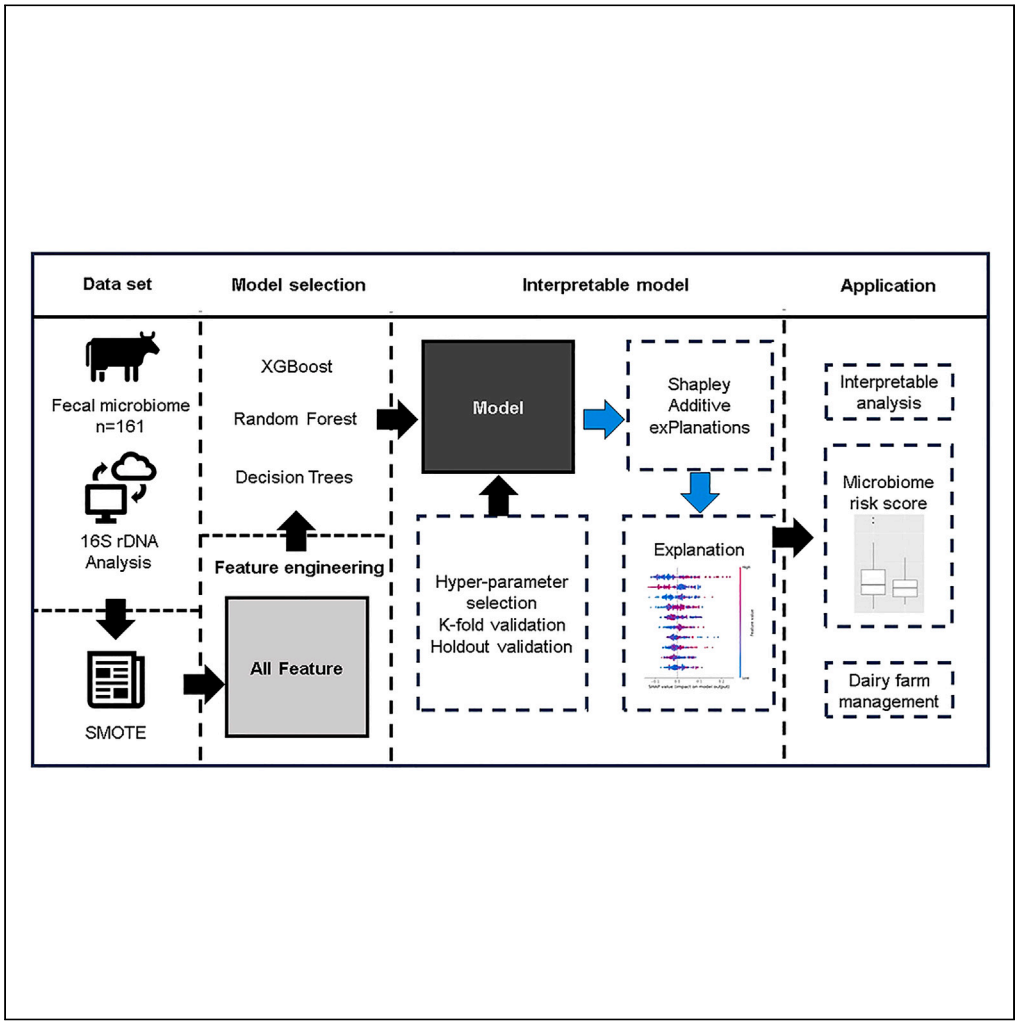
iScience

CellPress
OPEN ACCESS

Article

Interpretable machine learning reveals microbiome signatures strongly associated with dairy cow milk urea nitrogen



Qingyuan Yu, Hui Wang, Linqing Qin, Tianlin Wang, Yonggen Zhang, Yukun Sun

sunyukun@neau.edu.cn

Highlights

Modeling to predict dairy cow milk urea nitrogen using machine learning algorithms

Random forest algorithms screening for microbial markers of milk urea nitrogen

SHAP visualization model and build the MRS of milk urea nitrogen

Improving nitrogen utilization rate by intervening with individual information

Yu et al., iScience 27, 109955
June 21, 2024
https://doi.org/10.1016/j.isci.2024.109955

iScience

CellPress
OPEN ACCESS

Article

Interpretable machine learning reveals microbiome signatures strongly associated with dairy cow milk urea nitrogen

Qingyuan Yu,[1] Hui Wang,[1] Linqing Qin,[1] Tianlin Wang,[1] Yonggen Zhang,[1] and Yukun Sun[1,2,*]

SUMMARY

The gut microbiome plays an important role in the healthy and efficient farming of dairy cows. However, high-dimensional microbial information is difficult to interpret in a simplified manner. We collected fecal samples from 161 cows and performed 16S amplicon sequencing. We developed an interpretable machine learning framework to classify individuals based on their milk urea nitrogen (MUN) concentrations. In this framework, we address the challenge of handling high-dimensional microbial data imbalances and identify 9 microorganisms strongly correlated with MUN. We introduce the Shapley Additive Explanations (SHAP) method to provide insights into the machine learning predictions. The results of the study showed that the performance of the machine learning model improved (accuracy = 72.7%) after feature selection on high-dimensional data. Among the 9 microorganisms, *g__Firmicutes_unclassified* had the greatest impact in the model. This study provides a reference for precision animal husbandry.

INTRODUCTION

Dairy cattle are an important source of high-quality animal protein for human consumption.[1] Increasing milk protein production and improving feed efficiency have become the most coveted goals in the dairy industry.[2] Milk urea nitrogen (MUN) concentration can serve as a biomarker to assess how efficiently lactating dairy cows use nitrogen.[3] To a certain extent, monitoring MUN can determine whether the CP ratio of the diet is reasonable. The ability to digest protein feeds varies from cow to cow and is largely dependent on the different microbiota in the digestive tract.[4] Microbial data characteristics are composite, sparse and high-dimensional.[5] Key bacteria can be identified as biomarkers,[6] and accurate identification of biomarkers helps in pasture management.

In recent years, to unravel the complexity of the microbiome, researchers have turned to artificial intelligence. Owing to their powerful predictive and informative potential, machine learning (ML) and deep learning have emerged as key tools to advance microbiome research.[5] The use of random forest (RF) machine learning as a predictive tool for crop productivity based on soil microorganisms revealed a strong correlation between microorganism composition and crop yield, with *Actinomycetales* being the most influential taxon.[7] To reduce computational costs, 14 microbial features linked to oral diseases were chosen from the high-dimensional data for the prediction of oral health.[8] Machine learning offers superior predictive capabilities compared to traditional statistical modeling, but it does so at the expense of interpretability.[9] Providing interpretable predictions is more important than using black box models to provide accurate predictions for decision-making.[10,11] One can further investigate the variables in the explainable model to assess whether the finding is significant.[12]

To date, there is limited understanding of the relationship between feed conversion and gut microbes in dairy cows. Hence, it is necessary to incorporate both the compositional traits of microorganisms and indicators of feed conversion rate in order to gain a deeper understanding of the connection between microorganisms and feed conversion rate through machine learning. In the current study, there is a need to use gut microbial information to predict MUN through machine learning modeling to address the following questions: (1) Can gut microbial information be effective in predicting MUN when using the same feed (accuracy >70%)? (2) Can specific gut microbe traits serve as predictive markers for MUN and be utilized for the selection of cows with high feed conversion rates? (3) Is it feasible to identify and enhance MUN by intervening in significantly relevant individual data? The responses to these inquiries contribute to the characterization of microorganisms linked to feed conversion rates. Distinct groups of animals with varying feed conversion rates are selected to receive suitable daily feed amounts, thus facilitating precision ranching.

