



Model Stability Evaluation Under Random Seed Variation Using SSI and RRS

Nurul Chamidah¹, Azwa Abdul Aziz²

¹Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jakarta, Jakarta Selatan, Indonesia

²Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA), Terengganu, Malaysia

*nurul.chamidah@upnvj.ac.id

Abstract

Machine learning models are commonly evaluated using a single experimental run, implicitly assuming that performance remains stable under fixed conditions. However, many machine learning algorithms involve stochastic processes that may cause performance variability across different random seed configurations. We examine how sensitive tabular classifiers are to the choice of random seed. Multi-Layer Perceptron (MLP), Random Forest, Support Vector Machine (SVM), and XGBoost were each trained and tested under 30 seeds on three public benchmark datasets. To assess robustness, this study introduces two complementary stability measures: Stability Sensitivity Index (SSI) and Relative Robustness Score (RRS). Stability was measured on three metrics: Accuracy, F1-score, and ROC-AUC. The results demonstrate that model performance can vary substantially across random seed configurations; on the smallest dataset, the accuracy of a single run differed by up to 7.9 percentage points depending on the seed. XGBoost achieved the highest mean performance on the two large datasets and, in most cases, also the lowest variability there, whereas SVM was the most stable model on the small Breast Cancer dataset. MLP was generally the most seed-sensitive model on the large datasets, and Random Forest on the small one. F1-score varied considerably more than Accuracy and ROC-AUC, particularly under class imbalance. Furthermore, the smallest dataset generally exhibits greater variability than the larger datasets. These results argue for reporting seed-level variability, not only mean scores, as a routine part of model evaluation. The proposed SSI and RRS metrics can support stability-aware model selection and reproducibility-oriented evaluation in applications requiring consistent predictive performance.

Keywords: machine learning; model stability; performance variability; random seed; support vector machine; tabular data; XGBoost

How to Cite: N. . Chamidah and A. A. Aziz, "Model Stability Evaluation Under Random Seed Variation Using SSI and RRS", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 5, pp. 1490–1503, Oct. 2026, doi: [10.29207/resti.v10i5.7731](https://doi.org/10.29207/resti.v10i5.7731).
Permalink/DOL: <https://doi.org/10.29207/resti.v10i5.7731>

Received: April 20, 2026

Accepted: September 5, 2026

Available Online: October 16, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

Machine learning has become a fundamental approach for solving predictive problems across a wide range of domains, including healthcare, finance, marketing, and social sciences [1], [2]. With the increasing availability of structured datasets, supervised learning techniques applied to tabular data remain among the most widely used approaches for practical predictive modeling [3], [4]. Compared with other data modalities such as images or text, tabular datasets typically contain heterogeneous feature types and complex relationships between variables, requiring algorithms capable of handling structured inputs while maintaining predictive accuracy and interpretability [3], [5]. Consequently, numerous machine learning algorithms have been widely adopted for tabular prediction tasks, including

ensemble learning methods, kernel-based models, and neural networks [4], [6].

Random Forest, Support Vector Machines (SVM), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptrons (MLP) are among the algorithms commonly applied to tabular data. Random Forest employs an ensemble of decision trees, combining bootstrap aggregation with random feature selection to enhance predictive performance and help control overfitting [7]. XGBoost is an optimized implementation of gradient boosting that incorporates regularization, parallelization, and efficient tree construction to achieve high predictive accuracy and scalability for large datasets [8]. Support Vector Machines construct optimal separating hyperplanes in high-dimensional feature spaces and can model nonlinear relationships using kernel functions [9].

Meanwhile, Multilayer Perceptrons represent a class of feedforward artificial neural networks capable of learning complex nonlinear mappings through multiple hidden layers and nonlinear activation functions [10]. Their performance is usually summarized by single-number metrics such as accuracy or F1-score, each of which can be misleading when used alone [11]. The four algorithms were chosen to represent distinct machine learning paradigms: bagging-based ensembles, kernel-based learning, gradient boosting, and neural networks. This selection enables a comparative assessment of model stability across different algorithmic approaches.

Despite the rapid development of machine learning algorithms, an important aspect of model evaluation that has received comparatively limited attention is model stability. Many machine learning algorithms involve stochastic processes during training, including random parameter initialization, random data partitioning, and stochastic optimization procedures [12], [13]. These stochastic components are typically controlled using a parameter known as the random seed, which determines the sequence of pseudo-random numbers used during model training. In practice, the seed may control both how the data are partitioned and the internal randomness of the learning algorithm. Although random seeds are primarily used to ensure reproducibility, variations in seed values can lead to differences in training trajectories and consequently differences in model performance [13], [14]. Previous studies have shown that stochastic components in machine learning pipelines may influence model parameters and evaluation metrics across different experimental runs [14], [15].

In many empirical machine learning studies, model performance is commonly reported based on a single experimental run using a fixed random seed. While this approach simplifies experimentation and reporting, it may introduce bias in performance evaluation and limit the reliability of experimental findings [16], [17]. Performance metrics such as accuracy, F1-score, and area under the receiver operating characteristic curve (ROC-AUC) can vary depending on the specific random seed used during model training or dataset splitting [13], [18]. As a result, conclusions derived from a single experimental run may not accurately represent the true behavior or generalization capability of a machine learning model. Several studies have demonstrated that stochastic elements such as random cross-validation splits and model initialization can produce substantial variability in reported experimental results [15], [18], [19].

The issue of reproducibility has recently attracted growing attention in the machine learning community. Researchers have increasingly recognized the existence of a reproducibility challenge in machine learning research, where experimental outcomes may be sensitive to implementation details, hyperparameter configurations, and sources of randomness in training

pipelines [19], [20]. This challenge is further exacerbated by the increasing complexity of modern machine learning systems and experimental workflows [12], [20]. Consequently, improving the reliability, transparency, and reproducibility of machine learning experiments has become an important research direction in recent years [12], [21].

Although previous research has explored various aspects of machine learning evaluation, including cross-validation strategies, hyperparameter tuning, and statistical comparisons of algorithms [16], [17], [22], the influence of random seed variation on model stability remains relatively underexplored, particularly in studies involving tabular datasets [15], [19]. Most existing research focuses primarily on improving predictive accuracy rather than examining how robust model performance is to stochastic variations during training [12], [15]. Therefore, systematic empirical investigations are needed to better understand how sensitive widely used machine learning algorithms are to variations in random initialization and stochastic training processes [13], [15].

Recent research has increasingly emphasized the importance of reproducibility and robustness in machine learning experiments. Several studies have shown that machine learning results can be highly sensitive to experimental settings such as random initialization, dataset partitioning, and hyperparameter selection [15], [19], [20]. For example, Hutson [20] highlighted the growing concern of a reproducibility crisis in artificial intelligence, where experimental results may not be consistently replicated due to uncontrolled stochastic factors. Similarly, Gundersen and Kjensmo [12] discussed challenges in achieving reproducibility in artificial intelligence systems and emphasized the need for improved experimental reporting and evaluation protocols. More recent work by Pineau et al. [19] proposed guidelines for improving reproducibility in machine learning research, including clearer reporting of experimental configurations and repeated experimental evaluations. Furthermore, Bouthillier et al. [15] demonstrated that performance differences reported in machine learning benchmarks can often be influenced by experimental variance, suggesting that multiple runs are necessary to obtain reliable estimates of model performance. These findings highlight the importance of systematically analyzing stochastic variability in machine learning experiments.

