15 for the XgBoost and 0.23 for the logistic regression model, based on the absolute increase in mean AUC of the bootstraps. The 95% confidence interval of the difference in mean AUC between the best performing model with and without the addition of the static cow features, did not include zero for all three models. This improvement in model performance was considered significant. The addition of the BCS and locomotion score did not significantly improve the model. When comparing the mean AUC of the best model without the BCSs and locomotion scores to the best model with the BCS and locomotion scores, the 95% confidence interval of the mean difference included zero.



(a) Categorization based on a calcium cut-off value; hypocalcaemic cows with a calcium concentration $\leq$ 2.0 mmol/L in red and normocalcaemic cows with a calcium concentration $>$ 2.0 mmol/L in green.



(b) Categorization based on the linear mixed-effects model combined with the k-Means clustering method for dividing the calcium concentrations into two groups; normocalcaemic in blue and hypocalcaemic in orange with a correction for the day of blood sampling, parity and farm.

Fig. 3. Distribution of calcium concentrations per category (normocalcaemic and hypocalcaemic). Figure a is based on the cut-off value and figure b is based on the clustering method.

Table 2
Mean AUC with SD of the models evaluated on **the bootstraps of the validation set** for different combinations of input features and the calcium category to predict, based either on the clustering method or on the cut-off value of 2.0 mmol/L. The small set of static feature comprises the day of blood sampling compared to calving, parity and calving season. The complete set of static features contains next to the day of blood sampling compared to calving, parity and calving season also BCS and locomotion score. For the logistic regression model (Log Reg) upsampling was part of the grid search and was not extra evaluated using bootstraps, therefore only one mean AUC value per category is given.

Categorization method	Upsampling	Static features set	Log Reg AUC/SD		XgBoost AUC/SD		LSTM AUC/SD	
Cluster	–	–	0.51	0.082	0.58	0.066	0.63	0.059
Cluster	+	–	–	–	0.57	0.073	0.47	0.069
Cut-off	–	–	–	–	0.48	0.074	0.62	0.066
Cut-off	+	–	0.53	0.076	0.65	0.060	0.65	0.076
Cluster	–	Small	0.66	0.046	0.72	0.063	0.62	0.061
Cluster	+	Small	–	–	0.69	0.056	0.72	0.058
Cut-off	–	Small	–	–	0.79	0.061	0.76	0.055
Cut-off	+	Small	0.76	0.052	0.75	0.039	0.75	0.059
Cluster	–	All	0.69	0.055	0.68	0.059	0.75	0.055
Cluster	+	All	–	–	0.70	0.066	0.63	0.056
Cut-off	–	All	–	–	0.80	0.050	0.42	0.072
Cut-off	+	All	0.76	0.052	0.75	0.053	0.76	0.054

3.4. Performance of the final models

The model performance on the test set of the best hyperparameter configuration of each model is presented in Table 3. The AUC value was highest in the deep learning model, with the LSTM layer and all static features. The logistic regression model had the highest AP, but the difference is marginal. Table 3 also describes the accuracy, sensitivity and specificity of the best performing models with a decision threshold value of 0.5. The LSTM deep learning model had the highest sensitivity of 0.95 with a specificity of 0.44. The XgBoost had the highest specificity of 0.75 with a sensitivity of 0.51. Accuracy was also highest for the XgBoost model at this threshold. This value cannot be used to compare model performances as accuracy is threshold dependent.

Table 3
Performance of the best models tasked to predict the calcium categories of **the test set**. Sensitivity, specificity and accuracy were calculated using a decision threshold of 0.5 when predicting. Selection of the best model was based on the mean AUC values on **the bootstraps of the validation set** per model type.

Model	AUC	AP	Sensitivity	Specificity	Accuracy
Logistic regression	0.66	0.50	0.88	0.43	0.59
XgBoost	0.67	0.49	0.51	0.75	0.66
LSTM model	0.71	0.47	0.95	0.44	0.62

3.5. Feature importance

The results for the permutation feature importance are reported in Tables 4 and 5. For the best performing model built using only behavioural sensor data, walking time caused the biggest decrease in accuracy of 0.075 when permuted. Eating and standing were the least important, with a decrease of 0.03. The accuracy of the best performing model was most influenced by parity. The most important behavioural feature was rumination, proven by a decrease in accuracy of 0.012. In this model, walking, summer, eating, autumn and day of measurement relative to calving did not decrease the accuracy of the model when imputed, indicating no predictive power.

4. Discussion

4.1. What was achieved?

The proposed models predict the probability at parturition that a cow will be in the low calcium category post-partum, and thereby the risk of hypocalcaemia. To our knowledge, this is the first study that shows that prediction of the risk of hypocalcaemia with behavioural

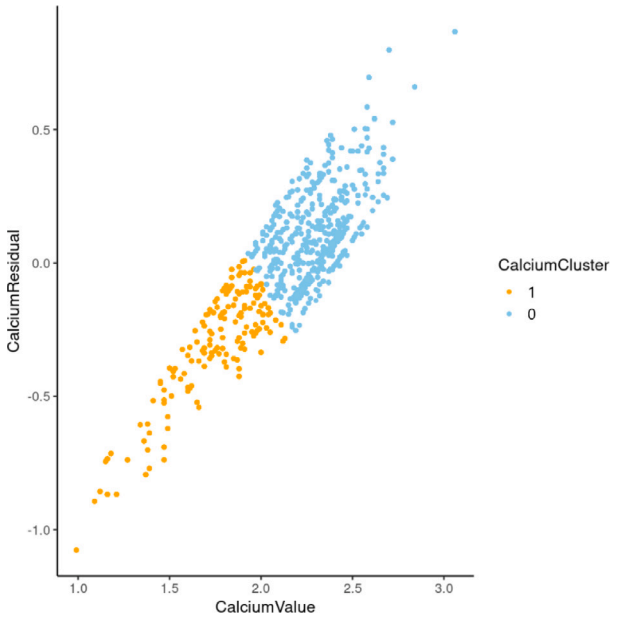


Fig. 4. The residuals of the calcium prediction by the linear mixed-effects model, plotted with their corresponding calcium concentration in mmol/L. The two clusters are determined by k-Means clustering and are visualized by a colour difference.

