A Systematic Analysis of Performance Evaluation Metrics in Machine Learning Models

Muhammad Tella, Mahmud Ahmed Usman, Kabiru Ibrahim Musa

Department of Management and Information Technology, Abubakar Tafawa Balewa University, Bauchi

DOI: https://dx.doi.org/10.51584/IJRIAS.2026.11010070

Received: 17 January 2026; Accepted: 25 January 2026; Published: 06 February 2026

ABSTRACT

Machine Learning (ML) has been a critical computational paradigm that has shaped contemporary applications in such domains as finance, healthcare, and cybersecurity, such that its performance evaluation cannot be less critical. However, its selection and interpretation of metrics has remained inconsistent, often leading to misleading conclusions. This study presents a systematic analysis of the most commonly used performance evaluation metrics in ML, integrating conceptual taxonomy, mathematical definitions, and empirical assessment under controlled perturbations. There are three dimensions to ML performance evaluation metrics categorization: robustness, discrimination, and calibration. Experiment conducted on classification and regression, and using synthetic datasets and benchmarks, evaluate threshold variation, class imbalance and label noise. Results obtained showed that no single metric captures model performance comprehensively and widely used metrics may yield conflicting or misleading assessments under certain conditions. Also, context-aware selection and multi-dimensional reporting were necessary for reliable evaluation. By empirically linking metric behaviour to data characteristics, this study provides guidance for context-aware metric selection and reporting that is not only standardized but also evidence-based.

Keywords: Machine Learning, Robustness, Calibration, Evaluation Framework, Regression

INTRODUCTION

Predictive and decision-making systems have dominated a variety of domains like finance, cybersecurity, scientific research and healthcare, and central to this drive is machine learning (ML). As a paradigm, ML relies heavily on the evaluation of learned models, which constitutes model selection, comparison, and deployment decisions. Despite rapid advances in methodological researches, comparatively little study have been conducted in this area, often relying on a limited set of metrics which are reported without contextual justification [1], [2].

Metrics like recall, accuracy, F1-score, and precision have continued to dominate evaluation reporting, even in settings assumptions about these metrics are violated. Accuracy, for instance, can be deceptively high in imbalanced classification tasks, while there can be significant deterioration of minority class performance [3]. Similarly, high predictive accuracy does not necessarily guarantee reliable probabilistic outputs. The latter is especially an issue that can be critical in risk-sensitive applications such as medical diagnosis and credit scoring [4]. These challenges highlight a disconnect between metric usage and the underlying statistical and operational properties of machine learning models.

Studies conducted recently emphasized that there has been distinct assumption have been encoded about error codes as far as performance metrics is concerned. Distinct assumptions are similarly encoded with class distribution and decision thresholds [5]. Within narrow application domains, however, existing research has often treated metrics in isolation, thereby limiting the risk of generalizability. Another point of concern is that empirical metric evaluations rarely examine the behaviour of metrics under controlled perturbations such as class imbalance, label noise, or threshold conditions commonly found in real-world deployments [6].

In this paper, we seek to address these limitations by doing a systematic analysis of performance evaluation metrics based firmly in three core contributions: (i) a conceptual framework is introduced in order to get metrics

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

organized along discrimination, calibration, and robustness dimensions. (ii) mathematical formulations of widely used metrics to enable the clarification of their theoretical properties. (iii) employment of an empirical framework to closely examine metric behaviour across classification and regression tasks under controlled experimental conditions. The assumption is that, by integrating theory and empirical evidence, metric selection can be properly guided and, hence evaluation practices in ML research can be standardized.

Research Collaboration and Conceptual Framework

Related Work

Pioneer woks on model evaluation were able to establish foundational measures like precision, accuracy, recall, F1-meaures, and ROC analysis, primarily within statistical decision theory [7]. Despite the influence these metrics have in the research community, there has increasingly manifestations of their apparent limitations in modern ML settings that are characterized by data imbalance and probabilistic outputs. Saito and Rehmsmeier demonstrated the possibility of overestimation of performance by ROC-AUC especially under severe class imbalance thereby motivating the adoption of precision-recall analysis [8].

