J. Dairy Sci. 104
https://doi.org/10.3168/jds.2021-20517

Detecting and predicting changes in milk homogeneity using data from automatic milking systems

D. Anglart,[1,2]* U. Emanuelson,[2] L. Rönnegård,[3,4] and C. Hallén Sandgren[1]
[1]DeLaval International AB, PO Box 39, SE-147 21 Tumba, Sweden
[2]Swedish University of Agricultural Sciences, Department of Clinical Sciences, PO Box 7054, SE-750 07 Uppsala, Sweden
[3]School of Technology and Business Studies, Dalarna University, SE-791 88 Falun, Sweden
[4]Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, PO Box 7023, SE-750 07 Uppsala, Sweden

ABSTRACT

To ensure milk quality and detect cows with signs of mastitis, visual inspection of milk by prestripping quarters before milking is recommended in many countries. An objective method to find milk changed in homogeneity (i.e., with clots) is to use commercially available inline filters to inspect the milk. Due to the required manual labor, this method is not applicable in automatic milking systems (AMS). We investigated the possibility of detecting and predicting changes in milk homogeneity using data generated by AMS. In total, 21,335 quarter-level milk inspections were performed on 5,424 milkings of 624 unique cows on 4 farms by applying visual inspection of inline filters that assembled clots from the separate quarters during milking. Images of the filters with clots were scored for density, resulting in 892 observations with signs of clots for analysis (77% traces or mild cases, 15% moderate cases, and 8% heavy cases). The quarter density scores were combined into 1 score indicating the presence of clots during a single cow milking and into 2 scores summarizing the density scores in cow milkings during a 30-h sampling period. Data generated from the AMS, such as milk yield, milk flow, conductivity, and online somatic cell counts, were used as input to 4 multilayer perceptron models to detect or predict single milkings with clots and to detect milking periods with clots. All models resulted in high specificity (98–100%), showing that the models correctly classified cow milkings or cow milking periods with no clots observed. The ability to successfully classify cow milkings or cow periods with observed clots had a low sensitivity. The highest sensitivity (26%) was obtained by the model that detected clots in a single milking. The prevalence of clots in the data was low (2.4%), which was reflected in the results. The positive predictive value depends on the prevalence and was relatively high, with the highest positive predictive value (72%) reached in the model that detected clots during the 30-h sampling periods. The misclassification rate for cow milkings that included higher-density scores was lower, indicating that the models that detected or predicted clots in a single milking could better distinguish the heavier cases of clots. Using data from AMS to detect and predict changes in milk homogeneity seems to be possible, although the prediction performance for the definitions of clots used in this study was poor.

Key words: dairy cow, clinical mastitis, clot, multilayer perceptron

INTRODUCTION

Milk intended for human consumption should be produced by healthy animals and be of acceptable quality, a condition that, under dairy conditions, is the responsibility of the farmer. Milk that is deviant in terms of color, smell, or homogeneity (i.e., abnormal milk) is not suitable for sale (Food and Drug Administration, 2017). To monitor the hygienic quality of the milk, prestripping before milking the cows and inspecting the foremilk for abnormalities is recommended (European Commission, 2004), in addition to monitoring the bulk tank SCC and bacterial count. Although not all abnormal milk necessarily originates from udders undergoing an inflammation (Rasmussen and Bach Larsen, 2003), changes in milk homogeneity such as clots, thick milk, or watery milk are generally established signs of clinical mastitis (**CM**; Giesecke and van den Heever, 1974; IDF, 2011). Therefore, milk inspection is a dual-purpose task, not only to ensure milk quality but also to identify cows with signs of illness. In a conventional milking parlor, milk inspection is commonly done by the milker when preparing the udder for milking, although implementation on farms is commonly less than 100% (Rodrigues et al., 2005; Wenz et al., 2007; Nielsen and Emanuelson, 2013). As no milker is present during milking in automatic milking systems (**AMS**), the

Received March 25, 2021.
Accepted May 20, 2021.
*Corresponding author: dorota.anglart@delaval.com

farmer needs to be supported by sensors that detect deviations and allow the system to issue alerts before milking potentially sick cows in order to prevent milk of unacceptable quality from ending up in the bulk tank.

Visual inspection for changes in milk homogeneity using commercially available inline filters has been suggested as a universal and objective method to define CM (Rasmussen, 2005; Claycomb et al., 2009; Kamphuis et al., 2013). The outcome of such inspection may be useful in identifying CM, but not in preventing milk changed in homogeneity from ending up in the bulk tank, because the filters are inspected when a milking has ended. Furthermore, the method requires manual labor and is thus not suitable for use in AMS. Hence, accurate predictions of milk homogeneity, for instance, using data generated by AMS as input to predictive models, would benefit farmers. To the best of our knowledge, no attempts have been made to generate such predictions.

