The reviewer’s comments are presented in black

The authors’ responses are highlighted in blue.

Reviewer's Responses to Questions

1. Are the objectives and the rationale of the study clearly stated?

Please provide suggestions to the author(s) on how to improve the clarity of the objectives and rationale of the study. Please number each suggestion so that author(s) can more easily respond.

Reviewer #1: Yes,  
More suggestions on my detailed comments to authors.

2. If applicable, is the application/theory/method/study reported in sufficient detail to allow for its replicability and/or reproducibility?  
  
Please provide suggestions to the author(s) on how to improve the replicability/reproducibility of their study. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #1: Mark as appropriate with an X:  
Yes [X] No [] N/A []  
Provide further comments here:  
  
I would recommend providing simulation and comparison codes as supplementary materials for the paper if you can.

Thank you for your suggestion, I have provided the access link to the code implementation in Section 5 (L776-778)

3. If applicable, are statistical analyses, controls, sampling mechanism, and statistical reporting (e.g., P-values, CIs, effect sizes) appropriate and well described?  
  
Please clearly indicate if the manuscript requires additional peer review by a statistician. Kindly provide suggestions to the author(s) on how to improve the statistical analyses, controls, sampling mechanism, or statistical reporting. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #1: Mark as appropriate with an X:  
Yes [X] No [] N/A []  
Provide further comments here:

4. Could the manuscript benefit from additional tables or figures, or from improving or removing (some of the) existing ones?  
  
Please provide specific suggestions for improvements, removals, or additions of figures or tables. Please number each suggestion so that author(s) can more easily respond.

Reviewer #1: some minor changes to tables and graphs would be beneficial in which I have explained in my detailed comments.

Thank you for your suggestion. I have addressed them accordingly and responded in the detailed comments.

5. If applicable, are the interpretation of results and study conclusions supported by the data?  
  
Please provide suggestions (if needed) to the author(s) on how to improve, tone down, or expand the study interpretations/conclusions. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #1: Mark as appropriate with an X:  
Yes [X] No [] N/A []  
Provide further comments here:

6. Have the authors clearly emphasized the strengths of their study/theory/methods/argument?  
  
Please provide suggestions to the author(s) on how to better emphasize the strengths of their study. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #1: There are rooms for improvements.  
  
More suggestions on my detailed comments to authors.

7. Have the authors clearly stated the limitations of their study/theory/methods/argument?  
  
Please list the limitations that the author(s) need to add or emphasize. Please number each limitation so that author(s) can more easily respond.

Reviewer #1: Not really, and this is one of the areas that they can improve upne greatly.

Thank you for your suggestion. In response to this request, we have compared our results with findings from the literature.

8. Does the manuscript structure, flow or writing need improving (e.g., the addition of subheadings, shortening of text, reorganization of sections, or moving details from one section to another)?  
  
Please provide suggestions to the author(s) on how to improve the manuscript structure and flow. Please number each suggestion so that author(s) can more easily respond.

Reviewer #1: No

9. Could the manuscript benefit from language editing?

Reviewer #1: No

Reviewer #1: This field is optional. If you have any additional suggestions beyond those relevant to the questions above, please number and list them here.

The authors of "Common Pitfalls in Evaluating Model Performance and Strategies for Avoidance" have conducted a much-needed and meticulous study addressing critical aspects often overlooked by researchers and practitioners in precision agriculture and the broader AI/ML research community. The paper's structure and content are commendably well-written and organized. However, two significant weaknesses undermine its potential impact as a seminal reference for years to come.

Simulation Realism and Data Complexity: One notable limitation is the simplicity of the simulations employed, which diverge considerably from real-world data scenarios. To enhance the robustness and applicability of the findings, it is strongly recommended that the authors either incorporate real datasets for model validation and comparison or simulate more complex data that includes co-linearity and non-linear interdependencies among features. By doing so, the arguments and conclusions drawn would gain greater validity and relevance to practical applications.

Thank you for your suggestion. We have included both more complex simulated data and a real-world dataset that we collected ourselves. The simulation procedure was also elaborated on in detail. The updated data description can be found from Lines 309 to 400.

Integration with Precision Agriculture Research: The paper would benefit significantly from deeper integration with existing research in precision agriculture and precision livestock farming. Currently, the literature citations throughout the text are somewhat subtle, making it challenging for less experienced readers to contextualize the paper within the broader domain. It is suggested that the authors systematically compare their results with findings from relevant studies in precision agriculture. This comparative approach would enrich the paper's content, providing a clearer perspective on how their proposed strategies and methodologies contribute to advancing the state of the art in agricultural modeling and decision-making.

We acknowledge the sparsity of the literature review. To address this, we have included more practical applications from the precision agriculture literature to compare with our findings in the simulation results. The updated comparison and discussion content begins at Line 540.

By addressing these two critical points, the authors can elevate their work to become a definitive and highly referenced resource within both the academic and practical realms of precision agriculture and AI/ML research.

more specific comments below:

title: please add "in agricultural studies" to your title

Thank you. We have added it to the title.

