Computers and Electronics in Agriculture 198 (2022) 107002

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Forecasting chronic mastitis using automatic milking system sensor data and gradient-boosting classifiers

John Bonestroo [a,b,c,*], Mariska van der Voort [c], Henk Hogeveen [c], Ulf Emanuelson [b], Ilka Christine Klaas [a], Nils Fall [b]

[a] DeLaval International AB, Gustaf De Lavals väg 15, 147 21 Tumba, Sweden
[b] Swedish University of Agricultural Sciences, Dep't Clinical Sciences, POB 7054, SE-750 07 Uppsala, Sweden
[c] Wageningen University and Research, Business Economics Group, Hollandseweg 1, 6706 KN Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Mastitis
Chronic
Sensor
Cow
Forecast

ABSTRACT

Although most of the losses due to mastitis per case in dairy production are estimated to be caused by clinical cases, subclinical cases, especially chronic, can also be problematic due to milk production losses and the risk of transmission of pathogens. Knowing which subclinical mastitis cases will become chronic at an early stage would be helpful in intervening in these cases. Automatic milking systems (AMS) can collect data on mastitis indicators such as conductivity, Somatic cell count (SCC), and blood in the milk for each milking. The aim of this study was to develop a sensor-based prediction model using SCC, conductivity, blood in the milk, parity, milk diversion, time interval between milkings, milk yield and DIM that forecasts the chronicity in subclinical mastitis cases after an initial increase in SCC. We used sensor data from 14 European and North American dairy farms (with herd sizes of lactating cows ranging from 55 to 638 cows and herd mean parities between 2.00 and 3.19) with an AMS and an online cell counter, measuring SCC. Typically, a threshold of 200,000 SCC/ml has been used to distinguish cows with subclinical mastitis from healthy cows. We used gradient-boosting trees and sensor data to forecast whether the SCC would decrease structurally below 200,000 SCC/ml in 50 days after the day at which the prediction was performed. Data from 30 and 15 days prior to the day where the forecast was made, was used. The model was trained on data from seven randomly selected dairy farms from the dataset and the data of the remaining seven dairy farms were used to estimate the predictive performance. These results were compared with two approaches that simulate how farmers would diagnose chronic mastitis with a simple prediction rule based on close-to-daily SCC (frequent sampling approach), and on less frequent monthly SCC (monthly sampling approach). We used accuracy, Matthew's correlation coefficient (MCC), and Area under the Curve (AUC) as metrics to assess the forecasting performance of the chronic mastitis prediction model. On average, the forecast model, using 30 days of sensor data prior to the day of prediction, outperformed the approaches according to the accuracy (chronic mastitis prediction model: 0.888, frequent sampling approach: 0.848, and monthly sampling approach: 0.865), MCC (chronic mastitis prediction model: 0.712, frequent sampling approach: 0.630, and monthly sampling approach: 0.552), and AUC metrics (chronic mastitis prediction model: 0.964 and frequent sampling approach: 0.941) metrics. The results also indicate that shortening the input requirement from 30 days of prior sensor data to 15 days has a limited effect on the performance of the model. Overall, this study shows that it is possible with a high accuracy to predict the future chronic mastitis status using past sensor data and machine learning models.

1. Introduction

Most of the economic losses due to mastitis in dairy production are estimated to be caused by clinical cases when estimated per case (Huijps et al., 2008). Nevertheless, subclinical cases, especially when they are chronic or long-term, can also be problematic due to milk production losses (Aghamohammadi et al., 2018) and the risk of transmission of pathogens (Swinkels et al., 2005). Subclinical mastitis is rarely treated during lactation on most dairy farms. However, some cases may develop into chronic subclinical mastitis, which can be defined as a case where

* Corresponding author at: Hollandseweg 1, Wageningen, the Netherlands.
E-mail address: John.bonestroo@delaval.com (J. Bonestroo).

https://doi.org/10.1016/j.compag.2022.107002
Received 19 November 2021; Received in revised form 20 April 2022; Accepted 21 April 2022
Available online 5 May 2022

Table 1
An overview of the variables used in the study.

Herd	No. milkings	Proportion milkings with milk diversion	Mean time between SCC Samples (in milkings)	No. lactations	Mean Parity	Proportion milkings with blood detected	Mean SCC (in 1000 cells per ml)	Mean STDCond uctivity	Mean IQRCond uctivity	Mean milk yield (in kg)	Mean time interval (in hours)
1	282,206	0.02	1.94	503	2.37	0.01	134.01	0.15	1.07	11.82	8.91
2	1,197,164	0.01	2.54	2,133	2.22	0.00	185.66	0.15	1.07	11.46	8.99
3	407,513	0.02	4.32	681	2.31	0.06	155.75	0.16	1.08	13.61	9.57
4	288,852	0.03	1.45	498	2.96	0.01	158.88	0.14	1.07	15.93	9.73
5	175,253	0.05	2.11	389	2.40	0.08	206.66	0.16	1.08	13.92	10.67
6	382,121	0.04	4.20	531	2.22	0.01	282.69	0.17	1.09	10.36	8.54
7	554,708	0.04	3.52	1,393	2.00	0.06	191.91	0.15	1.07	15.56	10.08
8	250,794	0.05	2.80	518	2.04	0.01	227.58	0.15	1.07	13.41	8.74
9	155,289	0.03	2.66	230	2.03	0.01	228.80	0.14	1.07	10.39	8.33
10	331,395	0.02	3.51	450	3.19	0.05	178.89	0.15	1.08	10.48	8.31
11	182,047	0.01	2.21	292	2.57	0.02	130.65	0.13	1.07	11.30	8.37
12	170,295	0.02	2.66	305	2.69	0.01	134.55	0.13	1.07	11.58	8.60
13	162,133	0.01	2.22	282	2.59	0.02	89.14	0.13	1.07	11.43	8.94
14	34,949	0.04	2.90	145	2.15	0.01	317.83	0.17	1.09	11.63	9.74

there is a long-term increased SCC that is not expected to cure spontaneously during lactation (Gonçalves et al., 2020). Chronic subclinical mastitis leads to prolonged periods of milk loss and increased risk of pathogen transmission. At the onset of and during subclinical mastitis, it would be useful to distinguish between cases that are expected to quickly cure spontaneously and cases that develop into chronic subclinical mastitis. In other words, it would be useful to forecast chronic subclinical mastitis so that early intervention (i.e. culling, early dry-off, or antibiotic treatment) is possible.

