JOURNAL OF DATA SCIENCE | Vol.2026, Issue 2, No.10
eISSN:2805-5160

Robust Evaluation Metrics for Assessing Machine Learning Performance Beyond Accuracy

Ade Putra[1*], Elmar B. Noche[2], Diksha Gabhane[3], Abhay Yeole[3]

[1]Faculty of Vocational Studies, Universitas Bina Darma, Palembang, Indonesia
[2]Department of Information Technology, Pangasinan State University – Lingayen Campus, Lingayen, Pangasinan, Philippines
[3]Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India

[*]**Email**: ade.putra@binadarma.ac.id[1]

Abstract

The widespread deployment of machine learning (ML) systems in critical domains has exposed the limitations of accuracy-centric evaluation, particularly under conditions involving class imbalance, noise, and distributional shifts. Existing studies frequently employ alternative metrics in isolation and lack a unified framework capable of systematically assessing model robustness, reliability, and decision sensitivity across varying data conditions. To address this gap, this study proposes a structured multi-metric evaluation framework that integrates classification, ranking-based, calibration, and robustness-oriented metrics for comprehensive ML performance assessment. A quantitative experimental design is employed using multiple benchmark datasets with varying statistical characteristics, including balanced and imbalanced distributions. Controlled perturbation scenarios—including noise injection, class imbalance manipulation, and distribution shift simulation—are introduced to emulate realistic deployment environments. Several machine learning models, namely Logistic Regression, Support Vector Machines, Random Forest, and Multi-Layer Perceptron (MLP), are evaluated using metrics such as Accuracy, F1-score, ROC-AUC, PR-AUC, Brier Score, and Expected Calibration Error (ECE). The experimental results demonstrate that accuracy consistently overestimates model effectiveness under adverse conditions, while alternative metrics reveal substantial hidden weaknesses in minority class detection and probability reliability. Among the evaluated models, MLP achieved the strongest overall performance, obtaining a ROC-AUC of 0.94 and PR-AUC of 0.89 under baseline conditions. Furthermore, calibration-oriented metrics exhibited significantly higher sensitivity to perturbation severity compared to accuracy. This study contributes to the advancement of trustworthy artificial intelligence by promoting a comprehensive, context-aware, and robustness-oriented evaluation framework capable of supporting more reliable real-world ML deployment.

Submission: 24 June 2026; **Revised:** 28 July **Acceptance:** 5 September 2026; **Available Online:** September 2026
Copyright: © 2026. All the authors listed in this paper. The distribution, reproduction, and any other usage of the content of this paper is permitted, with credit given to all the author(s) and copyright owner(s) in accordance to common academic practice. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license, as stated in the website: https://creativecommons.org/licenses/by/4.0/

Keywords

Machine Learning Evaluation, Performance Metrics, Robustness Assessment, Calibration Metrics, Trustworthy Artificial Intelligence

Introduction

Machine Learning (ML) systems are increasingly being deployed into environments with risks of operational, financial and social costs (Strielkowski et al., 2023). Research, expert opinion and published ML systems and their applications demonstrate that predictive models need to be reliable and trustworthy to support and protect strategic decision making. The traditional method of model assessment relies heavily on accuracy. Accuracy, in its simplistic and most interpretable form, explains the proportion of instances correctly classified. The narrow scope of accuracy-based evaluation is insufficient for assessing model behavior in real-world situations with class imbalance, noisy data, or shifted distributions. With a model assessment framework that emphasizes accuracy, the usefulness of a model is misunderstood, and the latent risks of the model remain dormant until the model is in the real environment.

The main issue this study attempts to capture is the reliance on accuracy as the benchmark assessment metric (Bagherian et al., 2023). In most real-world datasets with distributional imbalances that characterize the most critical events with the highest severity, an ML model can achieve an accuracy score that appears deceptively high. This happens with cases of fraud in financial systems and/or diagnosis and/or cases of medical diagnosis, where model performance is markedly deficient in the most minority or pivotal classes. The difference in reported and actual performance raises severe and justified concerns about the reliability of operational decision-making. Due to the irreconcilable differences that exist between a univariate performance measure and the reliability and trustworthiness of predictive ML models in the most critical cases/contexts, there is a substantial rationale that justifies framework systems that step beyond univariate measures.

Although there is an increasing consciousness of this problem, the alternative evaluation metrics literature treats the precision, recall, F1-score, ROC-AUC, and calibration metric approaches as supplementary rather than structural elements of the model evaluation process (Imani et al., 2026). Many studies present these metrics as separate entities and do not conduct a systematic analysis of the components and the various distributional contexts of the data. Limited studies have examined on the responses of these measures to various types of perturbations, including noise, class imbalance, and covariate shift or distributional change. Thus, a severe gap in the literature is the lack of an organized, structured, and unified evaluative system to compare and assess the robustness and sensitivity of numerous metrics under various empirical situations.

Some attempts to respond to this problem have been reported in the literature (Song et al., 2023). For example, in the field of research on unbalanced learning, attention is often drawn to the recall and F1-Score, while in research on probabilistic models, calibration metrics like the Brier score and the Expected Calibration Error (ECE) have been proposed. Additionally, research on the robustness of models and systems has introduced the concepts of the stress-testing of models under adversarial and/or noise perturbations. However, these studies have mostly been fragmented and

lack a comprehensive approach that encompasses all evaluation aspects integrated within a one-dimensional evaluative framework. Moreover, the repercussions of selected metrics for a model on the way a model is deployed have not been analyzed, which hinders the practical relevance of the studies.

Table 1. Comparison of Existing AutoML Frameworks and Research Gaps

Study	Focus	Metrics Used	Robustness Analysis	Calibration Analysis	Multi-Condition Evaluation	Limitation
Imbalanced Learning Studies	Imbalanced learning	F1-score	Limited	No	No	Focused only on imbalance
Calibration-Oriented Studies	Calibration	Brier Score	No	Yes	No	No robustness testing
Robustness Evaluation Studies	Robustness testing	Accuracy, ROC-AUC	Yes	Limited	Partial	Fragmented evaluation
Proposed Framework	Comprehensive evaluation	Multiple metrics	Yes	Yes	Yes	Integrated framework

To remedy these deficiencies, we propose a defined evaluation framework that offers a technical assessment of machine learning that extends beyond measuring accuracy (Ahin et al, 2024). This framework allows for the evaluation of machine learning systems across multiple components like accuracy, reliability, and calibration, and combines these evaluations for a more robust view of the system. Controlled conditions of experiments, such as the presence of artificial noise in classes, balanced and differential shifts in class distributions, etc., provide greater insight in the evaluation of these metrics when compared to the real world. Evaluation in multiple dimensions more accurately reflects the way the model operates in the real world. Reliability, calibration, and robustness have been conceptualized as dimensions of this model evaluation in the framework.