Despite these advances, several limitations remain. Most prior studies have primarily focused on demonstrating the existence of reproducibility challenges or recommending improved experimental practices [12], [19], [20]. However, relatively few studies have systematically quantified the sensitivity of machine learning models to random seed variation using interpretable and comparable stability measures [15], [19]. Furthermore, model evaluation continues to rely heavily on average performance metrics, while the

variability of performance across repeated runs is often underreported [15], [19], [21]. As a result, the robustness of machine learning models under stochastic experimental conditions remains insufficiently characterized, particularly for tabular classification tasks.

To address this research gap, this study investigates the effect of random seed variation on the stability of machine learning models applied to tabular classification tasks. Specifically, repeated experiments are conducted using multiple random seeds to evaluate the variability of model performance across several widely used machine learning algorithms, including Random Forest, XGBoost, Support Vector Machines, and Multilayer Perceptrons. By analyzing the distribution of performance metrics obtained from repeated experimental runs, this study aims to provide insights into the robustness and reliability of machine learning models when applied to structured datasets, as recommended in recent studies on experimental variance and reproducibility in machine learning research [15], [19], [21].

The contributions of this study are threefold. First, this work provides an empirical evaluation of the impact of random seed variation on the performance stability of several widely used machine learning algorithms. Second, the study introduces two stability measures, SSI and RRS, and uses them to analyze the variability of model performance through repeated experiments across multiple seeds, highlighting potential inconsistencies that may arise when model performance is evaluated using single-run experiments. Third, this research provides practical insights and recommendations for researchers and practitioners regarding more reliable experimental practices when evaluating machine learning models on tabular datasets.

The remainder of this paper is organized as follows. Section 2 describes the methodology used in this study, including datasets, machine learning models, and experimental procedures. Section 3 presents the experimental results and discusses model stability across different random seeds, together with the limitations of the study. Finally, Section 4 concludes the study.

2. Methods

This study adopts an empirical experimental approach to investigate the impact of random seed variation on the stability of machine learning models trained on tabular datasets. The primary objective is to evaluate how differences in random seed initialization influence model performance across repeated training runs.

In machine learning workflows, stochastic processes often occur during multiple stages of the training pipeline, including dataset partitioning, parameter initialization, and stochastic optimization procedures. These stochastic processes are typically controlled through a random seed parameter in order to ensure

experimental reproducibility. However, different random seed values may lead to variations in model training outcomes and evaluation results, potentially affecting the reliability of reported performance metrics. Previous studies have highlighted that stochastic components in machine learning experiments can introduce variance in benchmark results and influence experimental conclusions [15], [19].

To systematically examine this effect, the experiment is conducted by repeatedly training machine learning models using multiple random seed values while keeping all other experimental settings constant. The variability of model performance across repeated runs is then analyzed to evaluate model stability.

The overall experimental workflow of this study consists of several stages: data preprocessing, dataset splitting, model training, performance evaluation, and stability analysis. Each model is trained multiple times using different random seed values, and the resulting performance metrics are collected for further statistical analysis. The variability of these metrics is used as an indicator of model stability across different experimental runs.

2.1 Experimental Workflow

Figure 1 summarizes the protocol. Every dataset passes through the same preprocessing pipeline: standard scaling for numerical attributes and one-hot encoding for categorical ones, both fitted on the training portion only.

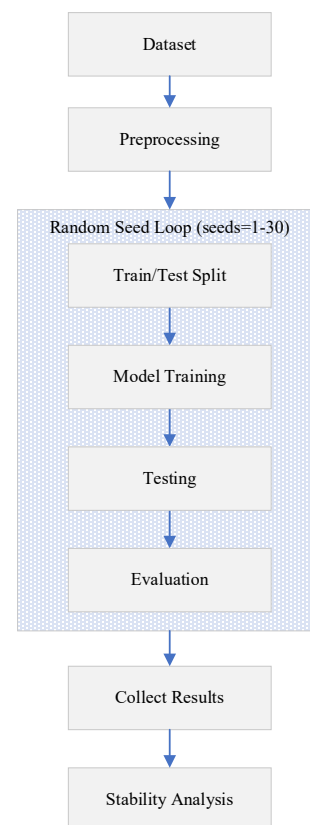


Figure 1. Experimental Workflow

A loop then runs over seeds $s = 1, \dots, 30$. In each iteration, the seed fixes the stratified 80:20 train-test split and, where the algorithm has one, its internal random state. The model is trained on the training portion and scored on the test portion, and its Accuracy, F1-score, and ROC-AUC are stored. The 30 scores for each model-dataset pair are then summarized with the mean, standard deviation, SSI, and RRS. Because preprocessing and hyperparameters are identical in every run, differences between runs can be traced to the seed alone.

First, the tabular datasets are collected and preprocessed. The preprocessing stage includes handling missing values, encoding categorical features into numerical representations, and constructing the feature matrix and target labels required for model training.

After preprocessing, the dataset is divided into training and testing subsets. The dataset splitting process is controlled by a random seed parameter to ensure reproducibility while allowing controlled variation across repeated experiments.

Next, machine learning models are trained using the training dataset. To analyze the influence of randomness, the training process is repeated multiple times using different random seed values. For each seed value, the model training procedure is performed independently while keeping all other parameters constant.

Following model training, predictive performance is evaluated using the testing dataset. Several evaluation metrics, including F1-score, accuracy, and area under the receiver operating characteristic curve (ROC-AUC), are recorded for each experimental run.

Finally, the performance results obtained from multiple random seeds are aggregated and analyzed. Statistical measures such as the mean and standard deviation of performance metrics are computed to assess model stability. A lower variance in performance indicates greater model stability, whereas higher variability suggests stronger sensitivity to random seed initialization.

This experimental workflow enables a systematic analysis of how stochastic elements in machine learning pipelines influence model performance across repeated experimental runs, which has been highlighted as an important consideration in reproducible machine learning research [15].

2.2 Datasets

This study employs publicly accessible tabular datasets commonly used for classification in machine learning research. The datasets were chosen according to three criteria: (1) public accessibility to support reproducibility, (2) structured features compatible with conventional machine learning algorithms, and (3) previous use in machine learning benchmarking studies. All datasets were obtained from the UCI Machine

Learning Repository, a widely used platform for benchmark datasets in machine learning research.

The datasets used in this study include the Adult Income Dataset, the Bank Marketing Dataset, and the Breast Cancer Wisconsin Dataset. These datasets represent different application domains, including socio-economic prediction, financial marketing analysis, and medical diagnosis. Table 1 summarizes the main characteristics of the datasets used in this study.

Table 1. Dataset Characteristics

Dataset	Instances Used		Features	Positive class	Task
Adult Income	48,842	45,222	14	24.8% (>50K)	Income classification
Bank Marketing	45,211	45,211	16	11.7% (yes)	Term deposit prediction
Breast Cancer Wisconsin	569	569	30	37.3% (malignant)	Tumor classification

The Adult Income dataset [23] contains demographic and employment-related attributes used to predict whether a person earns more than USD 50,000 per year from census data. The dataset includes both categorical and numerical attributes such as age, education level, occupation, and weekly working hours. We use the full version (adult.data and adult.test, 48,842 records) and remove the records with missing values, leaving 45,222.

The Bank Marketing dataset [24] comes from telephone campaigns of a Portuguese bank; the target is whether a contacted client subscribed to a term deposit, and positive cases form a small minority (11.7%). Its attributes include the duration of the last call, which is only known after the call ends; UCI recommends discarding it for realistic prediction, but it is retained here as in many benchmark studies, so absolute Bank scores should be read as optimistic.

The Breast Cancer Wisconsin dataset contains diagnostic features computed from digitized images of breast tissue samples. The classification task is to determine whether a tumor is benign or malignant based on these features. This dataset is commonly used for evaluating machine learning models in medical diagnosis tasks [25].