L 108: reference to this kind of research must also be given with caution as validity of using 12 features for predicting such a volatile trait (health) with such a small sample size is quite questionable. As a rule of thumb you must have 5 times independent samples as many as your parameter to have a valid prediction.

Thank you for pointing this out. We have replaced it with more suitable literature in L121 to L124.

L 119: in this case this authors have conducted a very neat nested cross validation and hyper-parameters tuning which worth to review/cite and it is in a very relevant topic to your study:  
https://www.sciencedirect.com/science/article/abs/pii/S0168169918309736

Thank you for the thoughtful help in providing relevant literature. We have included in the discussion section (L609 - 613)

L 151: i recommend to add another column to this table indicating each metrics' sensitivity to existence of outliers in data. for example we know RMSE is far more sensitive ( will be inflated ) when outliers exist in the data compared to MAE. Would be useful to see this comparison for other measures in the table.

Thank you for the suggestion. We agree that understanding robustness to outliers is essential when selecting an appropriate metric. When I attempted to implement the idea, I noticed a potential ambiguity in defining the criteria for “outliers.” To address this, we redesigned Study 4 (now renamed Experiment 4) and gradually introduced variance to the error to observe how MAE and RMSE respond differently to predictions with high error variance (i.e., outliers). Interestingly, I discovered that the robustness of MAE is not due to RMSE being overly sensitive to outliers; rather, it stems from MAE being less sensitive to error variance in the context of the bias-variance trade-off. The relevant discussion has been added from L677 to 682.

L 218: Table 2, I recommend use of self-explanatory terms instead of precision, recall, sensitivity, and specificity, please apply this rule throughout the entire manuscript.

also recommend please add False positive rate, False negative rate  and F2 score to this table as they are very important in some agricultural and  health diagnostic applications. It will increase your audience and citation in the long run.

Thank you for the suggestion, and we fully acknowledge the confusion caused by our original arrangement. In this revision, we have prioritized terminology that is more self-explanatory, such as TPR (True Positive Rate), TNR (True Negative Rate), FPR (False Positive Rate), and FNR (False Negative Rate). Additionally, we introduced the F-beta score to provide context for F2, which is a variant of F1, helping to clarify its application in specific use cases.

One aspect we intentionally retained is the use of the terms “precision” and “recall.” While we understand that recall is equivalent to TPR, these terms are widely and conventionally used together in AI/ML literature. Since one of our goals is to create material that is accessible and educational for agricultural scientists who may not be familiar with ML-specific terminology, we believe keeping “precision” and “recall” in this context will facilitate their understanding and bridge the two fields effectively.

We have updated the relevant sections to reflect these changes: Sections 1.6.2, 2.6, and 3.5. Furthermore, we have ensured that the discussion now provides a clearer connection between these metrics and their practical applications in precision agriculture, offering examples to help readers understand how and why these metrics are used. We hope this enhances the clarity and accessibility of the content for readers from diverse backgrounds

L 232: never use this ambiguous term. please use True positive rate (TPR) instead.

Thank you for your suggestion. Unless necessary, such as when introducing precision and recall in the AI/ML context, we prioritize using the terminology TPR throughout the manuscript.

L 243: well your conclusion here is not quite right as one can use false negative rate (FNR) and TNR to optimize for the negative predictions performance in applications that negative cases are costly to miss

Thank you for pointing this out. We completely agree that the original version focused solely on positive-focused metrics (such as TPR and recall), which limited the perspective. To address this, we have added negative-focused metrics (such as TNR and FNR) to provide a more balanced view. Additionally, we have corrected the incorrect statement in the original text. These updates aim to improve clarity and ensure a more comprehensive discussion of the metrics.

L 245: as mentioned above, use TPR, TNR instead.

Thank you for the suggestion. We have addressed them throughout the manuscript.

L 293: It would be interesting to see MAE among these metric as MAE is increasingly popular in literature due to its robustness to outliers.

Thank you for your suggestion, we have reported the result of MAE in Tables 4 and 5 in the section 3.1.

L 329: this is essentially problematic as if FS usually must be an outer loop of your training-testing pipeline or in your case the CV process. FS in real life example might end up with a different feature set in each fold if your data set is not big enough and hence comparing n models of n-fold cross validation (if they are contain different feature sets then) is basically comparing apples and oranges. I would be interested to hear your thoughts around this and possibly to see modify your text here.

We appreciate the concern that feature selection (FS) carried out inside each cross-validation (CV) fold can produce varying subsets of features across folds, thus yielding potentially “different models” in each fold. It is indeed true that, in practice, each CV fold may end up selecting a unique feature set—especially when the dataset is small or the number of candidate features is large. From a **model comparison** standpoint, this can appear to be an “apples vs. oranges” issue if one is trying to directly compare which exact feature set is “best.”

However, from a **performance estimation** perspective, the goal of CV is not to produce a single, fixed set of features at every fold, but rather to obtain an unbiased estimate of how the model, **including its feature selection procedure**, would generalize to new data. Incorporating FS within each fold is precisely the recommended approach to avoid data leakage and artificially inflated performance. If one were to select features once using the entire dataset (including the test folds) prior to CV, this would introduce bias because the **test data** would have influenced the selection process.