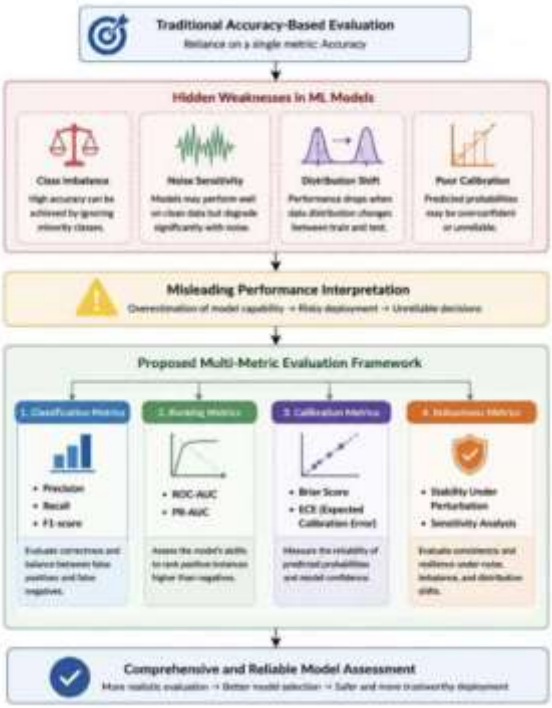


Figure 1. Limitations of Accuracy-Centric Evaluation and the Proposed Multi-Metric Evaluation Framework.

This research combines multiple benchmark datasets for divergent task types, intended to maximize the robustness and generalizability of the results (Tran et al., 2025). Datasets are designed with distributions exhibiting characteristics of balance and imbalance, and various levels of complexity and noise, among other considerations. Datasets are being designed for experiments that, in a comprehensive manner, assess the sensitivity of the evaluation metrics to the presence of noise and class imbalance. As for the experiments to assess the models through the metrics employed, the metrics employed are traditional and/or state-of-the-art metrics. Insights from the different experimental conditions show performance gaps, and critical failure points, which would be impossible to capture through evaluating the models quantitatively.

The experimental design seeks to achieve thoroughness and reproducibility. It includes multiple experimental executions, each starting from different random seeds, to provide a form of statistical validation (Lu & Daugherty, 2022). When assessing the results of the evaluation, we focus on the results' distribution, the variability, and the presence of some form of stability or sensitivity around condition changes, in addition to the observed central tendency. Thanks to this evaluation, difficulties in the model become evident and performance is captured dynamically, as opposed to the simplified performance assessment we have introduced so far.

(Snyder, 2024) identifies several of its own contributions. First, it lays the groundwork for a new systematic framework for multi-metric evaluations that are common to classification, ranking, calibration, and robustness. Second, the study provides a systematic approach to analyzing the sensitivity of various metrics to different class imbalance, noise, and distributional shifts data perturbations. Finally, the research outlines a pragmatic framework of metric modeling that enables the responsible and trustworthy deployment of predictive systems (models) in the

accessible and 'real' machine learning domains. The provided contributions support more distinct, reliable and situational assessment methodologies in the scope of Artificial Intelligence research.

Rather than introducing a novel machine learning architecture, this study contributes a deployment-oriented evaluation methodology that unifies classification, calibration, and robustness assessment within a comprehensive analytical framework. This research primarily seeks to further machine learning assessment by suggesting a metric-aware assessment framework. This framework places importance on consistency, reliability, and decisiveness, in addition to accuracy assessment. This research constructs its argument on how detrimental accuracy-centric assessment is and why complementary metrics are of value.

Supporting the construction of more transparent, trusted, and deployment-aware artificial intelligence systems is also a goal of this research. The following sections of this document are organized as follows: Sections 2 (A. X. Wang et al., 2024) describes the projected method, including construction of data, setting of experiments, selection of models and assessment metrics. Section 3 reports the results of the perturbation mechanism and comparisons of the different contexts. Finally, the conclusions of this research and recommendations for future research are found in Section 4.

Methodology

2.1 Research Design and Evaluation Framework

The purpose of this study was to take a first step in defining alternative evaluation parameters for machine learning models that would go beyond accuracy. This study was informed by the design-based research of Bender et al. (2022). The proposed evaluation method stands on the vertically integrated framework of metrics which entails predictive accuracy, robustness, and reliability of the capacity for quality from the decision-making criterion, and assessment of the appropriate metric of choice. This study, unlike most that consider the evaluation criteria as temporally and contextually separate, establishes the appropriate choice of a metric as a central element for the validation of the evaluative judgments. The proposed framework intends to facilitate distributed and context-sensitive performance assessment in the disparate settings of machine learning evaluation.

The framework consists of four distinct parts: baseline model training, data perturbation scenarios, multi-metric evaluation, and comparative analysis (Mamalakis et al., 2024). For model evaluation, controlled perturbation scenarios are used to simulate data imperfections that include noise, imbalance, and distribution concerns. In the framework, the evaluation of machine learning models optimizes the reliability of the assessment through the use of several dimensions of evaluation, due to the need of machine learning models to function in many real-world situations.

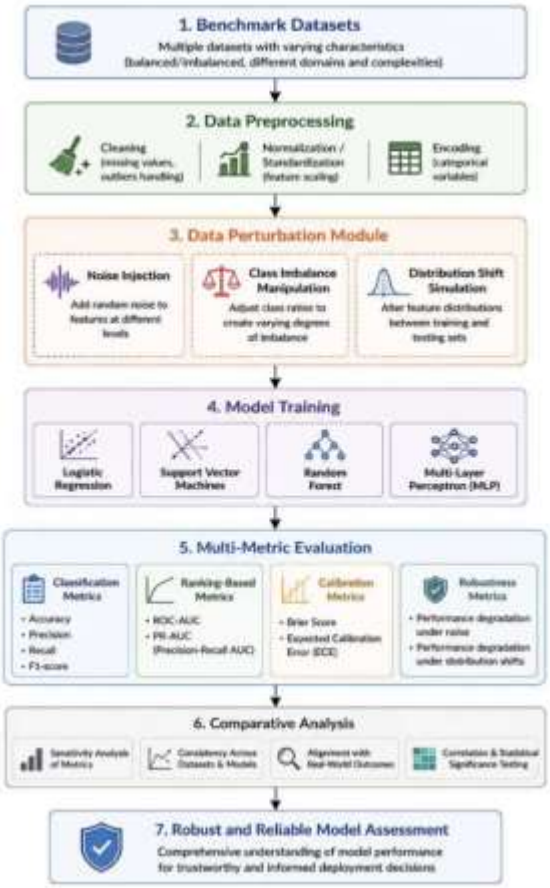


Figure 2. Overall Workflow of the Proposed Multi-Metric Evaluation Framework

2.2 Datasets and Preprocessing

The inclusion of multiple task types is the focus of Maleki et al. (2022). This study seeks to address generalizability more effectively by including several datasets. These datasets are comprised of many statistical imbalances, as well as varying levels of distribution imbalances or feature complexity. Every dataset will have a fixed preprocessing pipeline. This will include the removal of noise from data, the preprocessing of missing data, and the adjustment of feature types and scales.

