Optimized machine learning–collaborative filtering model for mastitis prediction in dairy cows

Jingzhu Wu[1], Yutong Liu[1], Yongjun Zheng[2,3*], Xiyan Yuan[1], Haoyu Wang[2], Shenghui Yang[2], Ning Guan[4,5], Xiaoyan Pei[4,5], Shu Li[6], Congming Wu[7*]

(1. *Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing 100048, China*;
2. *College of Engineering, China Agricultural University, Beijing 100083, China*;
3. *State Key Laboratory of Veterinary Public Health and Safety, Beijing 100193, China*;
4. *National Technology Innovation Center for Dairy, Hohhot 010110, China*;
5. *Inner Mongolia Yili Industrial Group Co. Ltd., Hohhot 010110, China*;
6. *Optimization of Inner Mongolia Animal Husbandry Co. Ltd., Hohhot 010000, China*;
7. *College of Veterinary Medicine, China Agricultural University, Beijing 100193, China*)

Abstract: Mastitis is a major disease affecting dairy cow health and milk production. This study established an integrated machine learning (ML) model combining herd- and individual-level data to achieve efficient and balanced prediction of clinical mastitis. Data were collected from 5284 lactating Holstein cows on two farms in southern and northern China. Five feature processing methods—recursive feature elimination (RFE), contrastive learning (CL), slopes and intercept, milk-conductivity ratio, and differences—were evaluated with four ML algorithms: Support vector machine (SVM), random forest (RF), XGBoost, and backpropagation neural network (BPNN). Among them, the XGBoost model with the milk-conductivity ratio feature achieved the best performance, with a sensitivity of 0.81 and specificity of 0.75. To further address the imbalance between sensitivity and specificity, collaborative filtering (CF) was introduced into the XGBoost model to incorporate both herd and individual cow information. The resulting XGBoost–CF model improved sensitivity to 0.83 and specificity to 0.87, enhancing the model's ability to identify both healthy and diseased cows. This integrated ML–CF framework provides an effective strategy for early mastitis prediction, offering practical support for intelligent dairy herd management and precision livestock farming.

Keywords: mastitis prediction, machine learning, feature selection, XGBoost, collaborative filtering
DOI: 10.25165/j.ijabe.20261901.10304

Citation: Wu J Z, Liu Y T, Zheng Y J, Yuan X Y, Wang H Y, Yang S H, et al. Optimized machine learning–collaborative filtering model for mastitis prediction in dairy cows. Int J Agric & Biol Eng, 2026; 19(1): 21–25.

1 Introduction

Mastitis is one of the three major diseases in modern dairy farming, causing inflammation of the mammary gland and leading to economic losses through reduced milk yield, inferior milk quality, and increased culling rates[1-4]. Current diagnostic methods, mainly based on visual inspection and milk physicochemical analysis[5-6], are subjective and inefficient, making them unsuitable for large-scale farms. Therefore, developing accurate and automated mastitis prediction models is essential for modern dairy herd management.

With the development of automatic milking systems, a large amount of real-time data can be used for disease prediction through ML[7-9]. Satola et al.[10] built ensemble ML models to identify subclinical mastitis with high sensitivity and specificity. Pakrashi et al.[11] developed a model to predict subclinical mastitis within seven days, reaching a sensitivity of 69.45% and specificity of 95.64%. Luo et al.[12] compared decision trees, RF, BPNN, and SVM, reporting decision trees performed best. Ebrahimi et al.[13] compared seven ML algorithms using milk conductivity data, noting high sensitivity but low specificity. Shi et al.[14] and Li et al.[15] confirmed that factors such as season, parity, and lactation stage influence mastitis prediction outcomes.