Despite the considerable number of studies on the sensor-based detection of clinical mastitis (Rutten et al., 2013; Jensen et al., 2016; Anglart et al., 2020), a smaller amount of research has been done into sensor-based detection of subclinical mastitis (Polat et al., 2010; Khatun et al., 2019). Subclinical mastitis is not commonly treated during lactation as it is not recommended (Krömker and Leimbach, 2017) and, therefore, subclinical mastitis detection may be regarded as less useful than clinical mastitis detection. However, prospective forecasting of chronic subclinical mastitis is now possible due to a clear definition of chronic subclinical mastitis (i.e. high inflammation indicators lasting longer than 4 weeks) (Bonestroo et al., 2021) and availability of data collected frequently from on-farm sensor systems.

Sensor systems can measure mastitis indicators such as conductivity, SCC (somatic cell count), and blood in the milk daily. Being more frequent than commonly performed monthly Dairy Herd Improvement (DHI) SCC sampling and testing, these high-frequency indicators could be used to obtain insight in udder health over time. The benefits of more frequent sampling would include a higher diagnostic performance to detect a case and potentially forecast the outcome of such a case. Therefore, both the occurrence of a sensor-based definition of chronic mastitis as well as the increased frequency of the measurement of udder health using AMS sensors make it possible to predict future chronic mastitis on an automatic basis thus providing the farmer with a useful tool to evaluate whether intervention is necessary.

The aim of this study was, therefore, to develop a sensor-based prediction model that forecasts the future subclinical chronic mastitis status based on past sensor data after an initial increase in SCC. The effect of using input data from a shorter period in the model on the predictive performance was explored using data from 30 days and 15 days prior to the moment of forecasting. The model based on SCC and using gradient-boosting trees, was compared to two approaches representing the performance achieved with simple prediction rules on monthly sampled data and daily SCC data alone.

2. Methods

2.1. Data

For this study, we used data from 14 herds from Belgium, Canada, France, Sweden, and the Netherlands, with herd sizes of lactating cows ranging from 55 to 638 cows. The data was retrieved from a central database of DeLaval International AB (Tumba, Sweden). Herds with an online cell counter (DeLaval OCC, DeLaval International AB, Tumba, Sweden) using and an automatic milking system (DeLaval VMS series, DeLaval International AB, Tumba, Sweden) were selected. The OCC was validated against laboratory SCC resulting in a high (0.82–0.86) correlation with laboratory SCC (Sørensen et al., 2016; Nørstebø et al., 2019) and measures the cells using ultraviolet fluorescence (Caja et al., 2016). The OCC is an add-on to the AMS that measures the commonly used SCC in the milk to assess the degree of udder inflammation.

Besides SCC data, the AMS collected data on the conductivity of the milk (in mS/cm) at quarter level, the occurrence of blood in the milk (using an RGB sensor) as well as milk yields (in kg). The data was recorded in different time intervals for different herds, but all herds started to record in 2016 or 2017 and the average time recorded per herd was 2.8 years, with a minimum of 1.4 years and a maximum of 4.2 years. The data was reported in a “per milking” frequency. An overview of the data can be seen in Table 1. SCC is not sampled every milking as it is dependent on a risk-based sampling algorithm. Nevertheless, it is on average sampled between 1.4 and 4.3 milkings and SCC samples were more likely to be taken in early lactation (i.e. before 30 DIM). Other missing values were far less prevalent and likely caused by the cow not entering the milking robot. This data included cow identification number, herd identification number, milk yield in kilograms, blood presence (binary variable indicating the presence of blood), SCC, DIM (days in milk), milk diversion (the action of diverting the milk away from the consumable milk bulk tank into a sink), 4 quarter mean conductivities throughout the milking, and parity (i.e. the lactation number of the cow).

We also calculated milk production rate (milk yield in kilograms divided by the time interval between milkings in hours, see Appendix A), standard deviation of quarter conductivities, interquarter ratio of conductivities (the highest quarter conductivity divided by the lowest quarter conductivity), and time interval between milkings in hours (16 variables in total). We selected cows for which we had the data from the start of the lactation (at least one milking reported in the first 10 DIM) as we wanted to have the start for all mastitis cases in the dataset. Furthermore, we removed all milking days that had a between-milking interval shorter than 3 and longer than 24 h because the milk-yield-based variables are misrepresented for milkings outside this range.