RESULTS

Gut microflora composition

In our investigation of intestinal flora composition, we utilized 161 dairy cows as our study subjects. These cows were categorized into two groups: the normal group (n = 72) and the high-concentration group (n = 89) based on their MUN concentration. Based on high-throughput

[1]College of Animal Sciences and Technology, Northeast Agriculture University, Harbin 150030, China
[2]Lead contact
*Correspondence: sunyukun@neau.edu.cn
https://doi.org/10.1016/j.isci.2024.109955

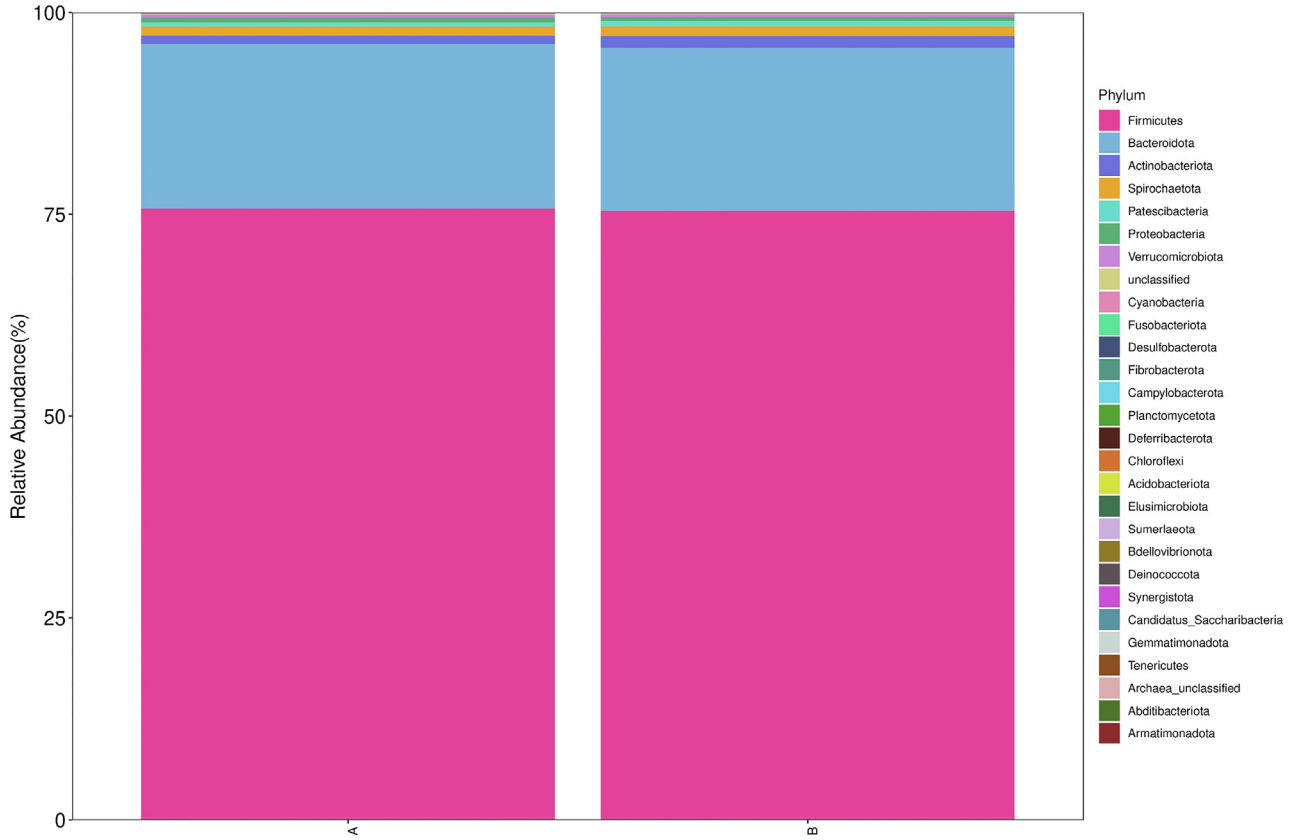


Figure 1. Gate-level relative abundance of gut microorganisms in 161 dairy cows

Different colors represent different species at the same level.

sequencing, the composition of each sample at the gate level was calculated. A total of 28 gates were identified (Figure 1). Among them, Firmicutes, Bacteroidota, Actinobacteriota, and Spirochaetota predominate (Relative abundance>1%). Differences in abundance indicate individual differences in gut microbes.

Differences in the ecological structure of the gut microbiota

We conducted both qualitative and quantitative analyses of the diversity and composition of fecal microbiota in the intestines of 161 dairy cows. Specifically, we assessed alpha diversity, which includes the Shannon indicator, Simpson indicator, Chao1 indicator, and Pielou_e indicator (Figure 2A). There was no significant difference in the four indicators (p values were 0.6, 0.99, 0.63, and 0.86). In the principal coordinate analysis (PCoA) of the microbiome at the genus level (Figure 2B), the results showed that the gut microbial composition was similar in the two groups A and B. The results of the PCoA showed that the gut microbial composition was similar in the two groups. Similar results were obtained using t-distributed stochastic neighbor embedding (T-SNE) analysis based on the abundance matrix (Figure 2C).

A machine learning approach to differentiating between normal and high-concentration groups of dairy cows

This study was conducted to investigate further whether ML could better differentiate between normal and high-concentration groups of cows (Figure 3). We employed abundance matrices at the genus level within the microbiome as inputs for the ML analysis. It is common to encounter sample imbalance in predictive modeling, where the model may have a bias toward the more prevalent categories within the training set, potentially affecting predictions. There were some differences between the normal and high concentration in the sample. To address this issue, we used the Synthetic Minority Oversampling Technique (SMOTE) algorithm to artificially generate additional samples and balance the data groups, resulting in a total sample size of 178.

For machine learning problems, data and features determine the upper limit of machine learning, while models and algorithms only approximate this upper limit. Feature engineering involves a series of processing steps applied to raw data to extract relevant features, which are then used as inputs for algorithms and models. RF serves as an efficient feature selection method for both discrete and continuous variables, offering a rapid means of ranking variable importance. In our study, we selected the top nine microorganisms based on their importance to be used as input features.

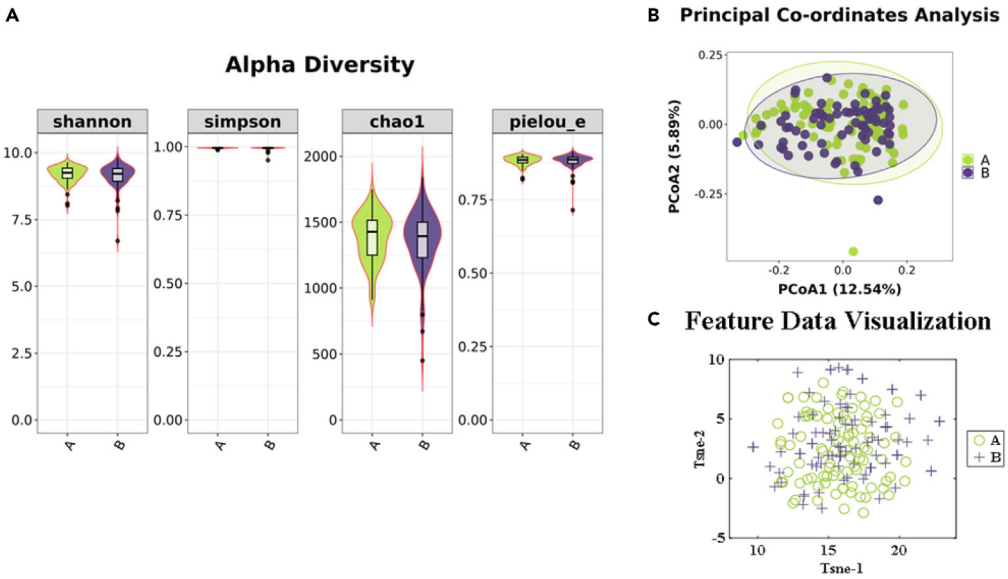


Figure 2. Analysis of differences in intestinal microbial biology in dairy cows of different MUN

(A) Shannon index, Simpson index, chao1 index, pielou_e index.