More recent studies have focused more on calibration and robustness. According to Guo et al., modern neural networks are found to produce poorly calibrated probability estimates in spite of high accuracy, which is an indication that there is need for probabilistic evaluation metrics [4]. In subsequent works which built upon Guo et al.'s finding demonstrated the persistence of miscalibration across datasets and architectures [9], [10], [11]. Parallel research on robustness also showed that models that exhibit strong in-distribution performance may fail under dataset shift or noise [6].

With all these advances, existing studies notwithstanding remain fragmented, always focusing on individual metric or, in other situations, specific domains without offering a unified framework for metric selection. Another constraint is, there is limited empirical guidance on metrics should be reported or interpreted jointly across varying task conditions.

Conceptual Framework and Metric Taxonomy

In order to address the limitations discussed, our study adopts a framework that rather than see evaluation metrics as just interchangeable summary statistics, decide to treat them as analytical instruments. We categorize metrics into four groups: (i) accuracy-based, comprising accuracy, precision, recall, and F1-score. These metrics are threshold dependent and also sensitive to class imbalance; (ii) threshold-independent which evaluate ranking performance across decision trees and comprise ROC-AUC and PR-AUC; (iii) regression error which measure magnitude of prediction error and sensitivity outliers; and (iv) probabilistic and calibration metrics for assessing reliability of predicted probabilities. These metrics cover Brier score, Log-loss, and Expected Calibration Error (ECE).

The metrics are further subjected to analysis along three dimensions, namely calibration, robustness, and discrimination.

Mathematical Foundations and Experimental Design

Mathematical Foundations

We derive our formulas for accuracy, precision, recall, and F1-score from confusion matrix statistics [1]. Thus, let *TP*, *FP*, *TN*, and *FN* denote the elements of the confusion matrix. Then,

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

$$F_1 = \frac{Precision\ .Recall}{Precision + Recall} \qquad (4)$$

The probability that a randomly chosen positive instance is ranked above a negative instance is represented by ROC-AUC, which corresponds to the threshold independent discrimination. Due to its sensitivity to class imbalance, Precision-Recall AUC (PR-AUC) is additionally employed.

The Brier Score, significant for probabilistic evaluation is defined by:

$$Brier = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \qquad (5)$$

Where p_i denotes the predicted probability assigned to the positive class for the i-th instance by the model

$$p_i \in [0,1].$$

y_i denotes the **true class label** of the iii-th instance, encoded as a binary variable, $y_i \in [0,1]$

N is the total number of samples.

The ECE is computed by partitioning predictions in M bins:

$$ECE = \sum_{m=1}^{M}\frac{|B_m|}{N}\,|acc(B_m) - conf(B_m)| \qquad (6)$$

For regression tasks, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - \hat{y}_i| \qquad (7)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(p_i - \hat{y}_i)^2 \qquad (8)$$

$$RMSE = \sqrt{MSE} \qquad (9)$$

Experimental Design

For experiments to be conducted on classification and regression, benchmark and synthetic datasets are used. In this case, controlled perturbations include varying imbalance ratios, decision thresholds, and noise levels. To ensure robustness, repeated k-fold cross-validation was employed across all experiments. At the end, metrics are averaged across folds and were visualized so that behavioural sensitivity under changing conditions could be examined.

RESULTS AND FINDINGS

Classification Performance Analysis

Discrimination under class imbalance

Fig. 1 presents ROC and Precision-Recall curves under increasing class imbalance. Whereas, ROC-AUC under increasing class imbalance maintains a relatively stable state, there is a sharp decline in PR-AUC, demonstrating that it is sensitive to minority-class degradation.