Before model training, several preprocessing steps were applied. Records with missing values (Adult only) were removed, and categorical attributes were converted into numerical representations using one-hot encoding. Numerical features were standardized to ensure consistent input ranges for machine learning algorithms.

After preprocessing, each dataset was divided into training and testing subsets. The splitting process was controlled using different random seed values in order to analyze the variability of model performance across multiple experimental runs. This approach allows the study to systematically evaluate the stability of machine learning models under different random initialization conditions.

2.3 Machine Learning Models

This study evaluates several widely used machine learning algorithms representing different modeling paradigms for tabular data classification. The selected models include neural networks, ensemble learning methods, margin-based classifiers, and gradient boosting algorithms. Evaluating models from different algorithm families allows the study to analyze whether model stability under different random seed configurations varies across different learning mechanisms [3], [4], [26], [27].

Let $x \in \mathbb{R}^p$ denote the feature vector containing p attributes and $y \in \{0,1\}$ denote the class label. The objective of a classification model is to learn a mapping function $f: \mathbb{R}^p \rightarrow \{0,1\}$ that predicts the class label y from the input feature vector x [4].

The machine learning models evaluated in this study include Multi-Layer Perceptron (MLP), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost).

An MLP passes the input through one or more fully connected hidden layers before producing an output [28]. Each neuron performs a weighted linear transformation followed by a non-linear activation function.

For a network with a single hidden layer, the hidden representation h is computed as Equation 1, where $x \in \mathbb{R}^p$ is the input feature vector, W_1 and b_1 are the hidden-layer weights and biases, and $\sigma(\cdot)$ is an element-wise non-linear activation function such as the Rectified Linear Unit (ReLU) [29].

$$h = \sigma(W_1 x + b_1) \quad (1)$$

The final prediction $\hat{y}$ is then computed as Equation 2, where W_2 and b_2 represent the weight matrix and bias vector of the output layer, respectively, and $\phi(\cdot)$ denotes the output activation function.

$$\hat{y} = \phi(W_2 h + b_2) \quad (2)$$

Model parameters are learned using backpropagation and gradient-based optimization methods [28]. Despite their flexibility, recent studies show that neural networks often underperform tree-based methods on tabular datasets [4], [26].

Random Forest [7] constructs an ensemble of K decision trees. Each tree is trained on a bootstrap sample of the training data and considers a random subset of features at each split. The final class prediction is determined by majority voting among the K trees, as expressed in Equation 3, where $T_k(x)$ denotes the prediction of the k -th decision tree for input x .

$$\hat{y} = \text{mode}(T_1(x), \dots, T_K(x)) \quad (3)$$

Random Forest reduces variance and improves generalization [7], and remains a strong baseline for tabular classification tasks [3], [27].

Support Vector Machine identifies a decision boundary that separates different classes by maximizing the margin between them [9].

$$w^T x + b = 0 \quad (4)$$

The decision boundary is defined in Equation 4, where $x \in \mathbb{R}^p$ denotes the input feature vector, $w \in \mathbb{R}^p$ denotes the weight vector that determines the orientation of the hyperplane, and b is the bias term that controls its position. The objective of SVM is to find the optimal values of w and b to maximize the margin between the classes.

Through kernel functions, SVM can model non-linear decision boundaries by implicitly representing input data in higher-dimensional feature spaces [9]. Despite its effectiveness in high-dimensional problems, SVM may be less competitive than ensemble learning algorithms when applied to large-scale tabular data [3], [4].

XGBoost is an efficient and scalable gradient boosting algorithm that iteratively builds decision trees, with each new tree aimed at reducing the prediction errors of the existing model [8]. The model prediction is defined in Equation 5, where $\hat{y}_i$ denotes the predicted output for the i -th instance, K represents the total number of trees, and $f_k(\cdot)$ denotes the prediction function of the k -th decision tree applied to input x_i .

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (5)$$

Equation 6 expresses the XGBoost objective function as the sum of a loss component and a regularization component. The term $l(y_i, \hat{y}_i)$ measures the discrepancy between the true label y_i and the predicted value $\hat{y}_i$, while $\Omega(f_k)$ penalizes model complexity to help reduce overfitting.

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (6)$$

XGBoost is widely recognized for its superior performance on structured or tabular datasets and has consistently outperformed many other methods in empirical benchmarks [3], [4], [8], [26], [27].

The sources of randomness associated with each machine learning algorithm are summarized in Table 2. These stochastic elements may influence the training process and lead to variability in model performance when different random seed values are used. For example, neural networks such as MLP rely on random weight initialization, while Random Forest incorporates randomness through sampling strategies during tree construction. In contrast, SVM and, with the default settings used here, XGBoost are deterministic for a given training set: their predictions change between runs only because the seed changes the train-test split. (The probability estimates of SVM, used for ROC-AUC, additionally involve a small unseeded internal cross-validation whose effect was about 0.0001 in our checks.) Their variability therefore reflects data partitioning alone, which must be kept in mind when comparing their stability with that of Random Forest

and MLP. By evaluating these models across multiple random seeds, this study aims to analyze how different sources of randomness affect model stability in tabular classification tasks. Therefore, repeated experiments

across multiple random seed configurations are conducted to systematically evaluate the variability of model performance.

Table 2. Machine learning models and sources of randomness

Model	Algorithm Category	Sources of Randomness
Multi-Layer Perceptron (MLP)	Neural Network	Random initialization of network weights and biases, and mini-batch shuffling during training
Random Forest (RF)	Bagging Ensemble	Bootstrap sampling of training instances and random feature selection at each tree split
Support Vector Machine (SVM)	Margin-based Classifier	None at model level (no random state set); predictions vary only through the random train-test split
XGBoost	Gradient Boosting Ensemble	None at model level with the settings used (no row or column subsampling); predictions vary only through the random train-test split

2.4 Experimental Procedure

To investigate the influence of random seed initialization on model stability, each machine learning model is trained repeatedly using different random seed values while keeping all other experimental parameters constant. The random seed controls several stochastic components in the training pipeline, including dataset splitting, model initialization, and internal randomization processes used by certain algorithms.

For each dataset, the data were divided into training and testing sets using an 80:20 stratified split. To evaluate the impact of random seed variation, the experiment was repeated 30 times using seed values ranging from 1 to 30. The same preprocessing pipeline (standard scaler for numerical features and one-hot encoder for categorical features) and model hyperparameter settings were maintained across all runs to ensure that observed differences were attributable solely to random seed variation.

The experimental configuration and model hyperparameters used in this study are summarized in Table 3. Only the random seed controlling data partitioning and model initialization (where applicable) was varied across experimental runs. The experiments were run in Python with scikit-learn 1.6.1 and XGBoost 3.1.3. Even with a fixed seed, MLP results can differ in the last digits between computing environments because of floating-point arithmetic.

Table 3. Model hyperparameter configuration

Model	Hyperparameter Configuration
Random Forest	n_estimators = 200, random_state = seed
XGBoost	n_estimators = 200, max_depth = 6, learning_rate = 0.1, eval_metric = 'logloss', random_state = seed
SVM	kernel = 'rbf', probability = True
MLP	hidden_layer_sizes = (100, 50), max_iter = 300, random_state = seed

Let $D = \{(x_i, y_i)\}_{i=1}^N$ denote a tabular dataset consisting of N instances, where $x_i \in \mathbb{R}^p$ represents the feature vector with p attributes and y_i represents the corresponding class label. The dataset is divided into two subsets: a training set D_{train} and a testing set D_{test} . The splitting process is controlled by a random seed parameter s , which determines the random sampling used during the partitioning process.

