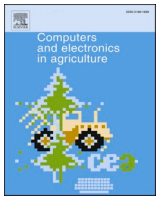
Computers and Electronics in Agriculture 228 (2025) 109698

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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



Development of individual models for predicting cow milk production for real-time monitoring

Jae-Woo Song [a], Mingyung Lee [b], Hyunjin Cho [b], Dae-Hyun Lee [a,c], Seongwon Seo [b,*], Wang-Hee Lee [a,c,*]

[a] Department of Biosystems Machinery Engineering, Chungnam National University, Daejeon 34134, South Korea
[b] Division of Animal and Dairy Sciences, Chungnam National University, Daejeon 34134, South Korea
[c] Department of Smart Agriculture Systems, Chungnam National University, Daejeon 34134, South Korea

ARTICLE INFO

Keywords:
Daily milk yield
Dairy cow
Individual modeling
Real-time monitoring
Smart livestock farming

ABSTRACT

Daily milk yield serves as a physiological indicator in dairy cows and is a primary target for prediction and real-time monitoring in smart livestock farming. This study attempted to develop an individual model for predicting daily milk yield and applied it to monitor the health status of dairy cows by designing a real-time monitoring algorithm. A total of 580 datasets were used for model development after data preprocessing and screening, which were subsequently used to develop the model by modifying the existing models based on nonlinear regression analysis. The developed model was then applied to short-term real-time monitoring of abnormal daily milk yields. The optimal model was able to predict the daily milk yield, with an R^2 value of 0.875 and a root mean squared error of 2.192. Real-time monitoring was designed to detect abnormal daily milk yields by collectively considering a 90% confidence interval and the difference between predicted values and expected trends. This study is the first to design a monitoring algorithm for daily milk yield from dairy cows based on an individual model capable of predicting the daily milk yield. This study expects that a platform will be necessary for highly efficient smart livestock farming, enabling high productivity with minimal inputs.

1. Introduction

The world's population has reached 8 billion and is expected to increase to 9.7 billion by 2050. Food demand is expected to increase by 70 %, demanding smart agriculture to produce high-quality agricultural products with reduced labor and effective management (Halachmi et al., 2019; Davis and White, 2020; Idoje et al., 2021; UN DESA, 2022). Accordingly, the increase in production costs due to rising management and maintenance expenses in livestock farming necessitates a shift from traditional livestock farming to smart farming. In dairy farms, smart livestock farming has been achieved with real-time monitoring and automated milk production using Internet of Things (IoT) technologies, such as Radio Frequency IDentification (RFID), cattle biometric and environmental sensors, automatic milking systems, AI-based data analysis (Stankovski et al., 2012; Akbar et al., 2020; Alonso et al., 2020; Unold et al., 2020; Farooq et al., 2022). For example, cow activity has been monitored using biosensors along with statistical analysis to provide robust heat alerts, while an RFID-based intake monitoring system has been employed with algorithms to detect abnormal intake associated with unhealthy conditions (Lee et al., 2023; Jung et al., 2024).

Milk yield is the most profitable productivity indicator in dairy farming, and serves as a fundamental physiological metric that can be measured and data-logged for all livestock farms (Edwards and Tozer, 2004; Lukas et al., 2009). Milk yield responds quickly to the health status of dairy cows and varies according to their health condition (Fleischer et al., 2001; Bareille et al., 2003). For example, the deterioration of health due to diseases can lead to a reduction in milk yield, which can also decline due to stress from high temperatures or humidity in cattle sheds (Fourichon et al., 1999). In addition, milk yield varies depending on the balance of nutrients in the feed, such as energy and protein, which reflects the feeding activity of dairy cows (Daniel et al., 2016). For this reason, even though other important factors, such as cattle behavior, body temperature, weight, and feed intake, exist, milk yield can be one of the most suitable indicators for daily monitoring of dairy cows due to its sensitivity to changes in health status or rearing environment. Moreover, daily milk production is high-accessible data

* Corresponding authors.
E-mail addresses: swseo@cnu.kr (S. Seo), wanghee@cnu.ac.kr (W.-H. Lee).

https://doi.org/10.1016/j.compag.2024.109698
Received 20 July 2024; Received in revised form 5 November 2024; Accepted 21 November 2024
Available online 26 November 2024

collected by most farms pursuing smart technology, making it a key focus. As the prediction and real-time monitoring of milk yield are essential elements in smart livestock farming for the effective management of dairy farms, model-based studies have been conducted to predict milk production (Adriaens et al., 2018; Nguyen et al., 2020; Ji et al., 2022). However, most studies have focused on predicting long-term milk production, whereas studies on the prediction of daily milk yield have been limited. This is because it has been challenging to apply owing to the need for data other than milk yield (such as calving, days in milk, age, weight, temperature, and humidity), requiring a series of measurable devices installed on dairy farms. Therefore, there is a need to develop models that can predict daily milk yield with high accuracy and precision using only yield data over time, which is ultimately applicable for establishing the real-time milk yield monitoring systems necessary for smart dairy farming (Adriaens et al., 2018; Zhang et al., 2020).

Statistical analysis is used to effectively process livestock data with large variations, resulting in modeling studies of lactation curves to characterize milk yields (Wood, 1967; Ali and Schaeffer, 1987; Wilmink, 1987; Mead, 2017). However, variations and patterns in data that appear differently for each cow suggest the need to develop individual models, because applying a statistical model to the entire population is challenging. In addition, statistical models developed based on the entire dataset are difficult to apply in real time, indicating the need for complementary approaches. In this study, we developed models that predict the daily milk yield of individual cows based on time-series lactation data using data processing techniques and statistical modeling. We then built a real-time monitoring algorithm for daily milk yield by relating the prediction results of the developed model to outlier detection using confidence intervals. This study advances prior research by developing individual-based models to predict daily milk production and introducing a novel approach for real-time monitoring method not previously attempted. We anticipate that this study will not only provide foundational economic data for the real-time prediction of milk productivity in dairy farms, but will also be utilized for monitoring cow health based on real-time tracking of milk yield fluctuations.

2. Materials and methods

2.1. Experimental animals

Data were obtained from four commercial farms in Boryeong, Chungcheongnam-do; Wonju, Gangwon-do; Gimcheon, Gyeongsangbuk-do; and Cheonan, Chungcheongnam-do, South Korea, for a total of three years, from January 2016 to December 31, 2018. A conventional milking parlor system (DeLaval International AB, Tumba, Sweden) was used at one farm located in Boryeong, Chungcheongnam-do, and a robotic milking system (Lely Industries NV, Maassluis, Netherlands) was used at the other three farms. The farms were geographically dispersed, each experiencing different environmental conditions, and milk yield data were collected over three years, accounting for seasonal factors to ensure minimal bias in the data collected.