Table 4
Feature importance of the best performing LSTM deep learning model with only sensor features, based on 50 permutations of each of the behavioural features. The importance is defined as the mean decrease in model accuracy based on the values of the validation set when comparing the original model with the permuted models. A bigger decrease means a higher reliability of the model to the feature.

Behavioural feature	Mean decrease in accuracy
Walking time	0.075
Rumination time	0.048
Lying time	0.043
Standing time	0.030
Eating time	0.030

Table 5
Feature importance of the best performing LSTM deep learning model. This model used behavioural sensor values but also parity, calving season and day of measurement relative to calving. The reported values are based on 50 permutations of each of the features. The importance is defined as the mean decrease in model accuracy based on the values of the validation set when comparing the original model with the permuted models. A bigger decrease means a higher reliability of the model to the feature. A negative result means the model accuracy does not decrease due to the permutation and this feature was therefore irrelevant for the model.

Input feature	Mean decrease in accuracy
Parity 1	0.136
Parity 3+	0.082
Parity 2	0.047
Rumination time	0.012
Winter	0.010
Standing time	0.008
Spring	0.003
Lying time	0.001
Walking time	−0.001
Summer	−0.003
Eating time	−0.005
Autumn	−0.005
Day of Calcium Measurement	−0.027

sensor data is possible. However, the best performing model still performs far from perfect for practical decision support. An AUC of 0.5 can be expected when a coin is tossed in order to predict whether a cow has hypocalcaemia or not, while a perfect model assigning each cow to the correct group would have an AUC of 1. With an AUC of 0.71 the model performs better than random, but still far from 1. In order to be able to use this model in practice, the model performance will have to increase.

The difference between traditional machine learning models and the deep learning model is that the LSTM can discover sequential patterns in the data. The applied machine learning models use flattened data, therefore the time dimension is lost. The deep learning model was the best performing model based on AUC. This finding could indicate that the temporal patterns in the sensor data, and not only the absolute occurrence of behaviour, differ between normocalcaemic and hypocalcaemic cows and that these patterns have a predictive value. A finding corresponding with Hendriks et al. (2020), who found that relative changes in daily and hourly daytime lying time in the two weeks before calving were negatively associated with the blood calcium concentration within 24 h after calving, in contradiction to the relative change in daily and hourly daytime steps, which were positively associated with the blood calcium concentration after calving. In this study, however, the differences in model performance between the three models were small and the conclusion that the deep learning clearly outperformed the other models cannot be drawn based upon the bootstrap results. However, our confidence in the superiority of the deep learning model remains steadfast, as it achieved the highest test set result of an AUC of 0.71 and demonstrated a substantial sensitivity of 0.95. The high sensitivity indicates few false negative values when predicting hypocalcaemia. This is important because false negatives can be especially dangerous since they could delay the detection of clinical hypocalcaemia and other transition diseases, thereby quickly lowering the user's confidence in the model (Petticrew et al., 2000).

4.2. Back to hypocalcaemia

Clinical hypocalcaemia impairs animal welfare, farm economics and has a long-lasting impact on transition success. It is associated with numerous postpartum health events including dystocia, retained placenta, ketosis, and mastitis (Curtis et al., 1983; Erb et al., 1985; Correa et al., 1990; Klerx and Smolders, 1997) In addition, CH affected cows produce less milk and have an increased time to pregnancy (Venjakob, 2018; Hostens et al., 2012; Pascottini et al., 2020; Probo et al., 2018). On the other hand, for subclinical hypocalcaemia, the effect on transition success is not as easily defined and depends on the day of calcium sampling, the duration of low blood calcium values and parity (Neves et al., 2018; McArt and Oetzel, 2023). Transient hypocalcaemia; only at 1 day after calving, does not lead to increased disease events and is associated with higher milk yield than normocalcaemic cows, while chronic or delayed SCH does lead to adverse events (McArt and Neves, 2020). It has been hypothesized that hypocalcaemia beyond 48 h after parturition is not caused by a primary problem of adapting to a new calcium demand, but rather by reduced feed intake or/and inflammation (Serrenho et al., 2021; Horst et al., 2021). This turns hypocalcaemia into a symptom rather than an individual disease, and could therefore be an indicator of an impaired transition period.

There is, however, no distinction made by the model between transient, chronic or delayed hypocalcaemia, since blood measurements were not taken at several fixed moments, but varied between the 48 h after calving and only one blood sample was taken from each cow. It is therefore plausible that cows were assigned to the risk group while in fact be a healthy, high producing cow with transient hypocalcaemia. The model also does not distinguish between subclinical and clinical hypocalcaemia, because there were too few clinically affected cows to train the model this distinction.

We believe that, for future research, it is important to change the way of calcium categorization. Multiple blood samples should be taken instead of only one calcium measurement. This makes it possible to differentiate between chronic, delayed and transient hypocalcaemia

and then train the model to predict clinical, chronic or delayed hypocalcaemia only. Moreover, from a practical point of view, it would be useful to differentiate between different forms of hypocalcaemia, since the best course of action for disease prevention could differ between variants.

Others have proposed predictive models for hypocalcaemia using different data. Ma et al. (2022) proposed a multivariate logistic regression model using blood analytes to predict the risk of subclinical hypocalcaemia. The authors report a very high model AUC of 0.90, suggesting high model performance, but the model did not evaluate using a test set and the results can therefore not be interpreted as predictions and the model is therefore likely overfitted. Besides, from each cow 2 blood samples were taken at different time points antepartum, making it a very labour-intensive method for prediction not suitable for practice. Using genomic information, Cavani et al. (2022) employed a multiple linear regression model to predict blood calcium concentration after calving. The authors reported a predictive correlation average of 0.463 ±0.056, 0.396 ±0.052, and 0.297 ±0.057 for blood calcium concentrations on day 1, day 2, and day 3 after calving, respectively. The strongest association was observed on day 1, indicating the highest predictability. Although the model by Cavani et al. (2022) did not achieve high accuracy, it demonstrated the potential of genomics for prediction.