In this study, a set of random seed values $S = \{s_1, s_2, \dots, s_n\}$ with $n = 30$ is used to conduct repeated experiments. For each seed $s_j \in S$, the dataset splitting and model training procedures are performed independently. This approach allows the variability of model performance due to random initialization to be systematically evaluated.

The experimental procedure can be summarized as follows: (1) Clean the dataset by removing records with missing values. (2) Select a random seed value s_j . (3) Split the dataset into training and testing sets using seed s_j . (4) Train the machine learning model M using the training dataset D_{train} . (5) Evaluate the trained model on the testing dataset D_{test} . (6) Record the resulting performance metrics. (7) Repeat the process for all seed values in S .

Let m_{ij} denote the performance metric obtained by model M_i when trained using random seed s_j . The performance results across multiple runs are then collected to analyze variability and stability.

This experimental design ensures that any variation in model performance can be attributed primarily to differences in random seed initialization rather than changes in model configuration or dataset preprocessing. By analyzing performance distributions across multiple runs, the study provides a more reliable assessment of model robustness compared to single-run evaluations.

2.5 Evaluation Metrics

To evaluate model performance, commonly used classification metrics are employed. These metrics measure different aspects of predictive performance and provide a comprehensive evaluation of model behavior [30], [31].

The first evaluation metric is accuracy. Accuracy measures the proportion of correctly predicted instances among the total number of samples. Accuracy can be seen in Equation 7.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

TP denotes true positive predictions, TN represents true negative predictions, FP indicates false positive predictions, and FN refers to false negative predictions.

Although widely used, accuracy may provide a misleading assessment of model performance on imbalanced datasets [30], [32].

The F1-score is the harmonic mean of precision and recall and is commonly used when class distributions are imbalanced, where $\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$ and $\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$. F1-score can be seen in Equation 8.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

It provides a balance between false positives and false negatives and is more informative than accuracy in skewed datasets [30], [32].

Finally, ROC-AUC summarizes how well the predicted scores rank positive instances above negative ones across all possible thresholds; 0.5 corresponds to random ranking and 1.0 to perfect separation [33].

2.6 Model Stability Analysis

To assess model stability, the performance metrics obtained from multiple runs with different random seeds are analyzed statistically. Instead of relying on a single evaluation result, the distribution of performance metrics across repeated runs is examined [12], [15].

The mean performance across all experimental runs is defined in Equation 9, where n denotes the number of experimental runs, and x_i represents the performance metric (e.g., accuracy) obtained in the i -th run. The mean provides an estimate of the expected model performance across repeated experiments [15].

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \quad (9)$$

The standard deviation (Equation 10) measures the spread of performance across runs; as is usual for a sample of runs, it divides by $n - 1$.

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2} \quad (10)$$

where σ denotes the standard deviation, x_i represents the performance value of the i -th run, and μ is the mean performance. Lower standard deviation reflects more consistent model performance across random seed configurations, whereas greater variability indicates higher sensitivity to seed changes [12], [15].

To explicitly quantify the sensitivity of a model to randomness, this study introduces the Stability Sensitivity Index (SSI), defined as the normalized variability of model performance.

$$SSI = \frac{\sigma}{\mu} \quad (11)$$

SSI can be seen in Equation 11, where σ represents the standard deviation of performance across runs and μ denotes the mean performance. This formulation is analogous to the coefficient of variation, a commonly used normalized dispersion measure in statistical analysis; seed-to-seed dispersion of this kind has also been used to assess the stability of fine-tuned language

models [34]. A lower SSI indicates higher stability, whereas a higher SSI reflects greater sensitivity to random seed initialization. SSI enables fair comparison between models with different performance scales by normalizing variability relative to mean performance.

While SSI provides a normalized measure of variability, it does not indicate whether the observed variability is practically acceptable. Therefore, this study introduces the Relative Robustness Score (RRS) to quantify model variability relative to a predefined tolerance threshold, as shown in Equation 12.

$$RRS = \frac{\sigma}{\tau} \quad (12)$$

σ denotes the standard deviation of model performance across repeated runs and τ represents an acceptable variability tolerance. In this study, τ was set to 0.01, corresponding to a 1% performance variation threshold. This threshold was selected to provide a practical reference for distinguishing acceptable and excessive performance variability.

The RRS provides an intuitive interpretation of robustness by expressing observed variability relative to the acceptable tolerance level. Lower RRS values indicate greater robustness and lower sensitivity to random seed variation, whereas higher RRS values indicate greater instability. For example, an RRS value below 1 indicates that the observed variability remains within the predefined tolerance threshold, while an RRS value greater than 1 suggests that variability exceeds the acceptable range.

While SSI expresses variability relative to mean performance, RRS assesses it against a practical tolerance criterion. These metrics offer complementary perspectives on model stability: SSI facilitates comparisons across models with different performance scales, whereas RRS indicates whether the observed variability falls within an acceptable range.

By reporting SSI and RRS alongside mean performance and standard deviation, this study provides a more comprehensive assessment of model stability across random seed configurations. This evaluation is relevant to ongoing concerns about reproducibility and robustness in machine learning research [12], [15], [19], [26], [35].

3. Results and Discussions

Tables 4–6 present the performance and stability of the evaluated models across three datasets. In addition to mean performance and standard deviation, the two stability measures defined in Section 2.6 are reported: the Stability Sensitivity Index (SSI) and the Relative Robustness Score (RRS). These metrics provide a more comprehensive assessment of model robustness under different random seed configurations. The SSI reflects the normalized variability of model performance, while the RRS provides a relative measure of robustness, where lower values indicate higher stability. In each

table, the highest mean and the lowest Std, SSI, and RRS within each dataset are shown in bold.

Table 4. Accuracy-based performance stability across datasets

Dataset	Model	Mean Acc	Std	SSI	RRS
Adult	SVM	0.853	0.0029	0.0034	0.29
Adult	RandomForest	0.850	0.0029	0.0034	0.29
Adult	XGBoost	0.871	0.0027	0.0031	0.27
Adult	MLP	0.817	0.0047	0.0058	0.47
Bank	SVM	0.905	0.0020	0.0023	0.20
Bank	RandomForest	0.906	0.0022	0.0024	0.22
Bank	XGBoost	0.909	0.0025	0.0027	0.25
Bank	MLP	0.885	0.0042	0.0047	0.42
Cancer	SVM	0.973	0.0137	0.0141	1.37
Cancer	RandomForest	0.958	0.0180	0.0188	1.80
Cancer	XGBoost	0.969	0.0178	0.0184	1.78
Cancer	MLP	0.974	0.0163	0.0167	1.63

Table 4 presents the accuracy-based performance stability of the evaluated machine learning models across the Adult, Bank, and Cancer datasets. In addition to mean accuracy, the analysis incorporates standard deviation, Stability Sensitivity Index (SSI), and Relative Robustness Score (RRS) to evaluate the robustness of each model under varying random seed configurations.

On the Adult dataset, XGBoost achieved the highest mean accuracy (0.871) while also exhibiting the lowest SSI (0.0031) and RRS (0.27) values among all models. This indicates that XGBoost not only provides superior predictive performance but also maintains strong robustness against random seed variations. SVM and Random Forest followed closely (SSI 0.0034 for both). In contrast, the MLP model obtained the lowest mean accuracy (0.817) and the highest instability, reflected by the largest SSI (0.0058) and RRS (0.47) values. These findings suggest that neural network-based approaches are more sensitive to random initialization on this large tabular dataset.

For the Bank dataset, all models demonstrated relatively low variability compared to the other datasets, lower than on Adult even though the two datasets are almost equal in size. Part of this difference follows from the higher accuracy level on Bank, since the sampling variability of a proportion shrinks as it approaches 1. SVM exhibited the highest stability, achieving the lowest SSI (0.0023) and RRS (0.20) values, although XGBoost achieved the highest mean accuracy (0.909). Random Forest also showed competitive stability with similarly low variability. These results indicate that both SVM and ensemble-based approaches maintain robust performance under repeated random seed configurations in large-scale tabular data.

