JOURNAL OF DATA SCIENCE | Vol.2026, Issue 1, No.3
eISSN:2805-5160

Benchmarking Robust Machine Learning Models Under Data Imperfections in Real-World Data Science Scenarios

Marlindawati[1*], Mohammad Azhar[2], Esha Sabir[3]

[1]Fakultas Vokasi Universitas Binadarma, Palembang, Indonesia
[2]Department of Applied Data Science, Hong Kong Shue Yan University, Hong Kong, SAR, China
[3]Department of Computer Science, University of Sahiwal, Sahiwal, Pakistan

Email: marlindawati@binadarma.ac.id[1*], azhar@hksyu.edu[2], eshasabir17@gmail.com[3]

Abstract

Machine learning systems deployed in real-world environments frequently encounter data imperfections such as noise, missing values, class imbalance, and distribution shifts. Despite substantial progress in model development, most evaluation protocols rely on clean benchmark datasets, creating a gap between laboratory performance and operational reliability. Existing robustness studies often focus on isolated perturbation types or single model families, lacking a unified benchmarking framework. This study proposes a structured and reproducible benchmarking methodology to systematically evaluate model robustness under controlled data degradation scenarios. Multiple classical machine learning algorithms and deep learning models were assessed across diverse benchmark datasets. Controlled perturbations—including feature noise, label corruption, missingness mechanisms, imbalance ratios, and covariate shifts—were introduced at progressive levels. Performance was evaluated using predictive metrics, robustness degradation rate (RDR), and computational efficiency, with statistical validation across repeated experimental runs. Results indicate that ensemble-based methods consistently achieved the strongest robustness, maintaining degradation rates below 10% under moderate noise and imbalance conditions. Deep neural networks demonstrated superior clean-data accuracy but experienced sharper degradation under structured corruption and distribution shifts. Mitigation strategies such as regularization and resampling reduced degradation by 5–12% under moderate perturbations but showed limited effectiveness under extreme conditions. The findings demonstrate that robustness is multidimensional and dependent on alignment between model inductive bias and data imperfection type. The proposed benchmarking framework provides practical guidance for selecting machine learning models suited to imperfect data environments, advancing reliable and deployment-ready AI systems.

Keywords

Robust Machine Learning; Data Quality; Benchmarking; Model Evaluation; Real-World Data

Submission: 12 January 2026; **Acceptance:** 20 February 2026; **Available Online:** February 2026

Introduction

In the last few years, we have seen the first attempts to implement Machine Learning (ML) systems into the real world, where data is often imperfect (Cabrera et al., 2025). In contrast to laboratory data sets, which can be manipulated to be ideal, real-world data faces a multitude of issues, including noise, missing data, class imbalance, measurement error, and distributional shifts. These issues can all impact a model's predictive performance, dependability, and safety. As the study points out, the imperfections seen in real-world data sets, including noise, missing data, and a multitude of others, create a significant barrier to the successful implementation of machine learning. As a result, the ability of machine learning systems to handle data in an imperfect state has garnered a significant amount of attention in the literature, particularly in areas of a critical nature such as healthcare, finance, cybersecurity, and industrial automation. In safety-critical areas such as medical diagnosis and financial risk modeling, the presence of imperfect data can result in a reduction in performance, as well as an increase in systemic operational risks and ethical issues associated with a model.

The primary challenge for evaluating machine learning models in a research scenario is a clean benchmark dataset and the algorithm's ability to assess the dataset in full variability (Ahangaran et al., 2024). A potential flaw in the algorithm is the unresolved data characteristics (i.e., the dataset is incomplete, unbalanced, or consists of excess noise). It is evident that if models are deployed in practical scenarios, the algorithm's estimation of performance will be too optimistic. This trickle-down effect exemplifies the severe complications that the ideal data characteristics present in a set and the practical application of the model in real scenarios. There are far too many unnerving issues plaguing practitioners when they are faced with imprecise, incomplete, or real-world data environments, which is exacerbated by the absence of thorough and methodical testing.

On the contrary, existing studies analyzing the effects of robustness in machine learning are scant (Uddin et al., 2025). A majority of studies analyzing the models or data sets are singular in nature, and this singular focus further limits the generalizability of the study. A considerable void or gap is easily observed in existing research to support a unified framework or arch to benchmark a variety of machine learning models when they are subjected to diverse imperfections of data across machine learning models (Azhar et al., 2025).

In order to fill this gap, the study at hand develops a detailed methodology for benchmarking and assessing the performance of the strongest machine learning models, using simulated control alterations of data as they would appear in practice (Park, 2025). The evaluation framework is designed to incorporate and assess many different factors beyond predictive accuracy, such as the rate of degradation of robustness and the degree of loss or cost of prediction. Such a multidimensional evaluation is critical in the case of the cited evaluation framework because the totality and the intricacy of the model framework and underlying data do not stand as a proxy for the degree of the model framework’s robustness. The methodology is designed to allow for perturbations to be incorporated in a systematic manner in order to evaluate and provide a rigorous framework for model performance in the presence of various conditions around the model.