4.3. Suggestions to improve model performance

As stated before, this study showed that it is possible to predict hypocalcaemia, but the AUC value is too low for practical implementation. Fortunately, there are multiple ways to improve model performance. The first element of a good performing model is high input and output quality. The behaviour as recorded by the sensors agrees with true behaviour, but with a range of error (Borchers et al., 2021; Nielsen et al., 2018). In combination with uncertainty in calcium measurements, BCS scoring and locomotion scoring, this results in an overall relatively low precision due to propagation of uncertainty. On top of this, 14.6% of the sensor values were missing from the dataset and had to be imputed. The imputation uses a mean value and does not take into account the values before and after the missing value. The imputation therefore causes a disruption in the sequential patterns of the sensor data, which makes them more difficult to analyse. In a study by Liseune et al. (2021) an improved way of missing value imputation was proposed using deep learning to fill in missing values based on the values observed in the same sensor sequence as well as the recorded values of the other features. This method led to a significant increase in model performance for a methodology-wise similar predictive model, and is therefore a promising method to use for model improvement.

A second important factor in model performance is the number of animals. As stated before, neural networks are prone to overfitting. A neural network quickly becomes very complicated compared to other models due to its many connections and weights. In the proposed model, many input variables were used; 5 different sensor features and 10 static features, complexing the model even more. The general rule is that as the complexity of a model increases, the noise of the training set is better memorized and the model performance on new data decreases (Alpaydin, 2020). Many measures were therefore taken when constructing the model in order to prevent overfitting. Although, another effective way to prevent overfitting is to increase the amount of training data. Besides, with 365 cows in the training set, there is a chance that the sample is not a correct representation of the population of cows in the Netherlands and that the models consequently cannot be generalized. Increasing the number of observations could therefore increase model performance and reliability, but unfortunately is also expensive and labour-intensive.

A possibility to assess model performance for representativeness would have been to leave one of the eight herds out and use the cows of this herd as a test set. It would then have been known how the model would perform on a new farm, and thus it would have said something about the generalizability of the model. This is a difficult trade-off because on the other hand to make the model as generalizable as possible, it is beneficial to train the model on cows from as many herds as possible. And therefore this method was not applied in this study.

The sensitivity of the LSTM deep learning model was high (0.95), which means the model performed well on identifying true positives as positive. For the XgBoost, however, the specificity was high (0.75), indicating a high performance on correctly identifying true negatives as negatives. Future research could combine these two models through ensemble learning, to benefit from the high sensitivity of the LSTM deep learning model but use the XgBoost to validate the samples qualified as negative.

A cow displays many kinds of behaviour, for instance social behaviour, which are not included in the model. Furthermore, a lot of different other variables are associated with the risk of hypocalcaemia and could therefore explain a part of the variability between cows. For instance, the weather is associated with the risk of hypocalcaemia (Roche and Berry, 2006) and the cow's diet (Thilsing-Hansen et al., 2002). In the future, these variables could be added to the predictive model to enhance its performance. At the same time, previous research suggested that sometimes a model with equal reliability can be made while using fewer features, provided that for each feature a correlation with milk fever was previously proved (de Mol et al., 2015).

It was already mentioned that differentiating between different forms of hypocalcaemia is useful. It is thereby necessary to define which calcium concentration is too low. Serrenho et al. (2021) already pointed out that the cut-off value of 2.0 mmol/L is dubiously evidence based. The clusters proposed in this study are an alternative to the cut-off value, as they correct for parity and day of measurement and have a more fluent context based cut-off. This clustering is not defined based upon a direct association with post-partum disease, and therefore could lack clinical relevance. Besides, it proved to be more difficult to predict the clusters, although the difference in model performance is minimal. Future research could use the association with the outcome of interest to define a better cut-off value. In addition to taking multiple blood samples from each cow on day 1, 2 and 4 after calving in order to differentiate between different hypocalcaemia variants (Serrenho et al., 2021; McArt and Oetzel, 2023). Both suggestion can be used to increase the clinical relevance of the prediction, and we believe that improving the calcium categorization will ameliorate the AUC values of the models the most.

4.4. Feature importance

DeGaris and Lean (2008) stated that age increases the risk of milk fever by approximately 9% per lactation. Furthermore, McArt and Neves (2020) suggested that the calcium concentration patterns differ between primiparous and multiparous cows. It is not surprising that parity was the most important feature jugged by the cross-validated permutation feature importance. For the behavioural features it depended on the model which feature was most important. Rumination time was important for both the model trained only on the sensor values as for the model which included static features. Contrarily, walking time was the most important for the 'only sensor' model but irrelevant for the model with static features. This finding suggests that the importance of features differs between model configurations, and further research is necessary to identify which behavioural feature is the most informative for the prediction of hypocalcaemia. In this study both a neck and a leg sensor were used, but because of the reason stated above it is not possible to say which provides more useful information and should be used in the future.

Lameness is a known risk factor for SCH (Neves et al., 2017) and can be quantified using the locomotion score. In addition, there is a known correlation between BCS and hypocalcaemia (Heuer et al.,

