Preprints.org

Technical Note

Not peer-reviewed version

# When Error Metrics Contradict: Clarifying RMSE and MAE in Machine Learning Evaluation

Walter Chen * and Kieu Anh Nguyen

Posted Date: 30 January 2026

doi: 10.20944/preprints202601.2392.v1

Keywords: RMSE; MAE; metrics; Random Forest; XGBoost

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2026 doi:10.20944/preprints202601.2392.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

*Technical Note*

# When Error Metrics Contradict: Clarifying RMSE and MAE in Machine Learning Evaluation

**Walter Chen * and Kieu Anh Nguyen**

Department of Civil Engineering, National Taipei University of Technology, 1 Sec 3 Chung-Hsiao E Rd, Taipei, 10608, Taiwan, ROC

* Correspondence: waltchen@ntut.edu.tw

**Abstract**

Root mean squared error (RMSE) and mean absolute error (MAE) are among the most widely used performance metrics in machine learning and scientific modeling. Although their mathematical relationship is well established, misunderstandings and misapplications of these metrics continue to appear in the literature. This technical note revisits the fundamental bounds relating RMSE and MAE and identifies a systematic error in a recently published paper in *Artificial Intelligence Review*, in which RMSE values are numerically smaller than the corresponding MAE values, a relationship that is mathematically impossible. Notably, these incorrect RMSE and MAE values are reported alongside other cited results within the same study that correctly satisfy the inequality RMSE $\geq$ MAE. In addition, supplementary experiments using two common and straightforward machine learning models, Random Forest and XGBoost, demonstrate that comparable or superior performance can be achieved in several of the same datasets used in the aforementioned paper without resorting to highly complex optimization frameworks. Collectively, these findings underscore the importance of verifying the correctness of basic performance metrics and of contextualizing claimed performance gains through transparent baseline comparisons in machine learning evaluation.

**Keywords:** RMSE; MAE; metrics; Random Forest; XGBoost

## 1. Introduction and Motivation

Quantitative evaluation plays a central role in the development and comparison of machine learning and optimization algorithms. Performance metrics such as the root mean squared error (RMSE) and the mean absolute error (MAE) are routinely employed to assess predictive accuracy, guide hyperparameter optimization, and rank competing models across a wide range of regression-based applications. Although widely used, these metrics are not interchangeable and are governed by fundamental mathematical relationships that restrict their admissible values.

In a recent study published in the leading artificial intelligence journal *Artificial Intelligence Review* [1], benchmark results were reported in which RMSE values are numerically smaller than the corresponding MAE values across multiple datasets (Tables 4–8). Under the standard definitions of RMSE and MAE, such a relationship is mathematically impossible. Importantly, this error is not confined to a single article. The same pattern can be identified in a series of algorithmic studies by the same authors published over the past six years [2–5], which primarily develop hybrid machine-learning frameworks optimized by nature-inspired and swarm-intelligence-based metaheuristic algorithms and validate their performance through extensive numerical comparisons.

The recurrence of this issue across multiple publications is therefore a matter of technical concern. In these studies, RMSE and MAE are not merely auxiliary descriptors but function as primary evaluation criteria and, in some cases, as objective functions within the optimization process itself [5]. Consequently, violations of basic metric relationships call into question the internal consistency of the reported numerical results and, by extension, the reliability of subsequent comparisons, statistical analyses, and claims of algorithmic superiority derived from them.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2026 doi:10.20944/preprints202601.2392.v1



This technical note focuses exclusively on the mathematical validity and theoretical relationship between RMSE and MAE. Because these metrics satisfy exact and easily verifiable inequalities, errors can be unambiguously identified without access to source code, raw data, or the need to execute complex algorithms. When such errors arise even at this fundamental level, they raise legitimate concerns regarding the numerical reliability and trustworthiness of other reported results in the same studies, particularly those derived from more complex optimization procedures or multi-stage learning frameworks that are not readily subject to simple analytical checks. Ensuring the correctness of basic performance metrics is therefore a necessary first step for establishing confidence in subsequent analyses and conclusions drawn from increasingly sophisticated machine learning algorithms.