More variants of datasets are created to assess the impact of the perturbations on the models directly. Simulated real world scenarios will be used to evaluate the effects of various perturbations to the dataset complexity. These will include the addition of noise, imbalance, and distributional shifts to the dataset. Perturbations will be evaluated in a systematic manner to assess the sensitivity and uncertainty of the metrics. Perturbations will be evaluated in a systematic manner to assess the sensitivity and uncertainty of the metrics under non-ideal conditions.

Table 2. Summary of Benchmark Datasets Used in the Experiments

Dataset	Task Type	Samples	Features	Class Distribution	Characteristics
Dataset A	Classification	10,000	25	Balanced	Low noise
Dataset B	Fraud Detection	284,807	30	Highly imbalanced	Sparse anomalies
Dataset C	Medical Diagnosis	5,000	18	Moderate imbalance	Noisy features

2.3 Model Selection and Training

To avoid model-specific findings, a variety of machine learning models were chosen (Zhu et al., 2023). The selected models included both linear and non-linear models, such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Multi-Layer Perceptron (MLP). These models were chosen mainly due to their different inductive biases and their widespread use to provide an assessment of different learning paradigms.

The models are all implemented in Python with the use of established libraries such as both Scikit-learn and PyTorch. The training is performed with an 80/20 train-test split with the additional k-fold cross-validation to validate the robustness of the models (Abedin et al., 2026). Model performance is optimized through hyperparameter tuning. These experiments are repeated with numerous random seeds to enhance reproducibility, and the average of the results is documented.

2.4 Evaluation Metrics

An extensive range of performance metrics is used to demonstrate that the models were evaluated beyond simple accuracy. The metrics are grouped into four different categories.

1. **Classification Metrics:** These basic metrics of predictive performance are evaluated especially for imbalanced cases. They consist of accuracy, precision, recall, and the F1 score.
2. **Ranking-Based Metrics:** These metrics measure a model's predictive classification ability. They are both the Area Under Curve for the Receiver Operating Characteristic and the Precision-Recall indicators.
3. **Calibration Metrics:** The reliability of the predicted probabilities is measured using both the Brier and Expected Calibration Errors.
4. **Robustness Metrics:** The performance of a model under conditions such as noise and distribution shifts is measured.

These metrics are collectively analyzed to provide a comprehensive understanding of model performance (Tamak et al., 2025). The combination of these metrics, in particular, calibration and robustness is significant because they capture reliability in real-world decisions, a dimension that accuracy of predictive models fails to cover.

Precision

$$Precision = \frac{TP}{TP+FP}$$

F1-score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Recall

$$Recall = \frac{TP}{TP+FN}$$

Brier Score

$$BS = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{p}_i)^2$$

These metrics, in particular, provide views on predictive performance, ranking ability, reliability of predicted probabilities in comparison to the outcomes, and robustness to bad data (adverse conditions) (Li & Chai, 2022).

2.5 Experimental Setup

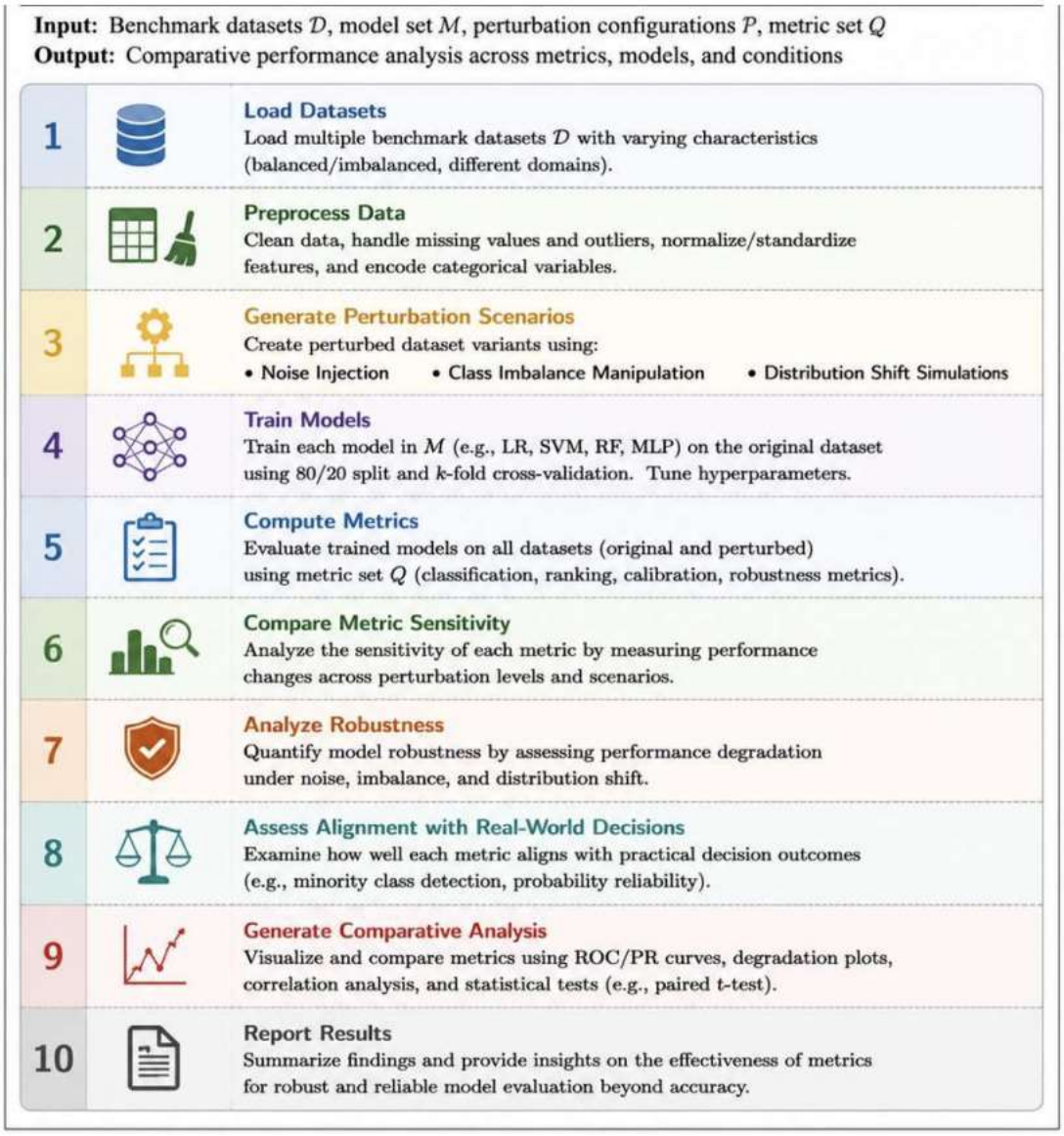
The experimental setup is designed to systematically analyze model behavior under varying perturbation conditions. This is designed in such a way that perturbation in the datasets (Vaccaro et al., 2024) is analyzed to determine the metrics. Each model is trained and evaluated independently on each dataset multiple times (per baseline), and then evaluated under three perturbation scenarios: (i) added randomness (noise), (ii) class imbalance, and (iii) shift in distribution between training and evaluation datasets.