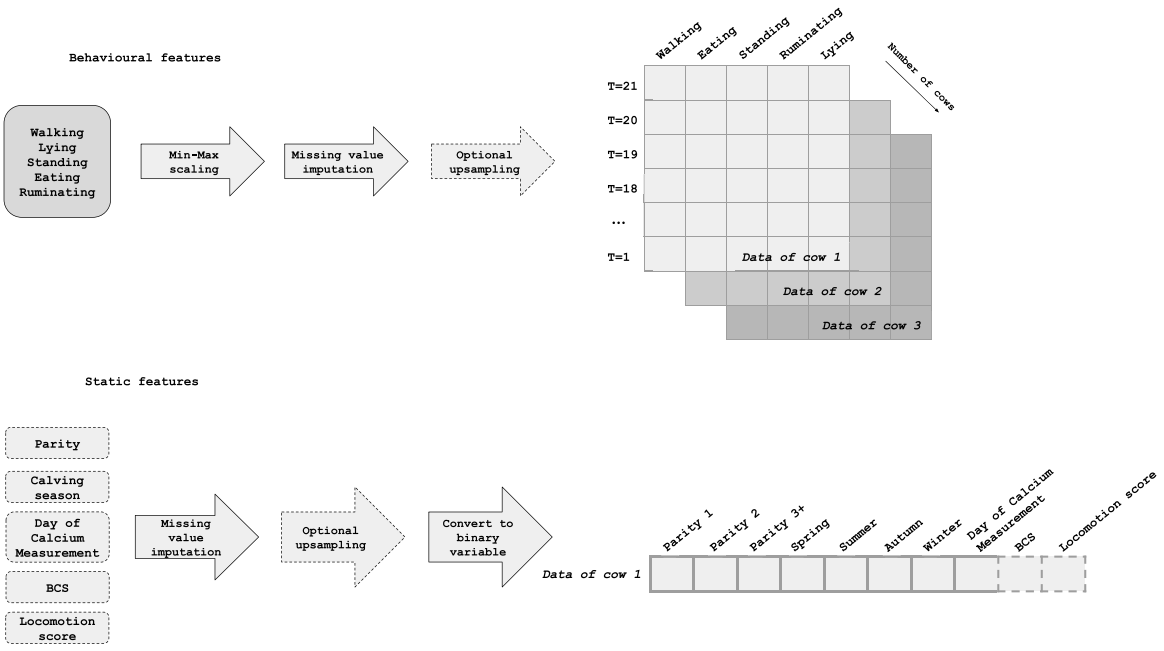


Fig. 5. Visualization illustrating the preprocessing steps applied to features following the train–validation–test split. Daily minutes spent per behaviour before calving were recorded and underwent scaling, imputation, and transformation into a 3D matrix. This matrix served as input for the LSTM deep learning model, and after flattening, for the logistic regression and XgBoost models. The *x*-axis represents days before calving (T), the *y*-axis displays various behaviours, and the *z*-axis corresponds to different cows. This process was repeated with upsampling to create a second matrix. Static features for each cow were documented, with imputation applied to missing values in BCS and Locomotion score. The categorical variables calving season and parity were converted to binary format. All variables were merged in an array encompassing all static features per cow. Notably, BCS and locomotion score are delineated with dashed boundaries, as they are unused in the small static features set. This process was also repeated with upsampling.

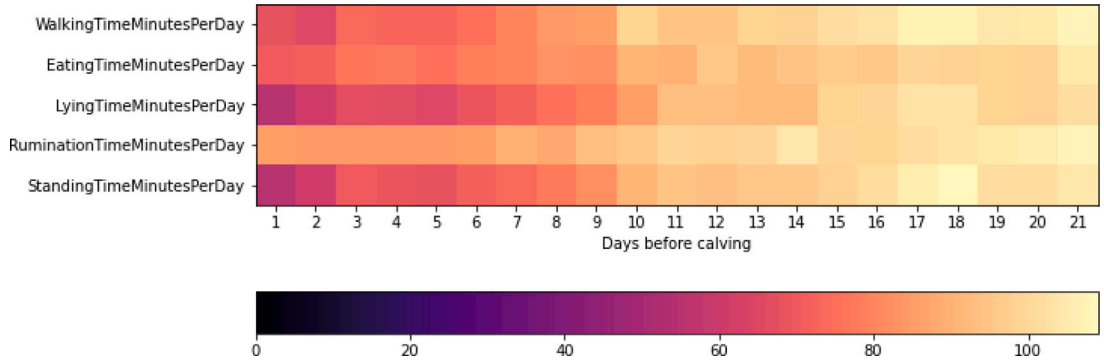


Fig. 6. Distribution of missing values per behavioural feature. In this study 609 cows participated, meaning that for each day and each behavioural feature 609 observations were recorded, wherefrom some were missing. The amount was visualized with a colour, a brighter colour means more missing values.

1999; Roche and Berry, 2006). One would therefore expect that adding BCS and locomotion score as input variables would improve model performance. It turned out that this was not the case, as there was no significant difference in mean AUC between the best performing model with BCS and locomotion score and the best performing model without. This contradiction could be explained by two phenomena. Firstly, as stated before, when increasing the complexity of the model, the noise of the training set is better memorized and the model performance on new data decreases. Adding two extra input features leads to many extra connections and weights, and therefore adds extra complexity, potentially causing overfitting. The second reason could be that the model already recognizes lameness through the sensor data for instance from the walking, standing and lying features and as a consequence the locomotion score does not provide extra information to use for the prediction. In this study the locomotion score and BCS were scored by the same trained veterinarian, but when implementing this model in practice there will not always be trained personnel to assess these scores. This makes it inconvenient for use in a future model.

4.5. Use in practice

The model predicts the probability for a cow to belong in the low calcium category. An ameliorated version of this model could serve as a tool to identify high-risk animals. A high-risk animal would be one with a high probability to fall within the low calcium category. As stated before, there is a known correlation between hypocalcaemia and transition diseases (Curtis et al., 1983; Erb et al., 1985; Correa et al., 1990; Klerx and Smolders, 1997). Early detection of high-risk animals could augment early detection of other associated diseases or underlying causes of reduced transition success. From a management perspective, this tool could serve as a method to keep track of calcium status of the herd. Nowadays, the only tool to evaluate total calcium concentration is to regularly take blood samples and measure calcium using quantitative analysis. A method not often applied due to costs, labour and an inability of on-farm testing (McArt and Oetzel, 2023). An improved model could provide insights in calcium status, evaluate preventive measures, review diet changes or be used as a screening tool.