## 2. Mathematical Properties of RMSE and MAE

Let $\{y_i\}_{i=1}^n$ denote observed values and $\{\hat{y}_i\}_{i=1}^n$ corresponding model predictions. Define the prediction errors as

$$e_i = y_i - \hat{y}_i.$$

The mean absolute error and root mean squared error are given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2}. \tag{1}$$

These two metrics are related by the well-known inequality

$$\mathrm{MAE} \le \mathrm{RMSE} \le \sqrt{n}\,\mathrm{MAE}, \tag{2}$$

which has been discussed extensively in the literature [6–8].

Under additional constraints, a tighter upper bound can be obtained. For least-squares fits, where the errors satisfy $\sum_{i=1}^n e_i = 0$, Willmott et al. [8] showed that the root-mean-square error is bounded by

$$\mathrm{RMSE} \le \sqrt{\frac{n}{2}}\,\mathrm{MAE}. \tag{3}$$

This upper bound is attained when the total absolute error is concentrated in two countervailing errors of equal magnitude and opposite sign. Equation (3) therefore represents a tighter, but condition-dependent, bound that applies specifically to least-squares error structures and does not contradict the general inequality in Eq. (2).

### 2.1. Lower Bound

The lower bound in (2) follows from a standard inequality between the arithmetic mean and the quadratic mean applied to the absolute errors. Specifically, applying the Cauchy–Schwarz inequality to the vectors $(|e_1|,\ldots,|e_n|)$ and $(1,\ldots,1)$ yields

$$\left(\sum_{i=1}^{n}|e_i|\right)^2 \le \left(\sum_{i=1}^{n}|e_i|^2\right)\left(\sum_{i=1}^{n}1^2\right) = n\sum_{i=1}^{n}e_i^2.$$

Dividing both sides by $n^2$ gives

$$\left(\frac{1}{n}\sum_{i=1}^{n}|e_i|\right)^2 \le \frac{1}{n}\sum_{i=1}^{n}e_i^2.$$

Taking square roots on both sides leads directly to

$$\mathrm{MAE} \le \mathrm{RMSE}.$$

Equality holds if and only if all absolute errors are identical, i.e., $|e_1| = \cdots = |e_n|$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2026 doi:10.20944/preprints202601.2392.v1



*2.2. Upper Bound*

The upper bound reflects the fact that, for a fixed MAE, the sum of squared errors is maximized when the absolute error is maximally concentrated in a single observation. Let

$$S = \sum_{i=1}^{n} |e_i| = n\,\mathrm{MAE}.$$

Since $|e_i| \geq 0$, the quantity $\sum e_i^2$ is maximized under the constraint $\sum |e_i| = S$ when one error equals $S$ and all remaining errors are zero. In this case,

$$\sum_{i=1}^{n} e_i^2 \leq S^2,$$

with equality attained if and only if $|e_j| = S$ for some $j$ and $|e_i| = 0$ for all $i \neq j$. Substituting this bound into the definition of RMSE yields

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \leq \sqrt{\frac{1}{n}S^2} = \sqrt{n}\,\mathrm{MAE}.$$

Together with the lower bound, Eq. (2) therefore provides inequalities that hold under the standard definitions of RMSE and MAE.

