Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

On the Stability and Reliability of Machine Learning Models Under Data and Randomness Variations

(A Comparative Study of Logistic Regression and Decision Tree Models Across Clean and Noisy Datasets)

Authors Name: Vishnupriya Shaji
Department of Computer Science
Dr. D. Y. Patil Arts, Commerce & Science College
Pune, India

Authors Name:Reetam Chakraborty
Department of Computer Science
Dr. D. Y. Patil Arts, Commerce & Science College
Pune, India

Abstract - Machine learning (ML) models are increasingly deployed in high impact domains such as healthcare, finance, and decision-support systems. While model accuracy is often emphasized, the stability and reliability of these models under small data perturbations and randomness variations remain under explored. This study investigates how minor randomness introduced through different random seeds and controlled noise injection affects the performance of commonly used supervised learning algorithms. Three datasets from distinct domains biological classification (Iris), financial behavior analysis, and medical diagnosis (Breast Cancer) are used to evaluate Logistic Regression and Decision Tree classifiers. Performance is assessed using accuracy, precision, and recall across multiple randomized runs. Experimental results demonstrate that while structured datasets such as Iris and Breast Cancer exhibit high robustness, real world noisy datasets such as finance data show significant sensitivity to randomness and noise. The findings highlight the importance of evaluating model stability alongside predictive performance, especially for real-world applications where reproducibility and trust are critical.

Keywords - Decision Tree, Logistic Regression, Model Stability, Noise Injection, Randomness, Reliability, Robust Machine Learning

1. INTRODUCTION

Machine learning (ML) techniques have become integral to a wide range of applications, including medical diagnosis, financial analysis, and intelligent decision-support systems. As these models are increasingly relied upon in critical real-world scenarios, questions surrounding their dependability have gained prominence. While predictive accuracy is commonly used as the primary evaluation metric, it alone does not adequately reflect whether a model can be trusted to behave consistently under varying conditions.

In real-world machine learning workflows, model outcomes are influenced by several sources of variability. Factors such as random weight initialization, data shuffling, train–test split selection, and stochastic learning processes can all introduce differences in model performance. Additionally, practical datasets often contain imperfections arising from measurement errors, subjective inputs, or incomplete information. Even small variations in data composition or randomness can therefore lead to noticeable fluctuations in results, potentially undermining confidence in model predictions.

These challenges raise an important concern regarding the reliability of machine learning systems: *can model predictions remain consistent when subjected to minor data perturbations or randomness?* Addressing this question is particularly important for high-impact domains where inconsistent outcomes may lead to incorrect decisions or reduced trust in automated systems. By systematically examining how controlled randomness and noise affect model performance across multiple domains, this study aims to provide empirical evidence on the stability and reliability of commonly used machine learning models.

2. BACKGROUND AND MOTIVATION

Model stability refers to the consistency of a model's predictions and performance when exposed to small variations in data or training conditions. Reliability, on the other hand, relates to whether a model's predictions can be trusted across repeated experiments. While ensemble learning and regularization techniques aim to improve robustness, many widely used baseline models are still sensitive to randomness.

Recent studies have shown that even with fixed hyperparameters, different random seeds can lead to different decision boundaries, especially for tree-based and stochastic algorithms. In sensitive domains such as healthcare and finance, this variability may result in inconsistent decisions, reduced interpretability, and ethical concerns.

Despite this, stability analysis is often overlooked in favor of performance optimization. This research is motivated by the need to complement performance evaluation with robustness analysis, thereby promoting more responsible and reliable machine learning practices.

3. RESEARCH PROBLEM AND OBJECTIVES

3.1 Problem Statement

Most machine learning studies prioritize predictive accuracy while neglecting the effects of randomness and minor data perturbations. As a result, models that perform well under a single experimental setup may behave inconsistently when exposed to slightly altered conditions. This raises concerns about the trustworthiness of machine learning models deployed in real-world environments.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

3.2 Research Objectives

The primary objectives of this study are:

1. To examine the effect of random seed variations on model performance, particularly with respect to accuracy, precision, and recall across multiple experimental runs.
2. To analyse the robustness of machine learning models under controlled noise injection, simulating realistic data imperfections and measurement inconsistencies.
3. To compare model stability across datasets from different application domains, including biological, financial, and medical datasets, in order to understand domain-specific sensitivity.
4. To assess differences in stability between linear and non-linear classification models, using Logistic Regression and Decision Tree classifiers as representative techniques.
5. To quantify performance variability using repeated experiments rather than single-run evaluations, emphasizing statistical consistency through mean and standard deviation measures.
6. To investigate whether high predictive accuracy necessarily indicates model reliability, by analysing the relationship between average performance and variability.
7. To highlight the need for more rigorous evaluation practices in machine learning research, particularly for applications where trust and reliability are critical.