(B) Two groups of microorganisms principal co-ordinates analysis (P CoA).

(C) T-sne visualization of dimensionality reduction.

Nine input variables for feature selection are used after feature engineering. The performance of each model is examined by K-fold (K = 10) cross-validation, i.e., decision tree, RF, XG Boost. As the number of validations increases, the RF model outperforms the decision tree and XG Boost. According to the performance of the model, we determined the most suitable RF algorithm to construct the final classification model (Figure 4A).

The RF algorithm's selection of 9 features demonstrates strong predictive capability when contrasted with the initial 684 input features. The model's accuracy was 61.4% when using all 684 features, but it improved to 72.7% when only the 9 selected features were used as input (9 features include:g__Firmicutes_unclassified, g__Dorea, g__Erysipelotrichaceae_UCG-007, g__Atopobiaceae_unclassified, g__UCG- 009, g__Eubacterium] _xylanophilum_group, g_Lactobacillus, g_Butyrivibrio, g_Phascolarctobacterium. The most important feature is g__Firmicutes_unclassified). There was a significant improvement in the model accuracy, and reducing 684 features to 9 feature inputs was able to reduce model complexity significantly. This reduces data storage space, saves model computation time, and facilitates data and model visualization. The model was externally validated with 80% accuracy using data that did not participate in the model. The dataset for external validation was 20 samples that did not participate in model training.

The RF algorithm consistently demonstrates high accuracy and is the most effective among the three algorithms for classifying gut microorganisms. Following the model-building process, model accuracy was further evaluated using a confusion matrix, which presents the

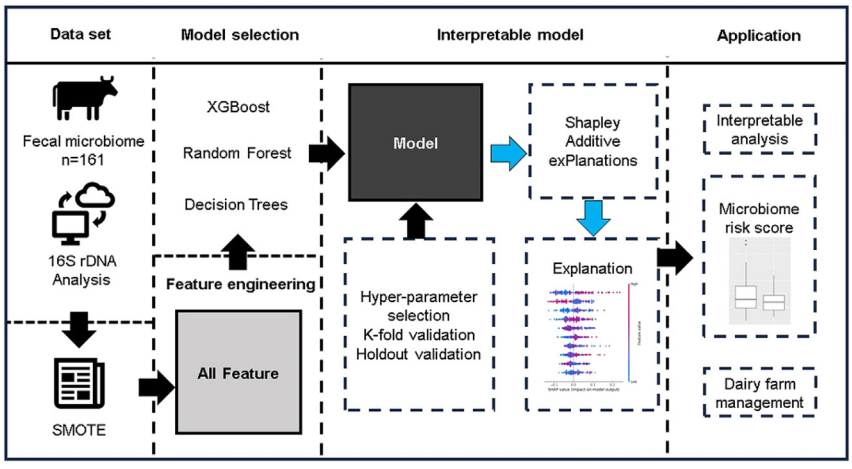


Figure 3. Machine Learning Framework Flowchart

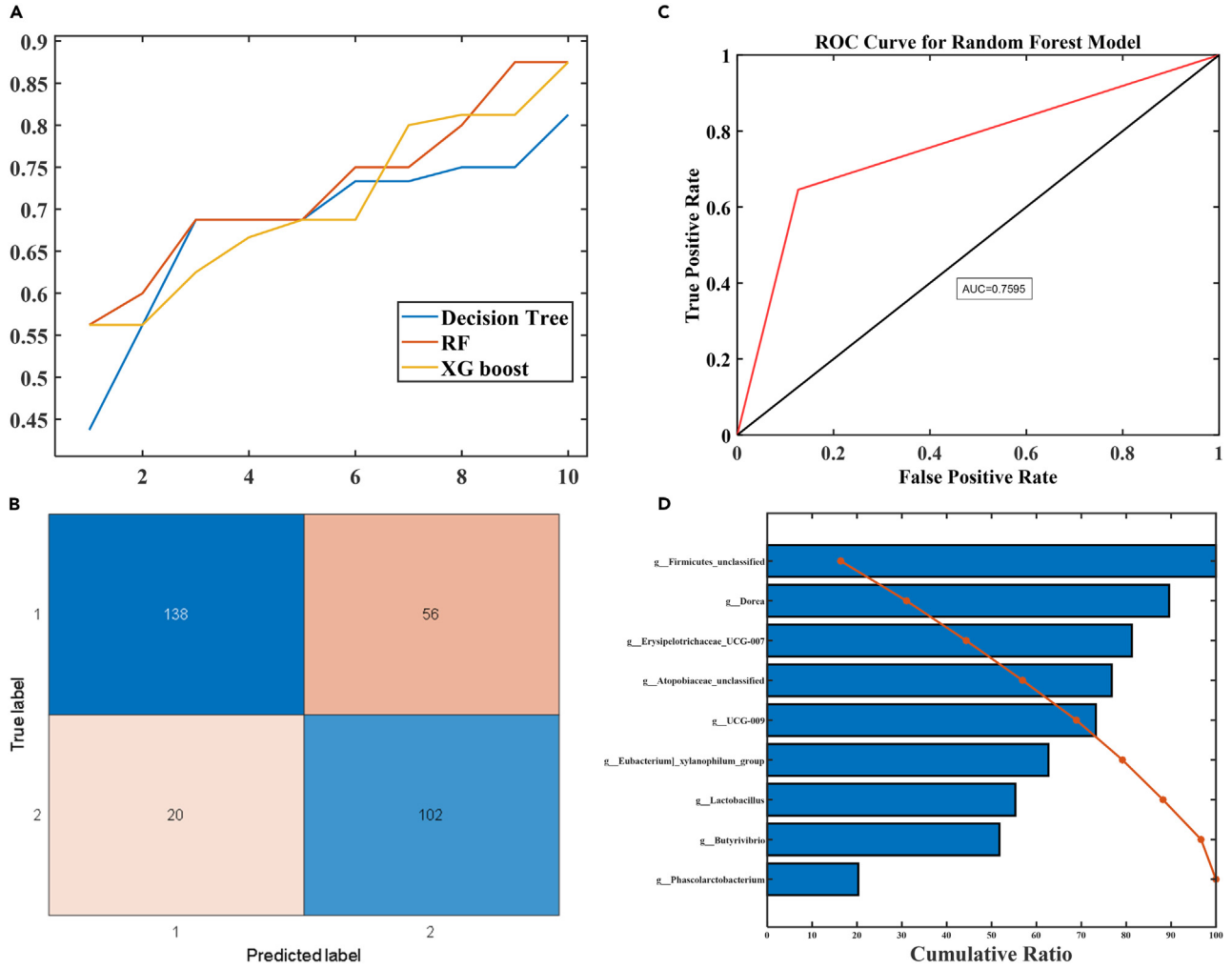


Figure 4. Model selection and prediction

(A) 10-fold cross validation.

(B) Confusion Matrix for Model Predictions.

(C) ROC curve.