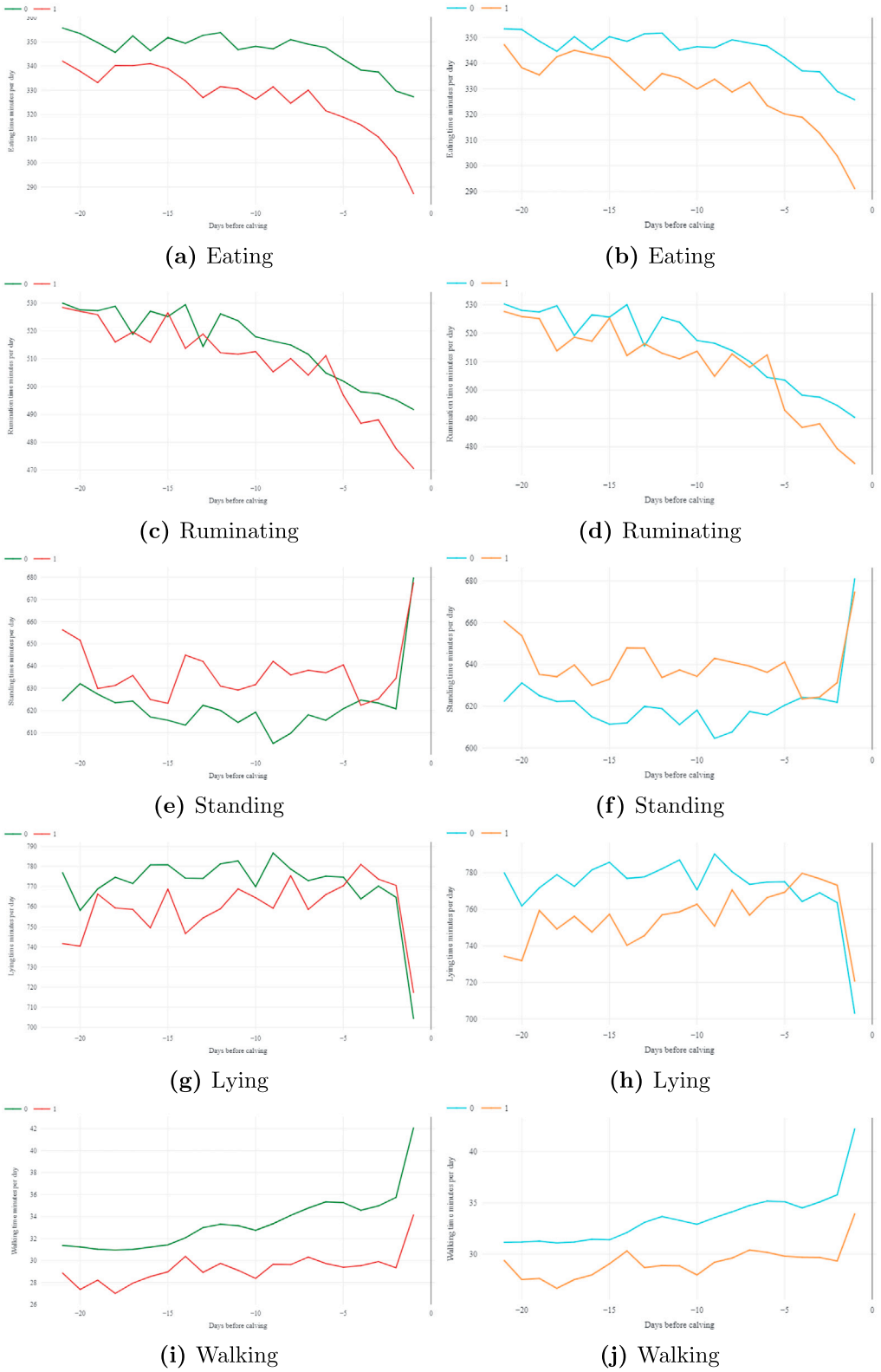


Fig. 7. Changes in average daily minutes spent on a behaviour per calcium categorization method during the 21 days before calving. On the left based on the cut-off value and on the right based on the cluster method. 1 = Hypocalcaemic (red/orange), 0 = Normocalcaemic (green/blue).

5. Conclusion

We were able to predict the risk of hypocalcaemia using behavioural sensor data and measured calcium concentrations, with an AUC value of 0.71 and an AP of 0.47. The behavioural patterns of the 21 days before calving contain valuable insights to predict hypocalcaemia after parturition, as do the static features: parity and calving season. The predictions of an ameliorated version of the model can be used to monitor herd calcium status and to identify animals at risk for transition diseases. Although there is still a long way to go to develop a model

suitable for widespread practical implementation, the proposed model provides a first step towards achieving that goal.

CRediT authorship contribution statement

Meike van Leerdam: Writing – review & editing, Writing – original draft, Software, Methodology. **Peter R. Hut:** Supervision, Investigation, Conceptualization. **Arno Liseune:** Software, Methodology. **Elena Slavco:** Software. **Jan Hulsen:** Conceptualization. **Miel Hostens:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Peter R. Hut reports financial support was provided by Boehringer Ingelheim Animal Health for the analysis of blood serum calcium.

Data availability

The research data, all the codes written for this paper and a figure to visualize the structure of the ontology can be found on https://github.com/Bovi-analytics/van-leerdam-et-al.

sensor_data.zip (Original data) (Github)

.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used ChatGPT (version 3.5) in order to provide suggestions for enhancing the writing flow of specific paragraphs. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Acknowledgements

The author would like to thank Boehringer Ingelheim for their support of the calcium measurements and Nedap Livestock Management (Groenlo, the Netherlands), especially Arnold Harbers, for providing the sensor data. The authors declare that they had no conflict of interest related to the study discussed in this manuscript.

Appendix

See Figs. 5–7.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: A system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation. pp. 265–283. http://dx.doi.org/10.48550/arXiv.1605.08695.

Alpaydin, E., 2020. Introduction to Machine Learning. MIT Press.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48. http://dx.doi.org/10.18637/jss.v067.i01.

Bauman, D.E., Currie, W.B., 1980. Partitioning of nutrients during pregnancy and lactation: A review of mechanisms involving homeostasis and homeorhesis. J. Dairy Sci. 63 (9), 1514–1529. http://dx.doi.org/10.3168/jds.S0022-0302(80)83111-0.

Borchers, M., Gavigan, S., Harbers, A., Bewley, J., 2021. An evaluation of a novel device for measuring eating, rumination, and inactive behaviors in lactating holstein dairy cattle. Animal 15 (1), 100008. http://dx.doi.org/10.1016/j.animal.2020.100008.