The objective of this study was to detect and predict visual changes in milk homogeneity (i.e., clots) using data regularly recorded by AMS. The multilayer perceptron (**MLP**), a classic feed-forward artificial neural network (**ANN**), was used. Artificial neural networks have potential to capture nonlinear relationships and interactions between predictors in a flexible manner and have previously been suggested for CM detection (e.g., Nielen et al., 1995; Sun et al., 2010; Ankinakatte et al., 2013), pathogen prediction (Heald et al., 2000; Hassan et al., 2009), and SCC prediction (Anglart et al., 2020). Similar to a milker, who uses faculties such as taste, smell, vision, and memory (Hillerton, 2000) to decide whether or not the milk should be discarded, ANN are designed to process information in a similar way, basing decisions on detected patterns and relationships in data and learning from them (e.g., Agatonovic-Kustrin and Beresford, 2000; Haykin, 2009).

MATERIALS AND METHODS

Data Collection

Data were collected at 4 commercial dairy farms located in Sweden (farms A and B) and in the Netherlands (farms C and D). The cows were milked using a total of 10 voluntary milking systems (DeLaval International AB). Farms were selected based on the availability of sensor equipment in addition to that available in the AMS, and all farms were equipped with online cell counters (**OCC**; DeLaval International AB). Data were collected between March 2017 and April 2018.

The cows involved in the data collection were mainly Holstein-Friesian. Data were collected for 2 mo at farms A and B and 3 mo at farms C and D. Each farm was visited on 3 occasions. On each visit, visual milk inspections (**MI**) of all cow quarters milked during 30 consecutive hours were performed. The 30 consecutive hours of MI are henceforth referred to as "periods." All cows milked in the AMS during the periods participated in the MI. The number of MI per cow thus depended on the number of AMS visits made by each cow. Table 1 summarizes details regarding data collection, number of AMS, number of cows and breed, parity, DIM, yearly milk production, and SCC at each farm.

The MI and subsequent scoring of the outcome was performed accordingly: a meshed filter, the mastitis detector (Vision 16 MastitisDetector; Ambic Equipment Ltd.), was used to obtain a representative sample of clots from each quarter during milking. Each filter was visually inspected for clots, gently rinsed with water to remove milk and foam, and filters with signs of clots were photographed. The MI were performed by the first author, who collected samples at all 4 farms and was responsible for training 2 support persons, one at farms A and B and another at farms C and D.

Table 1. Characteristics of each farm during data collection periods

Item	Farm A	Farm B	Farm C	Farm D
Start of data collection periods	March 9, 2017 March 22, 2017 April 11, 2017	August 15, 2017 September 5, 2017 September 27, 2017	February 1, 2018 March 24, 2018 April 4, 2018	February 3, 2018 March 26, 2018 April 6, 2018
Number of automatic milking systems	2	3	2	3
Number of cows	145	175	113	191
Number of cows in parity				
1	43	61	31	57
2	26	52	17	67
≥3	76	62	65	67
Milk production (kg/cow per year)	12,672	11,619	10,074	10,106
SCC (cells/mL)	285,000[1]	184,000[2]	236,000[1]	106,000[1]
Average number of milkings (cow/period)	3.0	3.2	3.8	3.3

[1]Arithmetic mean from monthly dairy herd improvement sampling.

[2]Bulk tank arithmetic mean.

Table 2. Definitions of scores and corresponding proportions of area covered with clots

Score	Defined as	Aggregate area of deposits on the filter
0	No signs	None
1	Trace	Diameter <3 mm
2	Mild case	Diameter ≥3 mm
3	Moderate case	Diameter ≥5 mm or approximately 10% covered
4	Heavy case	Between 10 and 50% covered
5	Very heavy case	More than 50% covered

The images showing signs of clots were scored when data collection was completed at all farms. The scoring scale provided by the filter manufacturer was modified to accommodate clot scoring at the quarter level. The scale ranged between 0 and 5, with 0 being defined as no signs of clots, 1 as trace, 2 as mild, 3 as moderate, 4 as heavy, and 5 as very heavy density of clots. An overview of the scores and definitions can be found in Table 2. Three assessors (2 veterinarians and an animal scientist) not involved in the data collection scored each image individually. A score was set as the quarter milking score (**QMS**) if at least 2 of the 3 assessors were in agreement; otherwise, the MI was removed from the data set. To assess scorer agreement, the Fleiss kappa (Fleiss, 1971) with 3 raters was computed using the "irr" package in R (Gamer et al., 2019).

Data Preparation

Predictor Variables. Data used in the analyses were extracted from the herd management system (DelPro, DeLaval International AB) of each farm during the 3 periods and covered the 30 h of each period and the 48 h before each period. The data contained cow-level information comprising AMS number, cow number, breed, parity, DIM, OCC value (cells/mL), mastitis detection index (unitless), milking duration (s), and the date and time of milking. The milking interval was calculated as the number of minutes between AMS milking visits. Quarter-level data comprised milk yield (kg), average milk flow (g/min), peak milk flow (g/min), electrical conductivity (mS/cm), expected milk production speed (kg/h), expected milk yield (kg), difference between expected and actual milk yield (kg), blood (mg/kg), attachment time (s), quarters set not to milk by farmer (yes or no), cups kicked off during milking (yes or no), and unmilked quarters reported by the system (yes or no). All variables except the AMS number and date and time of milking were used in model development.