2.2. Data acquisition

Daily volume of milk production was stored individually using Radio-Frequency Identification (RFID) tags for each cow, and a total of 307,053-time series data of daily milk production were obtained from 727 cows with a maximum number of days in milk (DIM) of 1091. As lactation characteristics vary depending on the parity of the cow, we classified the same cow with different parities into separate datasets. Consequently, the total number of datasets was 1,346, with an average parity of 2.46, which was used for data preprocessing.

2.3. Data preprocessing

2.3.1. Data screening

Owing to the reliance of statistical model development on the reliability of time-series data, we first conducted screening to identify individual cows with sufficient time-series data. From this perspective, we selected a dataset with milking data recorded within the first 30 days and a continuous record for over 200 days. A total of 580 datasets recorded milking data from 1 to a maximum of 728 days by considering individual cows, and parity was finally confirmed for further processing and modeling.

2.3.2. Outlier detection and removal

Outliers that could cause bias in the analysis and development of the regression model were removed from each dataset. Outlier detection was performed using the tsoutliers function in the forecast package developed for time-series data in R software (Hyndman and Khandakar, 2008; R Core Team, 2021). In this algorithm, outliers are detected by decomposing the time-series data into trend, seasonal, and residual components, and Tukey's fences are then used to detect abnormal observations in the residual components (Tukey, 1977; Cleveland et al., 1990). In addition, we considered data values recorded as zero during the milking period after milking initiation as outliers. All detected 5,445 outliers were removed, and the total number of 195,046 data points from 580 individual cows were ultimately used.

2.3.3. Data smoothing

To minimize the large variations and noise that hindered the analysis of daily milk yield, we performed data smoothing using five different methods: simple exponential smoothing, moving average, locally weighted scatterplot smoothing (Lowess), smoothing spline, and the Savitzky-Golay filter (Savitzky and Golay, 1964; Cleveland, 1979; Eilers and Marx, 1996; Hyndman, 2011; Hyndman and Athanasopoulos, 2018). Each smoothing method was independently applied to the entire dataset and executed using the corresponding function in Python with optimal parameter values suitable for subsequent model development (Dierckx, 1975; Press and Teukolsky, 1990; Seabold and Perktold, 2010; McKinney, 2011; Pedregosa et al., 2011; Virtanen et al., 2020) (Table 1). The parameters for each smoothing method were determined to minimize the fluctuation of actual data within the range suitable for regression analysis, while simultaneously achieving optimal performance when combined with the regression analyses, thereby facilitating future application in monitoring systems. Then, the optimal smoothing method was determined to maximize the performance of the model in predicting milk production, which is described later.

2.4. Model development

Total of six models were either employed or developed to predict the daily milk production for an individual cow based on smoothed data, and the best-performing model was selected. (Table 2). The Wood, Wilmink, and Dijkstra models were selected from existing lactation models due to their relatively simple structures, making them suitable for individual modeling that requires uniformity across cows (Wood, 1967; Wilmink, 1987; Dijkstra et al., 1997). These models were modified to improve fit while preserving the biological significance of each term within the models. In this study, we enhanced the regression performance by employing a simple transformation from a first-order term to a quadratic term while minimizing the increase in the number of parameters. In all existing models where multiple linear terms were present, we selected a linear term that improved the fitting performance and had parameters converging during the fitting process to convert it into a quadratic term. All regression analyses were performed using the curve_fit function in the Python scipy.optimize and sklearn. metrics packages (Vugrin et al., 2007; Virtanen et al., 2020). The models were compared in terms of R^2-value and root mean square error (RMSE) to

Table 1
Data smoothing methods coded in python.

Smoothing method	Parameter	Parameter value	Package & function
Simple exponential	'smoothing_level': Control the effect of previous observations and take a value between 0 and 1	0.1	statsmodels.tsa.holtwinters.SimpleExpSmoothing.fit
Moving average	'window': Size of the moving window	20	pandas.DataFrame.rolling
Lowess	'frac': Between 0 and 1. The fraction of the data used when estimating each y-value	0.1	statsmodels.nonparametric.smoothers_lowess.lowess
Spline	's': Balance between how well the spline curve fits the data and level of smoothing	1000	scipy.interpolate.UnivariateSpline
Savitzky-Golay	'window_length': The length of the filter window	40	scipy.signal.savgol_filter
	'polyorder': The order of the polynomial used to fit the samples	2	

Table 2
Models used to develop milking prediction models.

Lactation curve	Existing models	Modified models
Wood	$y = a \bullet t^{b} \bullet e^{-c \bullet t}$	$y = a \bullet t^{b} \bullet e^{-c \bullet (d \bullet t^{2} + e \bullet t)}$
Wilmink	$y = a + b \bullet e^{-k \bullet t} + c \bullet t$	$y = a + b \bullet e^{-0.05 \bullet t} + c \bullet t^{2} + d \bullet t$
Dijkstra	$y = a \bullet e^{b \bullet \frac{1 - e^{-c \bullet t}}{c} - d \bullet t}$	$y = a \bullet e^{b \bullet \frac{1 - e^{-c \bullet (d \bullet t^{2} + e \bullet t)}}{c} - f \bullet t}$

y: daily milk yield, **t**: time of lactation. The parameters a, b, c and d of existing models define the size and shape of the lactation curve. Wood model: a = level of production, b = rate of increase in milk yield reaching the peak, c = rate of decrease after the peak of milk yield. Wilmink model: a = level of production, b = rate of increase in milk yield reaching the peak, c = rate of decrease after the peak of milk yield, k value was set to 0.05 in this study. Dijkstra model: a = theoretical milk yield at the initial lactation, b = secretory cell proliferation rate at parturition, c = cell proliferation reduction rate, d = secretory cell apoptosis rate.

select the optimal model structure.

