**Associate Editor:**

Based on the reviewers' comments and my own reading, I recommend that authors address reviewers' comments carefully and revise the paper accordingly. The authors should significantly improve the quality of the paper, revise the abstract, introduction, main parts and conclusion of the paper, in terms of the presentations, logic, and format. The improvement on English writing and proofreading must be undertaken before resubmission.

We thank the associate editor for deciding in favor of a resubmission. The original version has been thoroughly revised to meet the recommendations of the four referees, improving the introduction, methodological section as well as the empirical application and conclusions.

….

When rewriting the paper we have followed the journal’s “Guide for Authors” and improved English usage by making it more formal. In particular, as emphasized by the fourth reviewer, we have avoided the use of the first-person plural, favoring the passive voice, third-person syntax, and the use of specific nouns.

**Reviewer #1: Contesta Juan con los cálculos de Ricardo y Victor.**

**Los resultados a obtener son: Logistic Regression, SVM, y la sensibilidad con SHAP,**

We thank the reviewer for carefully reading the manuscript and offering helpful comments and recommendations that have improved it. Below, we indicate how we address them in the new version. All modifications have been highlighted in yellow in the new text. Overall, the manuscript has been thoroughly revised to meet the recommendations of three additional referees, improving the introduction, methodological section as well as the empirical application and conclusions.

1. Provide a theoretical justification for the classification paradigm beyond intuition and visuals. Clarify the epistemological status of the probability estimates.

*Response*: TBD

2. Use a simpler and statistically appropriate classifier (e.g., logistic regression, decision tree) for the given data scale, or augment the dataset meaningfully with real observations.

*Response*: Apply and compare results with logistic regression. Ejemplo empírico. Annex… comentando que LR es paramétrico y SVM es no paramétrico. Citar a Efron sobre cuando lo paramétrico o no-paramétrico es adecuado. Cuando es ancha es mejor lo no pamétrico.

3. What does it mean operationally or policy-wise for a firm to have a "65% chance" of being efficient? Efficiency is not inherently probabilistic. The paper should critically discuss whether this probability reflects data uncertainty, model uncertainty, or intrinsic inefficiency fuzziness.

*Response*: TBD

4. The discussion on XAI and counterfactuals lacks depth and rigorous implementation. Sensitivity analysis and directional vectors are basic perturbation tools—not advanced XAI techniques.

*Response*: TBD

5. There is no benchmarking against alternative classifiers (e.g., SVMs, RFs), nor any discussion on model robustness.

*Response*: TBD

6. There is no concrete comparison or ablation study showing how variable importance vectors derived from SA outperform alternative approaches (e.g., SHAP, permutation importance, or linear approximations).

*Response*: Utilizar SHAP. Quizá se puede juntar con el 7.

7. Include ablation studies comparing XAI techniques and directional vectors to demonstrate the contribution of the proposed integration.

8. How does probabilistic classification relate to the fundamental axioms of production theory (convexity, monotonicity, etc.)?

9. What are the implications for efficiency measurement when probabilistic thresholds substitute for boundary estimation?

10. How does this affect traditional properties of DEA models such as translation invariance, unit invariance, or returns to scale?

**Reviewer #2:**

This manuscript details an extensive study by the authors on "Probability-based Technical Efficiency Analysis through Machine Learning". To enhance readability and build a knowledge base for future studies, I suggest the following comments and suggestions.

We thank the reviewer for carefully reading the manuscript and offering helpful comments and recommendations that have improved it. Below, we indicate how we address them in the new version. All modifications have been highlighted in yellow in the new text. Overall, the manuscript has been thoroughly revised to meet the recommendations of three additional referees, improving the introduction, methodological section as well as the empirical application and conclusions.

1. Kindly include a figure illustrating the methodology adopted in this manuscript.

*Response*: **Victor** Prácticamente el diagrama de flujos del algoritmo, que es el núcleo del graphical abstract. Pseudo Código, pseudo algorithm.

2. Please provide detailed information about the platform or environment used to execute the models.

*Response*: **Ricardo**

*Computational experiments were carried out on a workstation running Microsoft Windows 10 Pro (64-bit, version 10.0.19045) equipped with an Intel Core i7-9700K CPU @ 3.60GHz (8 cores), 16 GB RAM, and integrated graphics (Intel UHD Graphics 630). All analyses were conducted in R version 4.4.3 (2025-02-28, “Trophy Case”). The package Benchmarking (0.33) was employed for efficiency analysis, while caret (7.0-1) served as the main framework for training machine learning models. For model interpretability, the study relied on iml (0.11.4), shapviz (0.10.2), fastshap (0.1.1), and rminer (1.4.8).*

3. Kindly elaborate on the limitations of the current study.

*Response*: Enhance the last section to “Conclusions, **limitations** and future work”.