These studies confirm ML's potential but also highlight imbalanced sensitivity and specificity in existing models[16-18]. Moreover, individual variability among cows—such as differences in milk yield and physiological status—remains underexplored. To address these limitations, this study collected conductivity and milking data from 5284 Holstein cows on farms in northern and southern China, evaluated five feature processing methods with four ML algorithms, and developed an XGBoost–collaborative filtering (CF) hybrid model integrating herd- and individual-level

Received date: 2025-10-30 **Accepted date:** 2026-01-20
Biographies: Jingzhu Wu, Professor, research interest: smart sensing and food computing, Email: pubwu@163.com; **Yutong Liu**, Postgraduate, research interest: machine learning, Email:15765845833@163.com; **Xiyan Yuan**, Postgraduate, research interest: machine learning, Email: yuanxy1039@163.com; **Haoyu Wang**, Postgraduate, research interest: machine learning, Email: wangh@cau.edu.cn; **Shenghui Yang**, Professor, research interest: agricultural intelligent sensing and smart decision-making, Email: yshgxy@cau.edu.cn; **Ning Guan**, Engineer, research interest: research and development of dairy testing technology, Email: guanning1@yili.com; **Xiaoyan Pei**, Engineer, research interest: research and development of dairy testing technology, Email: peixiaoyan@yili.com; **Shu Li**, Engineer, research interest: research and development of dairy testing technology, Email: lishu_2010@126.com.
***Corresponding author: Yongjun Zheng**, Professor, research interest: agricultural engineering. College of Engineering, China Agricultural University, No.17 Qinghua East Road, Haidian District, 100083 Beijing, China. Tel: 010-62736385, Email: zyj@cau.edu.cn; **Congming Wu**, Professor, research interest: veterinary pharmacology and toxicology. College of Veterinary Medicine, China Agricultural University, No. 2 Yuanmingyuan West Road, Haidian District, 100193 Beijing, China. Tel: 010-62733378, Email: wucm@cau.edu.cn.

information for balanced and interpretable mastitis prediction.

2 Materials and methods

2.1 Data collection and preprocessing

Conductivity data were collected from two large-scale dairy farms in northern and southern China between January and December 2023, covering 5284 lactating Holstein cows (3730 northern, 1554 southern) with 1–4 samples per cow per day, totaling 3 441 182 records. Features included cow ID, parity, days in lactation, milk yield, conductivity, milk flow rates at 15-30 s, 30-60 s, and 60-120 s, average and peak flow rates, early milk yield, and milking duration. Clinical mastitis diagnosis was determined daily by farm veterinarians. The diagnosis was confirmed if at least one of the following signs was observed: (1) Abnormal milk: Visible changes such as clots, flakes, or discoloration; (2) Udder inflammation: Signs of swelling, heat, hardness, or pain upon palpation; (3) Systemic reactions: In severe cases, systemic signs such as elevated body temperature.

To improve data quality, preprocessing involved removing missing or non-positive values, filtering cows with parity 1–5 and lactation days 5–365, and eliminating outliers using the boxplot method. To determine the optimal prediction window, this study performed a comparative analysis on time windows of 0, 10, 20, and 30 days. Pearson correlation coefficient and mutual information entropy were utilized to evaluate the relationship between features and the mastitis label across these windows. The analysis revealed that feature correlations stabilized and reached a peak at the 20-day mark. Consequently, a 20-day window was selected for feature construction to achieve the best balance between information retention and noise reduction.

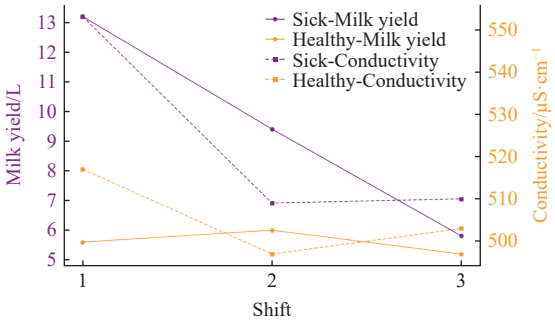
2.2 Feature processing and dataset construction

Five feature processing methods were applied to evaluate their impact on predictive performance.