Each cow received a farm-specific cow number, and the data sets from all farms were merged. Predictor variables corresponding to quarters set by the farmer as "not to be milked" in the AMS were considered faulty and removed. Numerical explanatory variables were normalized, with a mean value of 0 and a standard deviation of 1. Missing values were handled as follows: numerical variables with missing values were set to 0 (i.e., mean value), as suggested by Chollet (2017). Categorical variables with missing values were assigned an additional level indicating the missing value. Because OCC was considered a main predictor, the 17% missing values of OCC were imputed using random forest imputation (Stekhoven, 2013). The supportive variables for the imputation (i.e., AMS number, farm, cow number, parity, DIM, breed, composite udder milk yield, milking interval, and the date and time of the milking) were added to the imputation model. The OCC values were log-transformed. Factor explanatory variables such as cow number, parity, and breed were converted to dummy variables. Data from 3 milkings before the MI were used to create past-period variables (lags) for all predictor variables except cow number, DIM, parity, breed, and farm. Milking 0 was the milking of the MI, milking −1 was 1 milking before the MI, and so on.

Response Variables. To investigate the performance of models that detected single milkings containing clots as well as milkings with clots during a longer period (i.e., the 30-h periods), QMS were combined into 2 types of binary outcomes at the cow composite level as follows: cow milk class (**CMC**) and cow period class (**CPC**). The CMC was computed for each cow at each milking and was equal to 1 if any QMS $\geq$ 2; otherwise, it was equal to 0. The CPC was computed for each cow period by summing all QMS (1–5) for the cow, dividing the sum by number of quarters, and dichotomizing the resultant by setting a threshold such that periods when no quarter received a QMS $\geq$ 3 or periods when no quarter received a QMS $\geq$ 4, respectively, were set to 0, and thus excluded from the positive category, and all others were set to 1; these were labeled CPC.3 and CPC.4, respectively. Thus, each cow obtained 1 CMC for each milking and 2 CPC for each period, with a value of 0 corresponding to a negative outcome and a value of 1 to a positive outcome. Models with CMC as the response variable are henceforth referred to as CMC models, whereas models with CPC as the response variable are henceforth referred to as CPC models.

Test and Training Data. The data were divided into 70% training and 30% testing data using random sampling. A seed was used to obtain comparable results (i.e., testing and training data were the same for all models).

Creating the Model

Two CMC model variations were created as follows: CMC.D containing data from the same milking as the MI (i.e., a detection model) in addition to the past-period variables, and CMC.P excluding data from the same milking as the MI (i.e., a prediction model). For the response variables CPC.3 and CPC.4, predictors from the first MI of the period and from the 3 milkings before were included in the models.

The algorithm used in this study was the MLP, a classic feed-forward ANN. The implementation in Keras for R (Chollet, 2017) was used with the model sequential option. The MLP was constructed with 1 hidden layer using the default activation function [i.e., the rectified linear activation function (relu)]. Because the task was to identify milk samples with clots, the model was customized for a binary classification problem. Thus, the output layer was constructed with 2 units using the activation function softmax, an activation function that normalized the model output into a probability distribution. Binary accuracy was chosen as the metric, calculating the frequency of how often the predicted values equaled the actual values.

For the configuration of the learning process, ADAM (Kingma and Ba, 2015) was chosen as optimizer because this stochastic optimization method works well with little tuning of the hyperparameters. To prevent overfitting of the model, the weight regularization kernel regularizer l2 was used. Dropout between the layers was not used because it had a negative effect on detection performance and did not improve accuracy or loss.

Tuning the Hyperparameters

The number of units in the hidden layer was determined by running several CMC.D models, with 5 to 500 layers, that evaluated the accuracy and loss of each model. The CMC models were fitted with 10 epochs (the default number of times for full-forward and backward propagation) with a batch size of 32, and the validation split set to 0.2 (80% training, 20% testing), because these settings further lowered loss. Setting the regularizer option to 0.005 minimized the difference between validation loss and training loss. The parameters were again tuned for the CPC models, resulting in changing the number of epochs to 20 and setting the regularizer level to 0.05. Details regarding the settings of the models are summarized in Table 3.

Table 3. Model settings for models with the response variable setup cow milk class (CMC) and cow period class (CPC)

Item	CMC model	CPC model
Layer	1	1
Units	50	50
Regularizer	0.005	0.05
Epochs	10	20
Batch size	32	32
Validation split	0.2	0.2

Each model variation was run 10 times on the training data set. The performance of each of the 10 model runs was evaluated on the test data by comparing the predicted values of clots with the observed cases of clots (i.e., CMC and CPC) and by calculating the sensitivity (**Se**), specificity (**Sp**), positive predictive value (**PPV**), and negative predictive value (**NPV**). The results are presented for 1 representative run of each model in the evaluation (i.e., the result closest to the median Se and Sp over the 10 runs). To investigate the classification rate, a confusion matrix was created for each model. The predictions made by the CMC models were compared with the highest observed QMS within the same cow milking to obtain the misclassification rate of the QMS included in the CMC. All statistical procedures were carried out using R (http://www.r-project.org).

