Response Letter to Reviewers

**Subject:** Response to Reviewer Comments for "Empirical Analysis of Dataset Size Impact on Classification Performance in Precision Agriculture Using Machine Learning Models"

**Authors:** Khadija LECHQAR, Mohammed ERRAIS

Dear Editor and Reviewers,

We sincerely thank you for the constructive feedback on our manuscript. We have carefully addressed each comment by both reviewers. The manuscript has been revised to improve its scientific rigor, clarity, and contribution to the field. Below, we provide a detailed point-by-point response to each reviewer comment and describe the corresponding changes made to the manuscript.

# Response to Reviewer A

## Title Appropriateness

**Reviewer Comment:** *The current title is broad and lacks technical specificity. It should explicitly reference the methodological framework and models used. A more suitable title could be: "Empirical Analysis of Dataset Size Impact on Classification Performance in Precision Agriculture Using Machine Learning Models."*

**Our Response:** We have adopted the reviewer's suggested title.

**Changes Made:** Title changed from "Empirical study of the impact of dataset size on the performance of classification: application on precision agriculture" to "Empirical Analysis of Dataset Size Impact on Classification Performance in Precision Agriculture Using Machine Learning Models."

## Abstract Issues

**Reviewer Comment:** *The abstract does not sufficiently detail the methodology and results. It should explicitly state the number of datasets, the types of machine learning models used, the performance metrics evaluated, and the key empirical findings with numerical results. Additionally, conclusions such as "there is no correlation" should be supported with statistical analysis.*

**Our Response:** We have completely rewritten the abstract to address these concerns. The revised abstract now includes specific quantitative details and statistical validation of our findings.

### Changes Made:

 Added specific numbers: seven datasets, seven machine learning models, five performance metrics

Included numerical results: dataset size range (100-4,000 samples), statistical correlation coefficient (r

= 0.12, p > 0.05)

Added best-performing model results: Random Forest (88.48% accuracy), Decision Tree (85.37% accuracy)

Included statistical analysis supporting the "no correlation" conclusion Specified the two-phase methodology approach

## Related Work Section

**Reviewer Comment:** *A summary table should be included comparing prior research on dataset size and machine learning performance across different domains. The paper does not critically analyze gaps in the literature. The authors should clearly state what existing studies fail to address and how their work fills these gaps.*

**Our Response:** We have expanded the Related Work section (now Section 2) to address these concerns comprehensively.

### Changes Made:

 Added **Table 1** summarizing prior research with columns for study, domain, dataset size range, models used, and key findings

 Included a comprehensive **Section 2.3 Research Gap Identification** that explicitly states four critical gaps in existing literature.

 Added critical analysis of how each cited work relates to our study and its limitations  Clearly articulated how our work fills the identified gaps

## Discussion Section Missing

**Reviewer Comment:** *There is no explicit Discussion section that compares the obtained results with those in the Related Work section. The authors must contextualize their findings with existing research, discuss why their results diverge (if applicable), and explain any observed performance trends across different models. Justification is needed for the claim that dataset size does not impact classification performance. A statistical test should be conducted to validate this statement.*

**Our Response:** We have added a comprehensive discussion section that addresses all these points.

### Changes Made:

Added **Section 4.3 Statistical Validation of Findings** with detailed statistical analysis including:  Pearson correlation coefficients for all relationships

 ANOVA results comparing performance across size groups  Mann-Whitney U tests for non-parametric comparisons

Added **Section 4.5 Comparison with Existing Literature** that explicitly compares our findings with previous studies and explains consistencies.

Included **Section 4.4 Discussion of Key Findings** with three subsections:  Dataset Quality vs. Dataset Size analysis

 Model-Specific Performance Patterns explanation  Implications for Precision Agriculture

## Research Design Clarity

**Reviewer Comment:** *The research questions and hypotheses are not clearly stated. The study should explicitly define its research objectives in the introduction. The methodology section lacks detailed descriptions of dataset preprocessing, which are critical for reproducibility. The selection criteria for datasets are not clearly justified.*

**Our Response:** We have significantly enhanced the research design clarity throughout the manuscript.

### Changes Made:

 Added **Section 1.1 Research Objectives and Hypotheses** with three specific research questions and explicit hypotheses

 Added **Section 3.1 Dataset Selection and Characteristics** with clear selection criteria:  Representativeness of agricultural classification problems

 Sufficient samples for subdivision analysis

 Diversity in feature types and problem complexity  Balanced representation of agricultural domains

 Added **Section 3.3 Data Preprocessing** with comprehensive details:  Missing value treatment strategies

Feature scaling procedures



 Categorical encoding methods

 Outlier detection and treatment

## Results Interpretation

**Reviewer Comment:** *The claim that dataset size does not affect model performance contradicts widely accepted machine learning principles. A deeper exploration of dataset characteristics is needed. The discussion of why models like Decision Trees and Random Forest perform better is absent. Some metrics require interpretation beyond raw values. ROC curves should be included.*

**Our Response:** We have provided comprehensive interpretation and analysis of the results.

### Changes Made:

 Added **Section 4.4.1 Dataset Quality vs. Dataset Size** providing detailed explanation of why our findings don't contradict ML principles but rather highlight domain-specific considerations

 Added **Section 4.4.2 Model-Specific Performance Patterns** explaining why tree-based algorithms perform better:

 Interpretability alignment with agricultural domain knowledge

 Feature interaction handling capabilities

 Robustness to noise common in agricultural datasets

 Added statistical validation showing no significant correlations (all p-values > 0.05)

 While we acknowledge ROC curves would be valuable, space constraints limit their inclusion.