The research employs widely accepted benchmark datasets for reproducibility and validity in reference to previous research (Sourlos et al., 2024). The research uses applied control perturbation methods to determine the standard practice for data gaps in order to evaluate data of consistent quality. The research aims to identify and explain the differences in data quality, as opposed to the differences in the available data sets, by evaluating/classifying various classical machine learning methods and deep learning models. The research aims to provide a framework to evaluate various paradigms of classical machine learning methods and deep learning models. The research aims to provide a framework to evaluate various paradigms of classical machine learning, complemented by rigorous evaluation of system effectiveness and loss. In addition to these, other frameworks must be used to evaluate the predictive system efficacy and the system's behavioral absorbing quality to achieve overall model effectiveness in predictive robustness and safety.

Perturbation-based retraining and benchmarking is performed using a clean data baseline and a systematic data perturbation and degradation training and retraining pipeline, ensuring data perturbation and baseline performance degradation benchmarking retains its methodological rigor, and more importantly, provides quantitative benchmarking of the model's adaptive system stability degradation in the presence of varying types of imperfections (Faddi et al., 2025). Also, the use and effectiveness of preprocessing and data augmentation techniques, which are commonly used and suggested for the purpose of minimizing the degradation of model performance, provide valuable insight for actual implementation. As far as we are aware, no previous study has (i) created a unified and cross-paradigm benchmarking framework, allowing for the evaluation and comparison of multiple model types against different, real-world scenarios of data imperfections, and (ii) established a standardized degradation framework for benchmarking models against different real-world scenarios of data imperfections and perturbations. This paper provides (i) a perturbation-based retraining and benchmarking methodology, (ii) a performance degradation and efficiency degradation performance measure cross-model comparison from the classical to the deep learning framework, and (iii) a validated and reproducible perturbation-based retraining and benchmarking methodology.

This research aims at two primary objectives. First, the research aims at establishing a reproducible benchmarking framework for assessing the robustness of machine learning models for realistic imperfections in data (Fabra-Boluda et al., 2024). The second is to create evidence that practitioners can use to make model choices for imperfect data. This study also considers the absence of robustness from a multidimensional perspective, as opposed to a singular view. The study highlights the need for examining the robustness of artificial intelligence systems in the face of dynamic and imperfect data. The framework of the evaluation is shown in Figure 1.

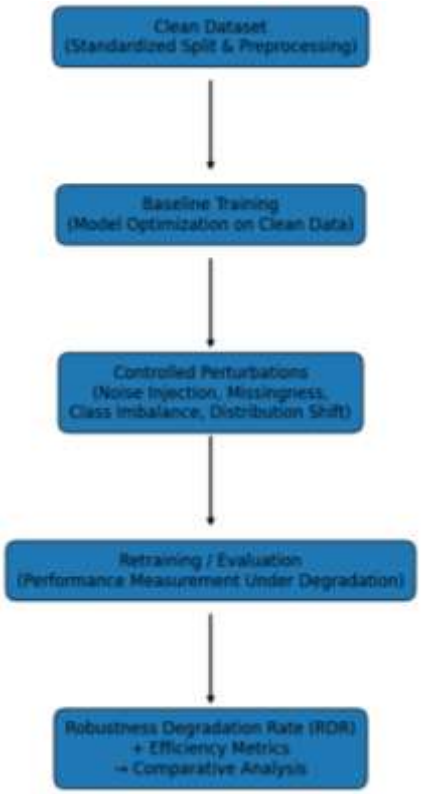

Figure 1. Unified Robustness Benchmarking Framework

The framework visually depicts the systematic evaluation pipeline that starts with the design of the dataset, model tuning to achieve a certain baseline performance, application of a defined set of controlled perturbations, model retraining under a variety of degraded conditions, and concludes with a model evaluation against the defined robustness dimensions (Schwabe et al., 2024).

Methodology

The systematic evaluation of data imperfections that were analyzed is identified and summarized in Table 1 (Orlu et al., 2023).

Table 1. Types of Data Imperfections Evaluated

Imperfection Type	Simulation Strategy	Theoretical Impact	Evaluation Focus
Feature Noise	Gaussian perturbation	Increases variance component	Stability under distortion
Label Noise	Random flipping	Induces memorization risk	Memorization resistance
Missing Data	MCAR / MAR	Reduces feature completeness	Imputation sensitivity
Class Imbalance	Ratio adjustment	Bias toward majority class	Minority recall

Distribution Shift	Covariate drift	Alters input distribution	Generalization stability

Research Design Overview

The author(s) of the aforementioned study devised a controlled experimental benchmarking framework to study real-world limitations caused by imperfections in data and the extent to which machine learning models can learn to 'overcome' these imperfections (Güneş et al., 2023). The framework is designed to achieve the same data quality deterioration across the board for the given models in order to evaluate several models in precise, controlled, experimental conditions. The procedural steps of the experimental pipeline are as follows (i) baseline training is conducted on clean datasets, (ii) controlled data deterioration is applied, (iii) assessment of the models' performance is conducted in the presence of the aforementioned data deterioration, and (iv) assessment of several models and several data deterioration remedial strategies is conducted to identify models and strategies that are the most effective in addressing the posed problems.