1) RFE iteratively eliminates less important features by training the model until a predefined number of features is retained[19].

2) CL reconstructs features by learning data representations that maximize similarity among similar samples while minimizing similarity among dissimilar ones[20].

3) Slopes and intercept capture daily fluctuations in milk yield and conductivity through least-squares fitting (Figure 1).



Note: Figure compares healthy and diseased cows across three daily milking shifts, showing that diseased cows exhibit stronger fluctuations in milk yield and conductivity.

Figure 1 Variations in three daily milking shifts for healthy and diseased cows

4) Milk-conductivity ratio accounts for the fact that clinical mastitis often causes a significant decrease in milk yield[21], while milk yield is also highly influenced by factors such as climate, diet, and lactation stage[22]. By combining milk yield with the more stable conductivity, this ratio stabilizes the feature and improves prediction performance.

5) Differences capture short-term changes in milk yield and conductivity, as mastitis affects these values. Fan et al.[23] indicated that using milking data from the past seven sessions yields optimal predictive results, while Bonestroo et al.[24] found little difference when reducing the input from 30 to 15 days. These findings confirm that incorporating multiple historical data points improves model accuracy. Since the original dataset lacks temporal features, this experiment constructs differences in milk yield and conductivity between the current day and the previous one or two days.

After constructing new features using the above five methods, a balanced dataset was generated by random undersampling. The original raw dataset (within the 20-day window) exhibited a severe class imbalance, containing 1 840 924 healthy samples compared to only 22 765 diseased samples (an approximate ratio of 80:1). This extreme disparity necessitated the undersampling strategy to prevent model bias. 80% of samples were used for training, 20% for testing, and remaining healthy data were retained as a historical dataset (Table 1).

Table 1 Sample distribution under different feature processing methods

Feature processing method	Training set	Test set	Historical dataset
RFE	36 424	9106	1 818 159
CL	26 800	8960	8960
Slopes and intercept	12 334	3084	628 760
Milk-conductivity ratio	36 424	9106	1 818 159
Differences	7354	1838	397 653

2.3 ML methods

To comprehensively compare and evaluate the impact of new features on model performance, four ML models were investigated: SVM, RF, XGBoost, and BPNN. All models were optimized using Bayesian Optimization (specifically the HyperBand algorithm) to efficiently navigate the high-dimensional parameter space, and 5-fold cross-validation was applied. The optimal hyperparameters for the best-performing XGBoost model are listed in Table 2. The experiments were implemented in Python 3.12.

Table 2 Optimal hyperparameters for the XGBoost model

Hyperparameter	Value	Description
n_estimators	300	Number of gradient boosted trees
max_depth	10	Maximum tree depth for base learners
learning_rate	0.1	Boosting learning rate (step size)
subsample	0.9	Subsample ratio of the training instances
colsample_bytree	0.9	Subsample ratio of columns when constructing each tree

2.4 ML-CF Model

Herd-level ML models predict cow diseases using data from multiple individuals but often suffer from false positives due to natural individual variability. To address this issue, a hybrid ML–CF model was developed, integrating herd-level features with individual cow information for more personalized mastitis prediction.

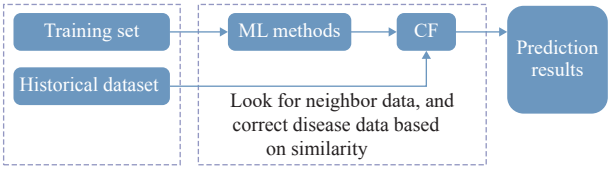
CF predicts unknown outcomes based on the behavior of similar entities. When applied to mastitis prediction, CF treats records of the same cow at different time points as distinct samples and adjusts ML outputs using their temporal similarity. Specifically, to quantify the similarity between the feature vector of the current prediction instance (A) and a historical healthy instance (B), this study employed the cosine similarity metric. The calculation

formula is defined as follows:

$$sim(A,B)=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\times\sqrt{\sum_{i=1}^{n}B_i^2}} \quad (1)$$

where, n represents the dimension of the feature vector.