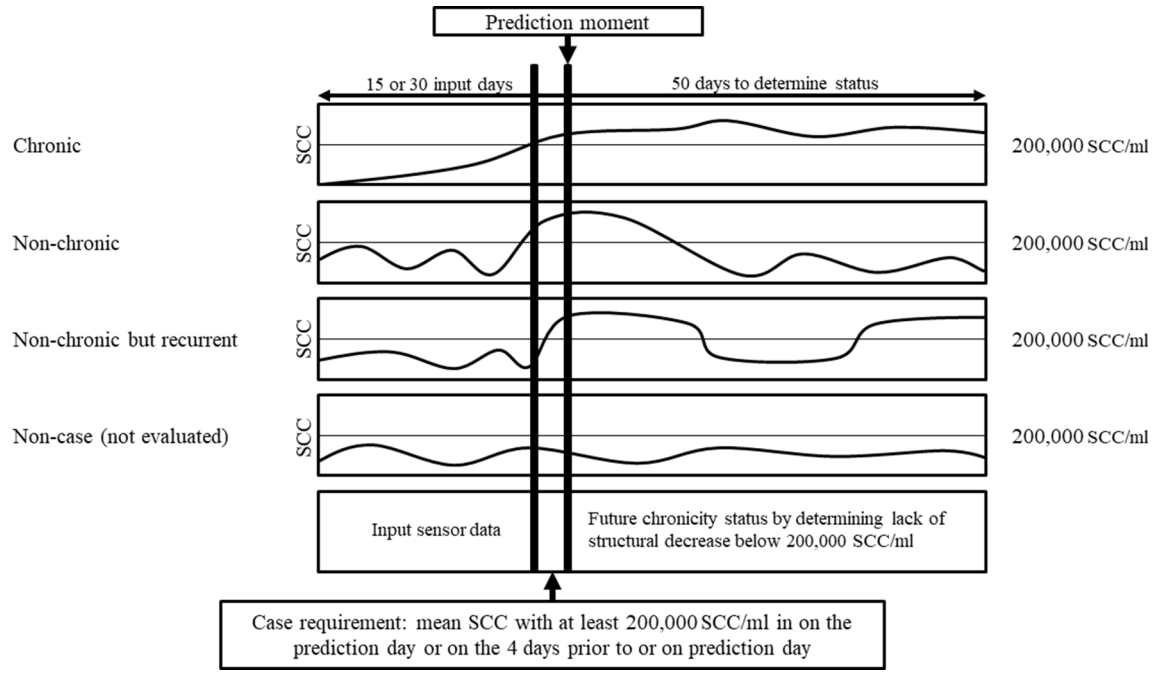

Fig. 1. Examples of the prediction task that was performed by the forecasting model where the future chronic mastitis label is created by determining whether the rolling 20-day mean of daily mean SCC decreased below 200,000 SCC/ml (0 = not chronic mastitis) or not (1 = chronic mastitis) at least once in the period from the prediction day to 50 days post the prediction day.

The combination of both steps removed a substantial share of cases with SCC equal or higher than 200,000/ml (222,607 to 148,172 high SCC observations).

2.2. Training and validation datasets

To create a training and a validation dataset, we randomly divided the herds in our dataset. Half of the herds were selected for the training set and the other half of the herds entered the validation set. Validation herds were identified as herd 1 to 7 beforehand while herds 8 to 14 were designated as training herds. The data from all the training herds were used to fit a prediction model all at once (i.e., the model was trained once using data from all herds), and data from the validation herds were used to test the model's performance.

2.3. Data pre-processing

All data processing and case prediction was performed in in Python 3.7. The data (i.e., milk yield, interquarter ratio of conductivity) from each milking per day was aggregated to a daily frequency using the mean, minimum, maximum, and standard deviation functions. These aggregated variables are the features used in the model. After the aggregation to a daily frequency, the daily mean, maximum, and standard deviation of quarter-level conductivity values (e.g., daily mean conductivity of the left-rear quarter) were aggregated to cow-level features. This aggregation was performed by calculating the mean over daily mean quarter conductivity values and the maximum over daily maximum quarter conductivity values. In addition, we also calculated the standard deviation over daily standard deviations of quarter conductivity values and the standard deviation over daily maximum quarter conductivity values. A description of the aggregated variables or features in the dataset with their calculation is given in Appendix A. All features had to be on cow level as we forecast chronic mastitis on cow level. The remaining quarter-level conductivity features were not included as input in the forecasting models as they were not reported on cow level. The daily maximum of SCC was transformed with the natural logarithm and defined as LnSCC. This was done as this feature was used to define the recovery.

2.4. Case definition

A prediction day (i.e., a day on which a prediction of a future state was made) was defined as a day in the lactation with at least a LnSCC higher than or equal to Ln(200) (International Dairy Federation, 2013) or having an LnSCC of such a level on 1 of the 4 days prior to the day. This would retain a prediction day when a SCC measurement was not done on the specific day but a mastitis case was ongoing. It is essential to mention that one mastitis case can have multiple prediction days as for each day of the episode, a forecast is performed. It would allow the farmer to monitor and get a forecast during an ongoing episode. For each day on which the future chronic mastitis status was forecasted, we used the data 30 days before the prediction day as input. Referring back to Appendix A, the feature values of each day during the last 30 days (e.g., MaxIQRConductivity on the 16th day before the prediction day) could be used by the forecasting method. Moreover, to derive the future chronic mastitis status for each prediction day, 50 days of data after the prediction day were needed (Fig. 1). Consequently, each day during lactation with 30 preceding and 50 successive days of data could be a prediction day, given that it had a recent increase in LnSCC equal to or above the 200,000 SCC/ml threshold. The need of preceding and successive data reduced the number of prediction days with SCC equal or higher than 200,000 SCC/ml from 148,172 to 93,383 prediction days with high SCC on day 0 (and increased it to 218,851 prediction days when prediction days with 200,000 SCC/ml during the previous 4 days were included).

2.5. Labelling of chronic mastitis cases

Filtering was used to determine a structural decrease in LnSCC below Ln(200) 1000 cells/ml (5.298 LnSCC). The future chronic mastitis status on a prediction day was labeled as not chronic if the rolling 20-day mean SCC decreased below Ln(200) 1000 cells/ml (0 = not chronic mastitis) at least once in the period from the prediction day to 50 days post the prediction day. The natural logarithm of SCC was used instead

Table 2
Herd descriptive statistics in terms of the number of cow lactations, the number of observations per input period.

Herd	Herd type	Cow lactations for 30-day input dataset	Prediction days for 30-day input dataset	Cow lactation for 15-day input dataset	Prediction days for 15-day input dataset
1.	Validation	261	14,806	270	15,881
2.	Validation	277	25,244	289	27,057
3.	Validation	214	14,976	228	16,586
4.	Validation	573	26,963	624	30,610
5.	Validation	137	9,204	142	9,997
6.	Validation	135	9,842	146	10,869
7.	Validation	140	8,898	144	9,472
8.	Training	371	22,907	413	24,953
9.	Training	341	13,430	360	14,774
10.	Training	297	30,105	308	32,555
11.	Training	217	12,047	237	13,523
12.	Training	201	19,485	207	20,621
13.	Training	157	10,046	165	11,028
14.	Training	36	898	46	1,254

of the untransformed SCC to make the recovery definition less sensitive to outliers and skewness. It was labeled chronic if no structural decrease occurred (1 = chronic mastitis). In other words, the label indicates whether the cow would recover (=0) or turn chronic (=1). When the SCC is consistently above Ln(200) 1000 cells/ml across the whole 50-day period in the future, it is chronic (the top example in Fig. 1), and when the SCC decreases structurally below Ln(200) 1000 cells/ml in the 50 days after a prediction day, it is not chronic (the second example from the top in Fig. 1). If a cow had an increase of SCC after a structural decrease in SCC, the cow was regarded as not chronic (the third example in Fig. 1). In these cases, it was impossible to determine whether the new increase in SCC was part of the initial episode or the start of a new episode. The 20-day rolling window was chosen to ensure that a case recovered long-term and not just for a few days. The 50 days post prediction day were chosen based on the approximate chronic cut-off of Bonestroo et al. (2020) of 4 weeks or approximately 30 days plus the rolling 20-day window (20 + 30 = 50 days). Because SCC is sampled using a risk-based sampling strategy in the OCC, SCC was not sampled every day. As such, we required at least 10 SCC measurements in the rolling 20-day window to calculate the rolling mean. If no rolling mean could be calculated at all, future chronic mastitis could not be determined, and the prediction day was discarded. In total over both the training and validation datasets, 56,817 (33,251 and 23,566 in the training and validation datasets) prediction days of the 218,851 (108,918 and 109,933 in the training and validation datasets) prediction days were identified as chronic, while 162,034 (75,667 and 86,367 in the training and validation datasets) were identified as not chronic prediction days.