Cavani, L., Poindexter, M.B., Nelson, C.D., Santos, J.E., Peñagaricano, F., 2022. Gene mapping, gene-set analysis, and genomic prediction of postpartum blood calcium in Holstein cows. J. Dairy Sci. 105 (1), 525–534. http://dx.doi.org/10.3168/jds.2021-20872.

Cavero, D., Tölle, K.H., Henze, C., Buxadé, C., Krieter, J., 2008. Mastitis detection in dairy cows by application of neural networks. Livestock Sci. 114 (2), 280–286. http://dx.doi.org/10.1016/j.livsci.2007.05.012.

Chapinal, N., De Passille, A., Rushen, J., Wagner, S., 2010. Automated methods for detecting lameness and measuring analgesia in dairy cattle. J. Dairy Sci. 93 (5), 2007–2013. http://dx.doi.org/10.3168/jds.2009-2803.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794. http://dx.doi.org/10.1145/2939672.2939785.

Chollet, F., et al., 2015. Keras. https://github.com/fchollet/keras.

Correa, M.T., Curtis, C.R., Erb, H.N., Scarlett, J.M., Smith, R.D., 1990. An ecological analysis of risk factors for postpartum disorders of Holstein-Friesian cows from thirty-two new york farms. J. Dairy Sci. 73 (6), 1515–1524. http://dx.doi.org/10.3168/jds.S0022-0302(90)78819-4.

Curtis, C., Erb, H., Sniffen, C., Smith, R., Powers, P., Smith, M., White, M., Hillman, R., Pearson, E., 1983. Association of parturient hypocalcemia with eight periparturient disorders in Holstein cows. J. Am. Veterinary Med. Assoc. 183 (5), 559–561.

de Mol, R., van Dijk, J., Troost, M., Sterk, A., Jorritsma, R., Hogewerf, P., 2015. 2. Early detection of metabolic disorders in dairy cows by using sensor data. In: Precision Livestock Farming Applications: Making Sense of Sensors To Support Farm Management. Wageningen Academic Publishers, pp. 274–280.

DeGaris, P.J., Lean, I.J., 2008. Milk fever in dairy cows: A review of pathophysiology and control principles. The Veterinary Journal 176 (1), 58–69. http://dx.doi.org/10.1016/j.tvjl.2007.12.029.

Drackley, J.K., 1999. Biology of dairy cows during the transition period: The final frontier? J. Dairy Sci. 82 (11), 2259–2273. http://dx.doi.org/10.3168/jds.S0022-0302(99)75474-3.

Ennett, C.M., Frize, M., Walker, C.R., 2001. Influence of missing values on artificial neural network performance. In: MEDINFO 2001. Ios Press, pp. 449–453.

Erb, H., Smith, R., Oltenacu, P., Guard, C., Hillman, R., Powers, P., Smith, M., White, M., 1985. Path model of reproductive disorders and performance, milk fever, mastitis, milk yield, and culling in Holstein cows. J. Dairy Sci. 68 (12), 3337–3349. http://dx.doi.org/10.3168/jds.S0022-0302(85)81244-3.

Ferguson, J.D., Galligan, D.T., Thomsen, N., 1994. Principal descriptors of body condition score in Holstein cows. J. Dairy Sci. 77 (9), 2695–2703. http://dx.doi.org/10.3168/jds.S0022-0302(94)77212-X.

Firk, R., Stamer, E., Junge, W., Krieter, J., 2002. Automation of oestrus detection in dairy cows: A review. Livestock Prod. Sci. 75 (3), 219–232. http://dx.doi.org/10.1016/S0301-6226(01)00323-2.

Garcia, R., Aguilar, J., Toro, M., Pinto, A., Rodriguez, P., 2020. A systematic literature review on the use of machine learning in precision livestock farming. Comput. Electron. Agric. 179, 105826. http://dx.doi.org/10.1016/j.compag.2020.105826.

Gertz, M., Große-Butenuth, K., Junge, W., Maassen-Francke, B., Renner, C., Sparenberg, H., Krieter, J., 2020. Using the XGBoost algorithm to classify neck and leg activity sensor data using on-farm health recordings for locomotor-associated diseases. Comput. Electron. Agric. 173, 105404. http://dx.doi.org/10.1016/j.compag.2020.105404.

Grummer, R.R., 1995. Impact of changes in organic nutrient metabolism on feeding the transition dairy cow. J. Animal Sci. 73 (9), 2820–2833. http://dx.doi.org/10.2527/1995.7392820x.

Gusterer, E., Kanz, P., Krieger, S., Schweinzer, V., Süss, D., Lidauer, L., Kickinger, F., Öhlschuster, M., Auer, W., Drillich, M., et al., 2020. Sensor technology to support herd health monitoring: Using rumination duration and activity measures as unspecific variables for the early detection of dairy cows with health deviations. Theriogenology 157, 61–69. http://dx.doi.org/10.1016/j.theriogenology.2020.07.028.

Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al., 2020. Array programming with NumPy. Nature 585 (7825), 357–362. http://dx.doi.org/10.1038/s41586-020-2649-2.

Hendriks, S.J., Huzzey, J.M., Kuhn-Sherlock, B., Turner, S.A., Mueller, K.R., Phyn, C.V.C., Donaghy, D.J., Roche, J.R., 2020. Associations between lying behavior and activity and hypocalcemia in grazing dairy cows during the transition period. J. Dairy Sci. 103 (11), 10530–10546. http://dx.doi.org/10.3168/jds.2019-18111.

Heuer, C., Schukken, Y., Dobbelaar, P., 1999. Postpartum body condition score and results from the first test day milk as predictors of disease, fertility, yield, and culling in commercial dairy herds. J. Dairy Sci. 82 (2), 295–304. http://dx.doi.org/10.3168/jds.S0022-0302(99)75236-7.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Horst, R., Goff, J., Reinhardt, T., 1994. Calcium and vitamin d metabolism in the dairy cow. J. Dairy Sci. 77 (7), 1936–1951. http://dx.doi.org/10.3168/jds.S0022-0302(94)77140-X.