The overall workflow of the ML–CF model is illustrated in Figure 2.



Note: The model first generates initial predictions using machine learning algorithms. These predictions are then recalibrated using collaborative filtering (CF) based on individual cow historical data, enhancing personalized prediction accuracy and reducing false-positive rates.

Figure 2 Workflow of ML–CF model

2.5 Evaluation metrics

Model performance was evaluated using sensitivity, specificity, precision, F1-score, and Matthews correlation coefficient (MCC), which together provide a comprehensive assessment of classification accuracy and balance. The calculation formulas for each metric are as follows:

$$\text{Sensitivity}=\frac{\text{TP}}{(\text{TP}+\text{FN})} \quad (2)$$

$$\text{Specificity}=\frac{\text{TN}}{(\text{TN}+\text{FP})} \quad (3)$$

$$\text{Precision}=\frac{\text{TP}}{(\text{TP}+\text{FP})} \quad (4)$$

$$\text{F1}=2\times\frac{\text{Precision}\times\text{Sensitivity}}{\text{Precision}+\text{Sensitivity}} \quad (5)$$

$$\text{MCC}=\frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \quad (6)$$

where, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Diseased samples were defined as positive and healthy samples as negative.

3 Results and discussion

3.1 Feature processing and model performance

Five feature processing methods were combined with four ML models to evaluate classification performance. For each combination, feature importance ranking was applied, and less informative variables were removed to optimize model performance.

Overall, ensemble-based models (RF and XGBoost) consistently achieved higher sensitivity, specificity, and F1-scores than SVM and BPNN, demonstrating stronger generalization capability. Specifically, the comparative analysis highlighted a critical trade-off: while BPNN achieved competitive sensitivity, it suffered from significantly lower specificity compared to tree-based ensemble methods. This indicates that BPNN was more prone to false positives on the imbalanced tabular dataset, whereas XGBoost provided a superior balance between detecting sick cows and minimizing false alarms. Among the feature processing methods, the milk–conductivity ratio yielded the best overall performance, improving sensitivity to 0.81 and specificity to 0.73. Further analysis of feature importance (based on random forest rankings) revealed that, in addition to the engineered milk-conductivity ratio, the top-contributing raw features were average flow rate and milk yield. This aligns with physiological expectations, as mastitis infection typically leads to a significant reduction in milk yield and alterations in flow dynamics due to inflammation and pain. However, an imbalance between sensitivity and specificity persisted due to individual cow variability (Figure 3).