*2.3. A Tighter Upper Bound Under Least-Squares Fit*

For least-squares fits, the errors satisfy the additional constraint

$$\sum_{i=1}^{n} e_i = 0, \tag{4}$$

which implies that the total positive and negative error magnitudes are equal. Let

$$S = \sum_{i=1}^{n} |e_i| = n\,\mathrm{MAE}, \qquad S_+ = \sum_{i:e_i>0} e_i, \qquad S_- = \sum_{i:e_i<0} (-e_i).$$

Then $S = S_+ + S_-$ and (4) gives $S_+ = S_- = S/2$. Since, for any nonnegative numbers with fixed sum $T$, the sum of squares is maximized when the mass is concentrated in a single entry, we have

$$\sum_{i:e_i>0} e_i^2 \leq S_+^2, \qquad \sum_{i:e_i<0} e_i^2 \leq S_-^2,$$

and therefore

$$\sum_{i=1}^{n} e_i^2 = \sum_{i:e_i>0} e_i^2 + \sum_{i:e_i<0} e_i^2 \leq S_+^2 + S_-^2 = \frac{S^2}{2}.$$

Substituting into the definition of RMSE yields

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \leq \sqrt{\frac{1}{n}\cdot\frac{S^2}{2}} = \sqrt{\frac{n}{2}}\,\mathrm{MAE},$$

which is the tighter upper bound reported by Willmott et al. [8]. Equality holds when the error mass is concentrated in two countervailing errors of equal magnitude and opposite sign, with all remaining errors equal to zero.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2026 doi:10.20944/preprints202601.2392.v1



## 3. Identification of Metric Errors in Reported Results

In the benchmark results reported by [1], RMSE values are shown to be consistently smaller than MAE values across five test datasets (Tables 4–8). Under the inequality in Eq. (2), such a relationship cannot occur if RMSE and MAE are computed according to their standard/correct definitions.

Under the standard definitions of RMSE and MAE, the inequality RMSE $\geq$ MAE must always hold. The reported results therefore indicate that the values labeled as RMSE are numerically incorrect and are not consistent with the definition of root mean squared error. The issue is not one of model quality but of metric interpretation: when fundamental mathematical relationships are violated, reported performance measures lose their diagnostic meaning, and subsequent comparisons across models or algorithms become unreliable.

It is worth emphasizing that similar RMSE–MAE errors can also be identified in several other publications by the same authors over the past six years [2–5]. However, as these studies were published in different journals and lie outside the scope of the present submission, they are not examined in detail here. Their recurrence nonetheless underscores the broader importance of carefully verifying the correctness of basic evaluation metrics before drawing conclusions from more complex analyses.

## 4. Consistency with Literature Results

It is noteworthy that the other studies cited alongside the results in [1] correctly satisfy the inequality RMSE $\geq$ MAE. This contrast demonstrates that the inequality RMSE $\geq$ MAE is consistently satisfied in the literature results when RMSE is computed according to its standard/correct definition, confirming that the reported violation cannot be attributed to dataset characteristics or benchmarking context. Indeed, under the standard definitions of RMSE and MAE, no dataset can intrinsically exhibit the property RMSE $<$ MAE, as this would contradict a fundamental mathematical inequality. When viewed together with similar errors documented across multiple additional publications by the same authors, the evidence suggests that the issue is systematic rather than isolated and is related to the computation and reporting of RMSE and MAE. This observation further underscores the importance of verifying basic metric relationships as an integral component of sound benchmarking practice.

## 5. A Straightforward and Low-Complexity Baseline: Random Forest and XGBoost as Reference Models

Beyond metric consistency, model evaluation should also consider the principle of parsimony. Occam's razor suggests that, all else being equal, simpler and more transparent models should be preferred over increasingly complex algorithmic frameworks, particularly when performance gains are marginal or inconsistent.

To examine this issue, we implemented two widely used and well-understood machine learning algorithms—Random Forest (RF) and Extreme Gradient Boosting (XGBoost)—as baseline reference models. Following the experimental protocol described in [1], all datasets were partitioned using the same data split: 63% for training, 27% for validation, and 10% for testing. No metaheuristic optimization or problem-specific tuning strategies were employed beyond standard practices commonly adopted in applied machine learning.