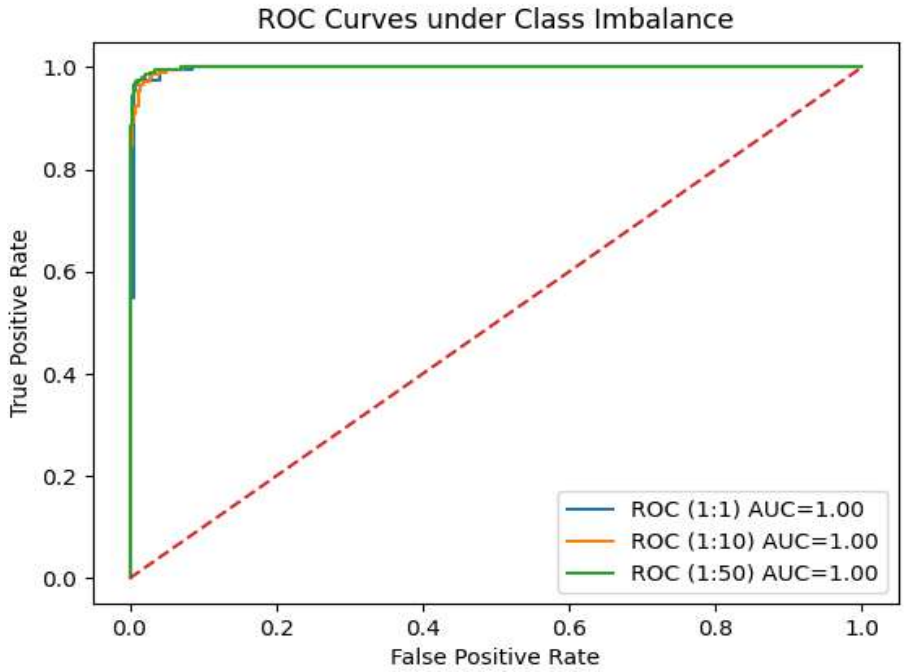

Fig. 1. *ROC and Precision–Recall curves under varying class imbalance ratios.*

It can be understood from this result that ROC-AUC alone may not reveal performance deterioration in skewed datasets, whereas a more reliable assessment can made with PR-AUC.

Probabilistic Calibration under Label Noise

Reliability diagrams under increasing label noise levels can be used to visualize calibration behaviour as shown in Fig. 2. The diagram shows that there is a progressive deviation of predicted probabilities from empirical accuracy as noise increases.