Horst, E.A., Kvidera, S.K., Baumgard, L.H., 2021. Invited review: The influence of immune activation on transition cow health and performance—A critical evaluation of traditional dogmas. J. Dairy Sci. 104 (8), 8380–8410. http://dx.doi.org/10.3168/jds.2021-20330.

Hostens, M., Ehrlich, J., Van Ranst, B., Opsomer, G., 2012. On-farm evaluation of the effect of metabolic diseases on the shape of the lactation curve in dairy cows through the MilkBot lactation model. J. Dairy Sci. 95 (6), 2988–3007. http://dx.doi.org/10.3168/jds.2011-4791.

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9 (3), 90–95. http://dx.doi.org/10.1109/MCSE.2007.55.

Hut, P.R., Hostens, M.M., Beijaard, M.J., van Eerdenburg, F.J.C.M., Hulsen, J.H.J.L., Hooijer, G.A., Stassen, E.N., Nielen, M., 2021. Associations between body condition score, locomotion score, and sensor-based time budgets of dairy cattle during the dry period and early lactation. J. Dairy Sci. 104 (4), 4746–4763. http://dx.doi.org/10.3168/jds.2020-19200.

Janiesch, C., Zschech, P., Heinrich, K., 2021. Machine learning and deep learning. Electron. Mark. 31 (3), 685–695. http://dx.doi.org/10.1007/s12525-021-00475-2.

Jensen, D.B., Hogeveen, H., De Vries, A., 2016. Bayesian integration of sensor information and a multivariate dynamic linear model for prediction of dairy cow mastitis. J. Dairy Sci. 99 (9), 7344–7361. http://dx.doi.org/10.3168/jds.2015-10060.

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6 (1), 1–54. http://dx.doi.org/10.1186/s40537-019-0192-5.

Kaneko, H., 2022. Cross-validated permutation feature importance considering correlation between features. Anal. Sci. Adv. 3 (9–10), 278–287. http://dx.doi.org/10.1002/ansa.202200018.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. http://dx.doi.org/10.48550/arXiv.1412.6980, CoRR.

Klerx, H., Smolders, E., 1997. Herd and cow random variation in models of interrelationships between metabolic and reproductive disorders in high yielding multiparous Holstein dairy cattle in The Netherlands. Livestock Prod. Sci. 52 (1), 21–29. http://dx.doi.org/10.1016/S0301-6226(97)00116-4.

Krogh, A., 2008. What are artificial neural networks? Nature Biotechnol. 26 (2), 195–197. http://dx.doi.org/10.1038/nbt1386.

Leary, N., Pembroke, A., Duggan, P., 1992. Single stable reagent (arsenazo III) for optically robust measurement of calcium in serum and plasma. Clin. Chem. 38 (6), 904–908. http://dx.doi.org/10.1093/clinchem/38.6.904.

LeBlanc, S., Lissemore, K., Kelton, D., Duffield, T., Leslie, K., 2006. Major advances in disease prevention in dairy cattle. J. Dairy Sci. 89 (4), 1267–1279.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E., Stoica, I., 2018. Tune: A research platform for distributed model selection and training. http://dx.doi.org/10.48550/arXiv.1807.05118.

Liboreiro, D.N., Machado, K.S., Silva, P.R., Maturana, M.M., Nishimura, T.K., Brandão, A.P., Endres, M.I., Chebel, R.C., 2015. Characterization of peripartum rumination and activity of cows diagnosed with metabolic and uterine diseases. J. Dairy Sci. 98 (10), 6812–6827. http://dx.doi.org/10.3168/jds.2014-8947.

Liseune, A., den Poel, D.V., Hut, P.R., van Eerdenburg, F.J.C.M., Hostens, M., 2021. Leveraging sequential information from multivariate behavioral sensor data to predict the moment of calving in dairy cattle using deep learning. Comput. Electron. Agric. 191, 106566. http://dx.doi.org/10.1016/j.compag.2021.106566.

Ma, X., Gao, C., Yang, M., Zhang, B., Xu, C., Yang, W., 2022. Characteristics and prediction of subclinical hypocalcemia in dairy cows during the transition period using blood analytes. Medycyna Weterynaryjna-Vet. Med.-Sci. Pract. 78 (1), 31–35. http://dx.doi.org/10.21521/mw.6607.

McArt, J., Neves, R., 2020. Association of transient, persistent, or delayed subclinical hypocalcemia with early lactation disease, removal, and milk yield in Holstein cows. J. Dairy Sci. 103 (1), 690–701. http://dx.doi.org/10.3168/jds.2019-17191.

McArt, J.A., Oetzel, G.R., 2023. Considerations in the diagnosis and treatment of early lactation calcium disturbances. Vet. Clin.: Food Animal Pract. 39 (2), 241–259. http://dx.doi.org/10.1016/j.cvfa.2023.02.009.

McKinney, W., 2010. Data structures for statistical computing in Python. In: van der Walt, S., Millman, J. (Eds.), Proceedings of the 9th Python in Science Conference. pp. 56–61. http://dx.doi.org/10.25080/Majora-92bf1922-00a.

Neves, R., Leno, B., Bach, K., McArt, J., 2018. Epidemiology of subclinical hypocalcemia in early-lactation holstein dairy cows: The temporal associations of plasma calcium concentration in the first 4 days in milk with disease and milk production. J. Dairy Sci. 101 (10), 9321–9331. http://dx.doi.org/10.3168/jds.2018-14587.

Neves, R., Leno, B., Stokol, T., Overton, T., McArt, J., 2017. Risk factors associated with postpartum subclinical hypocalcemia in dairy cows. J. Dairy Sci. 100 (5), 3796–3804. http://dx.doi.org/10.3168/jds.2016-11970.

Nielsen, P.P., Fontana, I., Sloth, K.H., Guarino, M., Blokhuis, H., 2018. Validation and comparison of 2 commercially available activity loggers. J. Dairy Sci. 101 (6), 5449–5453. http://dx.doi.org/10.3168/jds.2017-13784.

Overton, T., McArt, J., Nydam, D., 2017. A 100-year review: Metabolic health indicators and management of dairy cattle. J. Dairy Sci. 100 (12), 10398–10417. http://dx.doi.org/10.3168/jds.2017-13054.