4. Please present the highlights of this manuscript using bullet points.

*Response*: **JL**

5. Include the statistical improvement results comparing the proposed method with traditional DEA.

*Response*: Calcular la DDF sobre los datos originales y comparar los resultados. ¿Enseñarselo a él solo? ¿frase/anexo? ¿Nota a pie de página?

6. How was the data split performed? If it was randomized, please analyze the model's performance with various training data percentages (e.g., from 50% to 90%) to determine the optimal data split

*Response*: **Ricardo**

*The data splitting is random, but stratified within the levels of y when y is a factor. The function createDataPartition() of the package ‘caret’, called during the fitting phase, tries to balance the class distributions in classification tasks. Due to this, the new table reports the performance of NN and SVM when changing the number of folds. We tested CV=2 (50%), CV=5 (80%), and CV=10 (90%). For NN, the performance was very similar across folds, but CV=5 achieved the best Balanced Accuracy with the lowest standard deviation. For SVM, results were also close, but CV=2 showed higher variability.*

*There is a trade-off: with CV=2, the training set is small but the test set is large and well-represented, even for the minority class. With CV=10, the training set is large but the validation folds are small. In unbalanced datasets like ours, caret forces (if it is possible) at least one minority-class unit into each fold, which may lead to validation sets with too few observations per class, resulting in unstable performance metrics. For SVM, the overall performance was slightly lower than NN, but results improved as the class balance increased.*

*Two main conclusions can be drawn from this dataset. First, the number of folds does not substantially alter model performance; however, when test sets are small and may contain very few (or no) minority-class observations, using many folds is not advisable—rather, it is preferable to choose a number of folds that ensures more than one (ideally several) minority-class observations per fold. Second, improving class balance typically has a greater impact on performance than changing the number of folds, and the size of that impact depends on the ML technique: in our data, balancing hardly affects NN, but it clearly improves SVM.*

7. Kindly provide a graphical abstract summarizing the entire work. **Victor**

A screenshot of a computer

AI-generated content may be incorrect.

<https://www.elsevier.com/researcher/author/tools-and-resources/graphical-abstract>

Ejemplo de artículo con graphical abstract sacado de la revista

<https://www.sciencedirect.com/science/article/abs/pii/S0952197623015117>

A diagram of a house

AI-generated content may be incorrect.

A diagram of a memory processing process

AI-generated content may be incorrect.

*A white paper with black text

AI-generated content may be incorrect.*

*Response*: TBD

**Reviewer #3:**

We thank the reviewer for carefully reading the manuscript and offering helpful comments and recommendations that have improved it. Below, we indicate how we address them in the new version. All modifications have been highlighted in yellow in the new text. Overall, the manuscript has been thoroughly revised to meet the recommendations of three additional referees, improving the introduction, methodological section as well as the empirical application and conclusions.

1. The abstract includes broad claims about the advantages of the proposed method without providing specific metrics or evidence to support these claims. The abstract does not summarize key findings or results, such as specific improvements in efficiency assessments, which are crucial for conveying the significance of the research. The abstract could better highlight the practical implications and potential applications of the research findings in real-world settings.

*Response:* ***JL Hacer al final***

2. The introduction focuses heavily on the limitations of traditional DEA without adequately acknowledging existing solutions or advancements in the field, which may lead to a skewed perspective. The identification of gaps in the literature could be more specific and clearly articulated. Without precise gaps, the rationale for the proposed approach may appear weak.

*Response*: **Juan.** Stochastic DEA que utiliza el concepto de pababilidad, etc…

3. The introduction could better highlight the practical implications of the research, particularly in terms of how it can benefit decision-makers in various industries.

*Response:* ***JL Hacer al final.*** Nuevo proceso d bechmarking con probabilidad, eficiencia, etc.

4. The section states that 15 out of 97 firms are classified as efficient based on the additive DEA model. However, later it mentions that the NN model predicts 14 firms with a probability of efficiency exceeding 0.5. This slight discrepancy could suggest inconsistencies in classifications between the two methods.

*Response:* **Juan**. La técnica no tiene como propósito replicar el DEA, sino aprender el etiquetado del DEA. Busca generalizar.