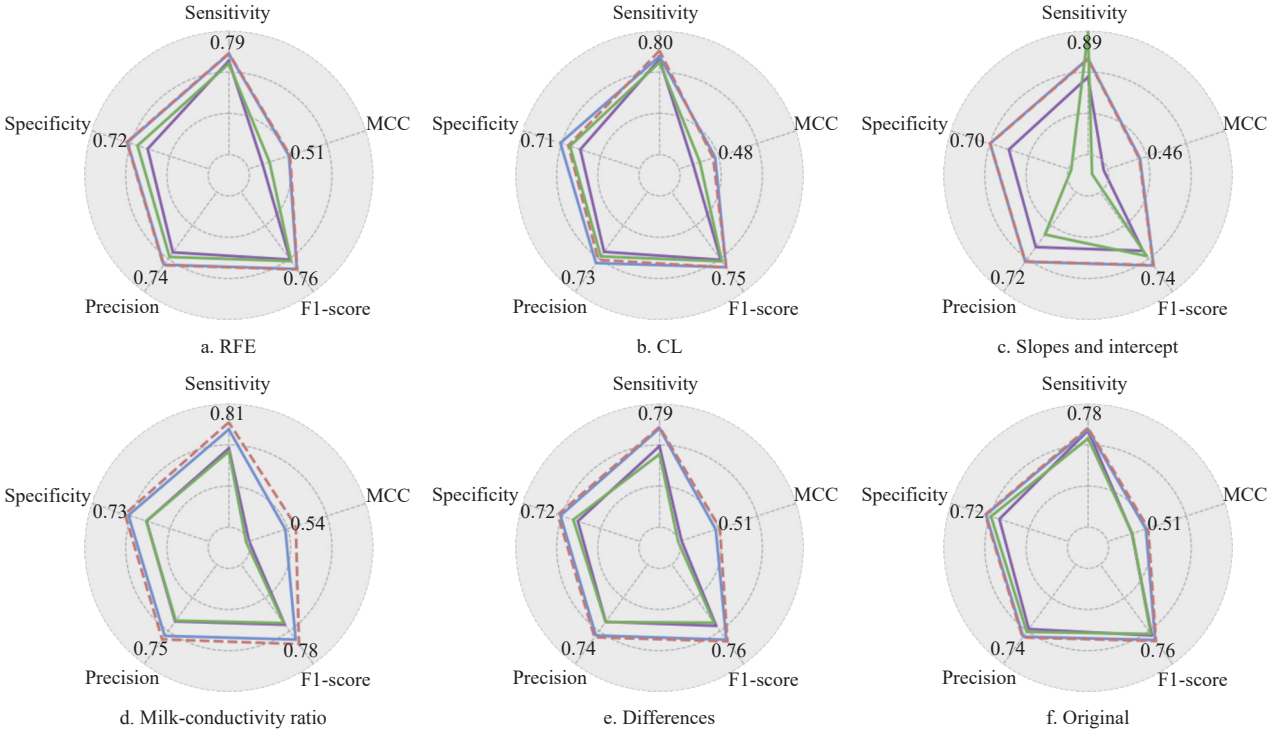
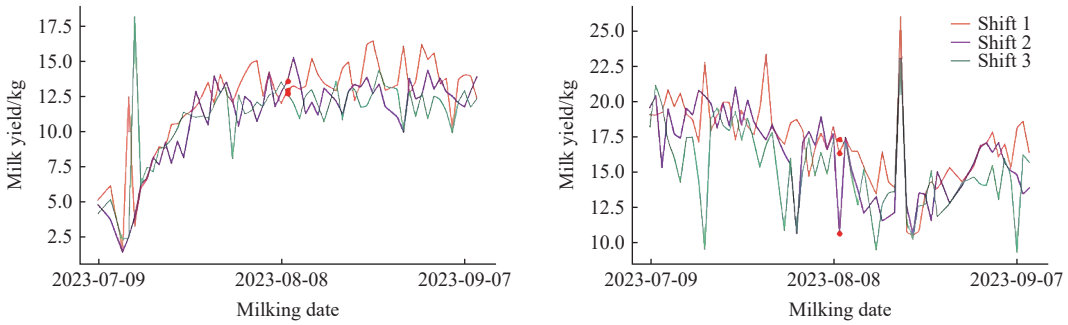


Figure 3 Comparison of test set prediction results under different feature processing methods

3.2 Individual variation and model limitations

Analysis of individual cow data in the dataset revealed substantial inter-individual variability among cows. For example, two cows were randomly selected from those diagnosed with mastitis by a veterinarian on August 9, 2023. The trends of milk yield within one month before and after diagnosis were plotted Figure 4, demonstrating clearly distinct production ranges. Similar variability was observed across other physiological and milking characteristics, which may lead to false-positive predictions and reduced model specificity. These findings underscore the necessity of incorporating personalized information into predictive models.



a. Cow A for different shifts of milk b. Cow B for different shifts of milk

Figure 4 Individual cow analysis at the dairy farm

3.3 XGBoost-CF model results

To address low specificity caused by individual variation, a collaborative filtering (CF) mechanism was integrated into XGBoost, forming the XGBoost–CF hybrid model. Initial predictions by XGBoost were recalibrated using each cow's historical data, reducing misclassifications. The XGBoost–CF model achieved a specificity of 0.87, an F1-score of 0.85, and an MCC of 0.70 (Table 3), demonstrating substantial improvement over traditional ML models and confirming the importance of combining herd-level and individual-level information.