RESULTS

Number of MI and Scoring

In total, 21,335 MI were performed on 5,424 milkings of 624 unique cows. The number of samples was 932 from 303 unique cows (Table 4), distributed according to 156 out of 553 cows in period 1, 149 out of 557 cows in period 2, and 138 out of 546 cows in period 3 having QMS ≥ 1. Of the collected samples, 30 quarter milkings were discarded due to failed sampling, unknown cow number, or missing image; therefore, 902 images were available for scoring. The scorer agreement based on 898 images (4 images received scores from only 2 scorers) was 0.72, indicating substantial agreement between the scorers (Landis and Koch, 1977).

The result of the scoring was that 379 images received a score of 1, 303 received a score of 2, 135 received a score of 3, 67 received a score of 4, and 8 received a score of 5 (Table 5). Seven images received a score of 0 (i.e., the corresponding quarters were considered to not have traces or clots) and were further considered quarters without clots. Three images were discarded due to scorer disagreement. Thus, 892 quarters with

Table 4. Number of milk inspections at the cow and quarter levels, images scored, and occurrence of clots or traces on the filter after each cow milking during all periods at all farms

Item	Total
Cow level	
Milk inspections	5,424
Unique cows inspected	624
Unique cows with clots or traces	321
Unique cows without clots or traces	303
Quarter level	
Milk inspections	21,335
Filters with clots or traces	932
Filters without clots or traces	20,403
Removed[1]	30
Images of filters with clots or traces	
Scorer disagreement	3
Scored as no clots or traces	7
Used in analysis	892

[1]Due to failed sample, unknown cow number, or missing image.

scores > 0 were available for analysis. Traces (QMS = 1) constituted 43%, mild cases constituted 34%, moderate cases constituted 15%, and heavy cases (QMS ≥ 4) constituted 8% of the 892 quarters. The prevalence of clots (QMS ≥ 2) in the total data set from all periods, including quarters without traces or clots, was 2.4%. The prevalence of CMC was 7%, CPC.3 was 16%, and CPC.4 was 7%.

Cow Milk Class

The results of all models are summarized in Table 6. The results of the 2 CMC models (i.e., for the detection and prediction of clots in a single milking) were very similar. The Se was 0.25 for the CMC.P model and 0.26 for the CMC.D model. The Sp was equally high for both models (Sp = 0.98). The PPV results were also similar (i.e., 0.53 for the CMC.D model and 0.47 for the CMC.P model), and the NPV was 0.95 for both model variations. Details on the 10 model runs that form the basis of the median values can be found in Supplemental Tables S1 and S2 (https://urn.kb.se/resolve?urn=urn:nbn:se:slu:epsilon-p-112047).

Table 5. Results of scoring of images with clots on filters for each farm

	Number of cases per score on each farm				
Score	A	B	C	D	Total
1	118	75	89	97	379
2	89	48	78	88	303
3	40	28	36	31	135
4	15	13	28	11	67
5	0	3	2	3	8

The test data set consisted of 1,610 cow milkings. The number of CMC classified as positive in the test data set was 112. The detection model, CMC.D, correctly classified 29 out of 112 positive cow milkings in the test data (i.e., as cow milkings when clots were observed in milk from at least 1 quarter). Most of the negative cow milkings in the test data (i.e., 1,472 out of 1,498) were correctly classified as milkings when no clots were observed in any quarter. The results of the prediction model, CMC.P, were similar (i.e., 28 out of 112 cow milkings in the test data were correctly predicted as cow milkings) when clots in at least 1 quarter were observed, and 1,467 out of 1,498 cow milkings in the test data were correctly predicted as cow milkings when no clots were observed in any quarter.

An analysis of the results of the CMC.D model in relation to the highest observed score within a cow milking contained in the CMC showed that the misclassification rate was lower for the CMC containing QMS ≥ 3 as the highest score, of which 43% (23 out of 54) of cases were correctly classified. The proportion of correctly classified CMC values was even higher for the CMC containing QMS ≥ 4 or QMS = 5, respectively, as the highest score, and 63% (8 out of 14) and 100% (2 out of 2) of these were correctly classified. The trend was similar for the CMC.P model (Table 7).

Cow Period Class

The CPC.3 model achieved a higher Se (0.23) than did the CPC.4 model (Se = 0.14). The Sp was extremely high for both models, but slightly higher for the CPC.4 model than the CPC.3 model (i.e., Sp of 1.00 and 0.98, respectively). Also, the PPV values for the CPC models were almost the same: 0.72 and 0.71 for CPC.3 and CPC.4, respectively. However, the NPV was 0.87 for the CPC.3 and 0.94 for the CPC.4 model (Table 6).