5. The text notes that balancing the classes improves precision but at the expense of sensitivity. If the balanced model significantly enhances overall predictive capability yet still results in lower sensitivity, this raises questions about the trade-offs associated with class balancing.

*Response:* **Ricardo/Victor** *TBD.* Meter una frase expliando este trade-off.

*Certainly, as noted, there is a trade-off between sensitivity or recall (the detection of true positives) and precision (the proportion of predicted positives that are correctly identified) in ML problems. If the primary objective is to detect all positive cases (units labeled as “efficient”), sensitivity becomes the key metric. However, this increases the risk that the optimal set of hyperparameters will overclassify units as efficient, thereby reducing precision. Conversely, if the goal is to achieve high precision, the resulting hyperparameters will classify only a few units as efficient—those with very high certainty—at the expense of lower sensitivity. Moreover, in our context, the datasets are usually unbalanced, which makes it more difficult for ML models to learn about the minority class because of the limited number of examples.*

*By balancing the datasets, the objective is to increase both sensitivity and precision simultaneously, since the model has more units to learn from. However, this outcome is not guaranteed, as it depends on several decisions made by the researcher and on the characteristics of the original dataset.*

*First, the choice of metric for hyperparameter selection is critical. Balanced Accuracy or F1 score account for both classes and are recommended for imbalanced datasets, but they may be less informative in more balanced scenarios. Evaluating multiple performance metrics simultaneously is therefore important to fully understand the model. To ensure consistency in our analysis, we evaluated all models using the same metrics and applied them in a fixed order in the case of ties. Nevertheless, alternative metrics such as precision (favoring stricter models) or F1 score (which combines sensitivity and precision) could also be considered.*

*Second, machine-learning algorithms differ in their propensity to overfit and in how they respond to class rebalancing. In our experiments, the neural network was comparatively stable under class imbalance and rebalancing—its behaviour depended mainly on the optimisation target and, in imbalanced datasets, on the number of cross-validation folds—whereas the SVM was more sensitive to class weights and sampling ratios. Increasing the effective weight of the minority class raised sensitivity for that class, typically at the cost of precision (more false positives).*

*Third, when the dataset is small and high-dimensional, pursuing near-perfect sensitivity for the positive class (here, “efficient”) is often counterproductive due to the curse of dimensionality—* *for example, when the number of inputs/outputs is large relative to the sample size, DEA tends to classify a large fraction of units as efficient because the frontier intersects many observations; this inflated efficiency is a small-sample/high-dimensional artefact rather than genuine technical efficiency. This amplifies label noise and encourages overfitting. In such settings, it is preferable to prioritise generalisation and robustness, monitoring Accuracy or Balanced Accuracy.*

*Fourth, the performance metrics reported in the tables are computed at a fixed 0.50 threshold: any unit with predicted probability > 0.5 is classified as efficient. Accordingly, the confusion matrix and derived metrics (precision, sensitivity, F1, etc.) are evaluated at that cutoff. However, as stated in the paper, in practice we do not regard a unit with a predicted efficiency of 0.51 as truly efficient. After model selection, the operational threshold is chosen by the analyst (e.g., 0.75, 0.85, 0.95) to reflect the desired confidence level and the relative costs of errors. This choice can reclassify borderline units—those that appeared efficient at 0.50—as weakly efficient or even assign them a non-zero inefficiency.*

*As a future direction,* *it would be interesting to examine how performance metrics change as the efficiency threshold varies for each trained model. For example, raising the threshold from 0.50 to 0.80 may sharply reduce sensitivity, as the model would detect fewer clearly efficient units, even if precision increases. Such an analysis would provide a more complete picture of each model’s performance across operating points and help ensure that model selection is robust to the choice of threshold.*

6. The results indicate that the unbalanced model showed the highest sensitivity but poor precision, while balanced models improved precision. If the unbalanced model's sensitivity leads to overclassification, how can it be deemed effective in practical scenarios?