2.6. Different input periods

Besides the default 30-day input period, we also pre-processed the data for a 15-day input period to evaluate the effect of different input periods on the forecasting performance. A shorter input period would allow to forecast chronic mastitis earlier in lactation and with less information. These different pre-processing steps resulted in 59,541 chronic mastitis cases and 107,702 healthy cases using a 30-day input period and 63,362 chronic mastitis cases and 118,808 healthy cases using a 15-day input period. A longer input period results in fewer cases to be forecasted as it requires more days with measured sensor data. Table 2 shows the number of cow lactations and prediction days per herd for both input period categories.

Table 3
The hyperparameter space that is explored in the random hyperparameter optimization by using 100 combinations of hyperparameters.

Hyperparameter	Distribution
Learning rate	Log uniform(0.01–1.0)
Minimum child weight	Uniform(1–20)
Maximum depth of each tree	Uniform(1–15)
Fraction of variables considered for each tree	Uniform(0.01–1.0)
Gamma	Log uniform(0.001, 0.5)
The number of trees	Uniform(50–200)

2.7. Gradient-boosting classification trees

We used the gradient-boosting trees algorithm as implemented in XGBoost (Chen and Guestrin, 2016) to create a prediction model that forecasts whether the cow would recover (=0) or turn chronic (=1), using all features in Appendix A for every input day (from the day of prediction to 29 days before the prediction). We chose gradient-boosting trees as it can deal with missing values, which can be frequent (Hogeveen et al., 2010). In addition, past work on clinical mastitis detection with boosting and bagged trees showed reasonable results (Kamphuis et al., 2010b; Kamphuis et al., 2010a) using similar sensor data. Essentially gradient-boosting trees (Friedman, 2001) use boosting with decision trees. In boosting, a combination of decision trees is made by sequentially building several decision trees from the data to minimize the classification error. In short, a first decision tree is fitted on the training data and a classification error is computed using the loss function (log loss function in our case). To minimize the classification error, the second decision tree uses the residual classification error of the first decision tree, and the third decision tree uses the residuals of the second decision tree, and so forth until the upper limit of the number of trees is reached. Each tree will give a prediction on a log(odds) scale. The final prediction of the model is the sum of the predictions of each decision tree multiplied by a pre-defined learning rate. Lastly, the final prediction in log(odds) is transformed to a prediction in probability by using a logit link function. Sequentially using the residual allows later decision trees to compensate for errors of the earlier decision trees. To address the class imbalance between the number of chronic cases and healthy cases, we set the positive class weight (scale_pos_weight in the Chen and Guestrin (2016)) to be equal to the ratio between the positive and the negative samples (i.e., chronic and not chronic cases) in the training dataset (33,251/75,667 (see Labelling of chronic cases) = 0.439 in the 30-day input period dataset and 0.427 in the 15-day input period dataset).

2.8. Hyperparameter optimization

Gradient-boosting trees have several hyperparameters (or settings) that can be optimized. An explanation of the specific hyperparameters can be found in Chen and Guestrin (2016). These hyperparameters cannot be directly estimated by the data as they have to be set before the learning process (i.e. the number of trees in a gradient-boosting trees classification model has to be set beforehand) and therefore require hyperparameter optimization. To determine the optimal hyperparameter set, we sampled 100 hyperparameter combinations from the distributions in Table 3. We used seven-fold random search cross-validation in the programming library scikit-learn (Pedregosa et al., 2011) to choose the hyperparameter combination.

In short, we separated the training dataset into seven folds based on the herd identification to ensure that every herd occupied one-fold. This separation was done to ensure that the herd-specific performance mimics the situation where the model is used in a new herd (Hogeveen et al., 2010) and it gives equal weight to each herd when determining the optimal hyperparameter combination as the average prediction performance is taken over all herd-specific AUC (see below) to calculate the final prediction score for a specific hyperparameter combination.

Table 4
The sensitivity, specificity, Matthew's correlation coefficient, accuracy, and Area under Curve (AUC) of the model predictions and the frequent and monthly sampling approaches over 7 validation herds using 30 days prior to the point of prediction as input.

Herd	Sensitivity	Specificity	Matthew's correlation coefficient	Accuracy	AUC
Model					
Herd 1	0.949	0.891	0.777	0.905	0.968
Herd 2	0.950	0.847	0.725	0.873	0.953
Herd 3	0.968	0.815	0.636	0.842	0.947
Herd 4	0.943	0.863	0.722	0.881	0.959
Herd 5	0.946	0.852	0.758	0.882	0.963
Herd 6	0.962	0.900	0.727	0.909	0.971
Herd 7	0.986	0.917	0.638	0.922	0.985
All herds	0.958	0.869	0.712	0.888	0.964
Frequent sampling approach					
Herd 1	0.894	0.873	0.710	0.878	0.946
Herd 2	0.901	0.809	0.639	0.832	0.929
Herd 3	0.965	0.728	0.538	0.769	0.930
Herd 4	0.820	0.850	0.609	0.843	0.913
Herd 5	0.916	0.773	0.645	0.818	0.934
Herd 6	0.948	0.849	0.635	0.863	0.962
Herd 7	0.904	0.935	0.635	0.933	0.974
All herds	0.907	0.831	0.630	0.848	0.941
Monthly sampling approach					
Herd 1	0.568	0.949	0.581	0.856	
Herd 2	0.520	0.928	0.502	0.825	
Herd 3	0.615	0.883	0.470	0.837	
Herd 4	0.607	0.928	0.565	0.856	
Herd 5	0.692	0.904	0.615	0.837	
Herd 6	0.614	0.944	0.572	0.896	
Herd 7	0.574	0.974	0.561	0.948	
All herds	0.599	0.930	0.552	0.865	

Next, 6 folds of data were used to train a gradient-boosting trees model with a specific set of hyperparameters that was randomly sampled from the distributions of the hyperparameter described in Table 3 (e.g., 0.02 learning rate, 5 minimum child weight, 2 maximum depth of each tree, 0.1 fraction of variables considered for each tree, 0.4 gamma, 75 number of trees). It could be any value as described by the distributions with a likelihood that is dependent on the type of distribution. To gather one hyperparameter combination, every hyperparameter distribution is sampled once. In total 100 hyperparameter combinations are sampled. This random search procedure has been proven to work well relative to a grid search (Bergstra and Bengio, 2012).