4. LITERATURE REVIEW

Understanding the stability and robustness of machine learning models under variations in data and randomness has been an area of growing interest, particularly as machine learning systems are increasingly deployed in real-world, high-risk domains.

1. Breiman (2001) highlighted the concept of model instability in decision trees, demonstrating that small changes in training data can lead to significantly different tree structures and performance outcomes. This work motivated ensemble methods such as Random Forests to reduce variance, but it also underscored the inherent sensitivity of certain algorithms to data sampling and randomness. https://doi.org/10.1023/A:1010933404324
2. Reimers and Gurevych (2017) investigated the effects of random initialization and random data splits on the reproducibility of neural network performance. Their study on natural language tasks showed that single-run performance reporting can be misleading, and that reporting average performance across multiple random seeds provides more trustworthy results. This reinforces the need for stability-aware evaluation practices across machine learning domains. https://aclanthology.org/D17-1184/
3. Raste et al. (2022) performed an empirical analysis of how randomness impacts model performance, revealing that variations due to random seed and train–test splits can contribute more to performance fluctuation than algorithmic differences. Their findings suggest that stability analysis should be an integral part of model evaluation, not just an afterthought. https://arxiv.org/abs/2206.12353
4. Zawia et al. (2022) proposed a robustness framework for evaluating classifier performance under data perturbations. While their primary application was in diagnosing metabolic disorders, the methodology — adding controlled noise and assessing statistical performance variation — directly aligns with the experimental approach in this study. Their work illustrates that model reliability can vary widely with realistic data perturbations. https://www.mdpi.com/2075-4426/12/8/1314
5. Balendran et al. (2025) conducted a scoping review of robustness concepts in healthcare machine learning, categorizing sources of variability including input noise, class imbalance, and domain shifts. Their analysis highlights that models performing well under ideal conditions may still fail in real clinical settings, emphasizing the importance of stability evaluation for trustworthy ML systems in sensitive applications. https://www.nature.com/articles/s41746-024-01420-1

5. METHODOLOGY

A. 5.1 Dataset Description

Three datasets from distinct domains were used:

1. **Iris Dataset:** A well-structured, low-noise biological classification dataset with four numerical features and three classes.
2. **Finance Dataset:** A real-world financial behaviour dataset containing demographic and investment preference features, characterized by categorical variables and inherent noise.
3. **Breast Cancer Dataset:** A medical diagnostic dataset used for binary classification of malignant and benign tumours, featuring numerical attributes. These datasets were chosen to represent varying levels of complexity, noise, and real-world relevance.

5.2 Data Preprocessing

Missing values were handled where necessary to maintain data integrity. Categorical features in the finance dataset were encoded using label encoding. Numerical features were standardized for Logistic Regression to ensure scale consistency. Train–test splitting was performed using different random seeds to introduce controlled randomness and enable stability analysis.

5.3 Noise Injection

To simulate minor data perturbations, Gaussian noise with small variance was added to numerical features. This process introduces realistic imperfections without significantly altering the underlying data

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

distributions, allowing evaluation of model robustness to small input variations.

5.4 Model Selection

Two supervised learning models were selected. Logistic Regression was used as a linear and interpretable baseline model, while Decision Tree was chosen as a non-linear model known for flexibility but higher sensitivity to data variations. This combination enables comparison of stability across different model types.

5.5 Evaluation Strategy

Experiments were repeated across 20 random seeds. Model performance was evaluated using accuracy, precision, and recall. Mean and standard deviation values were calculated to assess both average performance and variability. Comparisons were made between original and noise-injected datasets to analyse stability.

6. RESULTS AND ANALYSIS

6.1 Iris Dataset

Both Logistic Regression and Decision Tree models achieved high accuracy (≈95%) on the original Iris dataset. The introduction of noise resulted in negligible performance degradation, indicating strong robustness. Low standard deviation values across runs suggest high stability and reliability.