*Response:* **Ricardo/Victor** *TBD*

*As you mentioned, an unbalanced model may fail to correctly learn to identify efficient observations. To make the model account for the minority class, the metric optimized during cross-validation must reflect it, so that the model penalizes errors on the minority class more than on the majority class. Balanced Accuracy—i.e., sensitivity (true positives detected) and specificity (true negatives detected)—is one option to use during cross-validation.*

*In this way the model will consider the minority class, but it may still fail to learn correctly and end up predicting more units as belonging to the minority class than there actually are.*

*As I mentioned in the previous question, even if it predicts many units as efficient, many of them may be weakly efficient. The decision threshold is important, and the standard 0.5 is quite low to decide that a unit belongs to the efficient class.*

*If it still keeps classifying many units as efficient, using an unbalanced model can be useful when making changes in DMUs is very costly and changing the production process entails high risk. In this case, being more conservative can be a good option, although there are other methods to handle an imbalanced dataset—besides rebalancing with synthetic units or undersampling/oversampling—such as using repeated cross-validation.*

7. The section claims that firms need to make significant adjustments to reach efficiency thresholds. However, if some firms are already close to the minimum observed resource levels (like employees), this could conflict with the expectation of achieving efficiency through further reductions.

*Response:* ***José Luis*** *TBD*

8. The importance of 'Operating income' is stated as 50.8%, leading to key decisions in the model. If this variable heavily influences outcomes, yet the adjustments required for efficiency show minimal changes in operating income, this could indicate a contradiction in how input and output adjustments are linked to efficiency predictions.

*Response:* ***José Luis*** *TBD. Parece equivocado según las tablas 5 y 6*

9. The conclusion may not effectively summarize the main findings of the study, leaving readers without a clear takeaway. Also, it may not clearly relate back to the original research objectives stated in the introduction.

*Response: TBD. Al final*

**Reviewer #4:**

The authors proposed an integration of Data Envelopment Analysis (DEA) and Machine Learning (ML) to improve the robustness and interpretability of efficiency assessments for decision-making units (DMUs). The approach introduced a classification-based framework that reformulates DEA as a probabilistic classification task. Also incorporate Explainable AI (XAI) technique for sensitivity analysis and counterfactual explanations. Furthermore, the framework introduces a dynamic peer selection mechanism based on probability thresholds, offering adaptable and more informative benchmarking strategies. The authors employ data from food industries in Spain for applications of the proposed ML-DEA model.

However, authors have not convincingly presented their specific objectives, research findings and conclusions with some poor documentation.

We thank the reviewer for carefully reading the manuscript and offering helpful comments and recommendations that have improved it. Below, we indicate how we address them in the new version. All modifications have been highlighted in yellow in the new text. Overall, the manuscript has been thoroughly revised to meet the recommendations of three additional referees, improving the introduction, methodological section as well as the empirical application.

I would recommend for a major revision before consideration for acceptance in the journal after the following comments are addressed:

1. Abstract

We incorporate . . . It is advisable and appropriate to use passive voice, third person or specific nouns rather than we (pronoun) in academic writing.

*Response:* Following the referee’s suggestion in this first comment and additional ones below we have completely revised the manuscript to avoid the use of the first-person plural (we, ours, us, …) favoring the passive voice, the use of third-person grammar structures, as well as appropriate specific nouns.

2. Introduction

"However, despite its widespread adoption and commendable performance, traditional DEA approaches may encounter limitations in capturing the intricate patterns and structures characterizing involved processes and datasets (see, for example, a recent criticism in Sahil et al., 2025)." ?????? -

What was the criticism ? Kindly highlight or state at least one of such to support your claim. This will validate your statement above rather than "see, for example, a recent criticism"??

*Response:* ***Juan*** *TBD*

"However, there exist certain gaps that we believe the novel approach introduced in this paper can address. Before mentioning these gaps, we briefly review the main contributions relating ML and DEA. As we are aware, there are two predominant streams of research in the literature that explore their integration." It is advisable and appropriate to use passive voice, third person or specific nouns rather than "we" (pronoun) in academic writing.

"The first stream focuses on adapting existing ML techniques to ensure that the predictive function, typically representing a production function in our context, . . . "

It would be better to say "in this context" or "in the context of this study" "in this present study etc "

*Response:* Following these suggestions the new phrase read as follows: “The first stream focuses on adapting existing ML techniques to ensure that the predictive function, typically representing a production function in the efficiency measurement context,…”

"Methodological Innovation: As we are aware, for the first time in the literature, we propose a classification-based machine learning approach in the second stage of a DEA-ML hybrid framework, moving beyond the conventional regression-based techniques. In the first stage, we rely on the concept of Pareto-dominance to distinguish between two classes of DMUs: efficient and inefficient."

It is advisable and appropriate to use passive voice, third person or specific nouns rather than "we" (pronoun) in academic writing. - Kindly revised the Introduction section/manuscript and use appropriate "passive voice/third person/specific nouns" in presenting the authors views and reports. It is recommended that the authors remove the ALL the "We" and "Our" in this manuscript and use the appropriate reporting voice.

