Theriogenology

RLST-KNN: An Efficient Machine Learning Method for Prediction of Subclinical Ketosis of Dairy Cows Based on Imbalanced Data Processing Algorithm

--Manuscript Draft--

Manuscript Number:	
Article Type:	Original Research Article
Keywords:	Machine Learning; Dairy Cows; Prediction; Subclinical Ketosis; Imbalanced Data
Corresponding Author:	Yan Feng Northwest A&F University Xianyang, Shaanxi CHINA
First Author:	SHENGQUAN HU
Order of Authors:	SHENGQUAN HU
	Zhao Zhang
	zefeng Li
	Qiang Dong
	Yan Feng
Abstract:	Subclinical ketosis in dairy cows is one of the most common and prominent metabolic diseases affecting dairy production. Subclinical ketosis in dairy cows can cause loss of appetite, metabolic issues, and reduced milk production, leading to malnutrition and economic losses for producers. To reduce losses on farms, the development of a ketosis early prediction method using machine learning algorithms has become a research hotspot in recent years. However, In the process of using machine learning algorithms to establish a ketosis early prediction method, the issue of the data imbalance affecting the performance of methods needs to be addressed. To solve the problem, the paper proposed RLST-KNN method to establish a dairy cow ketosis prediction method. This method firstly utilized the Random Forest-Local Outlier Factor (RF-LOF) algorithm for imputing missing values. Then, the RLST-KNN method applied the Synthetic Minority Over-sampling Technique with the Tomek Links (SMOTETomeklinks) algorithm to enhance minority class data and achieve data balance. Finally, it used K-Nearest Neighbors (KNN) to predict subclinical ketosis. To verify the predictive performance of the RLST-KNN method this article compared the performance differences in ketosis prediction of five classifiers: logistic regression (LR), linear discriminant analysis (LDA), K-nearest neighbors (KNN), support vector machine (SVM), and naive Bayes (NB), both before and after balancing the dataset. We found that KNN had the best performance among the five classifiers. The experimental results indicated that the RLST-KNN algorithm performs excellently in predicting subclinical ketosis in dairy cows, achieving accuracy (ACC), F1-score, sensitivity (Sens), positive predictive value (PPV), negative predictive value (NPV), and AUC scores of 0.7501, 0.7486, 0.8946, 0.6471, 0.6436, 0.8961, and 0.8727, respectively. In addition, the RLST-KNN method achieved the highest performance in the early lactation period (three weeks postpartum). It demonstrates that RLST-KNN can predict ketosis in dairy cows during the peak period of subclinical ketosis occurrence.

Theriogenology
Author Agreement Form

This Form should be signed by all authors OR by the corresponding (or senior) author who can vouch for all co-authors. A scanned copy of the completed Form may be submitted online.

The authors confirm the following statements:

1. that there has been no duplicate publication or submission elsewhere of this work
2. that all authors have read and approved the manuscript, are aware of the submission for publication and agree to be listed as co-authors.

Author Name	**Signature**
Shengquan Hu.	Shengquan Hu
Zhao Zhang	Zhao Zhang
Ze feng Li	Zefeng Li
Qiang Dong	
Yan Feng	Yanfeng

Highlights

- A machine learning method based on data imbalance algorithm was proposed to predict subclinical ketosis in dairy cows.
- RF-LOF method is a new approach for the effective imputation of missing values
- RLST-KNN method is proposed to predict imbalanced subclinical ketosis data in dairy cows.
- RLST-KNN method achieves accurate prediction before the peak period of subclinical ketosis in dairy cows.

Date: 5-4-2025

The Editor Chief of the journal ‘*Theriogenology*’

Subject: Submission of manuscript.

Dear Editor,

We hereby submit the manuscript titled "RLST-KNN: An Efficient Machine Learning Method for Prediction of Subclinical Ketosis of Dairy Cows Based on Imbalanced Data Processing Algorithm" for review and potential publication in the journal ‘*Theriogenology*’.

The reasons for submitting this manuscript are as follows: This paper constructs an effective method for predicting subclinical ketosis in dairy cows based on imbalance data processing algorithms and improved missing value imputation methods.

Subclinical ketosis in dairy cows is a major metabolic disease that significantly impacts dairy production and is one of the most common issues in dairy cows. It can cause malnutrition and reduced milk output, leading to economic losses for producers. To mitigate these losses, developing a subclinical ketosis early prediction method using machine learning algorithms has become a key area of research. However, there are still some challenges like lacking good strategies for missing values imputing and data imbalance. To address these issues, this paper proposed the RLST-KNN method for predicting subclinical ketosis of dairy cows.

The RLST-KNN method utilizes the Random Forest and Local Outlier Factor (RF-LOF) algorithm to impute in missing values, applies the Synthetic Minority Over-sampling Technique with Tomek Links (SMOTETomeklinks) algorithm to achieve data balance, and finally uses the K-Nearest Neighbors (KNN) algorithm to classify diseased and healthy cows. The experimental results indicate that our proposed RLST-KNN method can achieve accuracy (ACC), F1-score, sensitivity (Sens), positive predictive value (PPV), negative predictive value (NPV), and AUC scores of 0.7501, 0.7486, 0.8946, 0.6436, 0.8961, and 0.8727, respectively.

We believe that the results and conclusions described in the current manuscript contribute significantly to further research on subclinical ketosis prediction, enabling farms to accurately predict ketosis in dairy cows. Therefore, the research will be of great interest to the readers of '*Theriogenology*'. We are pleased to submit this manuscript and request a review for potential publication in your esteemed journal.