Subsequently, sample predictions in the form of prediction probabilities were made, using the unused fold as a test fold. These probabilities were compared with the label (i.e., whether the case was going to be chronic or not in the future) using the area under the curve (**AUC**) metric. The AUC was implemented using the roc_auc_score function in scikit-learn (Pedregosa et al., 2011). The roc_auc_score function calculates the area under the receiver operating characteristic (**ROC**) curve. This procedure was repeated seven times, with each fold being the test fold once for every hyperparameter combination. The average AUC over the test folds was the final score of the hyperparameter combination.

The above procedure was done for 100 randomly sampled hyperparameter combinations. We chose the optimal hyperparameter combination where the model maximized the average AUC. This procedure was solely done to attain the optimal hyperparameter set on the training dataset. We trained a model with the optimal hyperparameters using all the training data from the training herds. This model was subsequently validated using data from the validation dataset. This resulted in the following hyperparameters for the model with 30 input days: 0.0544413 learning rate, 4 minimum child weight, 9 maximum depth of each tree, 0.2383 fraction of variables considered for each tree, 0.441697 gamma, and 142 number of trees. For the 15 input days model it resulted in 0.0606968 learning rate, 1 minimum child weight, 14 maximum depth of each tree, 0.488618 fraction of variables considered for each tree, 0.0315182 gamma and 135 number of trees.

2.9. Validation

The predictions of the model take the form of class probabilities. The predictions and the labels in the validation dataset were compared using the AUC, Matthew's correlation coefficient (MCC), accuracy, sensitivity, and specificity per herd. Accuracy, sensitivity, specificity, and MCC (Eq. (1), 2, 3, 4) were calculated for predicted future chronic mastitis statuses that were created using a threshold on the predicted probability that maximizes Youden's index ($\text{sensitivity} + \text{specificity} - 1$) per herd, weighting false positives and false negatives equally for each herd. Youden's index was estimated per herd by repeatedly calculating the sensitivity and specificity of all possible probability thresholds.

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}} \quad (1)$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (2)$$

$$Sensitivity = \frac{tp}{tp + fn} \quad (3)$$

$$Specificity = \frac{tn}{tn + fp} \quad (4)$$

where tp is the number of true positives, tn is the number of true negatives, fp is the number of false positives, and fn is the number of false negatives. AUC was estimated as indicated in the hyperparameter optimization section.

The predictive performance of the gradient-boosting trees classifier was compared to that of two default approaches or simple prediction rules, the monthly sampling approach (monthly sampling approach mimicking DHI sampling frequency, but using OCC data) and frequent sampling approach (using all OCC data available). These prediction rules were also applied on the same prediction days as the prediction model to get a chronic mastitis forecast based on SCC from the past 30 days.

Table 5
The sensitivity, specificity, Matthew's correlation coefficient, accuracy, and Area under Curve (AUC) of the model predictions, and the frequent approach over 7 validation herds using 15 days prior to the point of prediction as input. It was not possible to use the monthly sampling approach using the 15 day input.

Herd	Sensitivity	Specificity	Matthew's correlation coefficient	Accuracy	AUC
Model					
Herd 1	0.961	0.867	0.750	0.890	0.964
Herd 2	0.947	0.837	0.705	0.864	0.950
Herd 3	0.944	0.816	0.614	0.837	0.934
Herd 4	0.952	0.837	0.689	0.862	0.955
Herd 5	0.932	0.855	0.745	0.879	0.957
Herd 6	0.948	0.883	0.684	0.892	0.962
Herd 7	0.981	0.900	0.597	0.905	0.981
All herds	0.952	0.856	0.684	0.875	0.958
Frequent sampling approach					
Herd 1	0.902	0.874	0.713	0.880	0.946
Herd 2	0.895	0.844	0.672	0.856	0.935
Herd 3	0.938	0.760	0.545	0.790	0.921
Herd 4	0.739	0.882	0.589	0.851	0.899
Herd 5	0.912	0.800	0.665	0.834	0.935
Herd 6	0.867	0.885	0.633	0.883	0.946
Herd 7	0.879	0.938	0.632	0.934	0.971
All herds	0.876	0.855	0.636	0.861	0.936

- Monthly sampling approach: To mimic a monthly sampling frequency, the default approach used 2 SCC measurements in the preceding 30 days. This approach predicted future chronic mastitis to be present when the SCC was higher than 200,000 SCC/ml in the SCC sample before but closest in time to the prediction day (see Fig. 1) and the SCC sample furthest away in the preceding 30 days relative to the prediction day. If both SCC samples were equal to or higher than 200,000 SCC/ml, the prediction rule predicted chronic mastitis. Otherwise, no chronic mastitis was predicted. The monthly sampling approach mimics a situation where farmers use monthly SCC data of the previous month that is compared to the current month to determine chronic mastitis.
- Frequent sampling approach, this approach predicted future chronic mastitis when the number of days with 200,000 SCC/ml or higher prior to the prediction day was equal to or more than 13 days in the 30-day input period category (7 days in the 15-day input period category). Otherwise, no chronic mastitis was predicted. This threshold on the number of days was chosen to maximize Youden's index to forecast the future chronic mastitis status.

Comparing different approaches based on different metrics allowed us to determine whether the increase in predictive performance was due to more complex models or more frequent SCC samples. The AUC could not be computed for the monthly sampling approach as this approach results in a class prediction and not a continuous value. The differences in AUC, MCC, and accuracy between the model predictions and the default approaches were tested using Bonferroni-Holm-corrected paired t-tests on herd-specific performance measures (for 8 tests). Accuracy, MCC, and AUC were selected for the statistical tests. This decision was made as they take all classified cases into account, while specificity (no tp or fn) and sensitivity (no tn or fp) do not.