*Response:* As stated in our answer to the first comment, we have rewritten all sentences to avoid the use of first-person plurals (we, ours, us,…)

Also, "for the first time in the literature" is grammatically wrong. Kindly revise.

*Response:* Thank you. The first phrase of this paragraph: “As we are aware, for the first time in the literature, we propose a classification-based machine learning approach…”, has been changed to “This study introduces a classification-based machine learning approach into the literature…”

". . . we eliminate the risk of propagating errors inherent in continuous value predictions, thereby providing a clearer and more reliable distinction between the two classes. In the second stage, we apply classification models to predict the probability of being classified as efficient using the input and output variables."

"Reinterpreting DEA as a Classification Problem: Our approach allows reframing traditional DEA as a classification problem. First, the DEA technology differentiates the variable space between two regions: technically feasible and technically infeasible production processes."

"As a result, for any pre-defined level of probability, our method allows classifying DMUs into efficient or inefficient classes. Additionally, under the new paradigm, technical inefficiency measures can be interpreted as distances to the efficient boundary at a predefined probability threshold."

"Algorithm-Agnostic Approach for Robust Efficiency Assessments: A key advantage of our framework is its flexibility in algorithm selection. Unlike conventional DEA-ML models that rely on a specific regression technique, our method is not tied to a particular classification algorithm. This flexibility might eventually allow us to experiment with a wider range of ML models—including decision trees, Support Vector Machines, Neural Networks, and ensemble methods—ensuring that the results remain robust and consistent across multiple techniques. Nevertheless, for the sake of simplicity, in this seminal paper we focus on Neural Networks."

*Response:* As explained in our first answer, the use of the first-person plural has been avoided.

Also, in section 3. Integrating ML techniques for classification and Data Envelopment Analysis

"In this section, we perform the integration of machine learning techniques for classification tasks with Data Envelopment Analysis to enhance the measurement of technical efficiency. By combining the strengths of both methodologies, we aim to provide robust and insightful efficiency assessments of a set of DMUs. As aforementioned, while other ML classification methods could be considered, we focus on Neural Networks."

*Response:* Again, the use of the first-person plural has been corrected.

I recommend a major review on this manuscript's Introduction section/review of literature and justification of the research, incorporating the appropriate usage of words and sentence structure suitable for reporting academic research. What does the authors mean by "Nevertheless, for the sake of simplicity, in this seminal paper we focus on Neural Networks" which seminar paper?

*Response:* Thank you. What we intended to say is that the framework that we introduce in this study can use any ML method (either decision trees, Support Vector Machines, Neural Networks, and ensemble methods) as classifying algorithm, and that, given the impossibility of using all these methods in a single paper, we choose Neural Networks as⎯arguably⎯the most prominent ML classification technique. To prevent misunderstandings, we have rephrased the text in the following way: “Nevertheless, for the sake of simplicity, and given the impossibility of implementing all these models, this initial study focuses on Neural Networks as one of the most well-known and predominant ML classification techniques.…”.

Addendum: Many abbreviations were used in the Introduction Section, in which their full meaning would be necessary at the first mentioning or provide a Nomenclature section at the beginning.

*Response:* TBD. Thank you Una vez revisada la introducción, ver si está todo aclarado. Revisar en detalle y decirle que lo hemos hecho. At the first mentioning we have explained the acronym.

3. Background

In the equation (3) above what does "s.t" stands for? If it is "subject to" it should be written in full.

*Response:* Done.

Consider section 3.2 Probability-based efficiency analysis

"The core concept underlying our model is a multi-stage methodology aimed at enhancing efficiency assessment through the fusion of DEA and ML techniques."

It would better and appropriate to say, "The core concept underlying the proposed model is …."

*Response:* Thank you, we have made this change.

Consider section 4.3. Technical efficiency and benchmark peers

"We report in Table 4 the inefficiency values and benchmark peers for each firm using three thresholds corresponding to p = 0.75, p = 0.85 and p = 0.95."

This is better presented as "As reported in Table 4 . . . " - review the manuscript and remove "We", "us" and "our", and report these activities in the passive voice/third person/specific nouns

*Response:* Thank you. We follow this example throughout the text to replace the use of the first-person plural.