On the Cancer dataset, substantially higher variability was observed across all models. This dataset, which contains a relatively small number of samples, appears to amplify the impact of random seed selection. SVM achieved the best stability, with the lowest SSI (0.0141) and RRS (1.37), while MLP obtained the highest mean accuracy (0.974), marginally above SVM (0.973).

Random Forest and XGBoost exhibited the highest instability (SSI 0.0188 and 0.0184). With only 114 test instances, a single misclassification changes accuracy by about 0.009, and accuracies across seeds spanned up to 7.9 percentage points, so all RRS values exceed 1. The findings suggest that smaller datasets are more susceptible to seed-induced performance fluctuations.

Overall, the results demonstrate that model stability varies considerably across algorithms and datasets. XGBoost achieves the strongest predictive performance on the two large datasets, while SVM is the most stable model on Bank and Cancer. Furthermore, the results indicate that dataset size plays an important role in seed sensitivity, with the small Cancer dataset exhibiting substantially higher instability than the two large datasets.

Table 5 presents the stability analysis based on F1-score, which is particularly important for evaluating classification performance in datasets with class imbalance. Compared to accuracy, F1-score provides a more balanced assessment by simultaneously considering precision and recall. The inclusion of SSI and RRS based on F1-score therefore enables a more robust evaluation of model sensitivity under varying random seed configurations.

Table 5. F1-score-based performance stability across datasets

Dataset	Model	Mean F1	Std	SSI	RRS
Adult	SVM	0.668	0.0082	0.0122	0.82
Adult	RandomForest	0.676	0.0069	0.0102	0.69
Adult	XGBoost	0.717	0.0061	0.0085	0.61
Adult	MLP	0.627	0.0123	0.0197	1.23
Bank	SVM	0.464	0.0148	0.0319	1.48
Bank	RandomForest	0.500	0.0135	0.0269	1.35
Bank	XGBoost	0.560	0.0132	0.0236	1.32
Bank	MLP	0.502	0.0122	0.0243	1.22
Cancer	SVM	0.962	0.0191	0.0199	1.91
Cancer	RandomForest	0.943	0.0257	0.0272	2.57
Cancer	XGBoost	0.957	0.0254	0.0266	2.54
Cancer	MLP	0.963	0.0238	0.0247	2.38

On the Adult dataset, XGBoost achieved the highest mean F1-score (0.717) while also exhibiting the lowest SSI (0.0085) and RRS (0.61) values. This indicates that XGBoost maintains both strong predictive capability and relatively stable class discrimination performance across repeated runs. In contrast, MLP showed the lowest mean F1-score (0.627) and the highest instability, reflected by the largest SSI (0.0197) and RRS (1.23) values. Compared to the accuracy-based results in Table 4, all models show roughly three times higher SSI under F1-score, and MLP is the only model whose variability exceeds the 1% tolerance (RRS greater than 1).

For the Bank dataset, F1-score variability was much higher than accuracy variability for all models, and all RRS values exceed 1. In absolute terms (standard deviation) Bank remained less variable than Cancer, but relative to its much lower F1 level (SSI) it was about as variable. XGBoost achieved the highest mean F1-score (0.560) and the lowest SSI (0.0236), while MLP had the lowest standard deviation and RRS (0.0122 and 1.22) at

a lower mean (0.502). SVM was the least stable model under this metric (SSI 0.0319, RRS 1.48) and had the lowest mean F1-score (0.464). The disagreement between SSI and RRS for XGBoost and MLP illustrates that the two indices answer different questions: SSI relates variability to the performance level, whereas RRS compares it with a fixed tolerance. The relatively larger SSI and RRS values for F1-score compared to accuracy indicate that seed variation affects minority-class prediction performance more strongly than overall classification accuracy. This finding is particularly relevant because the Bank dataset exhibits class imbalance characteristics (11.7% positive cases).

On the Cancer dataset, substantially higher instability was observed across all models, consistent with the findings in Table 4. SVM maintained the best overall stability, achieving the lowest SSI (0.0199) and RRS (1.91) values, with a mean F1-score (0.962) almost equal to the highest one, obtained by MLP (0.963). In contrast, Random Forest and XGBoost exhibited considerably larger variability (SSI 0.0272 and 0.0266), indicating greater sensitivity to random seed initialization. The elevated instability observed in the Cancer dataset further supports the hypothesis that smaller datasets amplify the effects of random initialization and data partitioning.

Overall, the F1-score-based analysis confirms the stability patterns observed in the accuracy-based evaluation, with one exception: on Bank, SVM moves from the most stable to the least stable model. XGBoost achieves the strongest predictive performance on the large datasets, whereas SVM remains the most stable model on Cancer. Furthermore, the results suggest that F1-score is generally more sensitive to seed variation than accuracy for every model. This indicates that evaluating stability solely based on accuracy may underestimate the true variability of model behavior, especially in imbalanced classification settings.

Table 6. ROC-AUC-based performance stability across datasets

Dataset	Model	Mean AUC	Std	SSI	RRS
Adult	SVM	0.899	0.0035	0.0039	0.35
Adult	RandomForest	0.902	0.0031	0.0034	0.31
Adult	XGBoost	0.929	0.0024	0.0025	0.24
Adult	MLP	0.868	0.0042	0.0049	0.42
Bank	SVM	0.909	0.0045	0.0050	0.45
Bank	RandomForest	0.931	0.0034	0.0037	0.34
Bank	XGBoost	0.937	0.0029	0.0031	0.29
Bank	MLP	0.885	0.0055	0.0062	0.55
Cancer	SVM	0.995	0.0044	0.0044	0.44
Cancer	RandomForest	0.990	0.0093	0.0094	0.93
Cancer	XGBoost	0.993	0.0060	0.0060	0.60
Cancer	MLP	0.994	0.0053	0.0053	0.53

Table 6 presents the stability analysis based on ROC-AUC, which evaluates the discriminative capability of classification models independent of a fixed decision threshold. Compared to Accuracy and F1-score, ROC-AUC provides a threshold-independent assessment of model performance, making it particularly useful for examining the robustness of probabilistic predictions under different random seed configurations.

On the Adult dataset, XGBoost achieved the highest mean ROC-AUC value (0.929) while simultaneously exhibiting the lowest SSI (0.0025) and RRS (0.24) values. These results indicate that XGBoost not only produces strong discriminative performance but also maintains highly stable probability estimations across repeated runs. Random Forest and SVM demonstrated comparable stability with slightly larger variability, whereas MLP showed the highest sensitivity to random seed variation, reflected by the largest SSI (0.0049) and RRS (0.42) values. Although the instability observed in ROC-AUC is lower than that observed using F1-score, the results consistently indicate that neural network-based models are more susceptible to seed-induced fluctuations on this dataset.

For the Bank dataset, all models exhibited low variability. XGBoost achieved both the highest mean ROC-AUC (0.937) and the lowest SSI (0.0031) and RRS (0.29), followed by Random Forest (SSI 0.0037), while SVM (0.0050) and MLP (0.0062) were less stable. Compared to the F1-score analysis, the ROC-AUC results demonstrate noticeably lower variability across all models. This suggests that seed-induced fluctuations affect classification thresholds and class-specific predictions more strongly than the overall ranking quality of probabilistic outputs.

