COLLEGE OF AGRICULTURE AND LIFE SCIENCES
SCHOOL OF ANIMAL SCIENCES
VIRGINIA TECH

February 19, 2024

Dr. Paul Kononoff
Editor in Chief, Journal of Dairy Science

Dear Dr. Kononoff,

I am writing to express my sincere gratitude for the opportunity to submit a review paper to the Journal of Dairy Science (JDS). Although my initial submission was not accepted, I am appealing this decision, motivated by my commitment to enhancing the quality of machine learning (ML) research within the JDS community.

Addressing the Key Concerns

In the field of ML, reproducibility and comparability are key indicators of high-quality research and essential to the advance of ML-based applications in the dairy industry. However, a notable trend in recent JDS publications relying on ML approaches reveals a lack of detail necessary to meet these reproducibility and comparability standards. This issue is compounded by the rise of tools like ChatGPT, which allow ML code generation with minimal understanding of underlying models. This is akin to more "user-friendly" statistics packages which allow analysis of experimental data without attention to critical details like variance structures, etc. Although traditional graduate training has enabled editors and reviewers to have the background necessary to point out and address these traditional statistical challenges, few reviewers, editors, or researchers are formally cross-trained in dairy science and ML methods. As such, there is a need to support the JDS community in establishing publication standards for ML methods to avoid unnecessary and often unintentional errors in applying these methods. Although I have identified this need personally through my service as a reviewer to JDS, it was also identified by a survey of JDS editors conducted by Dr. Robin White in Spring of 2023 and led to my invitation to present at the NANP-JDS Workshop on Modeling Research Methods at the 2023 ADSA Annual Meeting. ML models are double-edged swords. They range from random forests to neural networks and offer powerful insights into non-linear relationships in data. Yet, their complexity can also lead to overfitting, and their evaluation can result in excessive confidence around their predictive power. The traditional statistical tests used in JDS papers, like ANOVA, cannot evaluate these complex models. This discrepancy underscores a significant gap in current ML research practices within the journal, and an area of opportunity for improving the rigor and quality of research published within JDS.

Given this background information, I am appealing to you to reconsider the decision regarding my submission (JDS.2023-24503, Invited Review: A Guide on Avoiding Common Pitfalls in Model Performance Metrics and Validation). I seek this appeal for two reasons. First, as highlighted above, there are tremendous opportunities to bring ML publication standards to the JDS community to support rigorous, quality, and impactful ML research in a time where the opportunities for error have never been higher. Second, the reviewer comments shared with me pose no technical errors in my

work, but rather suggest that the presentation of the work is not palatable to the JDS community. In the below sections, I will explain the approach I have proposed in the work, highlight its relevance to the JDS community through exploration of recent JDS publications, and propose how I can meaningfully respond to the well-justified reviewer comments that the presentation of the material should be adjusted to better align with the JDS audience needs, interests, and level of knowledge.

Proposing Solutions

There are two critical approaches that I present in the review. These approaches should be adopted in the publication of ML work to support reproducibility and comparability:

- Performance Metrics: Clearly defined metrics are essential for objective evaluation and comparison. For example, confusion can arise from the metric R^2 if it is not specified as referring to the squared Pearson correlation coefficient or the coefficient of determination. Clarifying the meaning of performance metrics in each study is crucial to ensure everyone interprets the results similarly (i.e., comparability).
- Model Evaluation: Evaluation is crucial to simulate unseen data and ensure reproducibility. The portion of data withheld for evaluation impacts the performance metrics, a detail often overlooked but crucial for accurate comparison. For instance, can we consider the model performance to be equal between two studies if they both report an R^2 of 0.85 as their prediction accuracy, but one study withholds 10% of the data for evaluation while the other withholds 20%? Providing some standards for reporting on the approaches to model evaluation, and for interpreting summaries of model evaluation are essential for both reproducibility and comparability of ML studies.

Although these principles are discussed in various statistical literatures, the concepts have not been effectively consolidated into a guide specifically for applying ML in dairy science. My review paper aims to fill this void by providing comprehensive guidelines supplemented by simulation experiments and practical examples from JDS publications.