With best regards,

Sincerely Yours,

Corresponding author:

Qiang Dong(ardour@126.com)

Yan Feng(yanfeng52@126.com)

RLST-KNN: An Efficient Machine Learning Method for Prediction of Subclinical Ketosis of Dairy Cows Based on Imbalanced Data Processing Algorithm

Abstract

Subclinical ketosis in dairy cows is one of the most common and prominent metabolic diseases affecting dairy production. Subclinical ketosis in dairy cows can cause loss of appetite, metabolic issues, and reduced milk production, leading to malnutrition and economic losses for producers. To reduce losses on farms, the development of a ketosis early prediction method using machine learning algorithms has become a research hotspot in recent years. However, In the process of using machine learning algorithms to establish a ketosis early prediction method, the issue of the data imbalance affecting the performance of methods needs to be addressed. To solve the problem, the paper proposed RLST-KNN method to establish a dairy cow ketosis prediction method. This method firstly utilized the Random Forest-Local Outlier Factor (RF-LOF) algorithm for imputing missing values. Then, the RLST-KNN method applied the Synthetic Minority Over-sampling Technique with the Tomek Links (SMOTETomeklinks) algorithm to enhance minority class data and achieve data balance. Finally, it used K-Nearest Neighbors (KNN) to predict subclinical ketosis. To verify the predictive performance of the RLST-KNN method this article compared the performance differences in ketosis prediction of five classifiers: logistic regression (LR), linear discriminant analysis (LDA), K-nearest neighbors (KNN), support vector machine (SVM), and naive Bayes (NB), both before and after balancing the dataset. We found that KNN had the best performance among the five classifiers. The experimental results indicated that the RLST-KNN algorithm performs excellently in predicting subclinical ketosis in dairy cows, achieving accuracy (ACC), F1-score, sensitivity (Sens), positive predictive value (PPV), negative predictive value (NPV), and AUC scores of 0.7501, 0.7486, 0.8946, 0.6436, 0.8961, and 0.8727, respectively. In addition, the RLST-KNN method achieved the highest performance in the early lactation period (three weeks postpartum). It demonstrates that RLST-KNN can predict ketosis in dairy cows during the peak period of subclinical ketosis occurrence.

Keywords: Machine Learning; Dairy Cows; Prediction; Subclinical Ketosis; Imbalanced Data

1. Introduction

Ketosis can have many harmful effects on dairy cows, such as reducing their appetite (Melendez and Serrano, 2024), lowering the first breeding rate (Rutherford et al., 2016), and increasing the likelihood of cows developing fatty liver disease (Yang et al., 2019). Therefore, as one of the most common metabolic disorders in dairy cows during the early lactation period, early detection and prevention of this disease are extremely important (Guliński, 2021).

Traditional methods for detecting ketosis have mainly determined whether dairy cows have ketosis by detecting ketosis-related substances in their blood, urine, or breath (Del Caño et al., 2023). Lei(Lei and Simões, 2021) pointed out that measuring the concentration of β-hydroxybutyrate in the blood is an effective method for the diagnosis of ketosis. Zhang (Zhang et al., 2021) found that there were various differences in urinary metabolites between cows with ketosis and healthy cows, such as higher levels of 3-hydroxybutyrate and acetone in the urine of cows with ketosis.

Qiao(Qiao et al., 2014) measured the concentration of acetone in the breath using gas chromatography-mass spectrometry and found that the concentration of acetone in the breath of dairy cows was significantly correlated with the concentration of ketone bodies in the blood and urine.

Ketosis can be divided into clinical and subclinical two types. Compared to dairy cows with clinical ketosis, cows with subclinical ketosis lack obvious symptoms, such as dry feces, foamy milk, and light-yellow urine (Huang et al., 2024). As a result, it is difficult to accurately detect subclinical ketosis using traditional methods. Additionally, there is a lag in predicting subclinical ketosis in cows. Although farmers can collect physiological data from cows on the farm, the analysis takes some time. Based on the physiological indicators and behavioral data of dairy cows, a subclinical ketosis prediction method can be established to proactively detect subclinical ketosis in cows, thereby alleviating the difficulties faced by farmers and scholars.

In recent years, scholars have gradually developed methods using machine learning to predict subclinical ketosis in dairy cows. Mandujano (Mandujano Reyes et al., 2021) proposed a detection model based on a full model selection approach with regression trees that can predict metabolic disorders in early transition dairy cows, which can provide guidance for using machine learning methods for subclinical ketosis prediction. Ferreira (Ferreira et al., 2024) compared the applications of three data fusion techniques(early fusion, late fusion, and cooperative learning) in the early detection of subclinical ketosis in dairy cows, and developed a real-time cloud computing system for detecting subclinical ketosis in dairy cows based on these three data fusion techniques. Wang (Wang et al., 2023) applied five machine learning algorithms (Extreme X-Boost, SVM, RF, KNN, and Artificial Neural Network) to a subclinical ketosis dataset with six indicators (parity, body condition score, dystocia score, daily rumination time, daily activity, and calving season) to predict the risk of subclinical ketosis. Satoła (Satoła and Bauer, 2021) developed a support vector classification (SVC) model that performed best at specific β-hydroxybutyrate (BHB) concentration thresholds, demonstrating the strong potential of SVC as a tool for detecting subclinical ketosis. Bauer (Bauer and Jagusiak, 2022) proposed a dairy cow subclinical ketosis detection model based on a Multilayer Perceptron network, which determines whether a cow has ketosis by analyzing the levels of BHB, Angiotensin-Converting Enzyme, and lactose in milk, as well as the ratio of fat to protein. They also proposed a Predictive Model Markup Language that could be used to describe the learning set, algorithms used in data mining applications, and related information.

However, there are currently no studies that focus on the imbalance phenomenon in the ketosis data of dairy cows. Imbalanced data refers to a situation where the class sizes in the dataset differ proportionally by a considerable margin. The presence of imbalanced data can negatively impact the training and prediction of machine learning models, as the model may tend to predict the category with a larger number of samples, leading to poor predictive performance for the minority class (Kaur et al., 2019). Therefore, to decrease the reduction in accuracy of the dairy cow subclinical ketosis prediction method caused by imbalanced data, this paper made the following contributions:

（1）A new missing value imputation method called the RF-LOF method was proposed, which could address the issue of traditional RF algorithms being easily influenced by outliers.

（2）A specialized prediction method for subclinical ketosis in dairy cows called RLST-KNN was proposed, which addressed the shortcomings of traditional machine learning algorithms in their inability to accurately predict the negative class in imbalanced datasets for subclinical ketosis in dairy cows. Additionally, it overcomes the issue of traditional ketosis prediction methods being unable to effectively handle missing values.

（3）Through sensitivity analysis, it was verified that the RLST-KNN method could effectively achieve ketosis prediction before the onset of the high-risk period for ketosis in dairy cows.