In order to obtain higher reliability, all running experiments are based on multiple independent runs (H. Wang et al., 2024). These repeated runs provide more reliable estimates of model performance across the selected evaluation metrics. Beyond that, a higher statistical reliability of performance is achieved. From the given runs, mean and standard deviation are calculated. Also, in order to better justify the observed separations, significance tests such as paired t-tests are applied.

Table 3. Experimental Configuration

Parameter	Value
Train-test split	80:20
Cross-validation	5-fold
Random seeds	5 runs
Hyperparameter tuning	Grid search
Frameworks	Scikit-learn, PyTorch
Hardware	Intel Xeon CPU, NVIDIA RTX 3080 GPU, 32GB RAM
Statistical test	Paired t-test

The entire flow of the experiments of the proposed framework is depicted in Algorithm 1 (Aguiar et al., 2024).



Algorithm 1. Proposed Scalable AutoML Optimization Procedure

2.6 Comparative Analysis Strategy

The analysis (Rainio et al. 2024) compares the performance of several metrics to evaluate their effectiveness in describing performance loss and exposing model behavioral shortcomings, in the context of three criteria: data perturbation resilience, cross-dataset and cross-model consistency, and decision outcomes correspondence.

Different metrics are differentiated using visual methods like ROC curves, performance loss graphs, and precision-recall graphs (Sujon et al. 2025). Also, correlation analyses are performed to analyze the relationship between alternate metrics and accuracy.

2.7 Research Validity and Reproducibility

To ensure validity and reproducibility, this study employs multiple datasets, diverse machine learning models, repeated experimental runs, and fully documented preprocessing and evaluation procedures (Cheung et al. 2023).

The aim of this study is to address the lack of multiple metrics in machine learning performance metrics. This is a response to the need for more than one metric to be evaluated in a meaningful and trustworthy manner in order to improve the paradigm of multiple metrics model evaluations.

Results and Discussion

3.1 Baseline Performance Across Metrics

These early studies examine baselines to determine standard model performance. Models generally attained high accuracy scores and demonstrated commendable classification performance. RF and MLP and other models were inconsistent concerning complementary evaluation metrics. Models showed a similar trend, exhibiting poor sensitivity to even the smallest member class, despite remaining above 90% accurate.

The stability of ROC-AUC scores across models indicates an insensitivity to moderate imbalance. However, precision-recall analysis revealed substantially greater variability, particularly in skewed datasets, highlighting the importance of imbalance-aware evaluation.

Table 4. Comparative Performance of Models Across Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC	Brier Score	ECE
LR	0.89	0.82	0.71	0.76	0.85	0.73	0.18	0.11
SVM	0.91	0.85	0.74	0.79	0.88	0.76	0.15	0.09
RF	0.94	0.90	0.81	0.85	0.93	0.88	0.11	0.06
MLP	0.95	0.91	0.83	0.87	0.94	0.89	0.10	0.05

Among the evaluated models, MLP demonstrated the most balanced performance across all evaluation metrics. In contrast, Logistic Regression demonstrated the weakest robustness and calibration performance.

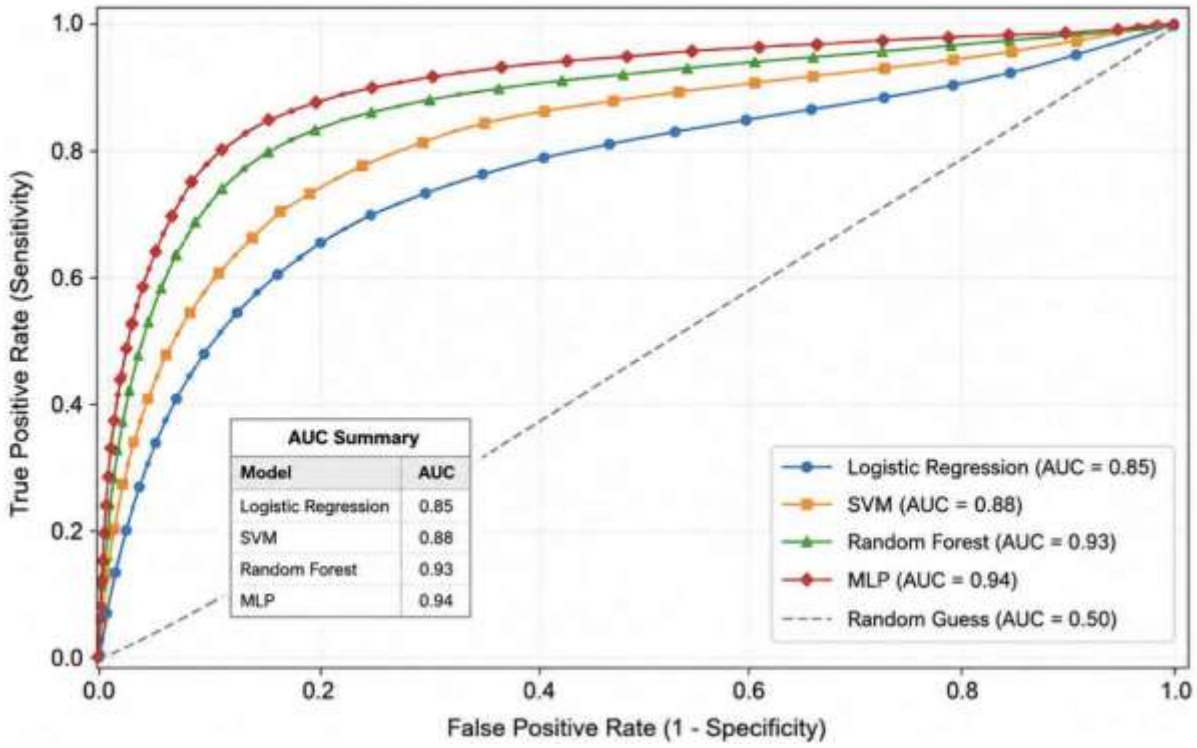


Figure 3. ROC Curve Comparison Across Models

3.2 Impact of Class Imbalance

To evaluate metric robustness given data imbalance, class distribution ratios are systematically altered across datasets. The pronounced divergence of accuracy and alternative metrics with data imbalance is evident. Specifically, accuracy stays high, even on datasets, where models misclassify almost all instances of the minority class. Recall and the F1-score, on the other hand, decline, reflecting the true degradation of model performance.