(D) Importance of features.

classification results in a clear 2x2 matrix format. The confusion matrix consists of four components: true positives, false positives, false negatives and true negatives. The results indicated that in group A, the precision rate was 87.3% and the recall rate was 71.1%, whereas in group B, the precision rate was 64.5% and the recall rate was 84.6 (Figure 4B). Additionally, the receiver operating characteristic (ROC) curve serves as a measure to assess the performance of a binary classification model. The area under the ROC curve is called the area under curve (AUC). The AUC values ranged from 0 to 1, with a larger AUC indicating a more effective model. The results of the model predictions were represented by the ROC curve, which showed an AUC of 0.7595 (Figure 4C).

Nine key microorganisms (genus level) were selected to predict MUN by downscaling the high-dimensional data through the RF algorithm. The feature importance in order is g__Firmicutes_unclassified, g__Dorea, g__Erysipelotrichaceae_UCG-007, g__Atopobiaceae_unclassified, g__UCG- 009, g__Eubacterium] _xylanophilum_group, g_Lactobacillus, g_Butyrivibrio, g_Phascolarctobacterium. The most important feature is g__Firmicutes_unclassified with 16.3% importance. The feature importance of the top four was 56.9%. The top four features were g__Firmicutes_unclassified, g__Dorea, g__Erysipelotrichaceae_UCG-007 and g__Atopobiaceae_unclassified. (Figure 4D).

SHAP-based interpretation of machine learning models

The SHAP model comes from a set of methods in cooperative game theory that have been shown to improve the interpretability of machine learning models. The SHAP utilizes the concept of Shapley value to compute the feature importance, which gives the contribution of each

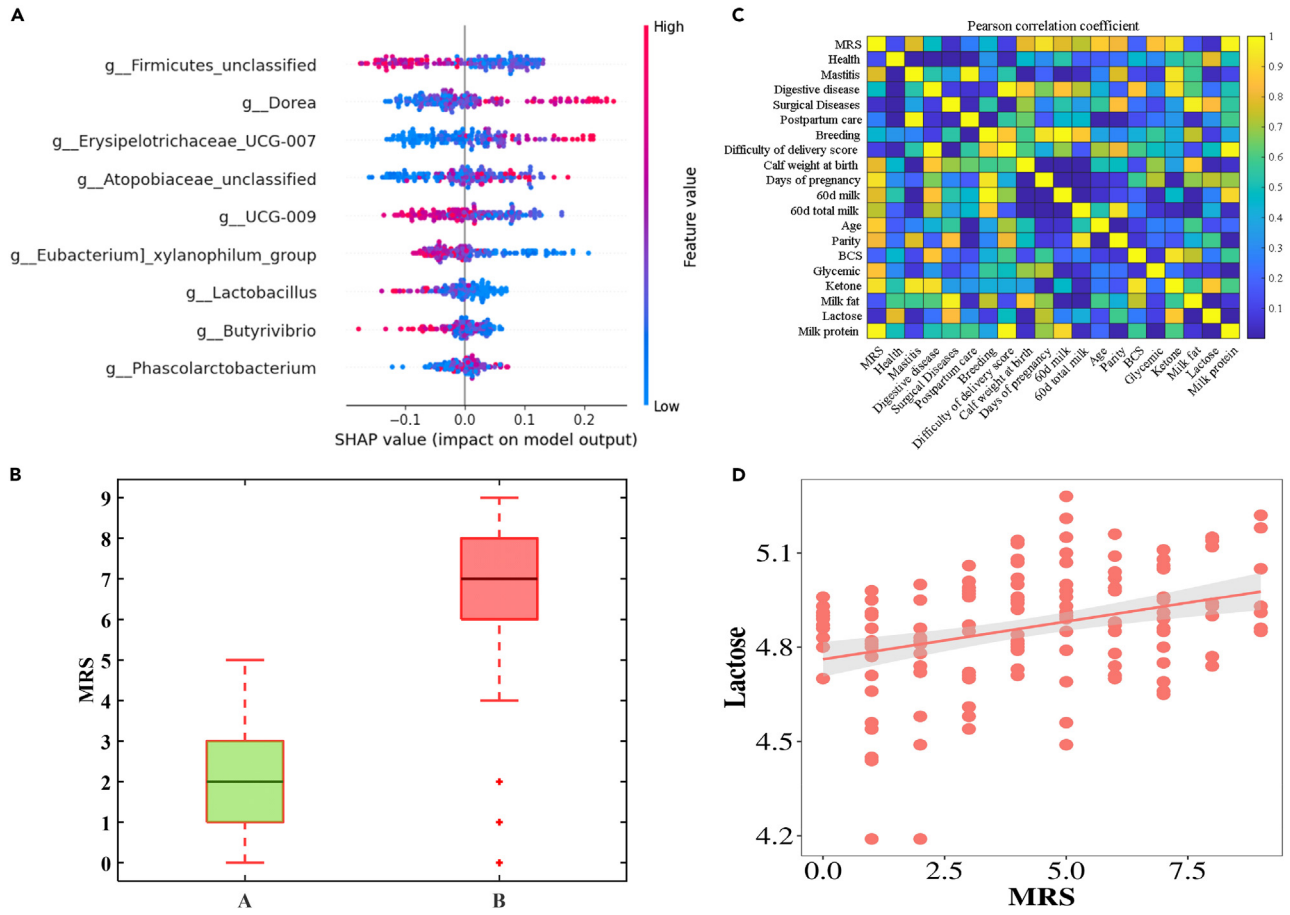


Figure 5. Model Application

(A) Model visualization (SHAP summary plot).

(B) Microbial Risk Score (MRS).

(C) MRS and individual information correlation analysis.

(D) Linear analysis of MRS and lactose.

feature to the output impact. The graphical presentation of the contribution of each feature to the prediction results makes the interpretation of the results clearer and easier to understand. Figure 5A shows the SHAP summary plot of the RF model trained with 9 input features. Each point in the SHAP summary plot corresponds to a sample in the dataset, and its color represents the value of the feature in the particular sample. As the value of the feature increases, the color of the point changes from blue to red. In the summary plot, a positive SHAP value indicates an increase in the influence of the feature on the predicted MUN, and a negative SHAP value indicates a decrease in the influence of the feature on the predicted MUN.

MRSs based on the SHAP approach represent the importance of each microbe for the classifier's decision for a sample, essentially providing more fine-grained information for each participant than RF feature importance. Of these, g__Firmicutes_unclassified has a greater impact on model predictions, with g__Dorea coming in second. The higher the g__Firmicutes_unclassified abundance, the higher the absolute SHAP value. The g__Firmicutes_unclassified is more suitable for the prediction of the normal group of MUN than the high-concentration group. The abundance of g__Firmicutes_unclassified in the normal group was higher than that in the high concentration group, and g__Firmicutes_unclassified can be fed to cows as a probiotic to maintain the stability of milk composition. Of the nine features, g__Phascolarctobacterium had the least impact on the model's predictions (Figure 5A).

To assess the risk associated with gut microbes, we calculated MRS with a score range of 0–9 based on microbiome features. When comparing the two groups, the mean MRS values were 2.2 for the normal group and 6.7 for the high-concentration group. In the normal group, about half of the scores fell within the range of 1–3, while in the high-concentration group, about half of the scores clustered between 6 and 8 (Figure 5B). This indicates that the MRS was higher in the high-concentration group compared to the normal group. Cows with lower MRS tend to be healthier, experience reduced nitrogen wastage, and contribute to improved economic outcomes for dairy farms.

Relationship between dairy cow production information and MRS

Consistent management practices for dairy cows, including feeding routines, ration formulations, and more, result in high and low MRS values that can be attributed to individual characteristics. Within the farm management system, we collect essential data for each cow, such as basic information, event occurrences, milk yield, and other relevant information.