2. Material and methods

2.1 Dataset

The original data were collected from the cow mastitis database of Afimilk (China) Agricultural Technology CO., Ltd. The herd contains more than 9000 Holstein dairy cows and is located in Dali County, Shaanxi Province (34°40’27” N, 110°7’34”E) in China. All cows are housed in a free-stall barn and fed a total mixed ration (TMR). The dataset includes 152,768 records of 5456 Holstein cows from February 2020 to March 2022, which contains 858 cows suffered from subclinical ketosis and 4,598 healthy cows. Cows suffered from subclinical ketosis have been set as the positive class and healthy cows as the negative class. The sample numbers of the negative class are significantly larger than that of the positive class, thus we consider this dataset to be imbalanced (Michelucci, 2024).

The dataset contains 24 features (5 numerical attributes and 19 categorical attributes), including number, lactation period, days in milk, number of ketosis episodes, milk yield on day 1-15 postpartum (record per day), first fat percentage, first protein percentage, first somatic cell count, first urea nitrogen level and days to first dairy herd improvement (DHI) postpartum.

2.2 Introduction to Indicators

Postpartum milk yield, fat percentage, protein percentage, somatic cell count (SCC), and urea nitrogen level are all indicators used to diagnose subclinical ketosis in dairy cows. Jeong (Jeong et al., 2018) found that cows with subclinical ketosis usually have lower milk production. Yang (Yang et al., 2019) found that cows with subclinical ketosis typically had a higher fat percentage and a lower protein percentage by collecting samples of plasma, milk, and feces. Cascone (Cascone et al., 2022) collected data from 1,588 lactating cows across 22 farms and found a high correlation among the subclinical ketosis status of the cows, and pointed out that cows with subclinical ketosis have higher SCC levels in their milk. Shin (Shin et al., 2015) collected blood samples from 213 cows at 1, 2, 4, 6, and 8 weeks postpartum, while dividing the cows into subclinical ketosis and non-ketosis groups. They found that the urea nitrogen levels in the blood of cows with subclinical ketosis were lower than those in the non-ketosis group. In summary, this paper selected a total of 19 indicators as inputs when constructing the method, including milk yield on days 1-15 postpartum, first fat percentage, first protein percentage, first SCC, and first urea nitrogen level.

2.3 Design of the prediction method

2.3.1 Framework design

Figure 1 presents the overall workflow for predicting the risk of subclinical ketosis in dairy cows. It is generally considered that data with a missing rate (MR) exceeding 50% - 70% is of low quality. To further ensure data usability, this paper deleted data with an MR over 50%. For data with an MR below 50%, this paper proposed a new imputation algorithm named RF-LOF. Subsequently, the SMOTETomeklinks algorithm was applied to balance the data and then divided into training sets and testing sets in a 7:3 ratio. In order to comprehensively consider the impact of imbalanced data on linear and nonlinear classifiers, this paper selected five representative commonly used algorithms, namely LR, LDA, KNN, SVM, and NB, to predict the subclinical ketosis and compare the performance differences of the classifiers before and after data balancing. Then we selected the optimal algorithm to design the subclinical ketosis prediction method for dairy cows.

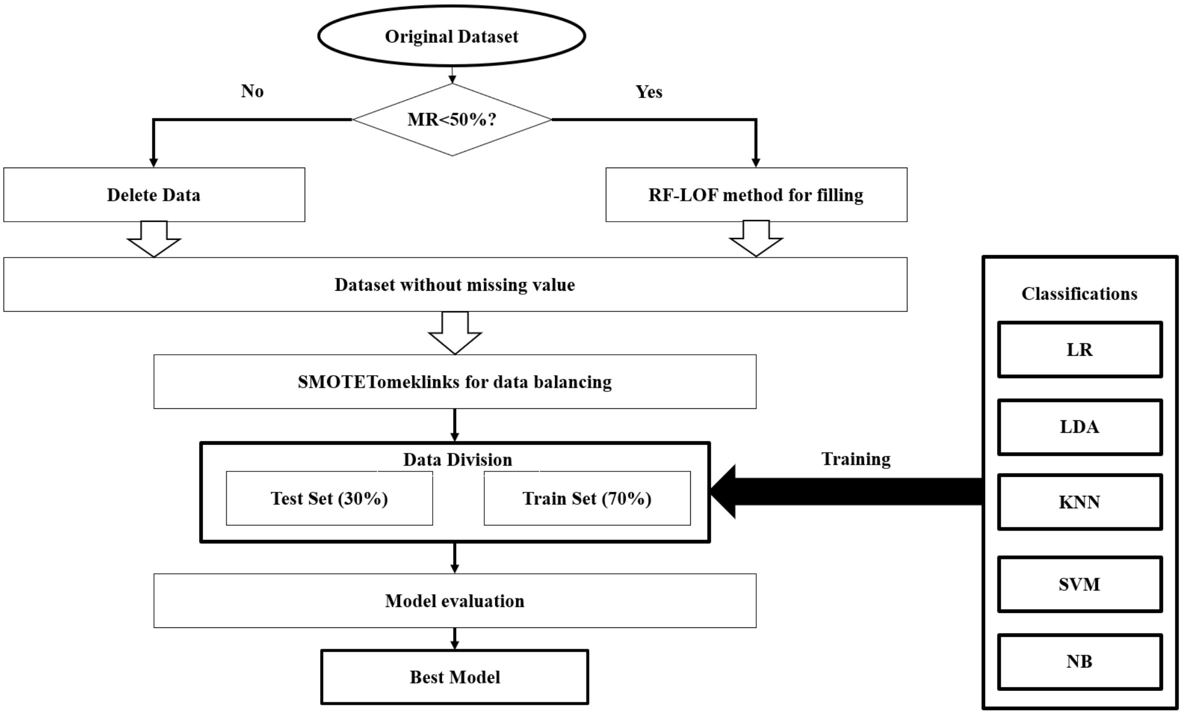

Figure 1 Overall Workflow

2.3.2 RF-LOF method for processing missing values

Considering that improper handling of missing values can lead to biased estimates, diminished statistical power, and invalid conclusions (Acock, 2005), we chose the RF method for data imputation from a range of machine learning techniques because RF can easily handle a mix of continuous and categorical variables without explicit data transformation (Leo and Adele, 2022). However, Outliers in datasets can reduce the accuracy of machine learning algorithms, including random forests (Alfian et al., 2023). Therefore, this paper improved the traditional RF algorithm by incorporating the LOF algorithm to clean outliers immediately after each iteration of the RF algorithm, using the cleaned results for the next iteration until convergence criteria are met. This approach can significantly enhance the accuracy of the RF algorithm in imputing missing values. The process of the RF-LOF method is shown in Figure 2.