On the Cancer dataset, higher variability was again observed across all models, consistent with the previous analyses. Nevertheless, ROC-AUC stability remained substantially better than F1-score stability. SVM achieved both the highest mean ROC-AUC value (0.995) and the lowest SSI (0.0044) and RRS (0.44) values, indicating highly robust discriminative performance despite the small dataset size. In contrast, Random Forest exhibited the largest instability (SSI 0.0094, RRS 0.93), followed by XGBoost (SSI 0.0060). These findings further support the observation that smaller datasets amplify the impact of random seed variation.

Overall, the ROC-AUC-based analysis confirms the general stability patterns identified in the previous evaluations while revealing lower overall variability compared to F1-score. XGBoost is the most stable model on Adult and Bank and SVM on Cancer, and XGBoost has the highest mean ROC-AUC on the two large datasets. Furthermore, the lower SSI and RRS values observed for ROC-AUC suggest that threshold-independent evaluation metrics are more stable under random seed variation than threshold-dependent metrics such as F1-score. This finding highlights the importance of evaluating model robustness using multiple complementary performance metrics rather than relying solely on accuracy-based assessments.

Notably, in terms of accuracy, all models exhibit substantially higher variability on the Cancer dataset compared to the Adult and Bank datasets. This observation indicates that smaller datasets are more sensitive to random seed variation, as minor differences in data partitioning and initialization can produce larger

fluctuations in model performance. In contrast, the two larger datasets demonstrate considerably lower variability across all evaluated models, suggesting that increased sample size contributes to more stable learning behavior and reduced sensitivity to stochastic factors.

Figure 2 illustrates the distribution of F1-scores across 30 random seed configurations for each dataset. On

Adult, XGBoost shows the narrowest distribution and MLP the widest; on Bank, all four distributions are wide relative to their means, with SVM the most dispersed relative to its level; on Cancer, SVM is the most concentrated, while the other three models show long lower tails, reaching F1 values below 0.88 in individual runs. These visual observations are consistent with the quantitative stability measurements presented in Tables 4–6.

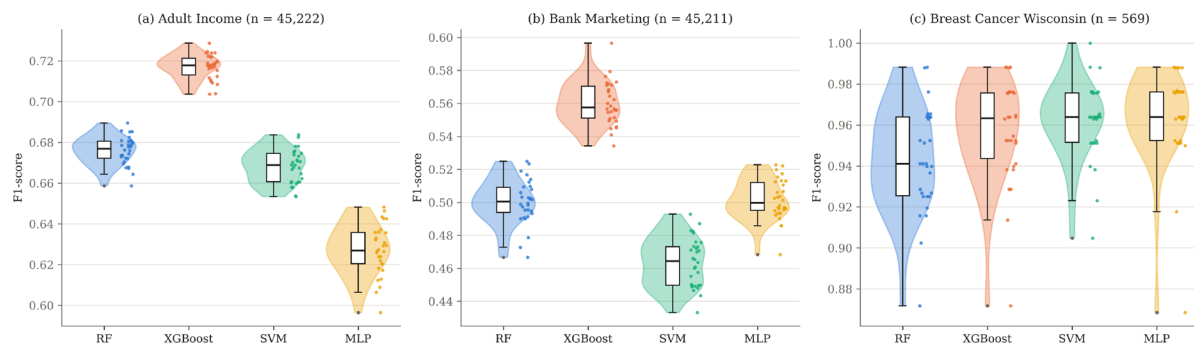


Figure 2. Distribution of F1-score across 30 random seed configurations for the Adult Income, Bank Marketing, and Breast Cancer datasets. The datasets contain 45,222, 45,211, and 569 samples (after removing records with missing values), respectively.

Across all datasets, several consistent patterns emerge. First, XGBoost generally achieves the highest predictive performance, particularly on the larger tabular datasets, confirming its effectiveness for structured data classification tasks. However, the highest predictive performance does not always coincide with the highest robustness: on Bank (accuracy) and on Cancer, SVM is more stable than the best-scoring model.

Second, the two models without model-level randomness in our configuration, SVM and XGBoost, have the lowest SSI in every dataset–metric combination (SVM in four and XGBoost in five of the nine). Their variability reflects the train–test split alone. Random Forest, whose bootstrap and feature sampling add a further source of randomness, stays close to them on the large datasets, which suggests that the split dominates its variability. MLP, in contrast, varies clearly more on Adult and Bank for accuracy and ROC-AUC, so its own initialization and optimization add substantial variance on top of the split.

MLP is the least stable model on Adult for all three metrics and on Bank for accuracy and ROC-AUC. It also stopped at the 300-iteration limit before converging in 22 of 30 Adult runs and 16 of 30 Bank runs, which may contribute to this variability; on Cancer, where it always converged, it was not among the least stable models. This finding suggests that stochastic gradient-based optimization and random weight initialization contribute substantially to variability in neural network performance.

Furthermore, the results reveal that stability patterns vary depending on the evaluation metric. For every dataset and model, F1-score exhibits higher variability

than Accuracy and ROC-AUC. Accuracy and ROC-AUC are of comparable stability on the two large datasets, while ROC-AUC is clearly the most stable metric on Cancer. This suggests that threshold-dependent metrics such as F1-score are more sensitive to seed-induced prediction fluctuations, particularly in imbalanced classification settings. In contrast, ROC-AUC, which evaluates ranking quality independently of a fixed decision threshold, appears less affected by random seed variation.

The findings indicate that model robustness depends on the algorithm used, the characteristics of the dataset, and the evaluation metric. In general, larger datasets tend to exhibit lower sensitivity to random seed variation, whereas smaller datasets may show greater performance variability. These observations underscore the need to incorporate stability analysis into machine learning evaluation, as mean predictive performance alone may not fully capture a model's reliability and reproducibility.

3.1 Key Finding 1: Predictive Performance Does Not Necessarily Imply Stability

The experimental results show that predictive performance and stability are distinct characteristics of machine learning models. On Adult, the most accurate model (XGBoost) is also the most stable, but on Bank (accuracy) SVM is more stable than XGBoost, and on Cancer SVM is more stable than MLP, which has the highest mean accuracy and F1-score.

These findings suggest that models with higher predictive accuracy are not necessarily the most reliable under repeated experiments. Therefore, evaluating machine learning models solely based on mean performance or single-run results may lead to

incomplete conclusions. Both predictive performance and stability should be considered jointly, particularly in applications requiring reproducibility and consistency. This observation supports recent recommendations emphasizing reproducibility-oriented evaluation rather than reliance on single-run performance estimates [15], [19].

3.2 Key Finding 2: Model Stability Depends on the Learning Mechanism and Its Sources of Randomness

The experimental results indicate that model stability is strongly influenced by the underlying learning mechanism of each algorithm. The two models without internal randomness in our configuration, SVM and XGBoost, show the lowest variability in every dataset–metric combination. This should be interpreted carefully: for these models the train–test split is the only source of variation, so their low variability partly reflects fewer sources of randomness rather than a more stable learning mechanism as such.

Random Forest also shows competitive stability, particularly on larger datasets such as Adult and Bank Marketing. Its ensemble aggregation mechanism appears to reduce sensitivity to stochastic fluctuations and improve robustness across repeated runs.

In contrast, MLP exhibits the highest instability on the two large datasets for most metrics. This behavior is likely caused by stochastic gradient optimization, random weight initialization, and the non-convex nature of neural network training, which can lead to different convergence outcomes under different random seeds; incomplete convergence within the iteration limit may also play a role.

These findings suggest that the learning characteristics of machine learning algorithms significantly affect robustness and reproducibility. Models relying heavily on stochastic optimization tend to be more sensitive to random seed variations, while margin-based and ensemble-based methods generally exhibit more stable behavior. Similar observations have been reported in previous reproducibility studies, which identified random initialization and optimization dynamics as important sources of experimental variability [15], [20].