Table 3 XGBoost-CF model test set results

Model	Sensitivity	Specificity	Precision	F1	MCC
XGBoost	0.81	0.75	0.75	0.78	0.54
XGBoost-CF	0.83	0.87	0.86	0.85	0.70

4 Conclusions

This study evaluated multiple feature processing methods and ML models for predicting clinical mastitis in dairy cows. To mitigate the imbalance between sensitivity and specificity caused by individual cow differences, ML models were integrated with CF. Data were collected from 5284 lactating Holstein cows on two large-scale farms in northern and southern China and preprocessed following veterinarian guidance. Five feature processing methods were applied to reconstruct the dataset, and models including RF, SVM, XGBoost, and BPNN were developed. The best-performing XGBoost model employed the milk–conductivity ratio feature construction and was combined with CF to form the final prediction framework.

Results indicated that the milk–conductivity ratio substantially enhanced model performance. Among the four ML models, XGBoost achieved superior results with a sensitivity of 0.81, specificity of 0.75, accuracy of 0.75, F1-score of 0.78, and MCC of 0.54. Incorporating CF further improved performance, increasing sensitivity to 0.83 and specificity to 0.87, thereby balancing the identification of both healthy and diseased cows.

Overall, this work provides a robust and interpretable ML–CF framework for early detection of clinical mastitis, offering practical guidance for precision dairy herd management. Future studies should extend the model to multi-farm datasets, integrate additional physiological and environmental variables, and explore real-time deployment in intelligent milking systems.

Acknowledgements

This research was supported by the National Key Research and Development Project of China (Grant No. 2016YFD0501304) and China Agriculture Research System of MOF and MARA (Grant No. CARS-36). The authors sincerely thank the veterinarians and staff of the participating dairy farms in southern and northern China for their assistance in data collection and expert guidance during data preprocessing.

[References]

[1] Chu M Y, Liu X W, Zeng X T, Wang Y C, Liu G. Research advances in the automatic detection technology for mastitis of dairy cows. Trans. Chin. Soc. Agric. Eng., 2023; 39(11): 1–12. (in Chinese)

[2] Chen L J. Cow mastitis and scientific prevention and control measures. China Anim. Health, 2023; 25: 40–41. (in Chinese)

[3] Ma R J, Du L, Ma W D, Zhao J H, Li Q C, Lei C Z, et al. Study on the general situation and prevention and control measures of subclinical mastitis. China Cattle Sci., 2023; 49(4): 47–50. (in Chinese)

[4] Ye W, Ma Z, Yu Y, Han B. Incidence status of mastitis in dairy cows and its prevention and treatment measures in China. Chin. J. Anim. Sci., 2023; 59(9): 343–348. (in Chinese)

[5] Wang A H, Yang L F. Causes, clinical symptoms, diagnosis and treatment of cow mastitis. Mod. Anim. Husb. Sci. Technol., 2023(10): 94–96. (in Chinese)

[6] Zhang Y, Shi Q, Zhou Q M, Feng W Y, Xu X, Wu X. Isolation, identification, drug sensitivity and pathogenicity of pathogenic bacteria in dairy cow mastitis. Heilongjiang Anim. Sci. Vet. Med., 2020; (23): 85–88, 167–168. (in Chinese)

[7] Liebe D M, Steele N M, Petersson-Wolfe C S, De Vries A, White R R. Practical challenges and potential approaches to predicting low-incidence diseases on farm using individual cow data: A clinical mastitis example. J. Dairy Sci., 2022; 105(3): 2369–2379.