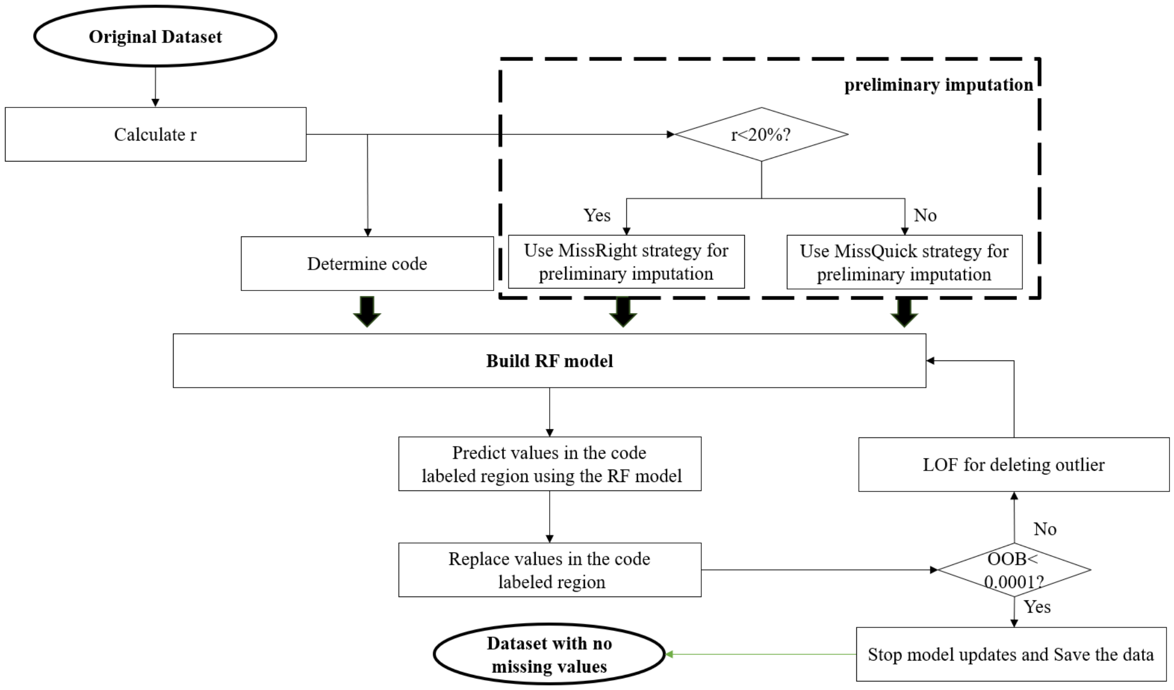
Original Dataset
Calculate r
preliminary imputation
r<20%?
Yes
No
Determine code
Use MissRight strategy for preliminary imputation
Use MissQuick strategy for preliminary imputation
Build RF model
Predict values in the code labeled region using the RF model
LOF for deleting outlier
No
Replace values in the code labeled region
OOB< 0.0001?
Yes
Dataset with no missing values
Stop model updates and Save the data

Figure 2 The process of the RF-LOF method

In Figure 2, MissQuick and MissRight are two pre-imputation methods, where the MissQuick method used mode or median for pre-imputation, while the MissRight method is based on iterative pre-imputation of neighboring samples. Out of bag (OOB) is a parameter used to evaluate the error of the RF model. The smaller it is, the higher the accuracy of the random forest model. The steps of the RF-LOF method are as follows:

(1) Calculate the total missing value rate, namely r.

(2) Determine an integer to identify each missing value, defined as code, which is usually a number that does not appear in the non-missing values.

(3) Choose a pre-filling scheme. If the r of the dataset is less than 20%, use the MissQuick method for pre-imputation. Otherwise, use the MissRight method for pre-imputation.

(4) Set the convergence condition: stop model updates when the OOB difference is less than 0.0001 or the maximum number of iterations is reached.

(5) Use the LOF algorithm to delete outliers immediately after each iteration of the RF algorithm. Then we use the cleaned results for the next iteration until convergence criteria are met.

2.3.3 SMOTETomeklinks for balancing data

The SMOTETomeklinks algorithm implements oversampling techniques to achieve a balanced distribution within the original training dataset(Sharma and Gosain, 2023). So this paper uses SMOTETomeklinks to improve the capacity of methods to accurately identify instances of the minority class. Compared to the traditional SMOTE algorithm, the SMOTETomeklinks algorithm can identify each Tomeklink for each samples (Zhou et al., 2021). In each instance that belongs to a Tomeklinked pair, the majority class sample is deleted.

2.3.4 Evaluation indexes

First, this paper used the imbalance ratio (IR) to measure the degree of data imbalance, and its formula is shown as Formula 1:

$$IR = \frac{N_{minority}}{N_{majority}} \quad (1)$$

where $N_{minority}$ represents the number of minority class data and $N_{majority}$ represents the number of majority class data.

In addition, this paper uses seven evaluation criteria to assess and compare the performance of predictive models, which are Acc, F1-Score, Sens, Spec, PPV, NPV, and AUC, where:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Sens = \frac{TP}{TP + FN} \tag{3}$$

$$Spec = \frac{TN}{TN + FP} \tag{4}$$

$$PPV = \frac{TP}{TP + FP} \tag{5}$$

$$NPV = \frac{TN}{TN + FN} \tag{6}$$

$$F1\ Score = 2 \times \frac{Sens \times Spec}{Sens + Spec} \tag{7}$$

$$G\text{-}mean = \sqrt{Sens \times Spec} \tag{8}$$

TP represents the true positive rate, TN denotes the true negative rate, FN indicates the false negative rate, and FP signifies the false positive rate.

In fact, the results of the SMOTETomeklinks algorithm may not always represent the minority class accurately, especially when dealing with highly imbalanced datasets or when the minority class exhibits complex boundaries. (Sharma and Gosain, 2023). This paper uses the silhouette score to evaluate the accuracy of the imbalanced data algorithm. The formula for the silhouette score of every sample is as follows,

$$s(i) = \frac{b(i) - a(i)}{max\left(a(i), b(i)\right)} \tag{9}$$

Where $\mathrm{a}(i)$ calculates the average distance between sample i and other samples within the same cluster. $\mathrm{b}(i)$ calculates the average distance between sample i and all samples in the nearest other clusters.