2.5. Monitoring algorithm development

After determining the best model structure, the optimal model was applied to the real-time short-term monitoring of daily milk production. Initially, we sought to determine the minimum number of days required to operate the model and experimented by gradually increasing the data from days 1 to 30. Moreover, we only selected a dataset with a start of DIM of two days or less and no missing DIM to reflect the characteristics of real-time monitoring and enhance the performance, resulting in a total of 389 datasets. Then, two models were developed based on distinct data utilization methods: one utilizing the data from the most recent 10 days and the other incorporating accumulated data beyond the initial 10 days. The former focuses on predicting values closer to observed measurements by exclusively utilizing recent data, whereas the latter aims to estimate trends in daily milk production using cumulative data. Both models were applied to predict daily milk production after 1, 3, 5, and 10 days to determine the most suitable model for real-time monitoring based on predictive performance. Finally, two monitoring approaches were concurrently implemented: the deviation range between the model utilizing cumulative data and the smoothed actual values, and the difference in predicted values between the model using recent data and the model using cumulative data. For range-based monitoring, we set a 90 % confidence interval to detect abnormal milk production, corresponding to the upper and lower 5 %. The predictive discrepancy between the two different models was conceived to predict instances where the predicted actual values were smaller than the expected trends in advance, which was further used to assess future abnormal values alongside the first monitoring method.

3. Results

3.1. Model development results

After applying the four smoothing methods to all datasets, we ran a regression analysis using three existing and three modified lactation curve models. The results of the regression analysis were summarized by R^2-value and root mean square error (RMSE). The lowest R^2-value was 0.702 for the combination of the spline and modified Dijkstra's methods, and the highest value was 0.875 for the combination of the simple exponential and modified Wilmink's methods. The RMSE showed the highest value (3.596) for the spline and Wood combination and the lowest value (2.1) for the lowess and modified Wilmink combinations. The combination of simple exponential smoothing and Lowess methods with the modified Wilmink model exhibited the best performance. Accordingly, a combination of simple exponential smoothing and the modified Wilmink model, which showed a high R^2-value and low RMSE while preserving the shape of the existing lactation curve relatively better, was used for monitoring (Fig. 1) (Table 3). By applying the selected optimal combination to the data for each cow, the best performance was found with an R^2-value of 0.996 and an RMSE of 0.436.

3.2. Data monitoring

The optimal combination of data smoothing and model structure was applied for the real-time prediction of daily milk production after 1, 3, 5, and 10 days (Fig. 2). The average absolute value of the error and RMSE calculated by the difference between the predicted value and the actual value according to the types of data utilization were the smallest for the one-day prediction, while they were drastically increased by the predicted date, with the highest for the ten-day prediction. Within the five-day prediction, the model with recent data resulted in a lower mean error and RMSE than when using cumulative data, while ten-day prediction showed higher and lower errors in the model with cumulative data (Table 4). In particular, one-day prediction using recent data showed a mean error of 0.321, indicating the best performance. However, the average error and RMSE in ten-day prediction increased by approximately three times compared to those in one-day prediction, suggesting poor performance. This might indicate that when an inflection point is present, predictions derived from recent data are more significantly affected than those generated using cumulative data. Consequently, the disparity between the two values tends to widen over time, resulting in a decline in predictive performance as the number of prediction days extends.

As a result, we determined that we can apply a one-day prediction to the real-time monitoring of daily milk production. For real-time monitoring, we set a 90 % confidence interval derived from smoothed data based on predicted values using accumulated data, and then detected cases where the actual value was smaller than the lower bound of the 90 % confidence interval or larger than the upper bound. Because cumulative data used all available data, while recent data focused on only the last 10 days, it was considered that cumulative data highlights trends, while recent data captures actual production changes, supporting

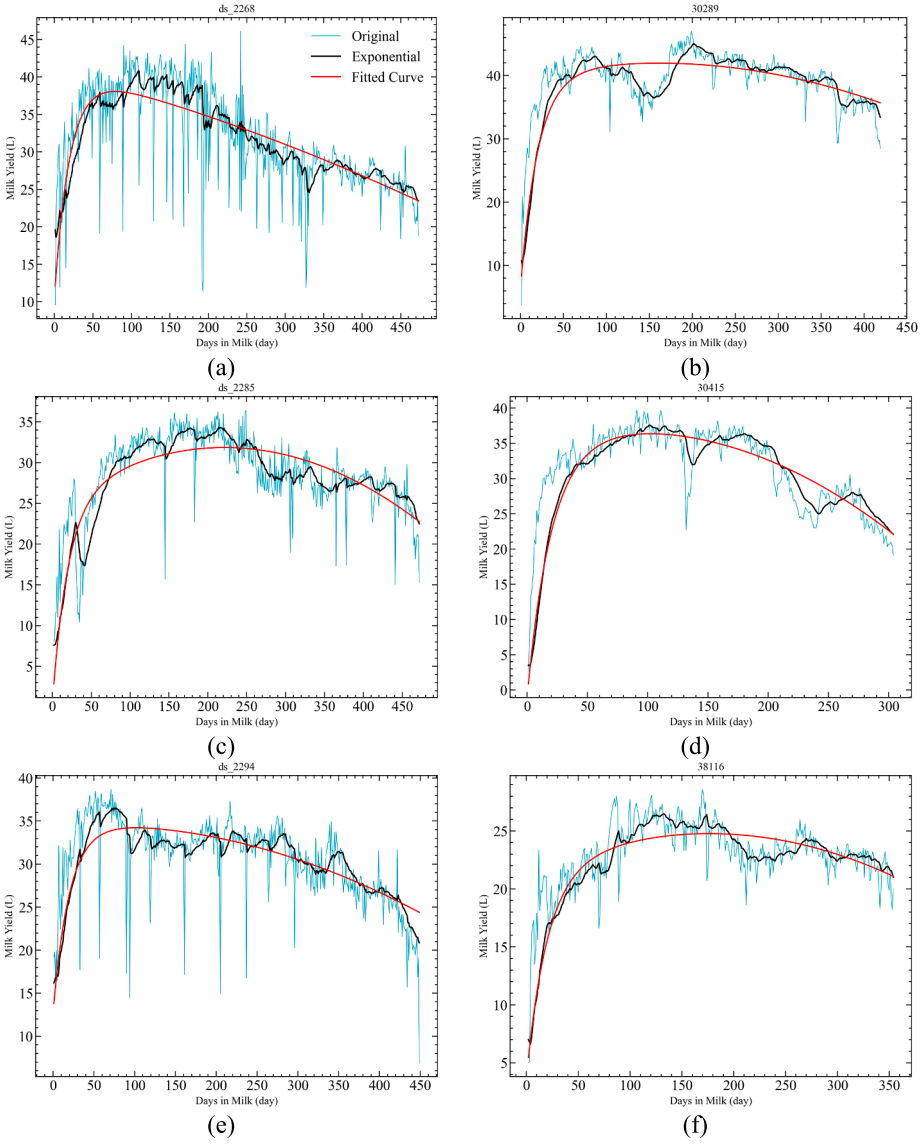


Fig. 1. Development of a daily milk production using the modified Wilmink model. Blue line: raw data, black line: data processed with simple exponential smoothing, and red line: prediction using the modified Wilmink model.