4. Tables and Figures

The Tables and Figures should be properly named/titled accordingly for clarity. Table 1 "upper panel" and "lower panel" are not appropriate (See Authors guide). It should be titled at the top of Table 1 - say Table 1a and 1b followed by the title. Similarly, Figures 6a and 6b should be titled underneath appropriately Fig. 6a and Fig. 6b, and not as Figure 6a (Top), nor Figure 6b (bottom).

Also, Tables presentation needs to be consistent based on Journal requirements (See Author's guide).

*Response:* Thank you. Following the reviewer’s advice we now use Table 1a and Table 1b, and Figure 6a and Figure 6b to identify the different sets of results and illustrations; i.e., any reference to upper o or lower panels and figures has been deleted.

5. Conclusions and future work

The authors should review and improve the presentation of this section, by first reducing the length and appropriately items the significance of and findings of the study and relevant implications.

More so, the authors should avoid using "We", "us" and "our", rather make use of the passive voice in presenting their research findings.

Furthermore, using First, Second… Fifth etc repeatedly on conclusion and future work can be avoided and addressed using bullets or paragraph and appropriate adverb to introduce each point.

*Response:* Following this advice we have rewritten the section on “Conclusions and future work” (now, “Conclusions, limitations and future research”), avoiding the use of the first-person plural and the numbering of the contributions of the study.

6. Performance metrics of proposed model

"To measure the performance of each specific balance proportion in the second stage, we rely on standard metrics commonly used in ML classification problems. We consider measures that not only pay attention to the two classes like 'Balanced accuracy' (the average of true predictions for each class), but also that are primarily focused on metrics related to the 'efficient' (minority) class of interest, such as 'Precision' (the proportion of positive predictions that are actually correct), 'Sensitivity' (i.e., the proportion of actual positives correctly identified) and, finally, 'F1-score' (the harmonic mean of 'Precision' and 'Sensitivity', balancing detection accuracy and reliability). Depending on the specific context and the metrics of interest, we evaluate model performance by ranking the results according to our chosen criteria."

The authors did not provide convincing explanation on the performance metrics formula nor any clear interpretations to which the efficacy of proposed model (ML-DEA) is benchmarked, to ascertain its robustness and predictive accuracy on DMU. What are the ranges of values, scholarly citations for acceptable F1-score, Precision and Sensitivity values?

*Response:* **Ricardo/Victor**TBD. Relacionado con el 5 y 6 del tercer evaluador. Esta parte se ha reescrito.

*This comment is important because it sheds light on the issue. There is no single threshold or specific metric value that guarantees a model is performing ‘correctly’. Acceptable values depend on the application, error costs (FP and FN), and class imbalance.*

*For example, in medical diagnosis (e.g., sepsis alerts) false negatives are extremely costly, so high ‘sensitivity’ is typically prioritized even at the expense of ‘precision’. In spam filtering the opposite applies: false positives (legitimate emails blocked) are more harmful, so ‘precision’ is optimized while some loss in ‘sensitivity’ is tolerated. In credit scoring, the trade-off between FP and FN is determined by business risk and safety constraints, leading to different ‘acceptable’ metric levels depending on the application. Moreover, class imbalance can bias standard performance metrics: two datasets that are otherwise similar but differ in minority-class prevalence can yield very different metric values—what is acceptable in one may indicate poor performance in the other. Jeni et al. (2013) propose mitigating this by reporting both the obtained and skew-normalized metrics, and by complementing ROC analysis with precision–recall curves to avoid masking effects under skew.*

*Within the proposed methodology, performance can be compared across multiple models, enabling selection of the best-performing alternative. Nevertheless, it is the investigator’s responsibility to decide which metrics to prioritize and what constitutes sufficient performance. This choice gives the analyst greater control, as they oversee the process and determine the final model selection. If results do not meet expectations, one can explore additional training procedures (e.g., repeated cross-validation), tune hyperparameters, or evaluate alternative machine-learning techniques.*

7. References

The references need to be presented properly, and DOIs clickable links are highly recommended per requirements and easy accessibility to validate.

*Response:* We have formatted the references following the instructions in “Guide for Authors” and added clickable DOIs.

8. Guide for Authors

I recommend that the authors carefully adhere to the EAAI guide while making the revisions.

https://www.sciencedirect.com/journal/engineering-applications-of-artificial-intelligence/publish/guide-for-authors

*Response:* We have formatted the manuscript following the instructions provided in the above link “Guide for Authors”.

We thank the referee for these useful comments and suggestions that have greatly improved the paper.