But in general, we tend to evaluate the quality of the algorithm by calculating the overall profile coefficient of the sample, and the formula for the overall profile coefficient is as follow,

$$S = \frac{1}{N}\sum_{i=1}^{N} s(i) \qquad (10)$$

The S ranges between -1 and 1, where a value closer to 1 indicates a better clustering effect, while a value closer to -1 indicates a poorer clustering effect (Pavlopoulos, 2024).

3. Results

To facilitate the comparison of performance between algorithms, this paper uses SVM as the base classifier for the experiments.

3.1 Parameter Determination

3.1.1 Parameters of the RF-LOF

To determine the key parameter of the RF-LOF algorithm, namely the LOF neighbor coefficient N, this paper takes different values of N to compare the interpolation effect of the RF-LOF algorithm in the original subclinical ketosis dataset under different N. The result is shown in Table 1.

Table 1 Comparison of SVM performance at different N

N	Acc	F1-Score	G-mean
5	0.8386	0.0086	0.0658
10	0.8191	**0.1986**	**0.3641**

20	**0.8399**	0.0511	0.1639
30	0.8361	0	0
40	0.8355	0	0

From the Table 1, we can see that, when the N is set to 10, SVM achieves the highest G-mean and F1 score. So we choose N=10 in RF-LOF method.

3.1.2 Parameters of the SMOTETomeklinks

To determine the optimal sampling ratio (R) and the number of neighbors (K) for the SMOTETomeklinks algorithm, this paper takes different values of R and K to compare the Acc, F1 score, and G-mean. The results are shown in Table 2 and Table 3.

Table 2 Comparison of SVM performance under different R

R	Acc	F1-Score	G-mean
2	0.7112	0	0
3	0.4520	0.5589	0.3963
4	**0.6527**	0.5986	**0.6463**
5	0.5427	**0.6724**	0.3713
6	0.5156	0.2265	0.3357

Table 3 Comparison of SVM performance under different K

K	Acc	F1-Score	G-mean
5	0.6527	0.5986	0.6463
7	0.5809	0.1438	0.1499
9	0.4410	0.5998	0.1950
11	0.4246	0.5931	0.0955
13	**0.6646**	0.6078	**0.6470**
15	0.5653	**0.6155**	0.5563

From the Table 2, we can see that, when the R is set to 4, SVM achieves the highest Acc and G-mean. From the Table 3, we can see that, when K is set to 13, SVM achieves the highest Acc and G-mean, So we choose R=4 and K=13 in SMOTETomeklinks method.

3.1.3 Parameters of the five classifiers

The key parameters for the five classifiers are taken from the default values of the relevant functions in Python 3.11, as shown in Table 3.

Table 4 Parameter settings for different classifiers

Method	Parameter
LR	Regularize terms=“l2”， Regularization intensity=1， Solver=“liblinear”
LDA	Solver=“lsqr”
KNN	Neighbour parameter k=5, Distance metrics="euclidean"
SVM	Penalty parameter C=1, Kernel functions="rbf"。 Kernel parameter=0.5
NB	Laplace parameter=1,

3.2 Imputing Missing Value

In the original dataset, samples of dairy cows with a missing rate greater than 50% were deleted due to low quality. At the same time, it was found that there were 4,887 missing records in the dataset, resulting in an overall missing rate of 3.1989%. The pre-imputation method for the RF-LOF algorithm is set to MissQuick. The key parameter N for the RF-LOF method is set to 10.

To compare the performance differences of various missing value imputation algorithms on subclinical ketosis data in dairy cows, this study used SVM as the base classifier. The performance of the RF-LOF algorithm was compared with two common missing value imputation methods (mean imputation and zero imputation) and the traditional RF missing value imputation algorithm in original subclinical ketosis dataset. The results are shown in Table 5.

Table 5 The performance of four missing value imputation methods

Method	Acc	F1-Score	G-mean
mean imputation	0.8234	0.0031	0.0830

zero imputation	0.8149	0.0319	0.1304
RF	**0.8240**	0.0068	0.0589
RF-LOF	0.8191	**0.1986**	**0.3641**

From Table 5, it can be observed that the RF-LOF algorithm has the highest F1-Score and G-mean, demonstrating the best performance among the four missing value imputation methods.

After using the RF-LOF algorithm for processing, a total of 704 outliers (LOF>1) and low-quality data points (MR > 50%) were cleaned, so we retained 4752 out of 5456 dairy cow data entries, all of which had their missing values imputed.

3.3 Imbalanced Data Processing

In order to compare the performance differences of various algorithms for handling imbalanced data on the subclinical ketosis dataset of dairy cows, this study used SVM as the base classifier. The performance of the SMOTETomeklinks algorithm was compared with three traditional imbalanced data algorithms (random oversampling, SMOTE, and ADASYN) in the subclinical ketosis dataset, and the results are presented in Table 6.

Table 6 The performance of four imbalanced data processing method

Method	Acc	F1-Score	G-mean
random oversampling	0.5350	**0.6701**	0.3446
SMOTE	0.6190	0.6244	0.6247
ADASYN	0.5346	0.2397	0.3406
SMOTETomeklinks	**0.6446**	0.6078	**0.6470**

From Table 6, it can be observed that the SMOTETomeklinks algorithm has the highest ACC and G-mean, demonstrating the best performance among the four imbalanced data processing methods.

Data balancing is achieved by SMOTETomeklinks and the number of positive and negative samples in the dataset is shown in Table 7.

Table 7 the number of positive and negative samples in the dataset

Item	Original	SMOTETomeklinks
Positive Class	858	2936
Negative Class	4598	3987
IR	0.1866	0.7363

According to Formula 10, the silhouette score of the dataset after processing with the SMOTETomeklinks algorithm is 0.0077. It is generally considered that a silhouette score greater than 0 indicates a better clustering effect(Wang et al., 2022).

3.4 Comparison of the Performance of Different Classifiers

To reflect the performance differences of the dairy cow subclinical ketosis prediction method before and after data balancing, this paper uses five classifiers on the datasets before and after imbalanced data processing, with the results shown in Table 8 and Table 9, respectively. The Receiver Operating Characteristic Curve (ROC) of the five classifiers before and after SMOTETomeklinks are shown in Figure 3 and Figure 4, respectively.