Table 3
Regression analysis results for each model after smoothing.

Smoothing method	Index	Wood	Wilmink	Dijkstra	ModifiedWood	Modified Wilmink	Modified Dijkstra
Simple exponential	R^2	0.816	0.821	0.868	0.864	**0.875**	0.781
	RMSE*	2.707	2.662	2.3	2.325	**2.192**	2.643
Lowess	R^2	0.8	0.798	0.844	0.842	0.862	0.795
	RMSE	2.598	2.601	2.22	2.239	2.1	2.383
Spline	R^2	0.705	0.703	0.737	0.735	0.755	0.702
	RMSE	3.596	3.595	3.364	3.361	3.216	3.485
Savitzky-Golay	R^2	0.709	0.706	0.742	0.743	0.758	0.703
	RMSE	3.496	3.508	3.246	3.248	3.145	3.398

* RMSE: Root mean square error.

distinct monitoring needs. Under this assumption, we examined instances in which the discrepancy between the predicted values derived from cumulative data and those derived from recent data exhibited a negative value (Fig. 3). These two monitoring criteria were set based on the assumption that daily milk production deviation from the overall trend could be an alert to be monitored for a cow, such as health-related problems. As a result of applying monitoring tests to selected 389 cows, 53,064 cases occurred when the difference in predicted values was

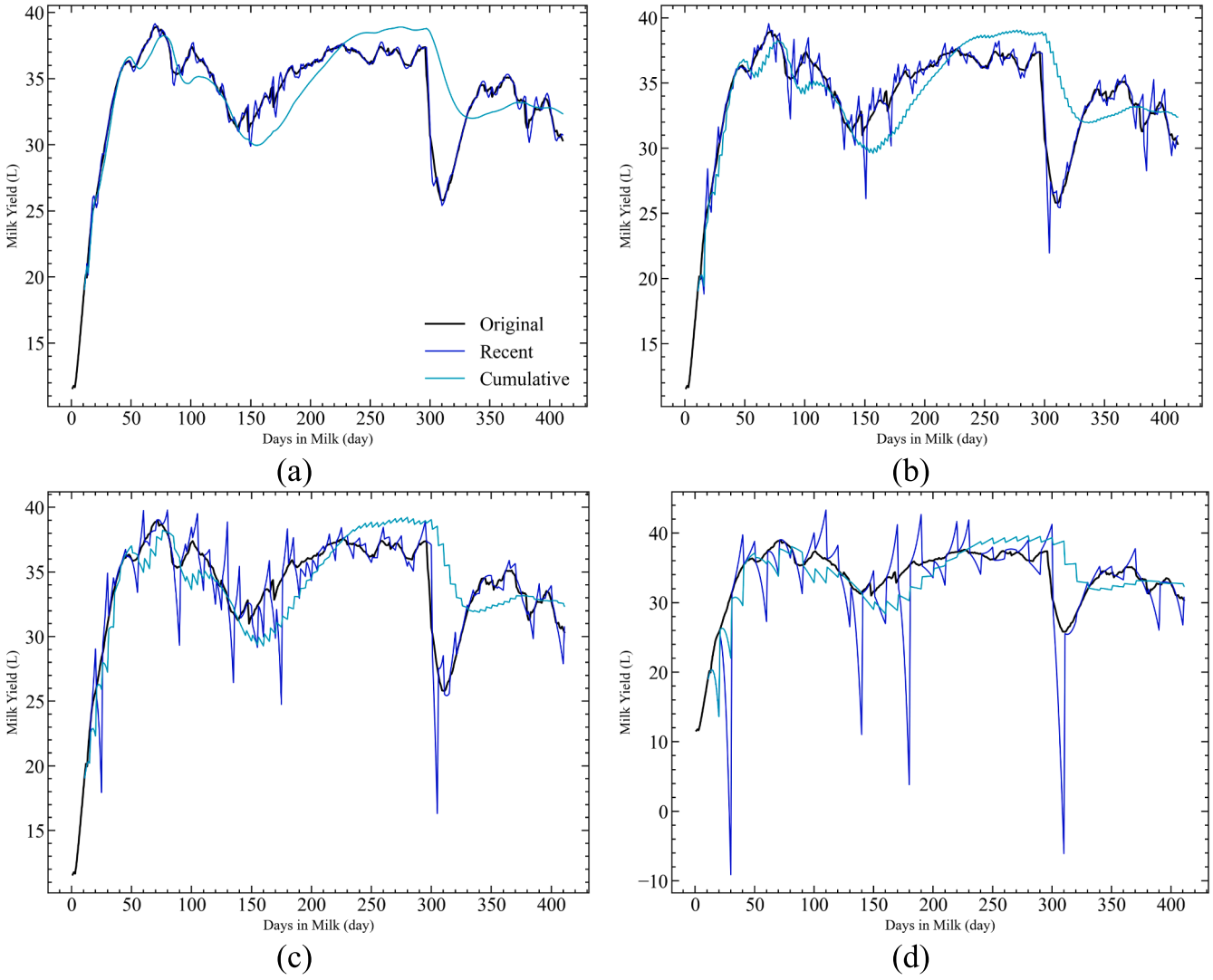


Fig. 2. Predictions by recent and cumulative methods of (a) 1 day, (b) 3 days, (c) 5 days, and (d) 10 days for a cow. Black line: raw data, blue line: prediction by using recent 10 days of data, and light blue line: prediction by using cumulative data.

Table 4
Model development results according to forecast days and method.

Method	Index	1 day	3 days	5 days	10 days
Accumulate	Mean Error	1.902	2.067	2.258	2.897
	RMSE	2.743	2.932	3.182	4.508
Recent	Mean Error	0.321	0.681	1.202	3.413
	RMSE	0.508	1.212	2.299	7.391‘

negative among the 129,945 data points, excluding the first 10 days. There was a total of 7,858 cases in which the actual data deviated from the monitoring standard, of which 4,505 were greater than the confidence interval and 3,353 were less. On average, for individual cows, approximately 136 negative values and 20 data points outside the monitoring standard were observed. Negative values occurred a maximum of 427 times and a minimum of 21 times, and the actual value outside the confidence interval occurred a maximum of 54 times and a minimum of twice.

4. Discussion

The large variation in milk production by cows and days demands individual prediction and monitoring owing to the low accuracy of population-level estimation. In addition, the collected data from different Such individual-based models enabled robust predictions with minimal bias by using parameters measured for each dairy cow from data that included geographical and seasonal variations. This study developed a model for predicting daily milk production and proposed a real-time monitoring method that is practically applicable for establishing smart livestock farming.