3.3 Key Finding 3: Ensemble-Based Methods Provide a Favorable Trade-off Between Performance and Stability

The experimental results show that ensemble-based methods, particularly Random Forest and XGBoost, provide a strong balance between predictive performance and stability. Across most datasets and evaluation metrics, these models achieve high predictive capability while maintaining lower variability than neural network-based approaches. XGBoost often achieves the highest predictive performance, whereas Random Forest demonstrates stable behavior on the large datasets but is the least stable model on Cancer.

Random Forest still exhibits moderate sensitivity to random seed variation because of bootstrap sampling and feature subsampling, whereas XGBoost, which uses no subsampling in our configuration, varies only through the data split.

The findings suggest that ensemble methods offer a practical compromise between predictive accuracy and reproducibility. On the large datasets, XGBoost combined the best mean performance with the lowest variability in most cases, making it well suited for applications requiring both effectiveness and consistency.

3.4 Key Finding 4: Dataset Characteristics Significantly Influence Stability

The experimental results indicate that dataset characteristics, particularly dataset size, strongly influence model stability under different random seed configurations. For accuracy, and for the standard deviation of F1-score, the larger Adult Income and Bank Marketing datasets exhibit lower variability than Cancer for every model, suggesting that larger sample sizes provide more stable data distributions and reduce stochastic effects during training.

In contrast, the Breast Cancer dataset shows substantially higher variability across most models and metrics. Due to its smaller size, slight changes in train–test partitioning can significantly affect learned decision boundaries and predictive outcomes, especially for the tree ensembles (Random Forest and XGBoost). With 114 test instances, each misclassified case moves accuracy by almost 0.9 percentage points.

The results also show that F1-score fluctuations are amplified both by small sample size (Cancer) and by class imbalance (Bank, 11.7% positives): relative to its mean, F1 on Bank is about as variable as on Cancer despite a sample nearly eighty times larger.

Overall, these findings suggest that model robustness should be evaluated together with dataset characteristics. Experimental results obtained from smaller datasets are more sensitive to random seed selection, highlighting the importance of repeated experiments to improve reliability and reproducibility in machine learning research. This finding is consistent with prior work showing that experimental variance becomes more pronounced when limited sample sizes amplify stochastic training effects [15].

3.5 Key Finding 5: Model Performance Is Strongly Dataset-Dependent

The experimental results demonstrate that no single machine learning model consistently outperforms others across all datasets and evaluation metrics. Instead, model effectiveness depends heavily on dataset characteristics and learning conditions. XGBoost generally achieves the best predictive performance on larger tabular datasets such as Adult Income and Bank Marketing, while on the smaller Breast Cancer dataset SVM and MLP reach almost identical mean

performance, with SVM being the more consistent of the two.

These findings suggest that model suitability is strongly influenced by factors such as dataset size, feature complexity, and class distribution. Larger datasets tend to favor ensemble-based methods like XGBoost, whereas smaller datasets benefit from lower-variance models such as SVM.

The results also indicate that model selection should not rely solely on predictive performance. Stability and reproducibility should also be considered, since models with slightly higher accuracy may exhibit greater sensitivity to random seed variation.

Overall, the findings highlight the importance of dataset-aware model selection, where models are evaluated based on both predictive performance and stability within the context of the target dataset and application domain.

3.6 Key Finding 6: Reporting Performance Variability Is Essential for Reliable Evaluation

The experimental results show that models with similar average predictive performance can exhibit substantially different levels of variability across random seed configurations. For example, on Cancer, MLP and SVM reach almost the same mean accuracy (0.974 and 0.973), yet SVM has a clearly lower standard deviation, SSI, and RRS (0.0137, 0.0141, and 1.37 versus 0.0163, 0.0167, and 1.63).

These findings highlight the limitations of evaluating machine learning models solely based on average performance from a single experimental run. Models with higher variability may produce inconsistent results depending on random seed selection, data partitioning, or initialization conditions.

The results further demonstrate that stability analysis provides important complementary information beyond traditional predictive evaluation. Metrics such as standard deviation, SSI, and RRS enable a more comprehensive assessment of model robustness by quantifying sensitivity to stochastic experimental factors.

Overall, the findings emphasize the importance of reporting both predictive performance and variability-related metrics in machine learning experiments to improve transparency, reliability, and reproducibility. This recommendation aligns with reproducibility guidelines proposed in recent machine learning research [12], [19].

3.7 Limitations and Future Work

Despite providing important insights into model stability under random seed variation, this study has several limitations.

First, only three binary tabular datasets were used, so it is untested whether the same patterns hold for time-series forecasting, natural language processing,

computer vision, graph learning, or large-scale foundation models. Future work should evaluate the proposed Stability Sensitivity Index (SSI) and Relative Robustness Score (RRS) on broader datasets with varying complexity, feature distributions, and imbalance characteristics. Since SSI and RRS are model-agnostic measures of performance variability, they can potentially be extended to more complex learning environments involving stochastic optimization, data augmentation, transfer learning, and distributed training. Such investigations would provide stronger evidence regarding the generalizability of the proposed stability framework across diverse machine learning paradigms.

Second, while this study used Accuracy, F1-score, and ROC-AUC, additional metrics such as precision, recall, MCC [11], balanced accuracy, and calibration measures may provide further insights into robustness, particularly for imbalanced datasets.

Third, the experiments used standard hyperparameter configurations without extensive optimization. In particular, the MLP frequently stopped at the 300-iteration limit before converging on the two large datasets, so part of its variability may reflect incomplete training. Since hyperparameter tuning may influence both predictive performance and stability, future studies should investigate the relationship between hyperparameter optimization and random seed sensitivity.

Fourth, although the proposed Stability Sensitivity Index (SSI) and Relative Robustness Score (RRS) provide simple and interpretable measures of variability, they may not capture all aspects of robustness. Future research may explore more advanced approaches such as confidence intervals, hypothesis testing, variance decomposition, and distribution-based robustness analysis.

Fifth, the number of randomness sources differs between models. In our configuration, SVM and XGBoost vary only through the train-test split, while Random Forest and MLP also depend on their own internal randomness; a design that varies the split and the model seed separately would allow these effects to be quantified independently. Moreover, even with fixed seeds, MLP results differed slightly between computing environments because of floating-point arithmetic, which limits exact reproducibility.

Finally, this study focused primarily on variability caused by the seed-controlled train-test split and model initialization. Other stochastic factors, including preprocessing, feature selection, cross-validation strategies, and optimization settings, were not explicitly analyzed. Future work may investigate the combined effects of these factors and explore methods for reducing variability, such as deterministic training settings or seed-agnostic optimization strategies.

Overall, the findings highlight the importance of incorporating stability analysis into machine learning

evaluation. Expanding the analysis to broader datasets, additional robustness metrics, and more comprehensive experimental settings may further improve the reliability and reproducibility of machine learning research.

4. Conclusions

We trained MLP, Random Forest, SVM, and XGBoost under 30 different seeds on three public tabular benchmarks to measure how much their scores change when nothing but the seed changes.

To facilitate a more comprehensive assessment of robustness, this study introduced two complementary stability measures, namely the Stability Sensitivity Index (SSI) and the Relative Robustness Score (RRS). The results demonstrate that model evaluation based on a single experimental run may produce incomplete conclusions, as performance can vary considerably under different random seed settings; on the Breast Cancer dataset, accuracy differed by up to 7.9 percentage points between seeds.