We have clarified in our revised manuscript (lines 596–604) an important **exception** for unsupervised FS procedures (e.g., the Successive Projections Algorithm). Because such methods do not leverage the target variable from the test set, they can, in principle, be safely applied to the entire dataset without causing data leakage. Still, this exception is strictly limited to feature selection techniques that are **truly unsupervised** (i.e., they do not use outcome information in any way). We have revised the text accordingly to emphasize this distinction and to ensure readers are aware of when it may be acceptable to conduct FS outside the CV loop.

L 358: please add MAE here as mentioned in previous comments

Thank you for your suggestion. We have included MAE in the simulation study. Please check the updated sections 2.5 and 3.4.

L400: this is really an irrelevant example and no sain researcher will attempt to simulate milk production across species!! A more realistic example would be stimulating milk production across multiple herds in a region and considering each herd as a block.

Thank you for pointing this out. We have removed the unrealistic example from the manuscript.

L 444: I would be interested to know how do you explain this with respect to the central limit theorem?

Thank you for pointing this out. Our simulation study indicates that when the number of folds K is smaller, the variance in the performance estimate decreases. Intuitively, this might seem counterintuitive since we typically think of more “repetitions” yielding more stable estimates.

However, the key mechanism here can be viewed through the lens of the Central Limit Theorem (CLT) and sample size. When K is smaller (say K=5), each test set is larger (20% of the data) compared to K=10 (10% of the data), or especially leave-one-out CV (where each test set is just 1 observation). According to the CLT, the variance of a sample mean is inversely proportional to the sample size. Thus, a larger test set in each fold generally yields a lower variance estimate for that fold. When these fold-specific estimates are averaged, the overall variance of the cross-validated performance can be lower as well.

L487: exactly because of this fact I am really keen for you to add MAE in your analysis

Thank you for the comment. We have revised Experiment 4 to specifically address the questions regarding the outlier robustness of MAE in L 677-682.

L552: as mentioned above, use self explanatory terms such as true negative rate (TNR), instead.

Thank you for the suggestion. We have addressed this concern throughout the manuscript.

L 587: I am still not convinced with this conclusion. as if one includes the block effect as one of the predictor features ( Herds in your example or HYS effect that classically is being used in animal breeding studies) then random CV would be the perfect choice.

Thank you for your comment. We agree that if the block effect (e.g., herd, year-season group) is fully observable and remains consistent within the model’s intended scope, then including it as a predictor in a **closed environment** makes random CV an appropriate choice. In such a scenario, all block levels that will appear in future predictions are already known, allowing us to estimate and incorporate these effects (e.g., herd ID) directly in the model. For example, if we only ever work with the same set of herds over time, using herd as a covariate eliminates the block-induced variation and yields an unbiased performance estimate under random CV.

However, many real-world applications operate in an **open environment** where new or partially unobserved blocks eventually appear. For instance, predicting maize production for a new year (e.g., 2025) cannot be handled simply by fitting a binary variable “year=2025” in the model if no historical data exist for that level of year. Instead, practitioners often decompose the environment into quantitative covariates (e.g., temperature, humidity, cumulative thermal time) that can be learned from past data and then extrapolated to new conditions. This approach, described in plant breeding studies (please check to see L625–646), addresses block effects that cannot be fully enumerated ahead of time. Consequently, when future blocks may differ from all prior examples, block-based CV (or a related strategy) can help avoid overestimating the model’s true generalization performance.

Also considering in real life when a new unseen block (herd) will be added to your test set in random CV you can set the herd effect to missing and then provide a first ( and fairly accurate) prediction for the new unseen cases then over the time they will be added to training set with known herd and phenotype and will provide much more accurate prediction for in future.

We appreciate this practical viewpoint. Indeed, one can initialize an unknown or unobserved block (e.g., a new herd) by assigning a missing or placeholder category and then provide a preliminary prediction. Over time, as data from the new herd accrue (including phenotypes and outcomes ***y***), the model can be retrained or updated to incorporate the now-known block effect, thereby improving future predictions. This incremental or online learning approach is a valid strategy in many real-world scenarios where new data become available over time.

However, for the purpose of **performance evaluation**—particularly when the model will face entirely new blocks **at prediction time**—it remains crucial to use a cross-validation scheme (e.g., block CV) that withholds entire blocks (i.e., new herd or season) during training. Random CV, even with missing or placeholder block levels, can inadvertently “see” partial signals from other blocks and thus yield over-optimistic performance estimates. By contrast, block-based CV provides a more realistic assessment of how the model will perform **before any new block data become available**, aligning with the scenario of predicting on genuinely unseen blocks.

Have you tried to include the block effect in your feature set and investigate the performance metrics?

Thank you for the suggestion. We have indeed considered including block (e.g., herd or season) as a feature. However, rather than using a single categorical variable for block, we decompose it into more granular features—such as temperature or humidity—to capture how each block differs. We then apply a block-based CV to investigate whether these decomposed variables effectively account for block-level variation. This approach allows us to evaluate whether the model truly generalizes when predicting entirely new blocks, rather than relying on an aggregated indicator variable that might not fully capture the nuances of an unseen environment.