[8] Naqvi S A, King M T M, Matson R D, Devries T J, Deardon R, Barkema H W. Mastitis detection with recurrent neural networks in farms using automated milking systems. Comput. Electron. Agric., 2022; 192: 106618.

[9] Tian H, Zhou X J, Wang H, Xu C, Zhao Z X, Xu W, et al. The prediction of clinical mastitis in dairy cows based on milk yield, rumination time, and milk electrical conductivity using machine learning algorithms. Animals, 2024; 14(3): 427.

[10] Satola A, Satola K. Performance comparison of machine learning models used for predicting subclinical mastitis in dairy cows: bagging, boosting, stacking, and super-learner ensembles versus single machine learning

models. J. Dairy Sci., 2024; 107(6): 3959–3972.
[11] Pakrashi A, Ryan C, Guéret C, Berry D P, Corcoran M, Keane M T, et al. Early detection of subclinical mastitis in lactating dairy cows using cow-level features. J. Dairy Sci., 2023; 106(7): 4978–4990.
[12] Luo W K, Dong Q, Feng Y. Risk prediction model of clinical mastitis in lactating dairy cows based on machine learning algorithms. Prev. Vet. Med., 2023; 221: 106059.
[13] Ebrahimi M, Mohammadi-Dehcheshmeh M, Ebrahimie E, Petrovski K R. Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep learning and gradient-boosted trees outperform other models. Comput. Biol. Med., 2019; 114: 103456.
[14] Shi Y L, Li W L, Tang Y J, Mi S Y, Xiao W, Liu L, et al. Studies on risk-assessment-model establishment and prediction of mastitis in Chinese Holstein cattle. Chin. J. Anim. Sci., 2021; 57(3): 84–90. (in Chinese)
[15] Li W L, Zhao T T, Da R, Shi Y L, Guo G, Wang Y C, et al. Application and optimization of dairy cow mastitis risk assessment system in Chinese Holstein. Chin. J. Anim. Sci., 2021; 57(10): 65–72.
[16] Bobbo T, Biffani S, Taccioli C, Penasa M, Cassandro M. Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. Sci. Rep., 2021; 11: 13642.
[17] Ozella L, Brotto R K, Forte C, Giacobini M. A literature review of modeling approaches applied to data collected in automatic milking systems. Animals, 2023; 13(12): 1916.
[18] Zhou X J, Xu C, Wang H, Xu W, Zhao Z X, Chen M X, et al. The early prediction of common disorders in dairy cows monitored by automatic systems with machine learning algorithms. Animals, 2022; 12(10): 1251.
[19] Zhou X Z, Wen H J, Zhang Y L, Xu J H, Zhang W G. Landslide susceptibility mapping using hybrid random forest with Geo Detector and RFE for factor optimization. Geosci. Front., 2021; 12(5): 101211.
[20] Zhang C S, Chen J, Li Q L, Deng B Q, Wang J, Chen C G. Deep contrastive learning: A survey. Acta Autom. Sin., 2023; 49(1): 15–39.
[21] Sun Y, Zhou G Y, Wu T B, Li Y L, Ji S Q, Zhang T. Recent research progress of cow mastitis in China. China Dairy, 2022(4): 43–51. (in Chinese)
[22] Zhai Y, Zhou B, Zhou F Z, Dai X, Liang Y, Zhang H R, et al. Analysis of factors affecting milk yield, conductivity, and activity level in Holstein cows. Chin. J. Anim. Sci., 2024; 60(6): 148–153. (in Chinese)
[23] Fan X, Watters R D, Nydam D V, Virkler P D, Wieland M, Reed K F. Multivariable time series classification for clinical mastitis detection and prediction in automated milking systems. J. Dairy Sci., 2023; 106(5): 3448–3464.
[24] Bonestroo J, van der Voort M, Hogeveen H, Emanuelson U, Klaas I C, Fall N. Forecasting chronic mastitis using automatic milking system sensor data and gradient-boosting classifiers. Comput. Electron. Agric., 2022; 198: 107002.