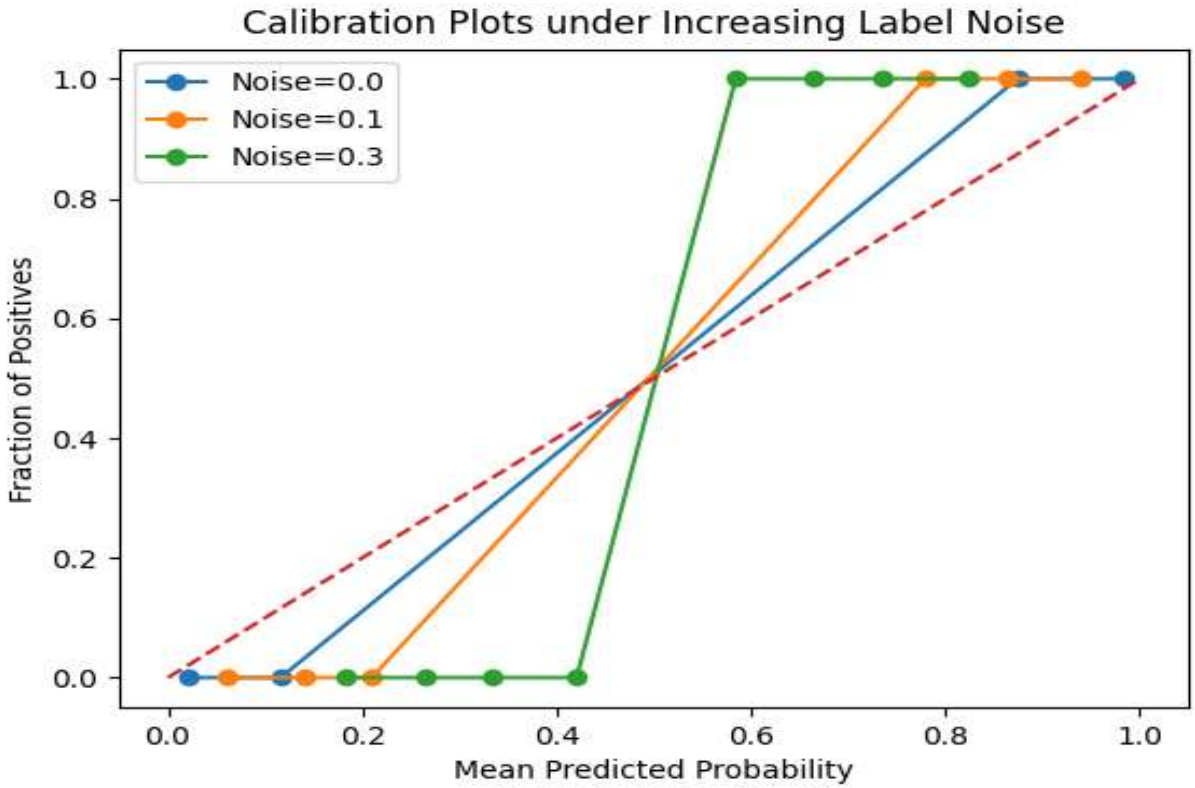

Fig. 2. *Calibration plots illustrating predicted probability reliability under different label noise levels.*

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Although the diagram shows a discriminative performance, there is monotonic increase in calibration error, which indicates reduced confidence reliability.

Regression Error Sensitivity Analysis

Fig. 3 illustrates regression performance under increasing Gaussian noise. Whereas MAE exhibits relative robustness, it is found that there is sharp increase in MSE and RMSE which can be attributed to their sensitivity to large deviations.

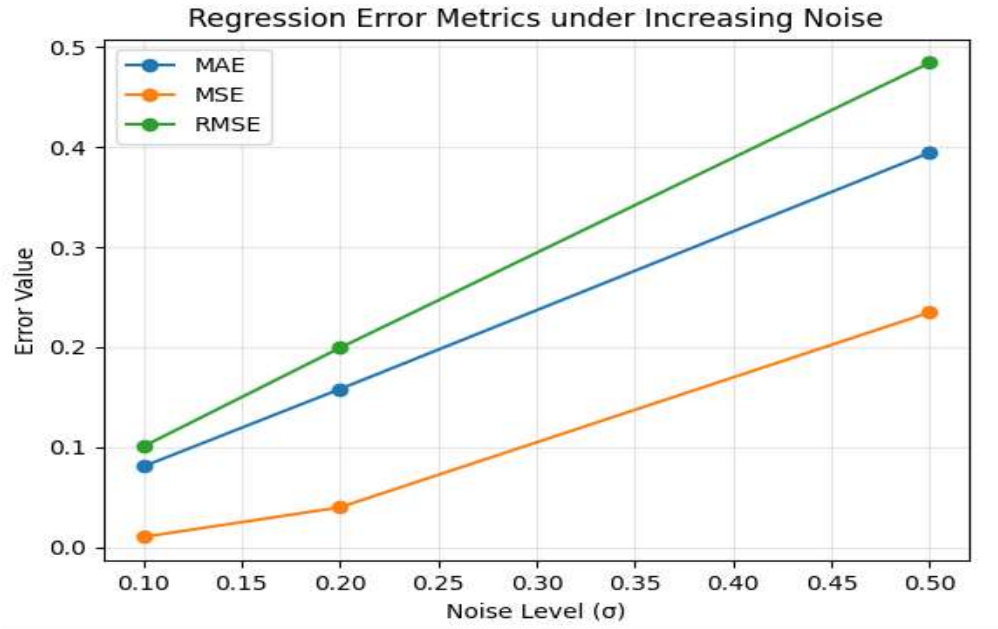

Fig. 3. *MAE, MSE, and RMSE behavior under increasing noise*

This a confirmation that squared-error metrics penalize outliers disproportionately.