3. Results

3.1. Using the previous 30 days as input to predict future chronic mastitis

A chronic mastitis forecasting model was trained and validated. Given automatically collected sensor data, it would allow the farmer to gain insight into the probable end of a case and to use sensor data to this end in a structural manner. Table 4 presents the sensitivity, specificity, MCC, accuracy, and AUC of the model predictions, frequent sampling approach, and monthly sampling approach. Overall, all performance measures varied slightly between the different validation herds. Nevertheless, the chronic mastitis prediction model outperformed the two approaches on all farms for almost all performance indicators. It can forecast chronic cases more accurately on almost all performance measures. More specifically, on accuracy (chronic mastitis prediction model: 0.888, frequent sampling approach: 0.848, and monthly sampling approach: 0.865), MCC (chronic mastitis prediction model: 0.712, frequent sampling approach: 0.630, and monthly sampling approach: 0.552), and AUC metrics (chronic mastitis prediction model: 0.964 and frequent sampling approach: 0.941) using 30 days of sensor data prior to the day of prediction. Using Bonferroni-Holm corrected paired t-tests, the differences between approaches and the model predictions were significant for AUC, accuracy, and MCC ($P < 0.05$) but not for accuracy ($P > 0.05$) when compared to the monthly sampling approach . The chronic mastitis prediction model also outperformed in terms of sensitivity (chronic mastitis prediction model: 0.958, frequent sampling approach: 0.907, and monthly sampling approach: 0.599) but the monthly sampling approach outperformed the other methods on specificity (chronic mastitis prediction model: 0.869, frequent sampling approach: 0.831, and monthly sampling approach: 0.930).

4. Using the previous 15 days as input to predict future chronic mastitis

Table 5 provides the sensitivity, specificity, accuracy, MCC, and AUC of the model predictions and frequent sampling approach. In this case, the monthly sampling approach could not be applied, as it requires at least 30 input days as it mimicked monthly DHI tests. In this case, the model outperformed the approach in sensitivity, MCC, accuracy and AUC but not in specificity. More specifically, the 15-day model outperformed the approach on all farms using the accuracy (prediction model: 0.875 and frequent sampling approach: 0.861), MCC (prediction model: 0.684 and frequent sampling approach: 0.636), and AUC metrics (prediction model: 0.958 and frequent sampling approach: 0.936). Using Bonferroni-Holm corrected paired t-tests, the differences between approaches and the model predictions were significant for AUC ($P < 0.05$), but not for MCC or accuracy ($P > 0.05$). Comparing the result with differing input periods, there does not seem to be a major difference between them. The decrease in performance, due to decreasing the input period from 30 days to 15 days, was limited.

5. Discussion

This is the first study that uses on-farm sensor data to predict future

chronic mastitis. We developed a prediction model and compared the performance to the monthly sampling approach and the frequent sampling approach. The significantly higher performance of the model compared to the performance of the approaches showed the potential value of future chronic mastitis prediction based on sensor data. The results show that this model would have value for farmers in forecasting chronicity because the approaches emulate how sensor data would be used without a sophisticated prediction model. This point is strengthened by the fact that farmers may not need to invest in extra sensor technology to gather these forecasts. Limited research has been published on future chronic mastitis forecasting. Bartel et al. (2019) created two chronic mastitis prediction models for healthy cows and unhealthy cows, respectively, using non-sensor DHI data and generalized additive models. They reported an AUC of 0.779 and 0.868 for the unhealthy cows and the healthy cow models, respectively. Although we built only one model to classify healthy and chronic cows, our reported AUC was larger. Nevertheless, these studies cannot directly be compared as they used a different future chronic mastitis definition using diverse types of data.

To keep the comparison between approaches fair, we have chosen an equal weighting between misclassification types by optimizing Youden's index. Preferring specificity to sensitivity would not result be a fair comparison to the monthly sampling approach. The consequences of misclassification differ between false positives and false negatives in the chronic mastitis forecasting. Chronic cases classified as not chronic have more time in the herd while more data is gathered that could lead to a correct prediction in the end, but this cow may infect other cows in this period. On the other side, a misclassification of a not chronic case as a chronic case leads to culling which can be costly. In that case, the farmer incurs unnecessary culling costs and unnecessary culling will likely lead to an unnecessary loss of life. One may argue that unnecessary culling is more costly than keeping a chronic cow in the herd for a longer period or vice versa and make the prediction algorithm cost sensitive to either class. However, our primary aim was to compare the model to the approaches and the monthly sampling approach does not allow us to adapt it and make it cost-sensitive apart from changing the prediction rule itself (e.g., from using the last 2 SCC measurements to the last 3 SCC measurements).

Several limitations constrained the study and its results. We based our future chronic mastitis definition on a long-term increased SCC without a period where SCC was below 200,000 cells/ml. Chronic mastitis itself is not well-defined in the literature. As SCC is a primary indicator for inflammation (International Dairy Federation, 2011) and DHI SCC has been used to indicate chronic mastitis in the past (St. Rose et al., 2003), we would argue that using SCC to operationalize chronic mastitis fits well. However, one could also have used conductivity to define chronic mastitis, but the conductivity thresholds of healthy versus sick cows are less well defined and accepted than SCC (Smith et al., 2001; International Dairy Federation, 2013), although work has been done to find thresholds for conductivity (Khatun et al., 2017). For the label, we recorded 50 days of SCC measurements after the day of prediction and determined whether, within the 50 days, there was a 20-day period where the mean SCC was lower than 200,000 SCC/ml. Another limitation was that we required more than 10 non-missing observations in the calculation of a rolling mean. Whether a value is missing may also be dependent on the sampling frequency. The sampling frequency is based on the mastitis risk assessment on the OCC sampling algorithm. This may cause structural missing values as it depends on the decision of the sampling algorithm and may bias the labels to be definable when the cow is indeed chronic and indefinable when a cow is not. Furthermore, it should be emphasized that the limitations do not make the model invalid from a practical perspective as farmers would detect chronic cases that they would not have detected (or detected later) without any additional cost.