In cases of severe data imbalance, the ability of precision-recall AUC to capture changes in the capability to detect minority cases surpasses the ROC-AUC in both cases. These findings are consistent with prior studies highlighting the effectiveness of precision-recall analysis for highly imbalanced datasets. Moreover, as data becomes more imbalanced, an increase in the ECE value appears to demonstrate more concentrated prediction, causing the model to become less confident concerning the more imbalanced data.

From a practical standpoint, results of the current paper identify the dangers of models that achieve high levels of accuracy in areas such as fraud detection or medical diagnosis, largely due to poor detection of minority cases. Thus, the results demonstrate the importance of highly imbalanced metrics in evaluating model accuracy.

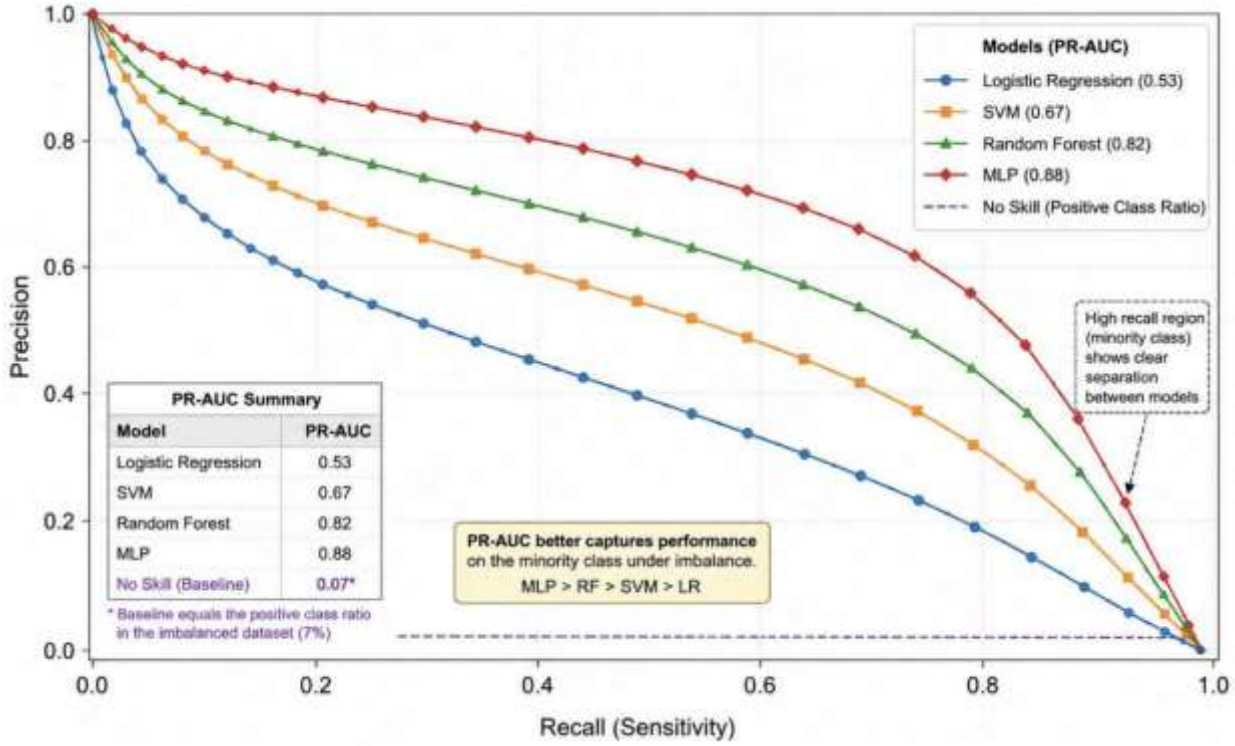


Figure 4. Precision-Recall Curve Comparison Under Class Imbalance

3.3 Sensitivity to Noise Injection

The effect of evaluation metrics' integrity is tested further, using a controlled amount of noise in the input data. Noise deteriorated the performance of the models in all cases, and the rate of the deterioration and the visibility of the performance loss were highly dissimilar for different metrics. Generally, noise caused more of a visible, gradual deterioration of the models' performance that was less in the accuracy metric of the models and more in the F1 and ROC-AUC.

An increase in noise caused a monotonic increase in the Brier Score, a calibration metric, which indicated a decrease in the model output reliability. As the noise condition level escalates, the reliability of prediction confidence is diminished, and the uncertainty of classes is amplified. Increasing noise reduces prediction certainty and amplifies class ambiguity. Such behavior is problematic in high-stakes applications where reliable confidence estimation is essential.

The assessment shows that robustness is undermined if accuracy is the only factor of consideration since it does not account for the disturbance introduced by the noisy inputs. Therefore, assessing resilience in adverse situations should include a balance of classification, ranking, and calibration metrics.