However, we provide comprehensive ROC-AUC numerical results in **Table 8**

## English Language and Clarity

**Reviewer Comment:** *There are multiple grammatical errors and unclear sentences throughout the paper. Inconsistent terminology should be standardized.*

**Our Response:** We have thoroughly revised the manuscript for language clarity and consistency.

### Changes Made:

 Completely rewrote problematic, introduction.

 Standardized terminology throughout

 Added transitions between sections.

## Formatting Compliance

**Reviewer Comment:** *The manuscript must adhere to the formatting guidelines outlined in the journal's author guidelines.*

**Our Response:** We have revised the manuscript to comply with Informatica journal formatting guidelines.

### Changes Made:

 Adjusted table formatting.

## References Quality Check

**Reviewer Comment:** *Some references do not directly relate to the field of precision agriculture or dataset size impact in machine learning. The authors should incorporate citations from relevant studies published in Informatica.*

**Our Response:** Some of references have to do with the relation between datasetsize and performance but in other fields.

## Dataset Details and Technical Feedback

**Reviewer Comment:** *The description of datasets is too brief. Data splitting strategy is not mentioned. Model hyperparameters are not disclosed. Statistical validation is required. Computational details are missing.*

**Our Response:** We have comprehensively addressed all technical concerns.

### Changes Made:

* Added Section 3.1 Dataset Selection and Characteristics with comprehensive dataset overview including source information
* **Added Section 3.2 Dataset Subdivisions Strategy** for data splitting strategy

Added Section 3.6 Experimental Setup with complete technical details:

 Hardware specifications (Windows 10, Intel Core i5 2.6 GHz, 8GB RAM)

 Software versions (Python 3.8.5, scikit-learn 0.24.2, etc.)  Validation strategy (5-fold cross-validation)

 Random seed for reproducibility (42)

Added **Section 4.3 Statistical Validation of Findings** with comprehensive statistical analysis Included computational efficiency considerations in the discussion

## Possible Incongruencies

**Reviewer Comment:** *Several incongruencies were identified regarding dataset subdivision, reference relevance, and numerical inconsistencies in tables.*

**Our Response:** We have addressed each identified incongruency.

### Changes Made:

 **Dataset subdivision clarity**: Added **Table 3** clearly showing which datasets were subdivided and how sampling was performed with stratification to maintain class distributions

 **Reference relevance**: Added contextual explanations for references [13] and [14], treated the relation between datasize and the performance of machine learning models in other domains

 **Numerical consistency**: Corrected all table formatting to ensure consistent decimal places (two decimal places throughout) and verified all calculations

 **Table reference accuracy**: Ensured all table references in the text are accurate and properly explain the analysis supporting conclusions

 **Dataset size documentation**: Clarified that subdivision was performed with stratified random sampling to maintain original class distributions, explaining any apparent discrepancies in subset sizes

# Response to Automatic Review Comments

## Methodology and Experimental Design

**Reviewer Comment:** *The methodology section lacks detailed description of preprocessing steps and justification for sub-dataset sizes.*

**Our Response:** We have added comprehensive methodological details as described in responses above, including **Section 3.3 Data Preprocessing** and **Section 3.2 Dataset Subdivision Strategy**.

## Analysis and Interpretation of Results

**Reviewer Comment:** *More nuanced discussion regarding performance variability is needed. Dataset quality definition is lacking. Statistical significance testing should be included.*

**Our Response:** These concerns have been addressed through the addition of **Section 4.3 Statistical Validation of Findings** and **Section 4.4 Discussion of Key Findings** with comprehensive analysis of performance variability and quality factors.

## Citations and Literature Review

**Reviewer Comment:** *Some citations appear to generalize or misalign with the context. Integration of Informatica publications is needed.*

**Our Response:** Some of the references treat the impact of datasize on performance of machine learning, but in other fields.

## Clarity and Detail for Replicability

**Reviewer Comment:** *Specific parameters and configurations are missing for replicability.*

**Our Response:** Added comprehensive technical details in **Section 3( including data preprocessing, and experimental setup)**

# Summary of Major Improvements

1. **Enhanced Scientific Rigor**: Added comprehensive statistical analysis with correlation coefficients, ANOVA, and non-parametric tests
2. **Improved Methodology**: Detailed preprocessing procedures, hyperparameter specifications, and experimental setup
3. **Expanded Literature Review**: Added critical gap analysis and comparison table of prior research
4. **Better Results Interpretation**: Added comprehensive discussion sections explaining findings and their implications
5. **Increased Reproducibility**: Provided complete technical specifications and parameter settings
6. **Enhanced Clarity**: Rewrote abstract, improved language throughout, and standardized terminology
7. **Stronger Contribution**: Clearly articulated how this work fills identified gaps in existing literature

We believe these revisions have improved the manuscript's scientific quality, clarity, and contribution to the field. The study now provides a rigorous empirical analysis with proper statistical validation and clear practical implications for precision agriculture applications.

We thank the reviewers for their valuable feedback and look forward to their assessment of the revised manuscript.

Sincerely,

Khadija LECHQAR and Mohammed ERRAIS Faculty of Sciences Ain Chock

Hassan II University, Casablanca, Morocco