Table 8 The performance of five classifiers without SMOTETomeklinks

Classifier	Acc	F1-Score	Sens	Spec	PPV	NPV	AUC
LR	**0.8387**	0	0	**0.9991**	0	0.8392	0.5485
LDA	0.8366	0.1003	0.0567	0.9857	**0.4333**	0.8452	**0.6810**
KNN	0.8323	0.0700	0.0393	0.9841	0.3214	0.8426	0.5838
SVM	0.8190	**0.1987**	**0.1397**	0.9490	0.3220	**0.8522**	0.6447
NB	0.8232	0.1250	0.0786	0.9657	0.3050	0.8486	0.6116

Table 9 The performance of five classifiers with SMOTETomeklinks.

Classifier	Acc	F1-Score	Sens	Spec	PPV	NPV	AUC
LR	0.5849	0.0091	0.0046	**0.9983**	0.6667	0.5847	0.5967
LDA	0.5873	0.0316	0.0162	0.9942	0.6667	0.5865	0.7109
KNN	**0.7501**	**0.7486**	**0.8946**	0.6471	**0.6436**	**0.8961**	**0.8727**
SVM	0.6446	0.6078	0.6620	0.6323	0.5819	0.7242	0.6840

NB	0.6051	0.1615	0.0914	0.9711	0.6229	0.6001	0.6867

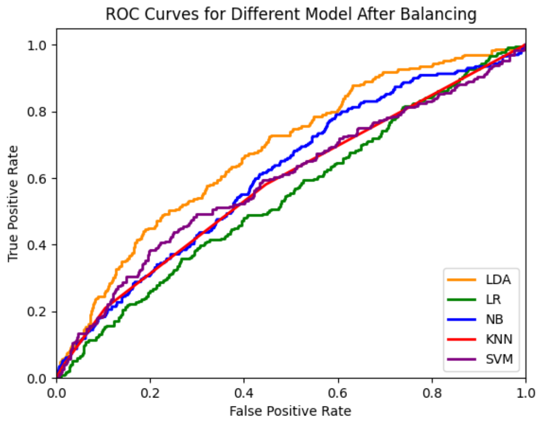

Figure 3 ROC Curves for Different Models Before Balancing

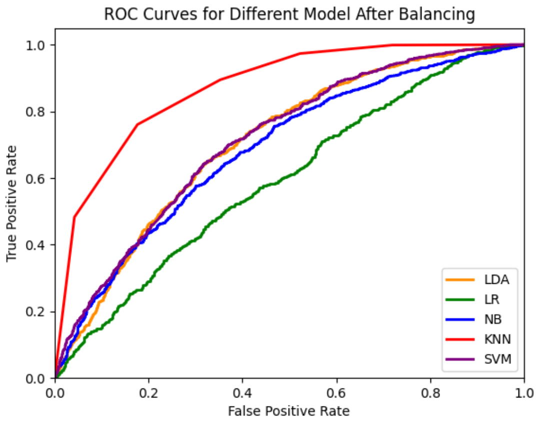

Figure 4 ROC Curves for Different Models After Balancing

The results indicate that after balancing the data, the Sensitivity, F1 Score, PPV, and AUC of the five models have all improved. Additionally, the ROC curves are more skewed towards the top left, which suggests that handling the imbalanced data for the ketosis dataset helps enhance the performance of the models. Based on these experimental results, we choose the KNN classifier, which performs the best, as the base classifier for the ketosis prediction model for cows.

Due to the processing of the method using the RF-LOF algorithm, SMOTETomeklinks algorithm, and KNN algorithm, we name this subclinical ketosis prediction method for dairy cows as the RLST-KNN method.

3.5 Sensitivity Analysis

To discuss the impact of different ketosis lactation periods on model performance, the metadata needs to be divided into multiple datasets based on varying ketosis lactation periods. Since the number of days of lactation of cows is one of the most important indicators for cows(Rodríguez-González et al., 2020), it is typically divided into early lactation, Mid lactation, and late lactation with intervals of

120 days(Van Knegsel and Kok, 2024). The datasets are divided into three subsets based on early lactation, mid-lactation, and late lactation. Due to the fact that dairy cows are at a high risk of suffering from subclinical ketosis during the three to six weeks after calving(Lei and Simões, 2021), In the sensitivity analysis, this paper further divides the dairy cows in early lactation into three categories: before three weeks, between three to six weeks, and after six weeks. The descriptive statistics of five subsets are shown in Table 10.

Table 10 Descriptive statistics of three subsets

Sub-DataSet	number of data	IR
Early lactation(before 3 weeks)	276	0.1694
Early lactation(3-6 weeks)	343	0.2035
Early lactation(after 6 weeks)	1261	0.0404
Mid lactation	2132	0.1351
Late lactation	1444	0.4606

Subsequently, the RLST-KNN method was used for ketosis prediction on the five datasets mentioned above. The results of sensitivity analysis are shown in Table 11.

Table 11 Results of sensitivity analysis

Sub-DataSet	Acc	F1 Score	Sens	Spec	PPV	NPV	AUC
Early lactation (before 3 weeks)	**0.9814**	**0.9904**	**1**	0.4333	**0.9811**	**1**	0.8833
Early lactation (3-6 weeks)	0.9215	0.9584	1	0.1667	0.9202	1	0.7805
Early lactation (after 6 weeks)	0.8796	0.9234	0.9965	0.5665	0.8602	0.9838	**0.9244**

Mid lactation	0.7443	0.7423	0.8515	**0.6626**	0.6579	0.8542	0.8398
Late lactation	0.7657	0.8544	0.9613	0.2734	0.7690	0.7373	0.7426

From Table 11, it can be observed that the RLST-KNN method achieved the highest ACC, F1-Score, Sens, PPV, and NPV in the early lactation(before 3 weeks) subset, while also obtaining the second-highest AUC. Therefore, the RLST-KNN method demonstrates its practical value by enabling timely predictions of ketosis in dairy cows even before the high-risk period for ketosis occurs.

4. Discussion

The proposed RLST-KNN method achieved excellent results in predicting dairy cow subclinical ketosis, reflected by high values of Acc, F1 Score, Sens, PPV, NPV, and AUC, which reached 0.7501, 0.7486, 0.8946, 0.6436, 0.8961, and 0.8727, respectively from Table 9. The RLST-KNN algorithm combines the advantages of three techniques: RF-LOF, SMOTETomeklinks, and KNN. First, the RF-LOF algorithm accurately filled in the missing values in the dataset. Then, the SMOTETomeklinks algorithm balanced the dataset, enhancing the model's generalization ability (Matharaarachchi et al., 2024). Finally, the KNN classifier is used to predict subclinical ketosis in cows.