Table 6. Sensitivity (Se), specificity (Sp), positive predictive value (PPV), and negative predictive value (NPV), from runs of models representing the median of 10 runs for the cow milk class (CMC) detection model using data from all milkings (CMC.D) and for the CMC prediction model excluding data from the milking of the current milk inspection (CMC.P), as well as for the cow period class (CPC) model with 2 different cow period class thresholds (CPS.3 and CPC.4)

Item	Se	Sp	PPV	NPV
CMC.D[1]	0.26	0.98	0.53	0.95
CMC.P[2]	0.25	0.98	0.47	0.95
CPC.3[3]	0.23	0.98	0.72	0.87
CPC.4[4]	0.14	1.0	0.71	0.94

[1]Seventh model run.
[2]Second model run.
[3]Sixth model run.
[4]Fourth model run.

Details on the 10 model runs forming the basis for the presented median values can be found in Supplemental Tables S3 and S4 (https://urn.kb.se/resolve?urn=urn:nbn:se:slu:epsilon-p-112047).

The test data set consisted of 486 cow periods. For CPC.3, 80 cow periods corresponded to the positive class, and for CPC.4, 35 cow periods corresponded to the positive class. The CPC.3 model correctly classified 25 out of 80 positive cow periods in the test data, whereas the CPC.4 model correctly classified 5 out of 35 positive cow periods in the test data. Both models were successful in classifying cow periods of a negative category. The CPC.3 model correctly classified 399 of 406 negative cow periods, whereas the CPC.4 model correctly classified 449 out of 451 negative cow periods.

DISCUSSION

In overall performance, the models displayed a high ability to distinguish cow milkings and cow milking periods without clots (i.e., high Sp), whereas their ability to detect cow milkings or cow milking periods with clots was lower (i.e., low Se). Thus, both CMC models correctly classified most of the cow milkings without observed clots (i.e., 98 out of 100 milkings were correctly classified or predicted). The performance of the CPC models in correctly classifying cow milking periods without observed clots was equally good, as only 4 out of 1,000 periods free from clots were wrongly classified as negative. A small number of false alerts (i.e., high Sp) is an important functionality for the farmer to trust the system (Mollenhorst et al., 2012). As the Sp of all model variations ranged between 98 and 100%, the results are promising.

It has been suggested that CM detection systems should have an Se of ≥80% if they are to identify an acceptable share of true cases (e.g., Hillerton, 2000; Hogeveen et al., 2010); this level is higher than the ≥70% Se recommended by the International Organization for Standardization (ISO, 2007). None of the 4 model variations reached the recommended performance in terms of Se. As the data was very unbalanced, down sampling (setting the 2 outcome classes to equal frequency), was evaluated using the CMC.D model. However, this did not improve the performance of the model.

The statistical measures Se and Sp are independent of the prevalence of observations of an event. Due to the low occurrence of clots in the data set, another way to evaluate the performance of the models is to calculate PPV and NPV because they also depend on the prevalence and are, from the user's perspective, more practical indicators. The PPV was overall moderate to high (50–70%), which, in practice, implies low false positive rates for the farmer when checking the cows classified as positive by the models. Low false positive rates have been shown to be important for farmers (Steeneveld et al., 2010; Mollenhorst et al., 2012), in fact, more important than actually finding all cases (Mollenhorst et al., 2012). For both CMC models, the PPV shows that clots will be found in approximately every second cow milking among cow milkings classified with a positive outcome (i.e., as having clots). As the PPV of the CPC models were even higher, farmers would find cows having periods of clots in 7 out of 10 cases classified as cow periods with a positive outcome, but the number of cases found at each milking would be lower because clots did not occur at every milking during a period.

In AMS, it is important that the detection system should send alerts of events in need of farmer action before a cow milking (Mollenhorst et al., 2012). The performance of the 2 CMC models indicated that prediction and detection performance were equally good, already giving information regarding an event of clots in milk at the previous milking. The presence of clots in 2 out of 3 consecutive milkings has been suggested to be included in the "gold-standard" definition of CM (Mein and Rasmussen, 2008; Kamphuis et al., 2013, 2016). The likelihood of a cow having a milking with clots decreases with shorter milking intervals and increases with longer milking intervals (Hallén Sandgren et al., 2021). Thus, the standard of considering clot presence in 2 out of 3 consecutive milkings as indicating

Table 7. The highest observed score within a cow milking versus the detected cow milk class (CMC.D) and the highest observed score within a cow milking versus the predicted cow milk class (CMC.P)

	CMC.D[1]						CMC.P[2]					
	Highest observed score						Highest observed score					
Detected or predicted	0	1	2	3	4	5	0	1	2	3	4	5
No clots	1,377	95	52	25	6	0	1,371	96	50	25	9	0
Clots	21	5	6	13	8	2	27	4	8	13	5	2

[1]Seventh run.

[2]Second run.

abnormal milk might be somewhat misleading for AMS with large variations in milking intervals. Possibly, the type of event that the CPC model captures (i.e., the weighted presence of clots during a certain period) could be a valuable tool in AMS when the milking interval differs.

The prevalence of clots (i.e., QMS $\geq$ 2) was higher than that found previously (Claycomb et al., 2009; Kamphuis et al., 2016). The results are hard to compare because clots were collected at the quarter level in this study, which likely resulted in more observations of clots than if samples had been collected at the cow composite level. A different clot density from that in samples collected at the cow composite level could also be expected.