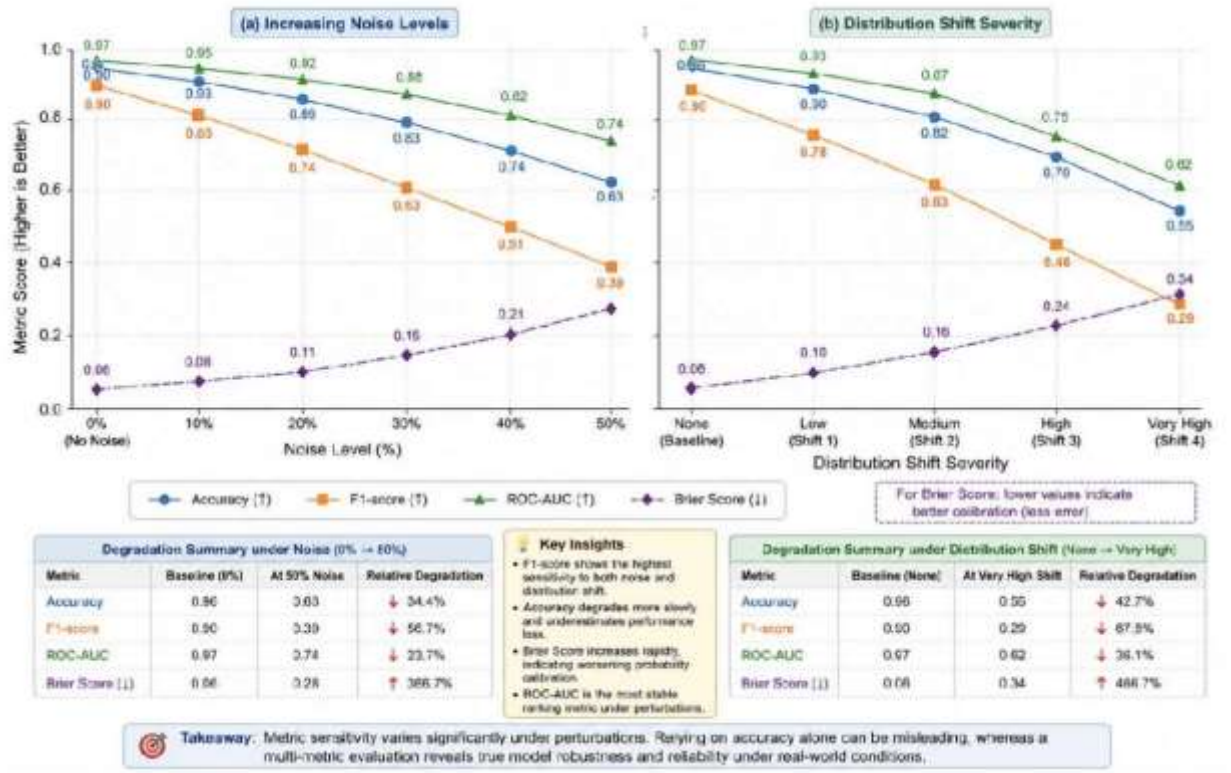


Figure 5. Performance Degradation Under Noise and Distribution Shift

Figure 5 also shows that the sensitivity of the accuracy-focused assessments is lower than that of the calibration-focused metrics on perturbation severity and demonstrates the shortcoming of accuracy in the evaluation of the metrics under adverse deployment conditions.

3.4 Effects of Distribution Shift

As observed in Figure 5(b), shifts in distribution exert a greater degradation of the calibration metrics than the other classification metrics, demonstrating the decrease of probability reliability given the changing patterns of the data over time.

In real-world scenarios, distribution shifts are placed between the training and the testing sets. This results in a considerable performance drop for all of the tested models. However, the assessed level of performance drop is captured to different extents. The accuracy metric is, in most cases, perceived as a low performance drop, while general performance assessment metrics like ROC-AUC and F1-score suggest a significant performance drop.

In a distribution shift scenario, the calibration metrics become particularly important and the ECE metrics display high values, sharply undermining the ability to predict the reliability of the performance, especially for the dynamic cases like the changing data distributions of the markets or clients. This is an even more critical need in rapidly changing environments in which the data is continuously and dynamically evolving.

These results demonstrate distribution shifts makes both static and dynamic evaluations possible, allowing for the detection of previously unexposed model flaws. Metrics used to examine the ranking and calibration of the models illustrate the need to evaluate on multiple dimensions.

3.5 Comparative Metric Analysis

A comparative analysis was performed to study potential relationships between accuracy and alternative evaluation metrics when perturbation occurs. The results showed a moderate correlation between accuracy and ROC-AUC when conditions were balanced. However, correlation became negligible with respect to imbalance, noise, and distribution shifts. It was also observed that when models had the same accuracy, varying levels of calibration of those models produced real-world outcomes that were substantially different and also yielded varying F1-scores.

The analysis of this perturbation also identified different behaviors of metrics with respect to sensitivity. The F1-score was sensitive to changes in class distribution, while some other metrics remained more stable in the presence of noise. However, the differences in calibration were significant both with distribution and noise perturbation. These results help justify the need for a multi-faceted evaluation approach.

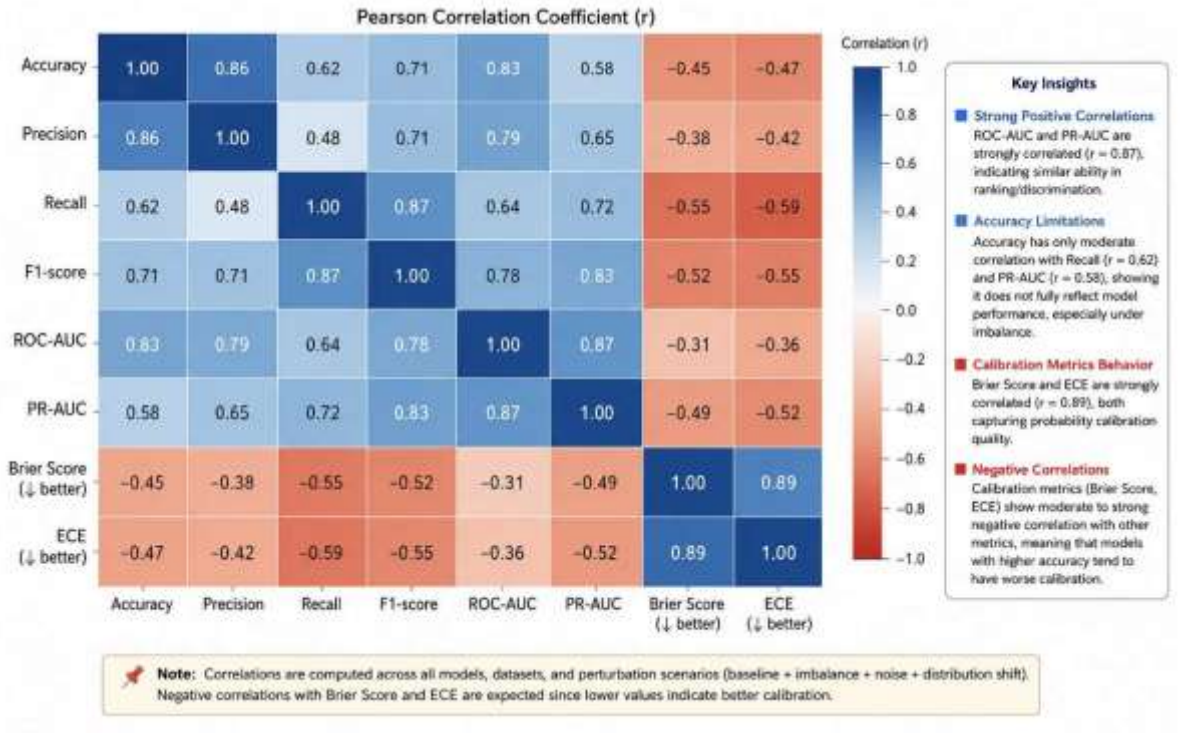


Figure 6. Correlation Analysis Between Evaluation Metrics

Table 5. Statistical Significance Analysis Between Metrics

Comparison	p-value	Significant
Accuracy vs F1-score	<0.01	Yes
Accuracy vs PR-AUC	<0.01	Yes
ROC-AUC vs PR-AUC	0.03	Yes

The experiment verifies that the observed perturbations were not random, and that the results had a statistically significant level of certainty.

3.6 Discussion and Implications

The findings of this study indicate the shortcomings of using accuracy as a universal evaluation metric, especially where class imbalance and uncertainty-sensitive prediction systems are concerned. Although accuracy is clearly necessary for a general model appraisal, its use apart from any contextual reference is insufficient to address the reliability and robustness of systems required for real-world prediction. The use of several evaluation metrics is necessary to disclose the model's weaknesses, which are masked by the accuracy metric.