Table 1 summarizes the comparative results between the SAPSO-based framework reported by Truong and Chou [1] and our RF and XGBoost implementations. For completeness, MAE, RMSE, and MAPE (Mean Absolute Percentage Error) are reported for each dataset.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2026 doi:10.20944/preprints202601.2392.v1



**Table 1.** Comparison of reported results from Truong and Chou [1] with Random Forest and XGBoost baselines using the same data split (63% training, 27% validation, 10% testing). Dataset sources are indicated for reference.

| Dataset | Study | MAE | RMSE | MAPE (%) |
|---|---|---|---|---|
| Dataset 1 [9] | Truong and Chou [1] | 1.2589 | 0.1531 | 4.8556 |
| | Random Forest (this study) | 0.9286 | 1.2842 | 3.5421 |
| | XGBoost (this study) | 1.1402 | 1.6684 | 4.3605 |
| Dataset 2 [10,11] | Truong and Chou [1] | 57.9853 | 6.9485 | 6.1588 |
| | Random Forest (this study) | 61.1470 | 83.6444 | 5.8287 |
| | XGBoost (this study) | 62.1046 | 90.1661 | 5.9244 |
| Dataset 4 [12] | Truong and Chou [1] | 3.7243 | 1.4297 | 12.3837 |
| | Random Forest (this study) | 2.7818 | 3.8750 | 12.2188 |
| | XGBoost (this study) | 2.5224 | 3.3226 | 11.9671 |

The results in Table 1 show that, despite their conceptual simplicity, Random Forest and XGBoost achieve comparable or superior performance in three out of the five datasets when evaluated using standard error metrics. In particular, both baseline models yield lower MAE and MAPE values than those reported in [1] for Dataset 1 and Dataset 4, without relying on a multi-stage stacking architecture or a sophisticated metaheuristic optimization scheme. RMSE values reported in [1] are not used for direct comparison here because they violate the fundamental inequality RMSE $\geq$ MAE and therefore cannot be interpreted as valid root mean squared errors under the standard/correct definition.

These findings do not invalidate the methodological contributions of the SAPSO-based framework. However, they raise a substantive question regarding the necessity and practical benefit of introducing a highly complex optimization pipeline when simpler, well-established machine learning models already provide competitive performance under the same data partitioning strategy. From a benchmarking perspective, such comparisons underscore the importance of including strong yet low-complexity baseline models when assessing the claimed advantages of newly proposed optimization algorithms.

## 6. Implications for Benchmarking and Model Evaluation

The results presented in this technical note have several broader implications for benchmarking and model evaluation in machine learning research. First, they demonstrate that even widely used and ostensibly simple error metrics such as RMSE and MAE require careful verification. When fundamental mathematical relationships between evaluation metrics are violated, the resulting performance measures lose their interpretability, rendering subsequent model comparisons and optimization claims unreliable.

Second, the additional analyses using Random Forest and XGBoost highlight the importance of including strong yet low-complexity baseline models in benchmarking studies. Under the same data partitioning strategy, these conventional machine learning methods achieve comparable or superior performance in a majority of the examined datasets, despite their substantially lower algorithmic complexity and absence of sophisticated optimization schemes. This finding underscores the need to contextualize reported performance gains by demonstrating that they cannot be readily attained by simpler and well-established models.

Taken together, these observations emphasize that rigorous metric validation and transparent baseline comparisons are not optional refinements but essential components of reproducible and credible empirical research in artificial intelligence.

## 7. Concluding Remarks

This technical note revisited the mathematical relationship between RMSE and MAE, identified systematic numerical errors in reported RMSE values that violate fundamental inequalities, and demonstrated the practical value of simple baseline models through additional empirical analysis. The purpose is not to challenge the scientific objectives or methodological creativity of prior studies, but to

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2026 doi:10.20944/preprints202601.2392.v1



highlight that reliable metric computation and meaningful baseline selection form the foundation of trustworthy model evaluation.