Collecting samples at the quarter level and throughout the milking might have had an effect on the proportion of very small deposits (e.g., traces), which were found in 42% of all milk samples with a score. Although the farmer would not likely detect or take notice of these cases, the traces might give extra information to the algorithm because observations with QMS = 1 seemed to accumulate to a larger extent among cows having clots in the milk (i.e. QMS $\geq$ 2) than in cows without clots (Hallén Sandgren et al., 2021) and they were therefore included in the CPC models. However, the size of a flake is hard to judge (Rasmussen, 2005), and many observations of traces might have been of single flakes. Furthermore, small deposits would also likely be captured by the milk filter before the milk is delivered to the bulk tank. Thus, individual milkings with single traces in quarters were considered a nonconcern for the farmer and were consequently excluded from the CMC models. The presence of small flakes has earlier been reported to be a poor indication of bacteriological infection (Giesecke and van den Heever, 1974), which might imply that QMS = 1 should also have been excluded from the CPC models to improve performance.

The scoring system used (i.e., the definition of what are considered as clots) could also affect results. The test of detection systems for changes in milk homogeneity (ISO, 2007) states that clots larger than 2 mm should be considered abnormal in both quarter and composite testing, whereas the density of clots in the filter, which might be a more appropriate way of judging the severity of the change in milk homogeneity, is not mentioned. Possibly, QMS $<$ 3 should also be considered "non-cases," also strengthened by the low repeatability of QMS 1 and 2 within period described in Hallén Sandgren et al. (2021). In one of the potential gold standard definitions investigated by Claycomb et al. (2009), low-density scores at the cow composite level were excluded, which increased the Se. Kamphuis et al. (2016) suggested presenting the prevalence of clots for all density scores, but also with the low-density scores excluded. Both CMC model variations showed a decreased misclassification rate for the CMC that incorporated the moderate and high scores and, subsequently, the high scores only (QMS $\geq$ 4), which indicated that the model would be able to distinguish severe from mild cases. Furthermore, the prevalence of high QMS (QMS $\geq$ 4) in the overall data set was very low (0.4%). Thus, both model training and testing were performed on data capturing of a very limited number of more severe cases, which might have affected the outcome. Single milkings with clots are not included in the definition of abnormal milk suggested by the ISO (2007). However, alerting of severe cases is one of farmers' top preferences in detection systems (Mollenhorst et al., 2012). Predicting single incidents of cow milkings with QMS $\geq$ 4, independently of their recurrence, would be an important first step in the development of algorithms that could predict clots. Watery milk cannot be detected using inline filters, and such cases may interfere with our results in terms of the misclassification of negative categories as positive.

The SCC is generally accepted as an important milk quality parameter (Politis and Ng-Kwai-Hang, 1988; Barbano et al., 2006; IDF, 2013) and reflects possible inflammation (Pyörälä, 2003). Consequently, SCC, or rather the California Mastitis Test, has been included in the testing of detection systems for changes in milk homogeneity (ISO, 2007). Kamphuis et al. (2008) demonstrated that inclusion of inline SCC improved the prediction performance of a CM model. In this study, the OCC was included as a predictor variable; therefore, information regarding SCC was included in all models. The downside of the MLP being a "black box" algorithm is that an estimate of each independent predictor variable cannot be obtained. This makes it hard to determine whether the performance would improve or worsen with the addition or removal of some of the predictor variables. It remains to be proven how applications including SCC data from OCC can best serve their intended purpose. For instance, combining clot occurrence with OCC values could result in a combined parameter that better reflects the milk quality and udder health status of the cow. Alternative methods for prediction that include parameter estimation, for example, generalized additive mixed models, and thus the possibility of evaluating the contribution of the different predictor variables, were considered. However, because the aim of the study was prediction performance rather than inference, and based on previous experiences with both generalized additive mixed models and MLP (Anglart et al., 2020), the MLP was chosen for this study.

Tuning the hyperparameters (i.e., creating the neural network structure and configuration) has a large effect on model performance (Larochelle et al., 2007; Smith, 2018). The hyperparameters were tuned manually to construct the MLP for each model. Manual tuning is one of the most common approaches for optimizing hyperparameters in neural networks (Bergstra and Bengio, 2012). In our manual tuning, we varied default settings in the Keras R package, keeping track of validation loss (i.e., minimizing the sum of errors). Each run through the network gave slightly different results. This was overcome by running each model several times during tuning as well as during the final model predictions. Grid search could also be an option for tuning the hyperparameters, although the method is more time consuming and not always the most effective in finding optimal settings (Bergstra and Bengio, 2012).

Because ANN learn through recognizing patterns in data, a very unbalanced data set could potentially cause the model to learn one class better than another. Furthermore, this is more likely when the data are noisy, as demonstrated by Murphey et al. (2004). In the current study, noise could be a consequence of clots with lower density scores not having a biological explanation that could be derived from the sensor information. This might explain the poor prediction performance for positive categories. The prediction and detection of clots collected by inline filters have not previously been investigated. Hence, the prediction target, based solely on the suggested clot size, might have been misconceived, incorporating density scores that were too low to be distinguished. These clot cases are probably not meaningful to detect from either a milk quality or CM detection point of view.