We have made several choices concerning the model choice as well as in training the model. In the hyperparameter optimization, we used random search in combination with herd-based cross-validation, and gradient-boosting trees on the training dataset. We used gradient-boosting trees as it tended to work well with tabular AMS sensor data (Kamphuis et al., 2010b) and natively supports missing values. Other models might perform better than gradient-boosting trees, but the aim of this study was to investigate the possibility of developing a future chronic mastitis prediction model and not to find the best performing model. In hyperparameter optimization, we could not do an exhaustive search for hyperparameters but the results in this paper still show the added value of a future chronic mastitis prediction model. We also used a herd-based split between training and validation datasets to avoid that the model learned herd-specific characteristics that might increase the predictive performance. This was done to mimic the performance of the algorithm when it is placed on a new farm. Different farm management strategies or pathogen populations might cause herd-specific associations. The performance of such a validated model might then be disappointing when the model is deployed on a new farm. Therefore, a more conservative herd-based cross-validation strategy should be preferred when testing on-farm detection or prediction models. However in a practical application, the current proposed model can be extended to be herd-specific by using part of the herd data to retrain the model partly (i.e., by applying transfer learning). Although interesting, this was outside the scope of this paper and hence was not performed.

We have trained the chronic mastitis prediction model using 15 and 30 days prior to the day of prediction as input. The predictive performance decreased when the number of input days decreased, however the differences were small (e.g. 0.964 to 0.958 herd-average AUC). This limited decrease in predictive performance with decreasing input period indicates that it might be possible to predict future chronic mastitis with only a small number of input days prior to the day of prediction. If a small number of input days are required, it becomes possible to predict future chronic mastitis early in lactation. Predicting chronic mastitis early in lactation is valuable as mastitis is most prevalent in early lactation (Nyman et al., 2007) and predicting which case may turn into a chronic case, with or without treatment, would be useful in decision making. Future research could investigate future possibilities of early lactation chronic mastitis forecasting.

6. Conclusion

We have developed a future chronic mastitis prediction model based on sensor data, which outperformed simple prediction rules that mimic current decision making based on monthly or more frequently sampled SCC data in predictive performance. Decreasing the input period from 30 to 15 days had only a limited effect on the predictive performance of the model. An accurate prediction of future chronic mastitis could indicate farmers of potentially chronic cows in the future, resulting in earlier and potentially more beneficial interventions such as treatment and more targeted culling. In the end, this supports sensor driven decision making towards less chronic cows.

CRediT authorship contribution statement

John Bonestroo: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Investigation, Validation, Writing – review & editing, Project administration. **Mariska van der Voort:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Investigation, Validation, Writing – review & editing, Project administration, Supervision. **Henk Hogeveen:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Investigation, Validation, Writing – review & editing, Project administration, Funding acquisition, Supervision. **Ulf Emanuelson:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Investigation, Validation, Writing – review & editing, Project administration, Funding acquisition, Supervision. **Ilka Christine Klaas:** Conceptualization, Methodology,

Table A1
Overview of the features and their definitions as used in the study.

Feature name	Explanation	Day mean	Std. dev.
DIM	The DIM of the day.	149.19	103.37
MeanYield	The mean of the milk yields from different milkings in kg on a day.	13.30	4.40
MaxYield	The maximum of the milk yields in kg from different milkings on a day.	15.33	4.98
MinYield	The minimum of the milk yields in kg from different milkings on a day.	11.37	4.35
STDYield	The standard deviation of the milk yields in kg from different milkings on a day.	2.40	1.84
TotalYield	The sum of the milk yields in kg from different milkings on a day	34.54	12.24
MeanIQRConductivity	The mean of the ratio between the quarter with the highest conductivity and the lowest conductivity for a milking over all milkings in mS/cm on a day.	1.08	0.08
MaxIQRConductivity	The maximum of the ratio between the quarter with the highest conductivity and the lowest conductivity for a milking in mS/cm over all milkings on a day.	1.10	0.10
MinIQRConductivity	The minimum of the ratio between the quarter with the highest conductivity and the lowest conductivity for a milking in mS/cm over all milkings on a day.	1.06	0.07
STDIQRConductivity	The standard deviation of the ratio between the quarter with the highest conductivity and the lowest conductivity for a milking in mS/cm over all milkings on a day.	0.03	0.04
MeanSTDConductivity	The mean of the standard deviation between the mean conductivities in mS/cm measured between the four quarters of all milkings on a certain day.	0.16	0.16
MaxSTDConductivity	The maximum of the standard deviation between the mean conductivities in mS/cm measured between the four quarters of all milkings on a certain day.	0.20	0.19
MinSTDConductivity	The minimum of the standard deviation between the mean conductivities in mS/cm measured between the four quarters of all milkings on a certain day.	0.12	0.14
STDSTDConductivity	The standard deviation over the standard deviation between the mean conductivities measured in mS/cm between the four quarters of all milkings on a certain day.	0.05	0.07
MeanTimeInterval	The mean time between milkings in hours on a day.	9.53	2.68
MaxTimeInterval	The maximum time between milkings in hours on a day.	11.00	2.95
MinTimeInterval	The minimum time between milkings in hours on a day.	8.16	2.80
STDTimeInterval	The standard deviation time between milkings in hours on a day.	0.12	0.14
MeanMilkRate	The mean milk production in kilograms per hour on a day.	1.46	0.45
MaxMilkRate	The maximum milk production in kilograms per hour on a day.	1.56	0.50
MinMilkRate	The minimum milk production in kilograms per hour on a day.	1.36	0.43
STDMilkRate	The standard deviation of milk production in kilograms per hour on a day.	0.12	0.14
LnSCC	The natural logarithm of the maximum SCC in 1000 cells/ml on a day	4.50	1.29
MeanSCC	The mean SCC in 1000 cells/ml on a day.	178.00	378.08
MaxSCC	The maximum SCC in 1000 cells/ml on a day.	223.41	471.33
MinSCC	The minimum SCC in 1000 cells/ml on a day.	138.97	324.94
STDSCC	The standard deviation of SCC in 1000 cells/ml on a day.	88.72	213.06
MeanBlood	The share of milkings that had a detection of blood in the milk on a day.	0.04	0.17
MaxBlood	The maximum of blood detections in the milk on a day (whether there was any blood in the milkings on a given day).	0.06	0.23
MinBlood	The minimum of blood detections in the milk on a day (whether there was any no blood milking on a given day).	0.02	0.15
STDBlood	The standard deviation of milkings that had a detection of blood in the milk on a day.	0.02	0.12
Treatment duration	A number indicating the start of a treatment where milk was diverted for several days in the future.	0.04	0.59
Parity	The parity of the cow at the time of milking.	2.50	1.53
MeanOverallConductivity	The mean of the daily mean quarter conductivities in mS/cm on a day.	4.63	0.43
MaxOverallConductivity	The maximum of the daily maximum quarter conductivities on a day.	4.96	0.59
STDOverallConductivity	The standard deviation of the daily standard deviations of quarter conductivities in mS/cm on a day.	0.07	0.08
STDMaxOverallConductivity	The standard deviation of the daily maximum quarter conductivities in mS/cm on a day.	0.16	0.19