6.2 Finance Dataset

In contrast, the finance dataset exhibited low overall performance and high variability across runs. Accuracy ranged from 16% to 66%, with noticeable performance fluctuations under noise injection. This indicates high sensitivity to randomness and highlights challenges associated with real-world, subjective data.

6.3 Breast Cancer Dataset

The breast cancer dataset demonstrated consistently high performance for both models, with accuracy above 91% even after noise injection. Slight increases in variance were observed for Decision Trees, but overall reliability remained strong.

7. EXPECTED OUTCOMES AND IMPLICATIONS

The expected outcomes of this research include:

1. Demonstration that structured datasets yield stable and reliable ML models.
2. Evidence that real-world noisy datasets are highly sensitive to randomness.
3. Validation of the need for repeated experiments and averaged metrics.
4. Increased awareness of model reliability as a critical evaluation dimension.

These findings emphasize that accuracy alone is insufficient for assessing model quality, especially in high-stakes applications.

8. CONCLUSION

This study empirically examined the stability and reliability of machine learning models under data and randomness variations across three domains. Results reveal that while traditional datasets like Iris and Breast Cancer are robust to minor perturbations, real-world financial data exhibits significant instability. The findings underscore the necessity of incorporating stability analysis into standard machine learning evaluation pipelines. Future work may extend this research by incorporating additional models, cross-validation strategies, and fairness-oriented robustness metrics.

9. REFERENCES

[1] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no.1, pp.5–32,2001. https://doi.org/10.1023/A:1010933404324

[2] O. Bousquet and A. Elisseeff, “Stability and generalization,” *Journal of Machine Learning Research*, vol. 2, pp.499–526,2002. http://www.jmlr.org/papers/volume2/bousquet02a/bousquet02a.pdf

[3] N. Reimers and I. Gurevych, “Reporting score distributions makes a difference in machine learning,” in *Proceedings of the 2017 Conference on Empirical Methods in NaturalLanguageProcessing*,2017. https://aclanthology.org/D17-1184/

[4] S. Raste, A. Agrawal, and S. Raskar, “On randomness in machine learning: an empirical study,” arXiv preprint arXiv:2206.12353,2022. https://arxiv.org/abs/2206.12353

[5] A. Zawia, Y. Liu, and D. Saha, “Robust machine learning under input perturbations: a framework and application in metabolic disease diagnosis,” *Journal of Personalized Medicine*, vol. 12, no. 8, 1314, 2022. https://www.mdpi.com/2075-4426/12/8/1314

[6] P. Balendran, S. Jain, and T. R. Singh, “Robustness in healthcare machine learning: a scoping review,” *npj Digital Medicine*, vol. 8, 14, 2025. https://www.nature.com/articles/s41746-024-01420-1

[7] I. Ben Braiek and F. Khomh, “Robustness in machine learning: a trustworthy AI perspective,” arXiv:2404.00897, 2024. https://arxiv.org/abs/2404.00897

[8] C. Wang, “Robustness and reliability of machine learning systems: a comprehensive review,” *Open Access Research Journal*, vol. 1, no. 1, pp. 12–30, 2023. https://www.opastpublishers.com/open-access-articles-pdfs/robustness-and-reliability-of-machine-learning-systems-a-comprehensive-review.pdf

[9] X. Zhang, Y. Chen, and Z. Liu, “Understanding the impact of data perturbation on classification model performance,” *Pattern Recognition*, vol. 117, pp. 107990, 2021. https://www.sciencedirect.com/science/article/pii/S0031320321002276

[10] S. Fort, H. Hu, and B. Lakshminarayanan, “Deep ensembles: a loss landscape perspective,” arXiv:1912.02757, 2019. https://arxiv.org/abs/1912.02757

[11] G. Hooker, “Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables,” *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 341–359, 2013. https://www.tandfonline.com/doi/abs/10.1080/10618600.2012.688069

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems, 2012*. https://papers.nips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[13] F. Pedregosa et al., “Scikit-learn: machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. http://jmlr.org/papers/v12/pedregosa11a.html

(This work is licensed under a Creative Commons Attribution 4.0 International License.)