The experimental results indicated that the RLST-KNN method achieved the highest performance, outperforming models built on the other four traditional classifiers. Additionally, this study compared the performance of the five classifiers before and after handling imbalanced data, finding that KNN performed better on the balanced dataset than on the imbalanced dataset, as evidenced by higher AUC and F1 scores for most classifiers after balancing. Comparing Tables 8 and 9, although the classifiers had higher accuracy on the unbalanced dataset, this was misleading, as the results were heavily biased towards the majority class, leading to very few correct predictions for the minority class. Specifically, while the five classifiers exhibited very high specificity and NPV on the unbalanced dataset, their sensitivity and PPV were close to 0, indicating an abnormal classification result. Furthermore, the experimental results have shown that the model performs well on the dataset from the first three weeks postpartum in terms of ACC, F1 Score, Sens, PPV, and NPV. It indicated that the method can predict subclinical ketosis in lactating cows early and accurately before the high-risk period for ketosis (three to six weeks postpartum), which has significant practical implications.

From the perspective of the experimental data, it is necessary to acknowledge some limitations of the method. Firstly, the dataset for dairy cow ketosis is relatively small, with only 5456 entries of dairy cow subclinical ketosis data, which means that the results may not be generalizable to a larger population of cows. Secondly, there are other important attributes that are not considered during the ketosis prediction method development, including lying time(Tucker et al., 2021) , feed data(Yameogo et al., 2008), and the composition and status of cow feces and urine (Zhang and Ametaj, 2017). Therefore, the method may not perform well in complex scenarios. Meanwhile, the RLST-KNN method is only applicable to subclinical ketosis in dairy cows and more experiments are needed to explore the method performance in clinical ketosis.

5. Conclusions

This paper presents a ketosis prediction method for dairy cows based on the RLST-KNN method. The RLST-KNN method mitigates the decline in prediction capability caused by missing values and data imbalance in other prediction methods.

Experimental results have shown that RLST-KNN outperforms traditional classifiers in terms of ACC, F1 Score, Sens, PPV, NPV, and AUC. In addition, the RLST-KNN method had the best performance before three weeks postpartum, indicating that the RLST-KNN method can accurately predict subclinical ketosis in dairy cows before the peak period of ketosis occurrence (three to six weeks postpartum), helping farms to proactively address the treatment of ketosis-related issues. In the future, it is essential to collect more ketosis data by gathering data from different farms with various breeds of cows or by incorporating additional key attributes such as lying time(Tucker et al., 2021), feed data(Yameogo et al., 2008), etc., to train a model capable of accurately detecting dairy cow ketosis in more complex scenarios.

Acknowledgments

We sincerely thank Afimilk (China) Agricultural Technology Co. for providing us with the dataset on dairy cow ketosis.

References

Acock, A.C., 2005. Working With Missing Values. J of Marriage and Family 67, 1012–1028. https://doi.org/10.1111/j.1741-3737.2005.00191.x

Alfian, G., Syafrudin, M., Fitriyani, N.L., Alam, S., Pratomo, D.N., Subekti, L., Octava, M.Q.H., Yulianingsih, N.D., Atmaji, F.T.D., Benes, F., 2023. Utilizing Random Forest with iForest-Based Outlier Detection and SMOTE to Detect Movement and Direction of RFID Tags. Future Internet 15, 103. https://doi.org/10.3390/fi15030103

Bauer, E.A., Jagusiak, W., 2022. The Use of Multilayer Perceptron Artificial Neural Networks to Detect Dairy Cows at Risk of Ketosis. Animals 12, 332. https://doi.org/10.3390/ani12030332

Cascone, G., Licitra, F., Stamilla, A., Amore, S., Dipasquale, M., Salonia, R., Antoci, F., Zecconi, A., 2022. Subclinical Ketosis in Dairy Herds: Impact of Early Diagnosis and Treatment. Front. Vet. Sci. 9. https://doi.org/10.3389/fvets.2022.895468

Del Caño, R., Saha, T., Moonla, C., De la Paz, E., Wang, J., 2023. Ketone bodies detection: Wearable and mobile sensors for personalized medicine and nutrition. TrAC Trends in Analytical Chemistry 159, 116938. https://doi.org/10.1016/j.trac.2023.116938

Ferreira, R.E.P., Angels de Luis Balaguer, M., Bresolin, T., Chandra, R., Rosa, G.J.M., White, H.M., Dórea, J.R.R., 2024. Multi-modal machine learning for the early detection of metabolic disorder in dairy cows using a cloud computing framework. Computers and Electronics in Agriculture 227, 109563. https://doi.org/10.1016/j.compag.2024.109563

Guliński, P., 2021. Ketone bodies – causes and effects of their increased presence in cows' body fluids: A review. Vet World 14, 1492–1503. https://doi.org/10.14202/vetworld.2021.1492-1503

Huang, Y., Zhang, B., Mauck, J., Loor, J.J., Wei, B., Shen, B., Wang, Y., Zhao, C., Zhu, X., Wang, J., 2024. Plasma and milk metabolomics profiles in dairy cows with subclinical and clinical ketosis. Journal of Dairy Science 107, 6340–6357. https://doi.org/10.3168/jds.2023-24496

Jeong, J.-K., Choi, I.-S., Moon, S.-H., Lee, S.-C., Kang, H.-G., Jung, Y.-H., Park, S.-B., Kim, I.-H., 2018. Effect of two treatment protocols for ketosis on the resolution, postpartum health, milk yield, and reproductive outcomes of dairy cows. Theriogenology 106, 53–59. https://doi.org/10.1016/j.theriogenology.2017.09.030

Kaur, H., Pannu, H.S., Malhi, A.K., 2019. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Comput. Surv. 52, 79:1-79:36. https://doi.org/10.1145/3343440

Lei, M.A.C., Simões, J., 2021. Invited Review: Ketosis Diagnosis and Monitoring in High-Producing Dairy Cows. Dairy 2, 303–325. https://doi.org/10.3390/dairy2020025

Leo, B., Adele, C., 2022. Using random forests v4 - Manual-Setting Up, Using, And Understanding Random Forests V4.