As machine learning and optimization methods continue to grow in complexity, ensuring clarity, mathematical correctness, and interpretability in performance assessment remains a shared responsibility of the research community. Attention to these fundamentals is crucial for sustaining cumulative progress and maintaining confidence in reported advances.

**Author Contributions:** The first author identified and formalized the theoretical issue, derived the mathematical results, and drafted the manuscript. The second author identified errors in the reported performance metrics and conducted the Random Forest and XGBoost baseline analyses. Both authors reviewed and approved the final manuscript.

**Funding:** This study was partially supported by the National Science and Technology Council (Taiwan) under Research Project Grant Numbers NSTC 114-2121-M-027-001 and NSTC 114-2637-8-027-014.

**Ethics approval and consent to participate:** Not applicable. This study does not involve human participants or animals.

**Consent for publication:** Not applicable.

**Data Availability Statement:** The datasets used in the comparison table are publicly available, and their sources are cited directly in the table.

**Materials availability:** Not applicable.

**Code availability:** The code used to generate the results in the comparison table using Random Forest and XGBoost is available from the author upon reasonable request.

**Acknowledgments:** This study was partially supported by the National Science and Technology Council (Taiwan) under Research Project Grant Numbers NSTC 114-2121-M-027-001 and NSTC 114-2637-8-027-014. During the preparation of this manuscript, the authors used ChatGPT 5.2 (OpenAI) for assistance with editing and polishing the writing. The authors have reviewed and revised the output and take full responsibility for the content of this publication.

**Conflicts of Interest:** The author declares that there are no competing interests.

## References

1. Truong, D.N.; Chou, J.S. Scientific approach to problem solving-inspired optimization of stacking ensemble learning for enhanced civil engineering informatics. *Artificial Intelligence Review* **2025**, *58*, 404.
2. Chou, J.S.; Truong, D.N.; Le, T.L.; Truong, T.T.H. Bio-inspired optimization of weighted-feature machine learning for strength property prediction of fiber-reinforced soil. *Expert Systems with Applications* **2021**, *180*, 115042.
3. Chou, J.S.; Truong, D.N. Multistep energy consumption forecasting by metaheuristic optimization of time-series analysis and machine learning. *International Journal of Energy Research* **2021**, *45*, 4581–4612.
4. Truong, D.N.; Chou, J.S. Fuzzy adaptive jellyfish search-optimized stacking machine learning for engineering planning and design. *Automation in Construction* **2022**, *143*, 104579.
5. Truong, D.N.; To, V.L.; Truong, G.T.; Jang, H.S. Engineering punching shear strength of flat slabs predicted by nature-inspired metaheuristic optimized regression system. *Frontiers of Structural and Civil Engineering* **2024**, *18*, 551–567.
6. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research* **2005**, *30*, 79–82.
7. Chai, T.; Draxler, R.R.; et al. Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific model development discussions* **2014**, *7*, 1525–1534.
8. Willmott, C.J.; Matsuura, K.; Robeson, S.M. Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment* **2009**, *43*, 749–752.
9. Golafshani, E.M.; Behnood, A. Application of soft computing methods for predicting the elastic modulus of recycled aggregate concrete. *Journal of cleaner production* **2018**, *176*, 1163–1176.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2026 doi:10.20944/preprints202601.2392.v1



10. Pham, T.A.; Ly, H.B.; Tran, V.Q.; Giap, L.V.; Vu, H.L.T.; Duong, H.A.T. Prediction of pile axial bearing capacity using artificial neural network and random forest. *Applied Sciences* **2020**, *10*, 1871.
11. Pham, T.A.; Tran, V.Q.; Vu, H.L.T.; Ly, H.B. Design deep neural network architecture using a genetic algorithm for estimation of pile bearing capacity. *PLoS One* **2020**, *15*, e0243030.
12. Nguyen, N.M.; Wang, W.C.; Cao, M.T. Early estimation of the long-term deflection of reinforced concrete beams using surrogate models. *Construction and Building Materials* **2023**, *370*, 130670.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.