Pascottini, O.B., Probo, M., LeBlanc, S., Opsomer, G., Hostens, M., 2020. Assessment of associations between transition diseases and reproductive performance of dairy cows using survival analysis and decision tree algorithms. Prev. Vet. Med. 176, 104908. http://dx.doi.org/10.1016/j.prevetmed.2020.104908.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. http://dx.doi.org/10.5555/1953048.2078195.

Petticrew, M., Sowden, A., Lister-Sharp, D., Wright, K., 2000. False-negative results in screening programmes: Systematic review of impact and implications. Health Technol. Assess. (Winchester, England) 4 (5), 1–120. http://dx.doi.org/10.3310/hta4050.

Probo, M., Pascottini, O.B., LeBlanc, S., Opsomer, G., Hostens, M., 2018. Association between metabolic diseases and the culling risk of high-yielding dairy cows in a transition management facility using survival and decision tree analysis. J. Dairy Sci. 101 (10), 9419–9429. http://dx.doi.org/10.3168/jds.2018-14422.

R. Core Team, t., 2013. R: A language and environment for statistical computing. http://dx.doi.org/10.1890/0012-9658(2002)083[3097:CFHIWS]2.0.CO;2.

Reinhardt, T.A., Lippolis, J.D., McCluskey, B.J., Goff, J.P., Horst, R.L., 2011. Prevalence of subclinical hypocalcemia in dairy herds. Vet. J. 188 (1), 122–124. http://dx.doi.org/10.1016/j.tvjl.2010.03.025.

Ribeiro, E., Lima, F., Greco, L., Bisinotto, R., Monteiro, A., Favoreto, M., Ayres, H., Marsola, R., Martinez, N., Thatcher, W., et al., 2013. Prevalence of periparturient diseases and effects on fertility of seasonally calving grazing dairy cows supplemented with concentrates. J. Dairy Sci. 96 (9), 5682–5697. http://dx.doi.org/10.3168/jds.2012-6335.

Roche, J.R., Berry, D.P., 2006. Periparturient climatic, animal, and management factors influencing the incidence of milk fever in grazing systems. J. Dairy Sci. 89 (7), 2775–2783. http://dx.doi.org/10.3168/jds.S0022-0302(06)72354-2.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536. http://dx.doi.org/10.1038/323533a0.

Sanders, A., Shearer, J., De Vries, A., 2009. Seasonal incidence of lameness and risk factors associated with thin soles, white line disease, ulcers, and sole punctures in dairy cattle. J. Dairy Sci. 92 (7), 3165–3174. http://dx.doi.org/10.3168/jds.2008-1799.

Serrenho, R.C., DeVries, T.J., Duffield, T.F., LeBlanc, S.J., 2021. Graduate student literature review: What do we know about the effects of clinical and subclinical hypocalcemia on health and performance of dairy cows? J. Dairy Sci. 104 (5), 6304–6326. http://dx.doi.org/10.3168/jds.2020-19371.

Soriani, N., Trevisi, E., Calamari, L., 2012. Relationships between rumination time, metabolic conditions, and health status in dairy cows during the transition period. J. Anim. Sci. 90 (12), 4544–4554. http://dx.doi.org/10.2527/jas.2011-5064.

Sprecher, D.e.a., Hostetler, D.E., Kaneene, J., 1997. A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. Theriogenology 47 (6), 1179–1187. http://dx.doi.org/10.1016/S0093-691X(97)00098-8.

Stangaferro, M., Wijma, R., Caixeta, L., Al-Abri, M., Giordano, J., 2016. Use of rumination and activity monitoring for the identification of dairy cows with health disorders. J. Dairy Sci. 99 (9), 7422–7433. http://dx.doi.org/10.3168/jds.2016-10907.

Steeneveld, W., Hogeveen, H., 2015. Characterization of Dutch dairy farms using sensor systems for cow management. J. Dairy Sci. 98 (1), 709–717. http://dx.doi.org/10.3168/jds.2014-8595.

Stygar, A.H., Gómez, Y., Berteselli, G.V., Dalla Costa, E., Canali, E., Niemi, J.K., Llonch, P., Pastell, M., 2021. A systematic review on commercially available and validated sensor technologies for welfare assessment of dairy cattle. Front. Vet. Sci. 8, 177. http://dx.doi.org/10.3389/fvets.2021.634338.

The Pandas Development Team, 2020. Pandas-dev/pandas: Pandas. http://dx.doi.org/10.5281/zenodo.3509134.

Thilsing-Hansen, T., Jørgensen, R., Østergaard, S., 2002. Milk fever control principles: A review. Acta Vet. Scand. 43 (1), 1. http://dx.doi.org/10.1186/1751-0147-43-1.

UNESCO, 2021. UNESCO recommendation on open science. pp. 1–34, Document code:SC-PCB-SPP/2021/OS/UROS.

Venjakob, P.L., 2018. Diagnosis and Prevalence of Periparturient Hypocalcemia and Associated Effects on Milk Production, Reproductive Performance and Health of Dairy Cows in Early Lactation (Ph.D. thesis). http://dx.doi.org/10.17169/refubium-1004.

Venjakob, P., Borchardt, S., Heuwieser, W., 2017. Hypocalcemia—Cow-level prevalence and preventive strategies in German dairy herds. J. Dairy Sci. 100 (11), 9258–9266. http://dx.doi.org/10.3168/jds.2016-12494.

Wathes, C.M., Kristensen, H.H., Aerts, J.-M., Berckmans, D., 2008. Is precision livestock farming an engineer's daydream or nightmare, an animal's friend or foe, and a farmer's panacea or pitfall? Comput. Electron. Agric. 64 (1), 2–10. http://dx.doi.org/10.1016/j.compag.2008.05.005.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. springer, http://dx.doi.org/10.1080/15366367.2019.1565254.

Wickham, H., Francois, R., Henry, L., Müller, K., et al., 2015. Dplyr: A grammar of data manipulation. p. p156, R package version 0.4 3.

Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., et al., 2016. Apache spark: A unified engine for big data processing. Commun. ACM 59 (11), 56–65. http://dx.doi.org/10.1145/2934664.