Case Studies

To provide more concrete examples, I have conducted a survey on six ML papers from the JDS. Five of them are top results by searching on Google Scholar with the keywords "journal of dairy science machine learning." The remaining one was chosen to represent the mentioned concern. In my analysis of the six papers, Becker et al. (2021) and Mota et al. (2021) stand out for their exemplary practices. Both studies effectively demonstrated model evaluation and clearly defined evaluation metrics, setting a standard for future studies. However, some papers did present issues:

- Ghaffari et al. (2019) (cited by 58): The model evaluation approach was not correctly implemented. Feature selection and hyperparameter tuning were omitted from the cross-validation process. Additionally, the risk of overfitting was heightened by the small sample size

(38 cows) relative to the number of predictors (170 metabolites). This type of oversight has been discussed in "The Elements of Statistical Learning" by Hastie et al. (2009, p. 247).

- Frizzarin et al. (2021a) (cited by 49): On page 7440, the same dataset was improperly used for both hyperparameter tuning (the number of factors in partial least squares regression) and external evaluation, potentially biasing the model. Furthermore, on page 7442, the use of the term "cross-validation data" is inappropriate.
- Brand et al. (2021) (cited by 38): In Table 1 on page 4985, the authors did not clarify how hyperparameters, such as retention rate and random selection rate, were determined during cross-validation. Moreover, the method of external evaluation to confirm the model's accuracy is not adequately described, with only a brief mention on page 4988 in the discussion section.
- Frizzarin et al. (2021b) (cited by 12): The study lacks any form of cross-validation. The method of selecting the validation set is neither random nor detailed, potentially skewing the results toward the chosen dataset.

I highlight these challenges not with intent to be critical of the work published in the journal, but to provide practical evidence of the need for resources for the JDS community to support best-practices in application of ML, ensuring that these errors are not propagated in future work.

Proposed Actions for Revision

Feedback on the initial submission highlighted two main areas for improvement: the abstract nature of the paper and the inclusion of unnecessary ML theory details. Again, these limitations are not technical inaccuracies in the work, but reflect the need for a different, more practical and nuanced means of presenting the information. To address these, I plan to collaborate with Dr. Robin White, a colleague at my institution and a key player in my involvement in the ADSA/JDS workshop from 2023, to provide more relatable dairy science examples and streamline the theoretical content. This revision will work to ensure the paper is more engaging and provides a more relatable guide for the JDS community. Toward that goal, I respectfully request reconsideration of my submission. I am convinced that with the opportunity to adjust the approach to presenting the content, the revised paper will significantly contribute to the rigor and clarity of ML research in dairy science.

I appreciate your consideration.

Sincerely,

Chun-Peng James Chen, Ph.D.
Assistant Professor of Animal Data Sciences
Virginia Tech, Blacksburg, VA, USA

References

Becker, C.A., A. Aghalari, M. Marufuzzaman, and A.E. Stone. 2021. Predicting dairy cattle heat stress using machine learning techniques. Journal of Dairy Science 104:501–524. doi:10.3168/jds.2020-18653.

Brand, W., A.T. Wells, S.L. Smith, S.J. Denholm, E. Wall, and M.P. Coffey. 2021. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. Journal of Dairy Science 104:4980–4990. doi:10.3168/jds.2020-18367.

Frizzarin, M., I.C. Gormley, D.P. Berry, T.B. Murphy, A. Casa, A. Lynch, and S. McParland. 2021a. Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods. Journal of Dairy Science 104:7438–7447. doi:10.3168/jds.2020-19576.

Frizzarin, M., T.F. O'Callaghan, T.B. Murphy, D. Hennessy, and A. Casa. 2021b. Application of machine-learning methods to milk mid-infrared spectra for discrimination of cow milk from pasture or total mixed ration diets. Journal of Dairy Science 104:12394–12402. doi:10.3168/jds.2021-20812.

Ghaffari, M.H., A. Jahanbekam, H. Sadri, K. Schuh, G. Dusel, C. Prehn, J. Adamski, C. Koch, and H. Sauerwein. 2019. Metabolomics meets machine learning: Longitudinal metabolite profiling in serum of normal versus overconditioned cows and pathway analysis. Journal of Dairy Science 102:11561–11585. doi:10.3168/jds.2019-17114.

Hastie, T., R. Tibshirani, and J.H. Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics. Springer.

Mota, L.F.M., S. Pegolo, T. Baba, F. Peñagaricano, G. Morota, G. Bittante, and A. Cecchinato. 2021. Evaluating the performance of machine learning methods and variable selection methods for predicting difficult-to-measure traits in Holstein dairy cattle using milk infrared spectral data. Journal of Dairy Science 104:8107–8121. doi:10.3168/jds.2020-19861.