Owing to the harsh environmental conditions and biological variability among individual cows, the data obtained in livestock farming not only exhibit large variations among individuals, but also significant noise and large variations over time within individuals. This factor poses a significant challenge to the implementation of data-driven smart livestock farming, highlighting the necessity of preprocessing the data (Pastell and Kujala, 2007; Riaboff et al., 2019; Rodriguez-Baena et al., 2020). In this study, we applied three types of data preprocessing; data screening, outlier detection, and data smoothing. Data screening is a necessary process to achieve data integrity by compensating for the missing data and lack of consistency in the data acquired due to the harsh environment of livestock farming (Jensen et al., 2018; Nørstebø et al., 2019; Chen et al., 2023). Data screening may need to be based on criteria that reflect the lactation characteristics of the cows. In this study, the integrity of data from the initial 30 days and continuous

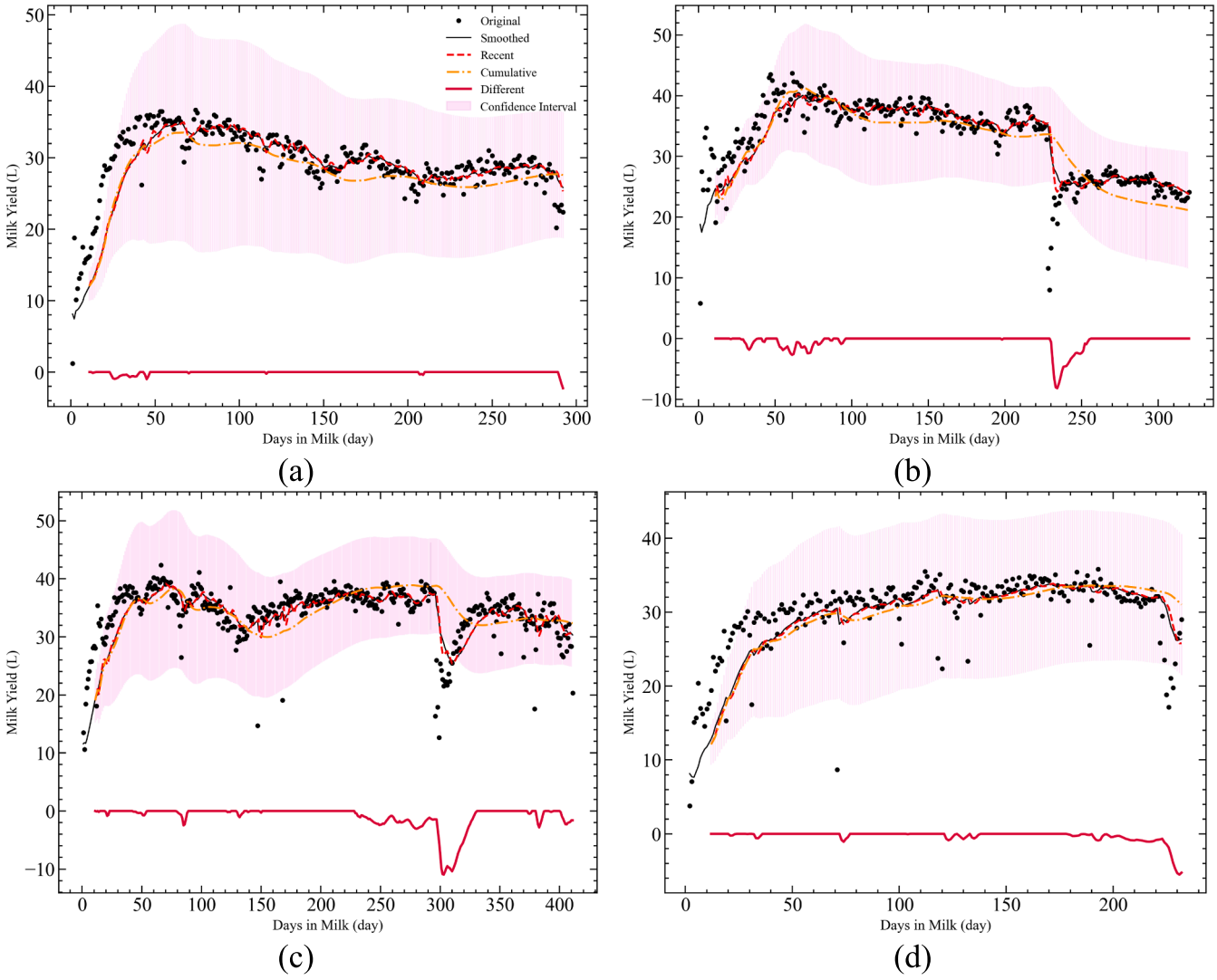


Fig. 3. Example of applying the developed model and criteria for monitoring individual cow milk production.

records over 200 days were used as criteria for data screening to reflect the characteristics of the lactation curve, including an increase in initial milk production with peak and data consistency (Wood, 1969; Ali and Schaeffer, 1987; Grossman and Koops, 2003). Outliers can be subjects of livestock monitoring, such as health anomalies or recording irregularities; however, they should be removed during model development to avoid biased results. Outlier removal is crucial, especially in livestock time-series data, which are influenced by the fluctuating environment and condition of the cows at the time of collection (Ismail et al., 2019; Bokde et al., 2023). High noise and daily variation are prevalent features of livestock data, but they pose challenges in discerning the underlying trends, which is the primary focus of predictive modeling (Friggens and Chagunda, 2005; Friggens et al., 2007). This means that a method for extracting the general trend of data is necessary, as shown by the data smoothing in this study. However, unlike screening or outlier detection, the criteria for data smoothing are not always clear and are often determined by analysts (Cavero et al., 2007; Huybrechts et al., 2014). Therefore, it is necessary to determine an appropriate configuration for data smoothing based on analysis objectives. For example, because this study attempted to develop a model for predicting daily milk production, we selected smoothing methods and configurations to achieve the highest predictive performance when combined with the model.