In conclusion, using data generated by AMS to detect and predict changes in milk homogeneity seems to be possible. Cow milkings without changes in milk homogeneity could be properly distinguished; however, the performance was poor at detecting cow milkings and cow milking periods with clots according to the definitions of clots used.

ACKNOWLEDGMENTS

Financial support for this study was provided by the Swedish Foundation for Strategic Research (SSF, Stockholm, Sweden). We thank the Kjell & Märta Beijer Foundation (Stockholm, Sweden) for funding Lars Rönnegård. We additionally thank the farmers participating in this study and a special thanks to Johan Östlund and Harry Tuinier for help with the data collection. Two of the authors (D. Anglart and C. Hallén Sandgren) are employed by DeLaval International AB. The authors have not stated any other conflicts of interest.

REFERENCES

Agatonovic-Kustrin, S., and R. Beresford. 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J. Pharm. Biomed. Anal. 22:717–727. https://doi.org/10.1016/S0731-7085(99)00272-1.

Anglart, D., C. Hallén Sandgren, U. Emanuelson, and L. Rönnegård. 2020. Comparison of methods for predicting cow composite somatic cell counts. J. Dairy Sci. 103:8433–8442. https://doi.org/10.3168/jds.2020-18320.

Ankinakatte, S., E. Norberg, P. Løvendahl, D. Edwards, and S. Højsgaard. 2013. Predicting mastitis in dairy cows using neural networks and generalized additive models: A comparison. Comput. Electron. Agric. 99:1–6. https://doi.org/10.1016/j.compag.2013.08.024.

Barbano, D. M., Y. Ma, and M. V. Santos. 2006. Influence of raw milk quality on fluid milk shelf life. J. Dairy Sci. 89:E15–E19. https://doi.org/10.3168/jds.S0022-0302(06)72360-8.

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

Chollet, F. 2017. Deep Learning with Python. Manning Publications Co.

Claycomb, R. W., P. T. Johnstone, G. A. Mein, and R. A. Sherlock. 2009. An automated in-line clinical mastitis detection system using measurement of conductivity from foremilk of individual udder quarters. N. Z. Vet. J. 57:208–214. https://doi.org/10.1080/00480169.2009.36903.

European Commission. 2004. Regulation (EC) N° 853/2004 of the European Parlamient and of the Council of 29 April 2004 laying down specific hygiene rules for on the hygiene of foodstuffs. Off. J. Eur. Union L 139:55.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76:378–382. https://doi.org/10.1037/h0031619.

Food and Drug Administration (FDA). 2017. Grade "A" Pasteurized Milk Ordinance (PMO) 1–426. Food and Drug Administration.

Gamer, M., J. Lemin, and I. Fellows Puspendra Sing. 2019. Irr: Various Coefficients of Interrater Reliability and Agreement. Accessed Oct. 10, 2020. https://CRAN.R-project.org/package=irr.

Giesecke, W. H., and L. W. van den Heever. 1974. The diagnosis of bovine mastitis with particular reference to subclinical mastitis: a critical review of relevant literature. Onderstepoort J. Vet. Res. 41:169–211.

Hallén Sandgren, C., D. Anglart, I. C. Klaas, L. Rönnegård, and U. Emanulesson. 2021. Homogeneity density scores of quarter milk in automatic milking systems. J. Dairy Sci. 104:XXX–XXX. https://doi.org/10.3168/jds.2021-20439.

Hassan, K. J., S. Samarasinghe, and M. G. Lopez-Benavides. 2009. Use of neural networks to detect minor and major pathogens that cause bovine mastitis. J. Dairy Sci. 92:1493–1499. https://doi.org/10.3168/jds.2008-1539.

Haykin, S. 2009. Neural Networks and Learning Machines. 3rd ed. Pearson Education.

Heald, C. W., T. Kim, W. M. Sischo, J. B. Cooper, and D. R. Wolfgang. 2000. A computerized mastitis decision aid using farm-based records: An artificial neural network approach. J. Dairy Sci. 83:711–720. https://doi.org/10.3168/jds.S0022-0302(00)74933-2.

Hillerton, J. E. 2000. Detecting mastitis cow-side. Pages 48–53 in Annual Meeting of the National Mastitis Council. National Mastitis Council.

Hogeveen, H., C. Kamphuis, W. Steeneveld, and H. Mollenhorst. 2010. Sensors and clinical mastitis—The quest for the perfect alert. Sensors (Basel) 10:7991–8009. https://doi.org/10.3390/s100907991.

International Dairy Federation (IDF). 2011. Suggested interpretation of mastitis terminology. Vol. 448. Bulletin: International Dairy Federation.

International Dairy Federation (IDF). 2013. Guidelines for the use and interpretation of bovine milk somatic cell count. Vol. 466. Bulletin: International Dairy Federation.