Software, Formal analysis, Data curation, Visualization, Investigation, Validation, Writing – review & editing, Project administration, Funding acquisition, Supervision. **Nils Fall:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Investigation, Validation, Writing – review & editing, Project administration, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was funded by an Industry PhD-program of the Swedish Government, reference number N2017/036895/SK and DeLaval International AB (Tumba, Sweden).

Potential conflicts of interest

John Bonestroo and Ilka C. Klaas are employed by DeLaval International AB. Mariska van der Voort, Nils Fall, Ulf Emanuelson, and Henk Hogeveen have no conflict of interest to report.

Appendix A. Overview of the features as used in the prediction model in this study

See Table A1.

References

Aghamohammadi, M., Haine, D., Kelton, D.F., Barkema, H.W., Hogeveen, H., Keefe, G.P., Dufour, S., 2018. Herd-level mastitis-associated costs on Canadian dairy farms. Front. Vet. Sci. 5, 100.

Anglart, D., Hallén-Sandgren, C., Emanuelson, U., Rönnegård, L., 2020. Comparison of methods for predicting cow composite somatic cell counts. J. Dairy Sci. 103, 8433–8442.

Bartel, A., E. Gass, F. Onken, C. Baumgartner, F. Querengässer, and M.G. Doherr. 2019. SCC predictions using generalized additive models: can they support mastitis management decisions? Page 24 in IDF mastitis Conference 2019, Copenhagen.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.

Bonestroo, J., van der Voort, M., Fall, N., Hogeveen, H., Emanuelson, U., Klaas, I.C., 2021. Progression of different udder inflammation indicators and their episode length after onset of inflammation using automatic milking system sensor data. J. Dairy Sci. 104, 3457–3473.

Caja, G., Castro-Costa, A., Knight, C.H., 2016. Engineering to support wellbeing of dairy animals. J. Dairy Res. 83, 136–147.

Chen, T., and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. Pages 785–794 in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Gonçalves, J.L., Kamphuis, C., Vernooij, H., Araújo Jr, J.P., Grenfell, R.C., Juliano, L., Anderson, K.L., Hogeveen, H., Dos Santos, M.V., 2020. Pathogen effects on milk yield and composition in chronic subclinical mastitis in dairy cows. Vet. J. 262, 105473.

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

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

International Dairy Federation. 2011. Suggested Interpretation of Mastitis Terminology (revision of Bulletin of IDF N° 338/1999). Brussels.

International Dairy Federation. 2013. Guidelines for the use and interpretation of bovine milk somatic cell counts (SCC) in the dairy industry. Brussels.

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, 7344–7361.

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

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

Khatun, M., Clark, C.E.F., Lyons, N.A., Thomson, P.C., Kerrisk, K.L., García, S.C., 2017. Early detection of clinical mastitis from electrical conductivity data in an automatic milking system. Anim. Prod. Sci. 57, 1226–1232.

Khatun, M., Thomson, P.C., Clark, C.E.F., García, S.C., 2019. Prediction of quarter level subclinical mastitis by combining in-line and on-animal sensor data. Anim. Prod. Sci. 60, 180–186.

Krömker, V., Leimbach, S., 2017. Mastitis treatment—Reduction in antibiotic usage in dairy cows. Reprod. Domest. Anim. 52, 21–29.

Nørstebø, H., Dalen, G., Rachah, A., Heringstad, B., Whist, A.C., Nødtvedt, A., Reksen, O., 2019. Factors associated with milking-to-milking variability in somatic cell counts from healthy cows in an automatic milking system. Prev. Vet. Med. 172, 104786.

Nyman, A.-K., Ekman, T., Emanuelson, U., Gustafsson, A.H., Holtenius, K., Waller, K.P., Sandgren, C.H., 2007. Risk factors associated with the incidence of veterinary-treated clinical mastitis in Swedish dairy herds with a high milk yield and a low prevalence of subclinical mastitis. Prev. Vet. Med. 78, 142–160.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Polat, B., Colak, A., Cengiz, M., Yanmaz, L.E., Oral, H., Bastan, A., Kaya, S., Hayirli, A., 2010. Sensitivity and specificity of infrared thermography in detection of subclinical mastitis in dairy cows. J. Dairy Sci. 93, 3525–3532.

St. Rose, S.G.S., J.M. Swinkels, W.D.J. Kremer, C.L.J.J. Kruitwagen, and R.N. Zadoks. 2003. Effect of penethamate hydriodide treatment on bacteriological cure, somatic cell count and milk production of cows and quarters with chronic subclinical *Streptococcus uberis* or *Streptococcus dysgalactiae* infection. J. Dairy Res. 70:387–394.

Rutten, C.J., Velthuis, A.G.J., Steeneveld, W., Hogeveen, H., 2013. Invited review: Sensors to support health management on dairy farms. J. Dairy Sci. 96, 1928–1952.

Smith, K.L., Hillerton, J.E., Harmon, R.J., 2001. Guidelines on normal and abnormal raw milk based on somatic cell counts and signs of clinical mastitis. National Mastitis Council, Madison, Wisconsin.

Sørensen, L.P., Bjerring, M., Løvendahl, P., 2016. Monitoring individual cow udder health in automated milking systems using online somatic cell counts. J. Dairy Sci. 99, 608–620.

Swinkels, J.M., Hogeveen, H., Zadoks, R.N., 2005. A partial budget model to estimate economic benefits of lactational treatment of subclinical *Staphylococcus aureus* mastitis. J. Dairy Sci. 88, 4273–4287.