Correlative models prioritize data fitting over the functional aspects of model components, necessitating structural changes in regression models to enhance their predictive performance (Motulsky and Christopoulos, 2004). However, even in correlation models, it is desirable to change the model structure while maintaining the functional aspects of the model components when the existing model is significant for these components (Archontoulis and Miguez, 2015). Accordingly, this study attempted to add a quadratic term to the existing linear term to improve the fitting capability of the lactation curve model. In the Wood and Wilmink models, by modifying the linear model to a nonlinear one, we were able to fit the declining pattern after peak milk production individually for each subject, which ultimately led to an improvement in model performance (rate (Wood, 1967; Wilmink, 1987). Conversely, in the case of the Dijkstra model, which was built based on the proliferation and death of mammary gland cells during lactation, the addition of a second-order term reduced the completeness of the existing structure, unlike the two models that clearly modified the terms, indicating a decrease in milk yield (Dijkstra et al., 1997).

Because the individual regression model uses the entire dataset, it is difficult to apply to real-time data. To address this, it is necessary to determine the appropriate size of the required dataset so that a model developed using partial data can achieve a performance similar to that of a model based on the entire dataset. However, the DIM of the data used in a real-time prediction model is challenging owing to large fluctuations and different data lengths. Consequently, if the data length for the model construction is excessively long or short, the discrepancy between

the predicted and actual values increases significantly. Practically, we examined the changes in model performance and prediction patterns by incrementally increasing the minimum number of days starting from 1 day, and it was set to 10 days, which could prevent model overfitting and overly biased predictions depending on the data quality and availability (Draper and Smith, 1998). Consequently, we predicted the next 1, 3, 5, and 10 days using the accumulated data after 10 days and the recent data from the most recent 10 days, which showed markedly different results depending on the number of prediction days. In particular, when predicting for 10 days, the accuracy was higher when using cumulative data than when using recent data. This suggests that recent models, which only reflect the trend of 10 days of data, perform better with fewer prediction days; however, as the number of prediction days increases, the degree of bias toward the upside or downside increases, resulting in a significant difference from the actual data. Therefore, predictions using cumulative data reflect the general trend of the lactation curve, whereas those using only recent data can be more accurate for short-term predictions than capturing general trends. Moreover, owing to the high variability of the data, even in individual cows, the prediction performance decreased as the prediction period extended. Therefore, in this study, we used a 1-day prediction model with the highest accuracy to monitor milk yield in individual cows. This is because detecting anomalies in milk yield monitoring is more suitable through high-performance short-term predictions than through low-performance long-term predictions.

Fluctuations in daily milk production, which can be caused by various factors, are closely related to cow health (Deluyker et al., 1991). For example, cows infected with mastitis show a decrease in milk yield starting 1–2 weeks before diagnosis, and immediately after diagnosis, they exhibit a significantly reduced milk yield compared with uninfected cows, suggesting that monitoring with a focus on the decrease in milk yield is necessary (Rajala-Schultz et al., 1999; Gröhn et al., 2004). Hence, in this study, we used a 1-day prediction model with the highest accuracy to monitor milk production in individual cows. This is because detecting anomalies in milk production monitoring is more suitable through high-performance short-term predictions than through low-performance long-term predictions. We set a criterion for detecting a decline in milk yield using a 90 % confidence interval, which minimizes the error of classifying normal data as outliers compared to 95 % and 99 % confidence intervals. There was a problem in that the rapidly increasing milk yield in the early stages was detected as an outlier for exceeding the upper bound of the confidence interval, even though it was a normal condition. Therefore, it was deemed appropriate to consider data falling below the lower bound of the confidence interval as abnormal milk yield caused by cows with health issues. In addition, based on the assumption that the predicted value (1-day prediction using recent data) is lower than the expected value (milk yield trend based on cumulative data), indicating a negative difference between them, we aimed to detect anomalies in milk yield. However, approximately 40 % of the 129,945 data points from 389 individual cows showed negative values with this criterion, suggesting that cross-checking by simultaneously using two criteria (the lower bound of the confidence interval and the negative value from the difference between predictions) might be required. Because a decrease in milk yield at a single point does not necessarily indicate a health issue in cows, it is anticipated that real-time monitoring can be achieved by detecting consecutive outliers based on these multiple criteria. To implement this approach practically, obtaining milk yield data from cows with verified health issues, such as mastitis, claudicant, ketosis, feed changes, and stress, and applying a monitoring system to establish robust criteria for identifying health problems will be necessary. Addressing this challenge is a key objective for future research as this study continues to develop. While this study was conducted using data from dairy farms in South Korea, which may introduce some limitations for farms in differing environmental conditions, we emphasize that the methodology developed here—individual modeling and model-based monitoring—can be effectively applied to other dairy farms as well.

5. Conclusion

Research on individual daily milk yield prediction and the development of real-time monitoring algorithms has not been previously conducted, despite being essential elements for efficient productivity management in dairy farms and smart livestock farming. To be practically utilized in dairy farms, this study developed daily milk yield prediction models for individuals along with monitoring algorithms. In particular, this study developed high-performance prediction models using only milk yield data, without the need for various measuring devices or associated data, enabling widespread applicability in many dairy farms. Still, there is a limitation that the monitoring system was not tested for a dairy cow with health problem due to data availability and animal ethical issue, suggesting future study on actual application of the algorithm on cows suffered from mastitis and other diseases. Through this research, in the future, we will be able to classify and provide early warnings based on the causes of yield fluctuations, such as mastitis, dystocia, ketosis, feed changes, and stress, etc. when collecting various cases (labeled data) of milk yield change patterns. Consequently, the developed real-time monitoring system detects milk yield abnormalities at short intervals, allowing early prediction of abnormalities in cattle health, nutritional status, and feeding management conditions. It is expected that immediate measures such as treating diseases, changing feed, and improving breeding environments will be able to contribute to increasing productivity and profits of farms. For example, by indexing the monitoring results based on changes in milk production and the degree of outliers, a system can be established, which allows dairy farm managers to easily utilize information about the condition of the cows and milk production levels.

CRediT authorship contribution statement

Jae-Woo Song: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mingyung Lee:** Writing – review & editing, Resources, Methodology, Investigation, Formal analysis, Data curation. **Hyunjin Cho:** Writing – review & editing, Resources, Investigation, Data curation. **Dae-Hyun Lee:** Supervision, Formal analysis. **Seongwon Seo:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Wang-Hee Lee:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization.

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.

Acknowledgements

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) and Korea Smart Farm R&D Foundation (KosFarm) through Smart Farm Innovation Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) and Ministry of Science and ICT (MSIT), Rural Development Administration (RDA) (Project No. 421022-04).

Data availability

Data will be made available on request.

References

Adriaens, I., Huybrechts, T., Aernouts, B., Geerinckx, K., Piepers, S., De Ketelaere, B., Saeys, W., 2018. Method for short-term prediction of milk yield at the quarter level to improve udder health monitoring. J. Dairy Sci. 101, 10327–10336. https://doi.org/10.3168/jds.2018-14696.