International Organization for Standardization (ISO). 2007. Automatic milking installations-Requirements and testing Installations de traite automatique-Exigences et essais. 1st ed. 20966: 2007.

Kamphuis, C., B. Dela Rue, G. Mein, and J. Jago. 2013. Development of protocols to evaluate in-line mastitis-detection systems. J. Dairy Sci. 96:4047–4058. https://doi.org/10.3168/jds.2012-6190.

Kamphuis, C., B. T. Dela Rue, and C. R. Eastwood. 2016. Field validation of protocols developed to evaluate in-line mastitis detection systems. J. Dairy Sci. 99:1619–1631. https://doi.org/10.3168/jds.2015-10253.

Kamphuis, C., R. Sherlock, J. Jago, G. Mein, and H. Hogeveen. 2008. Automatic detection of clinical mastitis is improved by in-line monitoring of somatic cell count. J. Dairy Sci. 91:4560–4570. https://doi.org/10.3168/jds.2008-1160.

Kingma, D. P., and J. L. Ba. 2015. Adam: A method for stochastic optimization. Pages 1–15 in 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.

Landis, J. R., and G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159–174. https://doi.org/10.2307/2529310.

Larochelle, H., D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. 2007. An empirical evaluation of deep architectures on problems with many factors of variation. Pages 473–480 in ACM Int. Conf. Proceeding Ser. Association for Computing Machinery.

Mein, G. A., and M. D. Rasmussen. 2008. Perfromance of system for automated monitoring of udder health: Would the real gold standard please stand up? Pages 259–266 in Mastits control- From Science to practice. Wageningen Academic.

Mollenhorst, H., L. J. Rijkaart, and H. Hogeveen. 2012. Mastitis alert preferences of farmers milking with automatic milking systems. J. Dairy Sci. 95:2523–2530. https://doi.org/10.3168/jds.2011-4993.

Murphey, Y. L., H. Guo, and L. A. Feldkamp. 2004. Neural learning from unbalanced data. Appl. Intell. 21:117–128. https://doi.org/10.1023/B:APIN.0000033632.42843.17.

Nielen, M., M. H. Spigt, Y. H. Schukken, H. A. Deluyker, K. Maatje, and A. Brand. 1995. Application of a neural network to analyse on-line milking parlour data for the detection of clinical mastitis in dairy cows. Prev. Vet. Med. 22:15–28. https://doi.org/10.1016/0167-5877(94)00405-8.

Nielsen, C., and U. Emanuelson. 2013. Mastitis control in Swedish dairy herds. J. Dairy Sci. 96:6883–6893. https://doi.org/10.3168/jds.2012-6026.

Politis, I., and K. F. Ng-Kwai-Hang. 1988. Effects of somatic cell count and milk composition on cheese composition and cheese making efficiency. J. Dairy Sci. 71:1711–1719. https://doi.org/10.3168/jds.S0022-0302(88)79737-4.

Pyörälä, S. 2003. Indicators of inflammation in the diagnosis of mastitis. Vet. Res. 34:565–578. https://doi.org/10.1051/vetres:2003026.

Rasmussen, M. D. 2005. Visual scoring of clots in foremilk. J. Dairy Res. 72:406–414. https://doi.org/10.1017/S0022029905000993.

Rasmussen, M. D., and L. Bach Larsen. 2003. Milking hygiene: new issues and opportunities from automatic milking. Ital. J. Anim. Sci. 2:283–289. https://doi.org/10.4081/ijas.2003.283.

Rodrigues, A. C. O., D. Z. Caraviello, and P. L. Ruegg. 2005. Management of Wisconsin dairy herds enrolled in milk quality teams. J. Dairy Sci. 88:2660–2671. https://doi.org/10.3168/jds.S0022-0302(05)72943-X.

Smith, L. N. 2018. A disciplined approach to neural network hyper-parameters: Part 1 – Learning rate, batch size, momentum, and weight decay. arXiv 1–21. https://arxiv.org/abs/1803.09820.

Steeneveld, W., L. C. van der Gaag, W. Ouweltjes, H. Mollenhorst, and H. Hogeveen. 2010. Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems. J. Dairy Sci. 93:2559–2568. https://doi.org/10.3168/jds.2009-3020.

Stekhoven, D. J. 2013. missForest: Nonparametic Missing Value Imputation using Random Forest. R Package version 1.4. Accessed Jun. 26, 2020. https://cran.r-project.org/web/packages/missForest/missForest.pdf.

Sun, Z., S. Samarasinghe, and J. Jago. 2010. Detection of mastitis and its stage of progression by automatic milking systems using artificial neural networks. J. Dairy Res. 77:168–175. https://doi.org/10.1017/S0022029909990550.

Wenz, J. R., S. M. Jensen, J. E. Lombard, B. A. Wagner, and R. P. Dinsmore. 2007. Herd management practices and their association with bulk tank somatic cell count on United States Dairy operations. J. Dairy Sci. 90:3652–3659. https://doi.org/10.3168/jds.2006-592.