The experimental findings show that predictive performance and stability are distinct model characteristics. XGBoost achieved the best mean performance on the two large datasets and was usually also the most stable there, whereas SVM was the most stable model on the small Breast Cancer dataset. The models without internal randomness in our configuration (SVM and XGBoost) were the most stable in every dataset-metric combination, which shows that the number of randomness sources must be considered when comparing stability. MLP was the most seed-sensitive model on the large datasets. The results also reveal that dataset characteristics, particularly dataset size and class imbalance, significantly influence stability: the small dataset produced the greatest variability in accuracy, and the imbalanced dataset produced F1-scores that were, relative to their level, as variable as on the small dataset. Applying SSI and RRS to Accuracy, F1-score, and ROC-AUC showed that they work as metric-independent summaries of model robustness, and that their rankings can differ when performance levels differ widely, as on Bank.

Taken together, the results show that reporting variability-related measures alongside average performance can improve the reliability, reproducibility, and interpretability of experimental results. From a practical perspective, the proposed SSI and RRS metrics provide a simple framework for identifying models that maintain consistent performance under stochastic experimental conditions. For benchmarking purposes, the findings suggest that model comparisons should be based on repeated experimental runs rather than single-run results. Furthermore, reproducibility-oriented evaluation protocols should report random seed configurations, performance variability, and stability measures in addition to average predictive performance.

Declarations

AI Usage Statement

The authors used ChatGPT (OpenAI) to assist with language editing, manuscript refinement, and the improvement of academic writing clarity. The authors designed the study, implemented the experiments, checked all reported results, and are responsible for the interpretation of the results and the scientific conclusions presented in this manuscript.

Author Contributions

NC: Conceptualization, methodology, formal analysis, data curation, investigation, writing - original draft, and writing - review and editing; AAA: Writing - review and editing, and critical revision of the manuscript. Both authors have read and approved the final version of the manuscript.

Funding Statement

This research did not receive specific funding.

Conflict of Interest Statement

The authors declare no conflict of interest.

Ethical Clearance

Not applicable. This study used publicly available datasets and did not involve direct recruitment of human participants or animal subjects.

Data and Code Availability

The datasets used in this study are publicly available through the UCI Machine Learning Repository. The experimental code and the result files supporting the findings of this study are available from the corresponding author upon reasonable request.

References

- [1] A. Esteva, A. Robicquet, B. Ramsundar, and others, "A guide to deep learning in healthcare," *Nat. Med.*, vol. 25, no. 1, pp. 24–29, 2019, doi: <https://doi.org/10.1038/s41591-018-0316-z>.
- [2] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012, doi: <https://doi.org/10.1145/2347736.2347755>.
- [3] V. Borisov, T. Leemann, K. Seßler, and others, "Deep neural networks and tabular data: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7499–7519, 2024, doi: <https://doi.org/10.1109/TNNLS.2022.3229161>.
- [4] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022, doi: <https://doi.org/10.1016/j.inffus.2021.11.011>.
- [5] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 18932–18943, 2021, doi: <https://doi.org/10.48550/arXiv.2106.11959>.
- [6] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *Journal of Machine Learning Research*, vol. 15, no. 90, pp. 3133–3181, 2014. [Online]. Available: <https://jmlr.org/papers/v15/delgado14a.html>
- [7] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: <https://doi.org/10.1023/A:1010933404324>.
- [8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*

- Mining, 2016, pp. 785–794. doi: <https://doi.org/10.1145/2939672.2939785>.
- [9] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: <https://doi.org/10.1007/BF00994018>.
- [10] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015, doi: <https://doi.org/10.1016/j.neunet.2014.09.003>.
- [11] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, p. 6, 2020, doi: <https://doi.org/10.1186/s12864-019-6413-7>.
- [12] O. E. Gundersen and S. Kjensmo, "State of the art: Reproducibility in artificial intelligence," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. doi: <https://doi.org/10.1609/aaai.v32i1.11503>.
- [13] S. Raste, R. Singh, J. Vaughan, and V. Nair, "Quantifying inherent randomness in machine learning algorithms," *arXiv preprint*, 2022, doi: <https://doi.org/10.48550/arXiv.2206.12353>.
- [14] D. Picard, "Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision," *arXiv preprint*, 2021, doi: <https://doi.org/10.48550/arXiv.2109.08203>.
- [15] X. Bouthillier, P. Delaunay, M. Bronzi, and others, "Accounting for variance in machine learning benchmarks," *Proceedings of Machine Learning and Systems*, vol. 3, pp. 747–769, 2021, doi: <https://doi.org/10.48550/arXiv.2103.03098>.
- [16] B. Bischl, M. Binder, M. Lang, and others, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 13, no. 2, p. e1484, Mar. 2023, doi: <https://doi.org/10.1002/widm.1484>.
- [17] O. Wilson, D. Schoeman, A. Bradley, and C. Clemente, "Practical guidelines for validation of supervised machine learning models in accelerometer-based animal behaviour classification," *Journal of Animal Ecology*, vol. 94, no. 7, pp. 1322–1334, Jul. 2025, doi: <https://doi.org/10.1111/1365-2656.70054>.
- [18] S. Bates, T. Hastie, and R. Tibshirani, "Cross-Validation: What Does It Estimate and How Well Does It Do It?," *J. Am. Stat. Assoc.*, vol. 119, no. 546, pp. 1434–1445, 2024, doi: <https://doi.org/10.1080/01621459.2023.2197686>.
- [19] J. Pineau, P. Vincent-Lamarre, K. Sinha, and others, "Improving reproducibility in machine learning research," *Journal of Machine Learning Research*, vol. 22, no. 164, pp. 1–20, 2021, doi: <https://doi.org/10.48550/arXiv.2003.12206>.
- [20] M. Hutson, "Artificial intelligence faces reproducibility crisis," *Science*, vol. 359, no. 6377, pp. 725–726, 2018, doi: <https://doi.org/10.1126/science.359.6377.725>.
- [21] R. Albertoni, S. Colantonio, P. Skrzypczyński, and J. Stefanowski, "Reproducibility of machine learning: Terminology, recommendations and open issues," *arXiv preprint*, 2023, doi: <https://doi.org/10.48550/arXiv.2302.12691>.
- [22] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012. [Online]. Available: <https://www.jmlr.org/papers/v13/bergstra12a.html>.
- [23] B. Becker and R. Kohavi, "Adult Data Set," 1996. doi: <https://doi.org/10.24432/C5XW20>.
- [24] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decis. Support Syst.*, vol. 62, pp. 22–31, 2014, doi: <https://doi.org/10.1016/j.dss.2014.03.001>.
- [25] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates," *Cancer Lett.*, vol. 77, no. 2–3, pp. 163–171, 1994, doi: [https://doi.org/10.1016/0304-3835\(94\)90099-X](https://doi.org/10.1016/0304-3835(94)90099-X).
- [26] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 507–520, 2022, doi: <https://doi.org/10.48550/arXiv.2207.08815>.
- [27] G. Somepalli, M. Goldblum, A. Schwarzschild, and others, "SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training," *arXiv preprint*, 2021, doi: <https://doi.org/10.48550/arXiv.2106.01342>.
- [28] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986, doi: <https://doi.org/10.1038/323533a0>.
- [29] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *International Conference on Machine Learning*, 2010. doi: <https://dl.acm.org/doi/10.5555/3104322.3104425>.
- [30] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, 2015, doi: <https://doi.org/10.1371/journal.pone.0118432>.
- [31] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, 2006, pp. 233–240. doi: <https://doi.org/10.1145/1143844.1143874>.
- [32] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: <https://doi.org/10.1109/TKDE.2008.239>.
- [33] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: <https://doi.org/10.1016/j.patrec.2005.10.010>.
- [34] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines," in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021, doi: <https://doi.org/10.48550/arXiv.2006.04884>.
- [35] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, 2023, doi: <https://doi.org/10.1016/j.patter.2023.100804>.