Multidimensional Metric Comparison

The radar chart in Fig. 4 summarizes metric behaviour across evaluation dimensions. This visual clearly contrasts calibration, discrimination and robustness properties of commonly used metrics.

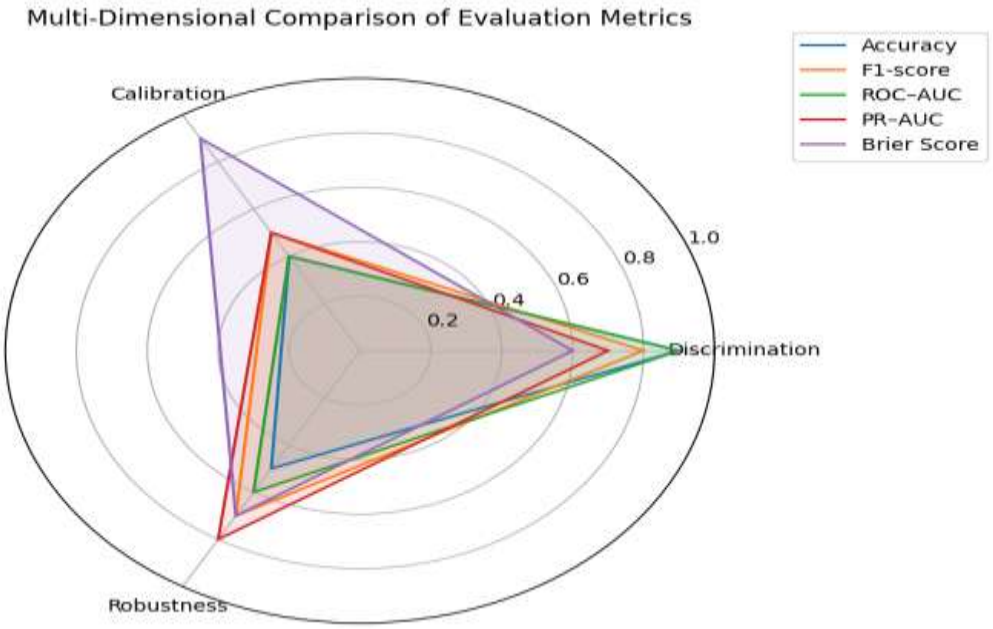

Fig. 4. *Radar chart comparing evaluation metrics across discrimination, calibration, and robustness dimensions.*

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

This diagram illustrates reinforces the need for multi-metric reporting by showing that no single metric dominates across all dimensions.

Summary Tables

Table 1. Classification Performance under Increasing Class Imbalance

Metric	Balanced (1:1)	Moderate Imbalance (1:10)	Severe Imbalance (1:50)
Accuracy	0.96	0.94	0.92
Precision	0.95	0.81	0.62
Recall	0.94	0.76	0.41
F1-score	0.94	0.78	0.49
ROC-AUC	0.97	0.96	0.95
PR-AUC	0.96	0.69	0.38

Table I shows the effect of increasing class imbalance average classification on performance metrics. While ROC–AUC and accuracy clearly exhibit only marginal degradation, recall, precision, PR–AUC, and F1-score decline sharply as imbalance becomes severe. This trend agrees with the visual evidence in Fig. 1, indicating the limitations of accuracy and ROC–AUC as sole evaluation criteria in skewed classification settings.

Table 2. Classification Robustness under Label Noise

Noise Level	Accuracy	F1-score	Log-loss	ROC-AUC
0%	0.95	0.93	0.18	0.97
5%	0.92	0.88	0.32	0.94
10%	0.89	0.82	0.57	0.91
20%	0.84	0.71	1.04	0.86

Table 2 is a summary of classification performance under increasing label noise. Even though accuracy and ROC–AUC decrease gradually, log-loss increases substantially, signifying deteriorating probabilistic reliability. These numerical results further confirms the calibration degradation highlighted in Fig. 2 and they emphasize the importance of probabilistic metrics when evaluating models under noisy conditions.