Pearson correlation analysis was performed between the MRS and the individual indicators (Figure 5C). The color-coding in the visualization corresponds to the *p*-value: yellower shades indicate larger *p*-values, while bluer shades indicate smaller ones. The results showed a significant association between surgical disease and lactose. Cows within the 0 to 60 days postpartum range were categorized as 1 if they had a history of surgical disease and 0 if they did not. MRS exhibited a positive correlation with both surgical disease ($p = 0.048$, r = 0.16) and lactose ($p = 0.046$, r = 0.13), as shown in Figure 5D. In animal production, it is possible to enhance MUN by intervening with lactose.

Markers of MUN concentrations of gut microorganisms at various stages

To further investigate whether there is an absolute abundance difference in the relative abundance of the nine microbial markers between 'normal concentration' and 'high concentration'. Log ratio and differential ranking techniques were applied. In Figure 6A, the logarithmic ratios (log2FC) between "normal concentration" and "high concentration" microorganisms for the 9 bacterial genera. The results showed that three microorganisms, g__Erysipelotrichaceae_UCG-007, g__Atopobiaceae_unclassified, and g__Dorea, had log2fc less than 0, which were −1.15, −0.46, and −0.32, respectively. The results showed that the log2FC of three microorganisms, g__Erysipelotrichaceae_UCG-007, g__Atopobiaceae_unclassified, and g__Dorea, were less than 0, with −1.15, −0.46, and −0.32, respectively. log2FC of the other six microorganisms were more than 0, with were 0.14, 0.15, 0.19, 0.23, 0.34, and 0.95. Stamp analysis was performed on the abundance of nine microbial species, five of which were significantly correlated (including g__Firmicutes_unclassified, g__Dorea, g__Erysipelotrichaceae_UCG-007, g__UCG- 009, g__Eubacterium] _xylanophilum_group) (Figure 6B).

DISCUSSION

The diet provided to the subjects on the farm had a crude protein content of 16.6%, which was sufficient to meet the nutritional needs of the cows. However, in 43% of the samples, the MUN concentration exceeded 18 mg/dL, suggesting that nearly half of the cows experienced nitrogen wastage. In addition, Portnoy et al.[13] measured MUN in 14 herds and found that MUN concentrations ranged from 11 to 23 mg/dL. In the measurement of MUN, Bittante[14] studied different breeds of dairy cows (Total 115819) and found that the mean value of MUN was 21.1 mg/dL. Currently, nitrogen wastage in dairy herds is more prevalent. Long-term effects of reduced nitrogen diets have been shown to improve health outcomes in animal models.[15] This study primarily examined the gut microbiome of dairy cows, utilizing an interpretable ML algorithm and gut microbial macrogenomics to identify microorganisms strongly associated with MUN. Additionally, a risk score was developed based on the ML algorithm. Due to the high dimensionality and complexity of macrogenomic data, traditional P CoA cannot distinguish between individuals in the normal and abnormal groups. On the other hand, t-SNE, although capable of preserving local features, faces limitations when mapping high-dimensional datasets to 2-3-dimensional space. To address this challenge, feature selection, as a dimensionality reduction technique, helps eliminate irrelevant, redundant, or noisy features, resulting in a reduced set of essential features from the original dataset.[16] This framework uses the RF algorithm for feature selection of 684 microbial features. Nine microorganisms were selected as model inputs based on feature importance and model performance. The RF algorithm can run efficiently on large datasets and is easily parallelized.

The microbial metabolism of nitrogen has a direct impact on how efficiently ruminants utilize N.[17] The extent of nitrogen metabolism varies across various genera of microorganisms. We conducted a Stamp analysis on the nine microorganisms selected by the RF algorithm. Five of the nine microbial signatures were significantly different, including g__Firmicutes_unclassified, g__Dorea, g__Erysipelotrichaceae_UCG-007, g__UCG- 009, and g__Eubacterium] _xylanophilum_group (Figure 6B). The interactions among microorganisms did not show significant effects at the level of individual microbial genera, but they were most accurately represented when considered in combination. In Figure 6A and 6B, The lower the abundance of g__Firmicutes_unclassified, g__UCG-009, and g__Eubacterium]_xylanophilum_group, the higher the abundance of g__Dorea, and g__Erysipelotrichaceae_UCG-007, and the higher the MUN of the cows, the worse the N wastage. All five microorganisms belong to Phylum Firmicutes. In Firmicutes, most of the genera are able to help plants access resources such as nitrogen, iron, and other mineral.[18] Currently, there is limited understanding of the levels and mechanisms of nitrogen metabolism in the nine microbial genera studied here. Various microbial genera may improve nitrogen utilization by regulating host health.[19,20] The literature showed that the genus g__Firmicutes_unclassified is the dominant contributor to systemic oxidative stress (OS) in postpartum dairy cows.[21] Furthermore, enhancing gut microbiota can alleviate postpartum systemic OS. In the investigation of osteoporosis, a significant difference in g__Firmicutes_unclassified was observed by the mouse model.[22] The g__Dorea was significantly associated with diarrhea and health in calves.[23] The g_Lactobacillus can enhance cattle immunity and improve feed conversion efficiency.[24]

Unlike traditional machine learning, interpretable machine learning not only provides the output of a model, but also explains the feature attribution mechanism between model inputs and outputs.[25] Interpretable models assist in identifying errors, leveraging domain-specific knowledge, and have the potential to enhance inference speed.[26] Nonlinear machine learning models like RF algorithms make it challenging to grasp the model's interpretability directly. Therefore, model-independent methods are always used to interpret the model. We interpret the results of the ML "black box" using SHAP, a popular interpretable machine learning feature attribution mechanism that applies