Mandujano Reyes, J.F., Walleser, E., Hachenberg, S., Gruber, S., Kammer, M., Baumgartner, C., Mansfeld, R., Anklam, K., Döpfer, D., 2021. Full model selection using regression trees for numeric predictions of biomarkers for metabolic challenges in dairy cows. Preventive Veterinary Medicine 193, 105422. https://doi.org/10.1016/j.prevetmed.2021.105422

Matharaarachchi, S., Domaratzki, M., Muthukumarana, S., 2024. Enhancing SMOTE for imbalanced data with abnormal minority instances. Machine Learning with Applications 18, 100597. https://doi.org/10.1016/j.mlwa.2024.100597

Melendez, P., Serrano, M.V., 2024. Update on ketosis in dairy cattle with major emphasis on subclinical ketosis and abdominal adiposity. Veterinary Medicine and Science 10, e1525. https://doi.org/10.1002/vms3.1525

Michelucci, U., 2024. Unbalanced Datasets and Machine Learning Metrics, in: Michelucci, U. (Ed.), Fundamental Mathematical Concepts for Machine Learning in Science. Springer International Publishing, Cham, pp. 185–212. https://doi.org/10.1007/978-3-031-56431-4_8

Pavlopoulos, J., 2024. Revisiting Silhouette Aggregation. https://doi.org/10.48550/arXiv.2401.05831

Qiao, Y., Gao, Z., Liu, Yong, Cheng, Y., Yu, M., Zhao, L., Duan, Y., Liu, Yu, 2014. Breath Ketone Testing: A New Biomarker for Diagnosis and Therapeutic Monitoring of Diabetic Ketosis. BioMed Research International 2014, 869186. https://doi.org/10.1155/2014/869186

Rodríguez-González, G.L., Bautista, C.J., Rojas-Torres, K.I., Nathanielsz, P.W., Zambrano, E., 2020. Importance of the lactation period in developmental programming in rodents. Nutrition Reviews 78, 32–47. https://doi.org/10.1093/nutrit/nuaa041

Rutherford, A.J., Oikonomou, G., Smith, R.F., 2016. The effect of subclinical ketosis on activity at estrus and reproductive performance in dairy cattle. Journal of Dairy Science 99, 4808–4815. https://doi.org/10.3168/jds.2015-10154

Satoła, A., Bauer, E.A., 2021. Predicting Subclinical Ketosis in Dairy Cows Using Machine Learning Techniques. Animals 11, 2131. https://doi.org/10.3390/ani11072131

Sharma, H., Gosain, A., 2023. Oversampling Methods to Handle the Class Imbalance Problem: A Review, in: Patel, K.K., Santosh, K.C., Patel, A., Ghosh, A. (Eds.), Soft Computing and Its Engineering Applications, Communications in Computer and Information Science. Springer Nature Switzerland, Cham, pp. 96–110. https://doi.org/10.1007/978-3-031-27609-5_8

Shin, E.-K., Jeong, J.-K., Choi, I.-S., Kang, H.-G., Hur, T.-Y., Jung, Y.-H., Kim, I.-H., 2015. Relationships among ketosis, serum metabolites, body condition, and reproductive outcomes in dairy cows. Theriogenology 84, 252–260. https://doi.org/10.1016/j.theriogenology.2015.03.014

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., Jerbi, K., 2023. Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. NeuroImage 277, 120253. https://doi.org/10.1016/j.neuroimage.2023.120253

Tucker, C.B., Jensen, M.B., De Passillé, A.M., Hänninen, L., Rushen, J., 2021. Invited review: Lying time and the welfare of dairy cows. Journal of Dairy Science 104, 20–46. https://doi.org/10.3168/jds.2019-18074

Van Knegsel, A.T.M., Kok, A., 2024. Consequences of Lactation Length Management for Health and Fertility in Dairy Cows, in: Gross, J.J. (Ed.), Production Diseases in Farm Animals. Springer International Publishing, Cham, pp. 571–586. https://doi.org/10.1007/978-3-031-51788-4_25

Wang, Haoran, Guo, T., Wang, Z., Xiao, J., Gao, L., Gao, X., Wang, Hongbin, 2023. PreCowKetosis: A Shiny web application for predicting the risk of ketosis in dairy cows using prenatal indicators. Computers and Electronics in Agriculture 206, 107697. https://doi.org/10.1016/j.compag.2023.107697

Wang, W., Yousaf, M., Liu, D., Sohail, A., 2022. A Comparative Study of the Genetic Deep Learning Image Segmentation Algorithms. Symmetry 14, 1977. https://doi.org/10.3390/sym14101977

Yameogo, N., Ouedraogo, G.A., Kanyandekwe, C., Sawadogo, G.J., 2008. Relationship between ketosis and dairy cows' blood metabolites in intensive production farms of the periurban area of Dakar. Trop Anim Health Prod 40, 483–490. https://doi.org/10.1007/s11250-007-9124-z

Yang, W., Zhang, B., Xu, C., Zhang, H., Xia, C., 2019. Effects of Ketosis in Dairy Cows on Blood Biochemical Parameters, Milk Yield and Composition, and Digestive Capacity. J Vet Res 63, 555–560. https://doi.org/10.2478/jvetres-2019-0059

Zhang, G., Ametaj, B.N., 2017. Ketosis Under a Systems Veterinary Medicine Perspective, in: Ametaj, B.N. (Ed.), Periparturient Diseases of Dairy Cows. Springer International Publishing, Cham, pp. 201–222. https://doi.org/10.1007/978-3-319-43033-1_10

Zhang, G., Mandal, R., Wishart, D.S., Ametaj, B.N., 2021. A Multi-Platform Metabolomics Approach Identifies Urinary Metabolite Signatures That

Differentiate Ketotic From Healthy Dairy Cows. Front. Vet. Sci. 8. https://doi.org/10.3389/fvets.2021.595983

Zhou, H., Yu, K.-M., Chen, Y.-C., Hsu, H.-P., 2021. A Hybrid Feature Selection Method RFSTL for Manufacturing Quality Prediction Based on a High Dimensional Imbalanced Dataset. IEEE Access 9, 29719–29735. https://doi.org/10.1109/ACCESS.2021.3059298

Declaration of Conflict of Interest

The authors declare there is no conflict of interest.