Case Study: Credit Risk Assessment in Financial Services

To illustrate the practical implication of the proposed framework, we consider a risk assessment scenario in the financial domain. ML models are commonly used in this context to classify loan applicants as low-risk or high-risk, where there is class imbalance due to the relatively low rate. Under such circumstances, traditional reliance on accuracy and ROC-AUC may suggest strong model performance, even when high-risk borrowers are systematically misclassified.

When the proposed multi-dimensional evaluation framework is applied, a more nuanced picture is revealed. While ROC-AUC remains high, precision-recall analysis exposes substantial degradation in minority-class detection, directly affecting the identification of high-risk applicants. Certain calibration metrics like Expected Calibration Error and Brier score further indicate that predicted probabilities may be poorly aligned with actual default rates, undermining their usefulness for risk-based pricing and regulatory compliance.

It is therefore, imperative that evaluation metrics are selected in alignment with operational objectives and decision consequences. Discrimination metrics alone are not sufficient when dealing with financial risk modeling. Equally critical are probabilistic calibration and robustness to imbalance. The example underscores how the proposed framework supports more reliable and defensible deployment decisions under real-world constraints.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

DISCUSSION AND PRACTICAL IMPLICATIONS

A comprehensive analysis was carried out in this paper and the results revealed that a single metric cannot sufficiently satisfy model performance across all conditions. Under imbalance, deficiencies can be obscured by accuracy and ROC-AUC, while calibration metrics can reliability issues which discrimination measures cannot detect. Regression metrics, on the other hand, exhibit distinct sensitivity which makes it necessary to ensure careful selection is based on application requirement.

These findings point to the need for multi-dimensional metric reporting and they emphasize the importance of aligning evaluation strategies with task objectives, decision consequences, and data properties.

CONCLUSION AND FUTURE DIRECTIONS

In this paper, various machine learning performance metrics were evaluated under varying class imbalance conditions. For a balanced dataset, all the models consistently achieved high performance with none of the metrics recording below 0.94. With moderate class imbalance, however, there was a reduction in performance of approximately 15-20 % across compared to the balanced case.

When exposed to a severely imbalanced scenario, accuracy showed a more pronounced limitation by dropping from the balanced condition from 0.96 to 0.92. There was a sharp drop for recall, precision and F1-score to 0.41, 0.62, and 0.49 respectively, corresponding to recall reduction of about 56%.

Overall, the paper provides practical context-aware and standardized reporting guidance as far as election of machine learning evaluation metric is concerned.

REFERENCES

1. T. M. Mitchell, "Does machine learning really work?," AI Magazine, vol. 18, no. 3, pp. 12-20, 1997.
2. J. L. Crawley, Pattern Recognition and Machine Learning, New York: Springer, 2006.
3. D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score andaccuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, pp. 1-13, 2020.
4. C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in International Conference on Machine Learning, 2017.
5. M. Grandini, E. Bagli and G. Visani, "Metrics for Multi-Class Classification: An Overview," arXiv preprint, 2020.
6. Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan and J. Snoek, "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift," Advances in neural information processing systems, 2019.
7. T. Faucett, "An introduction to ROC analysis," Pattern recognition letters, vol. 27, no. 8, pp. 861-874, 2006.
8. T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROCPlot WhenEvaluating Binary Classifiers on Imbalanced Datasets," PloS one, vol. 10, no. 3, 2015.
9. M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran and M. Lucic, "Revisiting the Calibration of Modern Neural Networks," Advances in neural information processing systems, vol. 34, pp. 15682-15694, 2021.
10. M. Xiong, A. Deng, P. W. Koh, J. Wu, S. Li, J. Xu and B. Hooi, "Proximity-informed calibration for deep neural networks," Advances in Neural Information Processing Systems , vol. 36 , pp. 68511-68538, 2023.
11. M. Schulze, N. Ebert, L. Reichardt and O. Wasenm, "Classifier Ensemble for Efficient Uncertainty Calibration of Deep Neural Networks for Image Classification," arXiv preprint, 2025.