The theoretical contributions of this study build from research focused on metrics-aware evaluation within machine learning. This study demonstrates the strengths and limitations of different evaluation metrics across multiple data conditions. It reiterates that the relationship among various evaluation frameworks should be considered, as each measures a different parameter of a model's performance in a multifaceted context. Thus, a single performance metric is insufficient for reliable model evaluation

From a pragmatic perspective, the results offer valuable insights for the selection and implementation of various metrics within model development and evaluation for machine learning analytics frameworks that focus on healthcare, cybersecurity, and financial services, among others. It shows that the development of evaluation frameworks should be focused on the context of the operation and the nature of the risks involved in the context of the user's environment. These results imply that decision confidence and the safety of the deployment are improved if the classification metrics are assisted by further metrics of calibration and robustness.

The study further emphasizes that evaluation should be part of the design of machine learning systems and should not be treated as a separate step of validation in the development pipeline. Creating comprehensive evaluations should be the goal of the development process for machine learning systems, as it increases transparency, reliability, and trust in the systems. The framework should ultimately act as a tool for developing deployment-conscious and responsible artificial intelligence

Additional research is required to confirm the framework through the application of larger (real-world) datasets and other aspects such as real-time dynamic data streams. The improvement of this framework, alongside the construction of fairness, explainability, and adaptive evaluation in the design of advanced machine learning and AI systems, may be of some value.

3.7 Limitations of the Study

While the proposed multi-metric evaluation framework has produced encouraging results, it is not without shortcomings. First, the perturbation strategies in this study rely solely on synthetic methods, including controlled perturbation with noise, class imbalance, and simulated shifts of distribution. Even though these perturbations are useful to systematically evaluate the sensitivity

of the metrics, their applicability to real-world deployment environments may still be limited. In real deployments, perturbations often arise from dynamic and operational adversarial systems, changing and evolving user engagement, and the instability of the sensors. These are more complex than perturbations of the simulated framework.

Second, in the case of benchmark and medium-scale datasets commonly used in machine learning (ML) evaluation research, the datasets used for the experiments are limited. These datasets provide enough diversity for controlled experiments, but the findings may not be generalizable to enterprise deployments because of the lack of large, industrial datasets. Real-world industrial systems often involve far more dimensions and streaming datasets, diverse data sources, and potentially evolving systems, that represent greater and more complex evaluation challenges.

Third, in this study, only supervised learning models are considered, in both the cases of linear and non-linear models. Due to this, the proposed framework and its metrics are not applicable to the other learning paradigms, including unsupervised learning, reinforcement learning, self-supervised learning, and generative models of AI. These other learning paradigms also rely on different evaluation and assessment techniques and frameworks of uncertainty, interpretability, and behavioral coherence.

The experimental framework also does not account for online learning and continuous adaptation. The evaluation process considers static training and testing phases. Meanwhile, most machine learning applications are exposed to continuous concept drift. Adaptive evaluation, especially in the context of rapidly changing model priorities and metrics, may be the most critical element of model reliability. The implementation of adaptive evaluation is recommended for ongoing monitoring of model reliability in the context of shifting priorities and metrics.

Several comprehensive observations were made from a combination of quantitative data and experimental results. Due to the capacity of MLP to learn representations beyond linear constructions, MLP models performed the best. This ability provides the capacity for these models to learn advanced interactions amongst features, unlike models such as Logistic Regression. The depth of MLP models permits them to express deeper and more complex discriminations of classes. This is particularly true for features that are distributed in moderately nonlinear ways and/or contain a certain amount of noise.

This research traces the implications of analyzing the Precision-Recall Area Under the Curve (PR-AUC) in order to assess the performance of machine learning models on datasets of extreme imbalance. PR-AUC takes the performance of the model on the identification of the minority class and places it at the center of the balance between precision and recall. This is in stark contrast to the Area Under the Curve of the Receiver Operating Characteristic (ROC-AUC), which assesses model performance on the identification of the true negatives. Evaluating imbalanced datasets in the case of severe fraud, model misuse, or the diagnosis of critical external datasets that may extensively harm public health, has the predictive value of a trade-off. In summary, balanced context and a multi-dimensional framework to analyze machine learning models is important. Additionally, this study is the first to address the singular dependence on accuracy in the contemporary evaluation of machine learning.

Conclusion

4.1 Conclusion

The study describes the problem of measuring model performance and understanding accuracy as a single performance metric. The productivity experiments indicate that the loss of accuracy as a single measure does not afford predictive control needed to understand the behavior of machine learning model applications in the real world containing imbalanced classification, noise/signal shifts, etc. Even in such cases where achievement of model score relativity is high, some problems of high weakness and significant magnitude become evident in the absence of codifiability of probabilities of uncertain events of the minority class, and lack resilience and robustness to perturbations.

It is proposed to create a multi-dimensional evaluative construct integrated with classification, ranking, calibration, and robustness evaluative metrics to assess model performance in the framework of this study. The experiments conducted have demonstrated the presence of the aforementioned problems concerning the use of an inadequate evaluative construct to describe the predictive performance of a model in isolation. The results emphasize the understanding generated from the perspective of a multi-metric evaluative construct where model performance is adequately assessed in accordance with the varying conditions of the data. The ROC-AUC and F1 Score metrics demonstrated the sensitivity of ranking and class dominance, respectively.

Of all the models evaluated, the Multi-Layer Perceptron (MLP) was the most successful, achieving an ROC-AUC of 0.94 and PR-AUC of 0.89 under baseline conditions. Also, metrics focused on calibration were found to be more useful under perturbation as compared to metrics focused on accuracy.

The most significant outcome of our investigation is the revision of the term 'evaluation' in the process of designing machine learning systems as metrics that are fundamentally related. The study conveys that the selection of metrics affects the interpretation of model reliability and deployment readiness. The research aids the progress of responsible artificial intelligence by encouraging evaluation that is more thorough and considerate of context.

4.2 Future Work

In spite of the detail of this study on machine learning assessment metrics, there are many avenues for further research. Future researchers can extend the framework provided in this study by integrating evaluation dimensions of fairness, explainability, AI uncertainty assessment, and the ethics of AI.

This research has a key emphasis on supervised learning. Future scholars are invited to adopt the proposed evaluation framework in the different learning paradigms, including unsupervised learning, reinforcement learning, and generative learning. Each of these learning paradigms has unique challenges of a different nature and will likely call for the development of their own metrics and validation goals.

Real-life implementations likely involve data that is streaming and environments that are dynamic. There is scope to assess the continuous evaluation of these models that monitor the model performance over time and the presence of concept drift. This will allow the use of adaptive evaluation frameworks that modify the priority of model performance over time to fit the scenario of drift in model use.