Akbar, M.O., Shahbaz Khan, M.S., Ali, M.J., Hussain, A., Qaiser, G., Pasha, M., Pasha, U., Missen, M.S., Akhtar, N., 2020. IoT for development of smart dairy farming. J. Food Qual. 2020, 4242805. https://doi.org/10.1155/2020/4242805.

Ali, T.E., Schaeffer, L.R., 1987. Accounting for covariances among test day milk yields in dairy cows. Can. J. Anim. Sci. 67, 637–644. https://doi.org/10.4141/CJAS87-067.

Alonso, R.S., Sittón-Candanedo, I., García, Ó., Prieto, J., Rodríguez-González, S., 2020. An intelligent Edge-IoT platform for monitoring livestock and crops in a dairy farming scenario. Ad Hoc Netw. 98, 102047. https://doi.org/10.1016/j.adhoc.2019.102047.

Archontoulis, S.V., Miguez, F.E., 2015. Nonlinear regression models and applications in agricultural research. Agron. J. 107, 786–798. https://doi.org/10.2134/agronj2012.0506.

Bareille, N., Beaudeau, F., Billon, S., Robert, A., Faverdin, P., 2003. Effects of health disorders on feed intake and milk production in dairy cows. Livest. Prod. Sci. 83, 53–62. https://doi.org/10.1016/S0301-6226(03)00040-X.

Bokde, N.D., Milkevych, V., Nielsen, R.K., Villumsen, T.M., Sahana, G., 2023. A novel approach for anomaly detection in dairy cow gas emission records. Comput. Electron. Agric. 214, 108286. https://doi.org/10.1016/j.compag.2023.108286.

Cavero, D., Tölle, K.H., Rave, G., Buxadé, C., Krieter, J., 2007. Analysing serial data for mastitis detection by means of local regression. Livest. Sci. 110, 101–110. https://doi.org/10.1016/j.livsci.2006.10.006.

Chen, S.Y., Boerman, J.P., Gloria, L.S., Pedrosa, V.B., Doucette, J., Brito, L.F., 2023. Genomic-based genetic parameters for resilience across lactations in North American Holstein cattle based on variability in daily milk yield records. J. Dairy Sci. 106, 4133–4146. https://doi.org/10.3168/jds.2022-22754.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. 74, 829–836. https://doi.org/10.1080/01621459.1979.10481038.

Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL: a seasonal-trend decomposition. J. Off. Stat. 6, 3–73.

Daniel, J.B., Friggens, N.C., Chapoutot, P., Van Laar, H., Sauvant, D., 2016. Milk yield and milk composition responses to change in predicted net energy and metabolizable protein: a meta-analysis. Animal 10, 1975–1985. https://doi.org/10.1017/S1751731116001245.

Davis, T.C., White, R.R., 2020. Breeding animals to feed people: the many roles of animal reproduction in ensuring global food security. Theriogenology 150, 27–33. https://doi.org/10.1016/j.theriogenology.2020.01.041.

Deluyker, H.A., Gay, J.M., Weaver, L.D., Azari, A.S., 1991. Change of milk yield with clinical diseases for a high producing dairy herd. J. Dairy Sci. 74, 436–445. https://doi.org/10.3168/jds.S0022-0302(91)78189-7.

Dierckx, P., 1975. An algorithm for smoothing, differentiation and integration of experimental data using spline functions. J. Comput. Appl. Math. 1, 165–184. https://doi.org/10.1016/0771-050X(75)90034-0.

Dijkstra, J., France, J., Dhanoa, M.S., Maas, J.A., Hanigan, M.D., Rook, A.J., Beever, D.E., 1997. A model to describe growth patterns of the mammary gland during pregnancy and lactation. J. Dairy Sci. 80, 2340–2354. https://doi.org/10.3168/jds.S0022-0302(97)76185-X.

Draper, N.R., Smith, H., 1998. Applied regression analysis (Vol. 326). John Wiley & Sons. https://doi.org/10.1002/9781118625590.

Edwards, J.L., Tozer, P.R., 2004. Using activity and milk yield as predictors of fresh cow disorders. J. Dairy Sci. 87, 524–531. https://doi.org/10.3168/jds.S0022-0302(04)73192-6.

Eilers, P.H., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties. Statist. Sci. 11, 89–121. https://doi.org/10.1214/ss/1038425655.

Farooq, M.S., Sohail, O.O., Abid, A., Rasheed, S., 2022. A survey on the role of iot in agriculture for the implementation of smart livestock environment. IEEE Access 10, 9483–9505. https://doi.org/10.1109/ACCESS.2022.3142848.

Fleischer, P., Metzner, M., Beyerbach, M., Hoedemaker, M., Klee, W., 2001. The relationship between milk yield and the incidence of some diseases in dairy cows. J. Dairy Sci. 84, 2025–2035. https://doi.org/10.3168/jds.S0022-0302(01)74646-2.

Fourichon, C., Seegers, H., Bareille, N., Beaudeau, F., 1999. Effects of disease on milk production in the dairy cow: a review. Prev. Vet. Med. 041, 1–35. https://doi.org/10.1016/S0167-5877(99)00035-5.

Friggens, N.C., Chagunda, M.G., 2005. Prediction of the reproductive status of cattle on the basis of milk progesterone measures: model description. Theriogenology 64, 155–190. https://doi.org/10.1016/j.theriogenology.2004.11.014.

Friggens, N.C., Ridder, C., Løvendahl, P., 2007. On the use of milk composition measures to predict the energy balance of dairy cows. J. Dairy Sci. 90, 5453–5467. https://doi.org/10.3168/jds.2006-821.

Grossman, M., Koops, W.J., 2003. Modeling extended lactation curves of dairy cattle: a biological basis for the multiphasic approach. J. Dairy Sci. 86, 988–998. https://doi.org/10.3168/jds.S0022-0302(03)73682-0.

Halachmi, I., Guarino, M., Bewley, J., Pastell, M., 2019. Smart animal agriculture: application of real-time sensors to improve animal well-being and production. Annu. Rev. Anim. Biosci. 7, 403–425. https://doi.org/10.1146/annurev-animal-020518-114851.

Huybrechts, T., Mertens, K., De Baerdemaeker, J., De Ketelaere, B., Saeys, W., 2014. Early warnings from automatic milk yield monitoring with online synergistic control. J. Dairy Sci. 97, 3371–3381. https://doi.org/10.3168/jds.2013-6913.