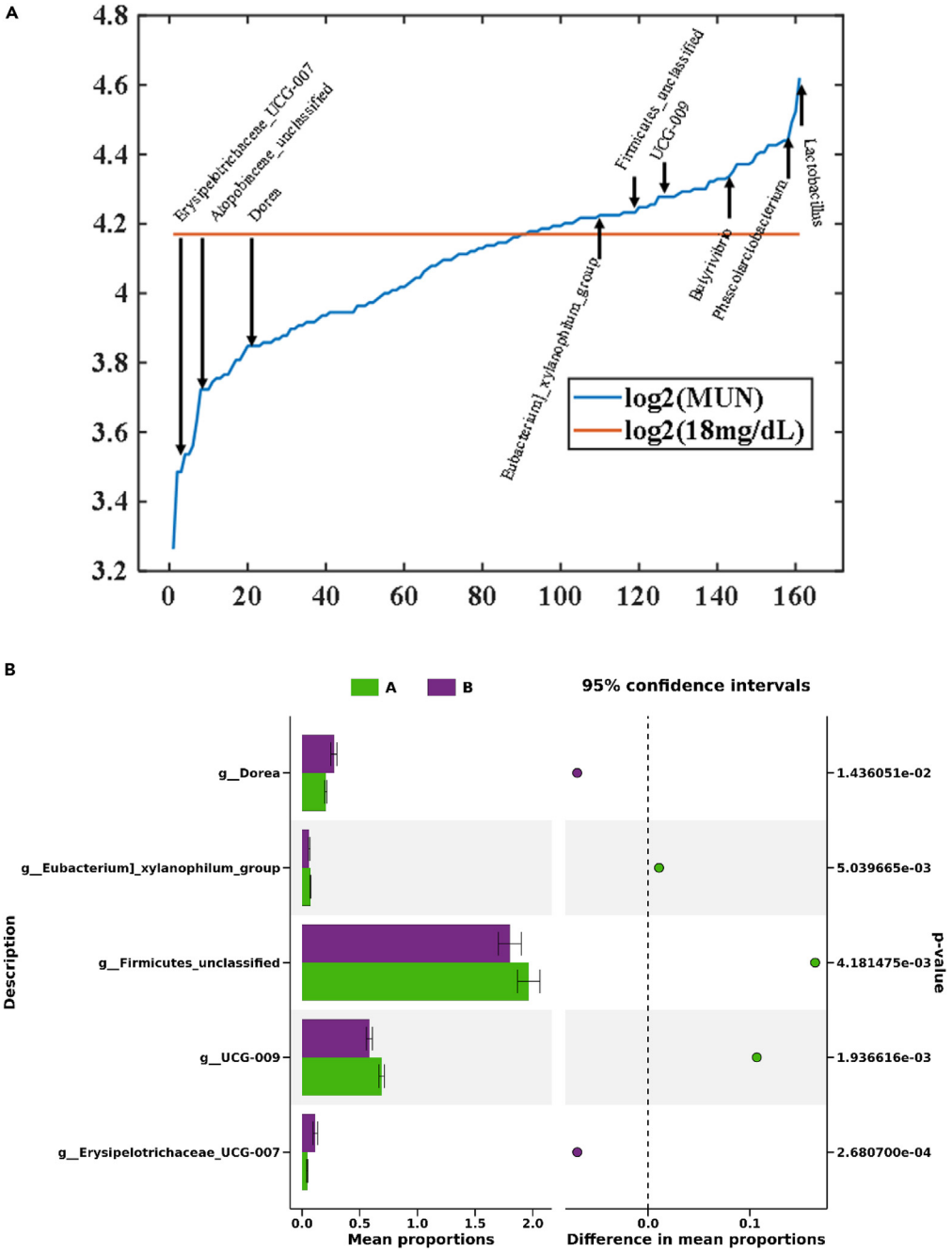


Figure 6. Analysis of nine microbial markers

(A) Nine microorganisms log2FC ranking. The horizontal coordinate is the number of samples and the vertical coordinate is log2 (MUN).

(B) Stamp analysis of abundance of 9 microorganisms.

game-theoretic concepts to measure the attribution of each factor to the output.[27] To further validate the model's accuracy and its relevance to animal production applications, we employed machine learning to identify significant variables that impact milk composition. We utilized an SHAP algorithm-based scoring system to facilitate the categorization and management of cattle. Notably, nitrogen wastage is more severe in dairy cows belonging to the high subgroups. Conversely, cows in the lower subgroups are not only more economically productive but also more environmentally friendly.

Influence on the concentration of MUN is multifaceted, due to the external environment the cow is in, feeding practices, dietary nutrients, and other factors alike. Our primary focus was on collecting individual information regarding the cows, and within this

dataset, we found a significant association between lactose levels and MRS scores. Lactose concentrations tend to remain relatively stable, and there is limited available literature concerning their relationship with other production traits in dairy cows.[28] There might be a correlation between lactose levels and the overall health of the cows. Lactose has recently been proposed as a potential indicator trait for udder health and metabolic status in dairy cows.[29,30] The healthier the cow, the more stable the milk composition. Moreover, MRS was significantly linked to surgical disorders, with the prevalent surgical disorders in this dairy farm primarily being hoof diseases in cows. It is worth noting that there was limited documentation and quantification of the severity of these hoof diseases, making it challenging to apply treatments like lactulose.

Conclusion

Our approach combined interpretable ML modeling with gut microbiology investigations in dairy cows. This study ML framework was effective in predicting MUN (accuracy = 72.7%). We identified 9 gut microorganisms strongly linked to MUN levels and developed MUN risk assessment models using SHAP values, providing valuable tools for dairy farmers to optimize their production practices. This methodology serves as a reference for achieving precision animal husbandry with a focus on individualized approaches.

Limitations of the study

Current measurements of MUN vary between laboratories; assuming an insignificant influence of this factor, we excluded it from our consideration.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Sample collection
  - Sample measurement
  - Grouping
  - DNA extraction, amplification, and sequencing
  - Bioinformatics analysis of the gut microbiota
  - Machine learning process framework
  - Analysis of potential factors for predicting milk composition

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109955.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development (2021YFD1300503-2); National Center of Technology Innovation for Dairy (2022-1); National Key Research and Development Program (2022YFD1602305).

AUTHOR CONTRIBUTIONS

Conceptualization, Q.Y. and Y.Z.; methodology, Q.Y. and Y.S.; software H.W.; investigation, L.Q. and T.W. writing– original draft, Q.Y.; writing– review and editing Y.S.; supervision Y.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 26, 2023
Revised: April 10, 2024
Accepted: May 8, 2024
Published: May 10, 2024