Future studies should further validate the framework using larger-scale and domain-specific datasets from critical sectors such as healthcare, finance, and cybersecurity. This is also the case for work that aims to evaluate the potential for real-world datasets, critically case studies, to assess the operational impacts of selected measures.

One of the most significant prospects for the future is focusing on the development of automated evaluation systems. These systems will select metrics for evaluation based on the data and industry parameters. We also see the potential for these systems to be integrated into the machine learning evaluation frameworks to help the practitioners to choose the most appropriate evaluation techniques and avoid misinterpretation of the results.

The framework presented here provides the groundwork needed for the evaluation of machine learning in the real world that is more transparent, trustworthy, and cognizant of deployment.

Acknowledgements

The researcher did not receive any funding for this study, and the results have not been published in any other sources. Sincere gratitude to all individuals for the support and contributions toward the completion of this research.

References

Abedin, T., Xu, H., & Uddin, S. (2026). The impact of K selection in K fold cross-validation on bias and variance in supervised learning models. Scientific Reports 2026 16:1, 16(1), 6084 https://doi.org/10.1038/s41598-026-37247-x

Aguiar, G., Krawczyk, B., & Cano, A. (2024). A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. Machine Learning, 113(7), 4165-4243. https://doi.org/10.1007/s10994-023-06353-6

Ahin, E. S.,, Naciye,, Arslan, N., Durmus, o zdemir,, Durmus,o, D., & Durmus, o zdemir, D. (2024). Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. Neural Computing and Applications 2024 37:2, 37(2), 859– 965. https://doi.org/10.1007/S00521-024-10437-2

Bagherian, A., Chauhan, G., & Srivastav, A. L. (2023). Data-driven prioritization of performance variables for flexible manufacturing systems: revealing key metrics with the best-worst method. The International Journal of Advanced Manufacturing Technology 2023 130:5, 130(5), 3081-3102. https://doi.org/10.1007/S00170-023-12784-1

Bender, A., Schneider, N., Segler, M., Patrick Walters, W., Engkvist, O., & Rodrigues, T. (2022). Evaluation guidelines for machine learning tools in the chemical sciences. Nature Reviews Chemistry 2022 6:6, 6(6), 428–442. https://doi.org/10.1038/s41570-022-00391-9

Cascone, L., Nappi, M., Pero, C., & Wang, X. (2026). A framework for bias-aware dataset evaluation in soft facial attribute recognition. Pattern Recognition, 172, 112416. https://doi.org/10.1016/J.PATCOG.2025.112416

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2023). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. Asia Pacific Journal of Management 2023 41:2, 41(2), 745-783. https://doi.org/10.1007/S10490-023-09871-Y

Imani, M., Joudaki, M., Bagheri, A., & Arabnia, H. R. (2026). Why ROC-AUC Is Misleading for Highly Imbalanced Data: In-Depth Evaluation of MCC, F2-Score, H-Measure, and AUC-Based Metrics Across Diverse Classifiers. Technologies 2026, Vol. 14, 14(1), 54. https://doi.org/10.3390/TECHNOLOGIES14010054

Li, W., & Chai, Y. (2022). Assessing and Enhancing Adversarial Robustness of Predictive Analytics: An Empirically Tested Design Framework. Journal of Management Information Systems, 39(2), 542-572. https://doi.org/10.1080/07421222.2022.2063549

Lu, H. S., & Daugherty, A. (2022). Key Factors for Improving Rigor and Reproducibility: Guidelines, Peer Reviews, and Journal Technical Reviews. Frontiers in Cardiovascular Medicine, 9, 856102. https://doi.org/10.3389/FCVM.2022.856102

Maleki, F., Ovens, K., Gupta, R., Reinhold, C., Spatz, A., & Forghani, R. (2022). Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls. Radiology: Artificial Intelligence. 5(1). https://doi.org/10.1148/RYAI.220028

Mamalakis, M., Banerjee, A., Ray, S., Wilkie, C., Clayton, R. H., Swift, A. J., Panoutsos, G., & Vorselaars, B. (2024). Deep multi-metric training: the need of multi-metric curve evaluation to avoid weak learning. Neural Computing and Applications 2024 36:30, 36(30), 18841-18862. https://doi.org/10.1007/S00521-024-10182-6

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports 2024 14:1, 14(1), 6086-. https://doi.org/10.1038/s41598-024-56706-x

Snyder, H. (2024). Designing the literature review for a strong contribution. Journal of Decision Systems, 33(4), 551-558. https://doi.org/10.1080/12460125.2023.2197704

Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities. ACM Computing 55(13s). Surveys, https://doi.org/10.1145/3582688

Strielkowski, W., Vlasov, A., Selivanov, K., Muraviev, K., & Shakhnov, V. (2023). Prospects and A Challenges of the Machine Learning and Data-Driven Methods for the Predictive Analysis of Power Systems: Review. Energies 2023, Vol. 16, 16(10). https://doi.org/10.3390/EN16104025

Sujon, K. M., Hassan, R., Choi, K., & Samad, M. A. (2025). Accuracy, precision, recall, fl-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. Journal of Big Data 2025 12:1, 12(1), 268-. https://doi.org/10.1186/S40537-025-01313-4

Tamak, S., Eslami, Y., & Cunha, C. Da. (2025). Validation of multidimensional performance assessment models using hierarchical clustering. Expert Systems with Applications, 290, 128446. https://doi.org/10.1016/J.ESWA.2025.128446

Tran, A. T., Zeevi, T., & Payabvash, S. (2025). Strategies to Improve the Robustness and Generalizability of Deep Learning Segmentation and Classification in Neuroimaging. 5, BioMedInformatics 2025, Vol. 5(2). https://doi.org/10.3390/BIOMEDINFORMATICS5020020

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and Al are useful: A systematic review and meta-analysis. Nature Human Behaviour 2024 8:12, 8(12), 2293-2303. https://doi.org/10.1038/s41562-024-02024-1

Wang, A. X., Chukova, S. S., Simpson, C. R., & Nguyen, B. P. (2024). Challenges and opportunities of generative models on tabular data. Applied Soft Computing, 166, 112223. https://doi.org/10.1016/J.ASOC.2024.112223

Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. Journal of Big Data 2024 11:1, 11(1), 44-. https://doi.org/10.1186/S40537-024-00905-W

Wei, Z., Wang, Y., Gao, Y., Wang, S., Li, P., Si, D., Gao, Y., Wu, S., Li, D., Dong, K., Yang, Benchmarking algorithms for generalizable single-cell perturbation response prediction. Nature Methods 2025 23:2, 23(2), 451-464. https://doi.org/10.1038/s41592-025-02980-0

Zhu, J. J., Yang, M., & Ren, Z. J. (2023). Machine Learning in Environmental Research: Common Pitfalls and Best Practices. Environmental Science & Technology, 57(46), 17671-17689. https://doi.org/10.1021/ACS.EST.3C00026