Hyndman, R.J., 2011. Moving Averages. In: Lovric, M. (Ed.), International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_380.

Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting: principles and practice. OTexts.

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. J. Stat. Softw. 27, 1–22. https://doi.org/10.18637/jss.v027.i03.

Idoje, G., Dagiuklas, T., Iqbal, M., 2021. Survey for smart farming technologies: challenges and issues. Comput. Electr. Eng. 92, 107104. https://doi.org/10.1016/j.compeleceng.2021.107104.

Ismail, Z.H., Chun, A.K.K., Razak, M.I.S., 2019. Efficient herd–outlier detection in livestock monitoring system based on density–based spatial clustering. IEEE Access 7, 175062–175070. https://doi.org/10.1109/ACCESS.2019.2952912.

Jensen, D.B., van der Voort, M., Hogeveen, H., 2018. Dynamic forecasting of individual cow milk yield in automatic milking systems. J. Dairy Sci. 101, 10428–10439. https://doi.org/10.3168/jds.2017-14134.

Ji, B., Banhazi, T., Phillips, C.J., Wang, C., Li, B., 2022. A machine learning framework to predict the next month's daily milk yield, milk composition and milking frequency for cows in a robotic dairy farm. Biosyst. Eng. 216, 186–197. https://doi.org/10.1016/j.biosystemseng.2022.02.013.

Jung, J.M., Kim, D.H., Cho, H., Lee, M., Jeong, J., Lee, D.H., Seo, S., Lee, W.H., 2024. Multi-algorithmic approach for detecting outliers in cattle intake data. J. Agric. Res. 15, 101021. https://doi.org/10.1016/j.jafr.2024.101021.

Lee, W.H., Lee, M., Lee, D.H., Jung, J.M., Cho, H., Seo, S., 2023. A statistical method to standardize and interpret the activity data generated by wireless biosensors in dairy cows. J. Agric. Sci. 161, 678–685. https://doi.org/10.1017/S0021859623000576.

Lukas, J.M., Reneau, J.K., Wallace, R., Hawkins, D., Munoz-Zanzi, C., 2009. A novel method of analyzing daily milk production and electrical conductivity to predict disease onset. J. Dairy Sci. 92, 5964–5976. https://doi.org/10.3168/jds.2009-2066.

McKinney, W., 2011. pandas: a foundational Python library for data analysis and statistics. Python High Perform. Scientif. Comput. 14, 1–9.

Motulsky, H., Christopoulos, A., 2004. Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting. Oxford Univ. Press. https://doi.org/10.1093/oso/9780195171792.001.0001.

Nguyen, Q.T., Fouchereau, R., Frenod, E., Gerard, C., Sincholle, V., 2020. Comparison of forecast models of production of dairy cows combining animal and diet parameters. Comput. Electron. Agric. 170, 105258. https://doi.org/10.1016/j.compag.2020.105258.

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. https://doi.org/10.1016/j.prevetmed.2019.104786.

Pastell, M.E., Kujala, M., 2007. A probabilistic neural network model for lameness detection. J. Dairy Sci. 90, 2283–2292. https://doi.org/10.3168/jds.2006-267.

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

Press, W.H., Teukolsky, S.A., 1990. Savitzky-Golay smoothing filters. Comput. Phys. 4, 669–672. https://doi.org/10.1063/1.4822961.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available online: https://www.R-project.org.

Rajala-Schultz, P.J., Gröhn, Y.T., McCulloch, C.E., Guard, C.L., 1999. Effects of clinical mastitis on milk yield in dairy cows. J. Dairy Sci. 82, 1213–1220. https://doi.org/10.3168/jds.S0022-0302(99)75344-0.

Riaboff, L., Aubin, S., Bédère, N., Couvreur, S., Madouasse, A., Goumand, E., Chauvin, A., Plantier, G., 2019. Evaluation of pre-processing methods for the prediction of cattle behaviour from accelerometer data. Comput. Electron. Agric. 165, 104961. https://doi.org/10.1016/j.compag.2019.104961.

Rodriguez-Baena, D.S., Gomez-Vela, F.A., García-Torres, M., Divina, F., Barranco, C.D., Daz-Diaz, N., Jimenez, M., Montalvo, G., 2020. Identifying livestock behavior patterns based on accelerometer dataset. J. Comput. Sci. 41, 101076. https://doi.org/10.1016/j.jocs.2020.101076.

Savitzky, A., Golay, M.J., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047.

Seabold, S., Perktold, J., 2010. Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science. https://doi.org/10.25080/Majora-92bf1922-011.

Stankovski, S., Ostojic, G., Senk, I., Rakic-Skokovic, M., Trivunovic, S., Kucevic, D., 2012. Dairy cow monitoring by RFID. Sci. Agric. 69, 75–80. https://doi.org/10.1590/S0103-90162012000100011.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley Publishing Company, Reading, MA, USA.

UN DESA., 2022. United Nations Department of Economic and Social Affairs, Population Division. World Population Prospects 2022: Summary of Results. UN DESA/POP/2022/TR/NO. 3.

Unold, O., Nikodem, M., Piasecki, M., Szyc, K., Maciejewski, H., Bawiec, M., Dobrowolski, P., Zdunek, M., 2020. In: IoT-Based Cow Health Monitoring System. Springer International Publishing, Cham, pp. 344–356. https://doi.org/10.1007/978-3-030-50426-7_26.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.

H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Vugrin, K.W., Swiler, L.P., Roberts, R.M., Stucky-Mack, N.J., Sullivan, S.P., 2007. Confidence region estimation techniques for nonlinear regression in groundwater flow: three case studies. Water Resour. Res. 43. https://doi.org/10.1029/2005WR004804.

Wilmink, J.B.M., 1987. Adjustment of test-day milk, fat and protein yield for age, season and stage of lactation. Livest. Prod. Sci. 16, 335–348. https://doi.org/10.1016/0301-6226(87)90003-0.

Wood, P.D.P., 1967. Algebraic model of the lactation curve in cattle. Nature 216, 164–165. https://doi.org/10.1038/216164a0.

Wood, P.D.P., 1969. Factors affecting the shape of the lactation curve in cattle. Anim. Sci. 11, 307–316. https://doi.org/10.1017/S0003356100026945.

Zhang, W., Yang, K., Yu, N., Cheng, T., Liu, J., 2020. Daily milk yield prediction of dairy cows based on the GA-LSTM algorithm. In: 2020 15th IEEE International Conference on Signal Processing. ICSP, 1, pp. 664–668. https://doi.org/10.1109/ICSP48669.2020.9320926.