REFERENCES

1. Xue, M.Y., Xie, Y.Y., Zhong, Y., Ma, X.J., Sun, H.Z., and Liu, J.X. (2022). Integrated meta-omics reveals new ruminal microbial features associated with feed efficiency in dairy cattle. Microbiome *10*, 32. https://doi.org/10.1186/s40168-022-01228-9.
2. Mizrahi, I., and Jami, E. (2018). Review: The compositional variation of the rumen microbiome and its effect on host performance and methane emission. Animal *12*, s220–s232. https://doi.org/10.1017/S1751731118001957.
3. Huhtanen, P., Cabezas-Garcia, E.H., Krizsan, S.J., and Shingfield, K.J. (2015). Evaluation of between-cow variation in milk urea and rumen ammonia nitrogen concentrations and the association with nitrogen utilization and diet digestibility in lactating cows. J. Dairy Sci. *98*, 3182–3196. https://doi.org/10.3168/jds.2014-8215.
4. Cabezas-Garcia, E.H., Krizsan, S.J., Shingfield, K.J., and Huhtanen, P. (2017). Between-cow variation in digestion and rumen fermentation variables associated with methane production. J. Dairy Sci. *100*, 4409–4424. https://doi.org/10.3168/jds.2016-12206.
5. Hernández Medina, R., Kutuzova, S., Nielsen, K.N., Johansen, J., Hansen, L.H., Nielsen, M., and Rasmussen, S. (2022). Machine learning and deep learning applications in microbiome research. ISME Commun. *2*, 98. https://doi.org/10.1038/s43705-022-00182-9.
6. Duvallet, C. (2018). Meta-analysis generates and prioritizes hypotheses for translational microbiome research. Microb. Biotechnol. *11*, 273–276. https://doi.org/10.1038/s43705-022-00182-9.
7. Hao-Xun, C., Haudenshield, J.S., Bowen, C.R., and Hartman, G.L. (2017). Metagenome-Wide Association Study and Machine Learning Prediction of Bulk Soil Microbiome and Crop Productivity. Front. Microbiol. *8*, 519. https://doi.org/10.3389/fmicb.2017.00519.
8. Yan, Y., Bao, X., Chen, B., Li, Y., Yin, J., Zhu, G., and Li, Q. (2022). Interpretable machine learning framework reveals microbiome features of oral disease. Microbiol. Res. *265*, 127198. https://doi.org/10.1016/j.micres.2022.127198.
9. Son, H., Hyun, C., Phan, D., and Hwang, H.J. (2019). Data analytic approach for bankruptcy prediction. Expert Syst. Appl. *138*, 112816. https://doi.org/10.1016/j.eswa.2019.07.033.
10. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x.
11. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2021). Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. https://doi.org/10.48550/arXiv.2103.11251.
12. Meske, C., and Bunde, E. (2020). Using Explainable Artificial Intelligence to Increase Trust in Computer Vision. Proceedings of the International Conferenee on Artificial Intelligence in Hu-nan-Computer. https://doi.org/10.48550/arXiv.2002.01543.
13. Portnoy, M., Coon, C., and Barbano, D.M. (2021). Performance evaluation of an enzymatic spectrophotometric method for milk urea nitrogen. J. Dairy Sci. *104*, 11422–11431. https://doi.org/10.3168/jds.2021-20308.
14. Bittante, G. (2022). Effects of breed, farm intensiveness, and cow productivity on infrared predicted milk urea. J. Dairy Sci. *105*, 5084–5096. https://doi.org/10.3168/jds.2021-21105.
15. Solon-Biet, S.M., Walters, K.A., Simanainen, U.K., Mcmahon, A.C., Ruohonen, K., Ballard, J.W.O., Raubenheimer, D., Handelsman, D.J., Le Couteur, D.G., and Simpson, S.J. (2015). Macronutrient balance, reproductive function, and lifespan in aging mice. Proc. Natl. Acad. Sci. USA *112*, 3481–3486. https://doi.org/10.1073/pnas.1422041112.
16. Ma, X.A., Xu, H., and Ju, C. (2023). Class-specific feature selection via maximal dynamic correlation change and minimal redundancy. Expert Syst. Appl. *229*, 120455. https://doi.org/10.1016/j.eswa.2023.120455.
17. Tan, P., Liu, H., Zhao, J., Gu, X., Wei, X., Zhang, X., Ma, N., Johnston, L.J., Bai, Y., Zhang, W., et al. (2021). Amino acids metabolism by rumen microorganisms: Nutrition and ecology strategies to reduce nitrogen emissions from the inside to the outside. Sci. Total Environ. *800*, 149596. https://doi.org/10.1016/j.scitotenv.2021.149596.
18. Hashmi, I., Bindschedler, S., and Junier, P. (2020). Firmicutes. In Beneficial Microbes in Agro-Ecology, pp. 363–396. https://doi.org/10.1016/b978-0-12-823414-3.00018-6.
19. Honerlagen, H., Reyer, H., Abou-Soliman, I., Segelke, D., Ponsuksili, S., Trakooljul, N., Reinsch, N., Kuhla, B., and Wimmers, K. (2023). Microbial signature inferred from genomic breeding selection on milk urea concentration and its relation to proxies of nitrogen-utilization efficiency in Holsteins. J. Dairy Sci. *106*, 4682–4697. https://doi.org/10.3168/jds.2022-22935.
20. Zhu, J., Xie, H., Yang, Z., Chen, J., Yin, J., Tian, P., Wang, H., Zhao, J., Zhang, H., Lu, W., and Chen, W. (2023). Statistical modeling of gut microbiota for personalized health status monitoring. Microbiome *11*, 184. https://doi.org/10.1186/s40168-023-01614-x.
21. Gu, F., Zhu, S., Hou, J., Tang, Y., Liu, J.X., Xu, Q., and Sun, H.Z. (2023). The hindgut microbiome contributes to host oxidative stress in postpartum dairy cows by affecting glutathione synthesis process. Microbiome *11*, 87. https://doi.org/10.1186/s40168-023-01535-9.
22. Qiao, X., Li, X., Wang, Z., Feng, Y., Wei, X., Li, L., Pan, Y., Zhang, K., Zhou, R., Yan, L., et al. (2024). Gut microbial community and fecal metabolomic signatures in different types of osteoporosis animal models. Aging *16*, 1192–1217. https://doi.org/10.18632/aging.205396.
23. He, Z., Ma, Y., Yang, S., Zhang, S., Liu, S., Xiao, J., Wang, Y., Wang, W., Yang, H., Li, S., and Cao, Z. (2022). Gut microbiota-derived ursodeoxycholic acid from neonatal dairy calves improves intestinal homeostasis and colitis to attenuate extended-spectrum β-lactamase-producing enteroaggregative Escherichia coli infection. Microbiome *10*, 79. https://doi.org/10.1186/s40168-022-01269-0.
24. Fu, C., Shah, A.A., Khan, R.U., Khan, M.S., and Wanapat, M. (2023). Emerging trends and applications in health-boosting microorganisms-specific strains for enhancing animal health. Microb. Pathog. *183*, 106290. https://doi.org/10.1016/j.micpath.2023.106290.
25. Min, C., Wen, G., Gou, L., Li, X., and Yang, Z. (2023). Interpretability and causal discovery of the machine learning models to predict the production of CBM wells after hydraulic fracturing. Energy *285*, 129211. https://doi.org/10.1016/j.energy.2023.129211.
26. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine. Learning. https://doi.org/10.48550/arXiv.1806.00069.
27. Lundberg, S., and Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. https://doi.org/10.48550/arXiv.1705.07874.
28. Miglior, F., Sewalem, A., Jamrozik, J., Bohmanova, J., Lefebvre, D.M., and Moore, R.K. (2007). Genetic Analysis of Milk Urea Nitrogen and Lactose and Their Relationships with Other Production Traits in Canadian Holstein Cattle. J. Dairy Sci. *90*, 2468–2479. https://doi.org/10.3168/jds.2006-487.
29. Haile-Mariam, M., and Pryce, J.E. (2017). Genetic parameters for lactose and its correlation with other milk production traits and fitness traits in pasture-based production systems. J. Dairy Sci. *100*, 3754–3766. https://doi.org/10.3168/jds.2016-11952.
30. Ebrahimie, E., Ebrahimi, F., Ebrahimi, M., Tomlinson, S., and Petrovski, K.R. (2018). A large-scale study of indicators of sub-clinical mastitis in dairy cattle by attribute weighting analysis of milk composition features: highlighting the predictive power of lactose and electrical conductivity. J. Dairy Res. *85*, 193–200. https://doi.org/10.1017/S0022029918000249.
31. Gou, W., Ling, C.W., He, Y., Jiang, Z., Fu, Y., Xu, F., Miao, Z., Sun, T.Y., Lin, J.S., Zhu, H.L., et al. (2021). Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated With Type 2 Diabetes. Diabetes Care *44*, 358–366. https://doi.org/10.2337/dc20-1536.
32. Kauffman, A.J., and Stpierre, N.R. (2001). The relationship of milk urea nitrogen to urine nitrogen excretion in Holstein and Jersey cows. J. Dairy Sci. *84*, 2284–2294. https://doi.org/10.3168/jds.S0022-0302(01)74675-9.
33. Logue, J.B., Stedmon, C.A., Kellerman, A.M., Nielsen, N.J., Andersson, A.F., Laudon, H., Lindström, E.S., and Kritzberg, E.S. (2016). Experimental insights into the importance of aquatic bacterial community composition to the degradation of dissolved organic matter. ISME J. *10*, 533–545. https://doi.org/10.1038/ismej.2015.131.
34. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. *37*, 852–857. https://doi.org/10.1038/s41587-019-0209-9.

35. Warton, D.I., Wright, S.T., and Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. Methods Ecol. Evol. *3*, 89–101. https://doi.org/10.1111/j.2041-210X.2011.00127.x.

36. Ding, Y., Fan, L., and Liu, X. (2021). Analysis of feature matrix in machine learning algorithms to predict energy consumption of public buildings. Energy Build. *249*, 111208. https://doi.org/10.1016/j.enbuild.2021.111208.

37. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique (AI Access Foundation). https://doi.org/10.1613/jair.953.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Dairy cow fecal samples	This paper	N/A
Milk samples	This paper	N/A
Critical commercial assays		
16S rDNA sequencing	Lian Chuan Biotechnology	Project code: LC-P20230227009
Microbiological analysis.	Lian Chuan Biotechnology	https://www.omicstudio.cn/home
Deposited data		
Raw high-throughput sequencing data	This paper	NCBI: PRJNA1098610
Individual dairy cow information	This paper	N/A
Measurement of milk composition.	This paper	N/A
Nutritional composition of dairy cow diets.	This paper	N/A
Software and algorithms		
PyCharm	Python Software Foundation	https://www.python.org
MATLAB 2021b	Drawing software	https://ww2.mathworks.cn
Digital Intelligent Dairy Cattle Mobile Platform	Cattle Farm Management System	https://www.yimucloud.com
Decision Tree	This paper	https://data.mendeley.com/datasets/cr5th59dmn/1
Random Forest	This paper	https://data.mendeley.com/datasets/cr5th59dmn/1
XG Boost	This paper	https://data.mendeley.com/datasets/cr5th59dmn/1
SHAP	Yan et al.[8]	https://doi.org/10.1016/j.micres.2022.127198
MRS	Gou et al.[31]	https://doi.org/10.2337/dc20-1536
Other		
Milk composition analyzer (UL 40AC)	You Chuang Technology	N/A
Milk urea nitrogen meter (HLD-21)	Ha Deluo Technology	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Qingyuan Yu (Yuqy3512@163.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The microbiological raw data from this study have been uploaded to NCBI (BioProject ID: PRJNA1098610).
- The code was uploaded to Mendeley Data (https://data.mendeley.com/datasets/cr5th59dmn/1).
- Individual information on dairy cows is available from the corresponding author upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Holstein cows from a local farm in Shang Zhi City, Heilongjiang Province (China) were included in this study. One hundred sixty-one healthy dairy cows with parity 2–4 at 60 days postpartum were selected as research objects. The study was approved by Northeast Agricultural

University Animal Care and Use Committee (Harbin, China) (protocol code NEAUEC20220225) and were in accordance with the recommendations of the academy's guidelines for animal research.

METHOD DETAILS

Sample collection

The cows were distributed over two cattle house and fed the same feed. The diet formulation and nutrient composition are shown in the table below. Feces were collected twice in the morning and evening, mixed and stored at −20°C. The farm uses a rotary milking machine and milks three times a day. The milk samples in this study were collected after the medicated bath and before serving the cup. The collection was at 7:00 every morning (the first milking of the day).

Diet formulation and nutrient composition	
Items	
Diet formulation (Kg)	
Lucerne	2.7
Green fodder	23
Beet meal	1.45
Corn flour	4.05
Cornflakes	3.6
Soybean meal	3.1
Puffed soybean	0.8
Canola meal	2.15
Cottonseed	1.1
Fat powder	0.3
Met	0.03
Purella	0.03
Premix feed	0.6
Water	3.5
Total	46.41
Nutrient composition (%)	
NDF	28.19
ADF	17.09
DM	91.86
CP	16.60
Fat	5.58

Sample measurement

Milking machines were used to identify individual animals, and their numbers were recorded along with the collection of samples. Within 24 h, the milk samples were tested. Milk samples were measured by a milk composition analyzer (UL 40AC) from You Chuang Technologies and a milk urea nitrogen meter (HLD-21) from Ha Deluo Technologies. Feces are collected through the rectum and are used to test for intestinal microorganisms. Samples were subjected to 16S rDNA high-throughput sequencing for microbe-host relationships.

Grouping

According to the National Standard for Food Safety (GB 19301-2010), the optimal range of MUN concentrations is 8–14 mg/dL, and concentrations of 15–18 mg/dL are considered critical, at which point attention should be given. Exceeding 18 mg/dL can lead directly to negative effects.[32] We divided the cows into two groups based on MUN concentrations (as shown in Figure S1).

Normal group

A. Normal MUN (MUN = 8–14,15-18 mg/dL)

High concentration group.

B. High MUN (MUN>18 mg/dL)

Lactating cows were selected 60 days after calving according to the pasture management system. A total of 161 samples were collected, in which fecal samples corresponded to milk samples. Out of these, 20 randomly selected data points were set aside for external validation of the model.

DNA extraction, amplification, and sequencing

The microbiome's total DNA was extracted from fecal samples using the CTAB method, and the quality of the DNA extraction was assessed via agarose gel electrophoresis, while the DNA was quantified using an ultraviolet spectrophotometer. Primers for V3-V4 fragments were selected for PCR amplification and sequenced using a NovaSeq 6000 sequencer.[33] The DADA2 plugin was invoked in the QIIME2 platform for filtering and noise reduction data to generate ASV feature tables.[34] Then further species annotation, diversity analysis, etc. were performed.

Bioinformatics analysis of the gut microbiota

For this study, the statistical analysis was done in the R.4.4 environment. We qualitatively and quantitatively analyzed differences in microbial diversity and composition of the gut microbiota at different feed conversion rates. PCoA analysis was employed as an unconstrained dimensionality reduction method to investigate the similarities or differences in the composition of sample communities. PCoA is based on distances other than Euclidean distances, and identifies potential principal components affecting the differences in the composition of sample communities through dimensionality reduction.[35] t-SNE is a nonlinear dimensionality reduction algorithm for downscaling high-dimensional data to 2 or 3 dimensions and visualization.

Machine learning process framework

We designed an ML framework based on a tree-model to address the data imbalance related to gut microbes and feed conversion. Microorganisms highly correlated with the feed conversion ratio were selected as biomarkers. Tree-structured ML models usually have better prediction performance than individual models (e.g., SVMs), because most tree models employ integrated learning. They can effectively handle nonlinear and high-dimensional data, and they provide valuable feature importance information for feature selection.[36]

Data features can directly affect the predictive performance of our model. The better the selected features, the better the final model performance obtained. In high-dimensional microbial datasets, some unnecessary features, can reduce model prediction accuracy and increase model complexity. Feature engineering is the process of using knowledge about the data domain to create features that enable machine learning algorithms to achieve optimal performance. The ability to simplify the model can improve model accuracy and reduce overfitting. The SMOTE algorithm is processed on the dataset to synthesize a small number of samples to solve the problem of sample imbalance.[37] The algorithms of the modeling framework are all done in a Python 3.8 environment, based on a machine-learning library (Sklearn).

In order to reduce the dimensionality of the data and save computational time, we use RF to downscale the data. Random Forest (RF) is a kind of integrated learning method based on Bagging, that can deal with high dimensional data, and it is not easy to overfit. Considering the accuracy of the model and the number of features, 9 key features were selected according to the feature order (as shown in Figure S2).

For model prediction, Decision Tree, Random Forest, and XGBoost were chosen, and all of the above models are supervised machine learning methods. Cross-validation of the individual models using K-fold (K = 10) reveals that Random Forest predictions are more accurate. In the Appendix, the model error is lowest when the number of trees is at 83 (as shown in Figure S3). The accuracy did not improve as the number of trees continued to increase.

In this study, predictions were modeled using random forests. The dataset was divided into a training set (80%) and a validation set (20%). There are 141 samples in total. The training set is used for model training by providing input features and targets so that the model can learn the mapping relationship between features and targets. The test set is used to evaluate the final performance of the model. After modeling, the SHAP values were applied to the interpretation of the black box model. SHAP relates game theory to local explanations and represents the only possible consistent and locally accurate method of attributing additive features according to expectation.[27] Calculating Shapley values for each feature in various models allows for the interpretation of their predictions, enhancing trust and transparency in machine learning models.

Analysis of potential factors for predicting milk composition

Calculate MRSs by utilizing feature selection after evaluating the SHAP values of each feature, as described by literature.[31] We established a milk composition risk score using the 9 identified microorganisms (ranging from 0 to 9). The formula for calculating MRS is as follows:

$$MRS_i = \sum_{j=1}^{n} Sij$$

where MRS is the microbial risk score for individual i, and Sij is the jth microbial SHAP score for individual i. Correlation analysis was used to analyze the interrelationships between MRS and basic information about individual cows.