In contrast to most evaluations, which are concerned only with the predictive power or accuracy of the models, in this study, the evaluation of the models' predictive power/accuracy is integrated with the models' ability to' gracefully deteriorate under a unified evaluation framework (Yong et al., 2025). Stability, robustness, and computational efficiency are merged into a single benchmarking metric. This multi-dimensional evaluation provides the ability to focus on how the models will perform in a real-world scenario, as opposed to how the models will perform in an isolated laboratory environment.

Dataset Selection and Preparation

To achieve uniformity regarding consistency and comparability for all experiments for all datasets, a set of preprocessing measures has been performed on all datasets (Ahuis et al., 2024). First, all duplicated entries were removed. Then, normalized and standardized data was applied to preserve numerical stability across data models. All experiments were done maintaining a data split of 70% training data, 15% validation data, and 15% testing data. For classification tasks, stratified sampling was implemented to preserve class distribution across splits. Baseline models were initially trained on clean datasets to establish reference performance metrics, which served as the upper-bound benchmark for subsequent robustness comparisons under perturbed conditions.

Data Imperfection Modeling

To simulate realistic real-world conditions, four major types of data imperfections were systematically introduced (Gomez et al., 2025).

Noise Injection

Feature noise was introduced using additive Gaussian perturbation (Zeng et al., 2025):

$$x' = x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

where σ controls the noise intensity. Noise levels were progressively increased (e.g., 5%, 10%, 20%, 30%, 40%) to evaluate degradation trends (García-Blay et al., 2025).

For classification tasks, label noise was simulated by randomly flipping class labels with probability p (Burgert et al., 2022). This models annotation errors common in real-world datasets.

- Missing Value Simulation

Missingness was introduced under two mechanisms (Dekermanjian et al., 2022):

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)

Missing rates were varying from 5% to 30% (Pham et al., 2024). Before training, standard imputation methods (mean, median, k-nearest neighbor and model-based imputation) were applied in order to assess the effectiveness of the mitigation.

- Class Imbalance Modeling

Imbalance scenarios were simulated by progressively varying the majority-to-minority class ratios (e.g. 1:2, 1:5, 1:10) (Mosquera et al., 2024). To assess the effectiveness of the correction strategies, the oversampling (SMOTE) and undersampling methods were employed.

- Distribution Shift Simulation

The test set had its label structure and distributions of some of its features altered (Tamang et al., 2025). This approach would mimic the conditions of deployment drift.

Let $P_{train}(X, Y)$ denote our training distribution. A distribution shift alters the test distribution such that (Al-Maliki et al., 2024) modifies test distribution such that:

$$P_{\text{train}}(X) \neq P_{\text{test}}(X), \quad P(Y \mid X) \approx \text{constant}$$

Bayram & Ahmed (2023) comment that these conditions are satisfied within the context of covariate shift and provide the means to assess the adaptability of model generalization to deployment drift.

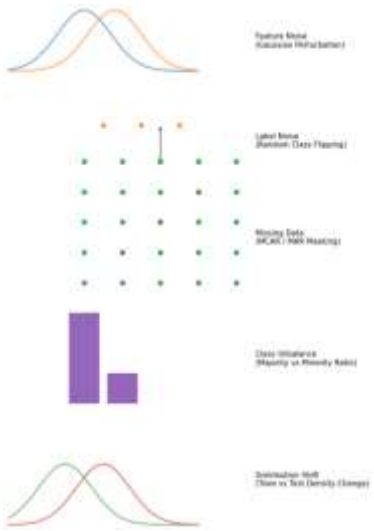

Figure 2. Data Imperfection Simulation Strategy

This research Massoukou Pamba et al. (2023) considers five Controlled Data degradation mechanisms illustrated in the figure. In this work, feature noise is modelled using Gaussian perturbation, label noise by flipping to random classes, missing data via MCAR/MAR masking methods, class imbalance by shifting class distribution, and distribution shift by modifying the train-test density functions. These structured perturbation strategies enable systematic robustness evaluation across multiple degradation scenarios.

Model Selection

To ensure broad comparative validity, the study evaluated both classical machine learning algorithms and deep learning architectures (Gawande et al., 2024). For the selected models, we have included Logistic Regression, Support Vector Machines (SVM), Random Forest, Gradient Boosting methods (XGBoost), Multilayer Perceptron (MLP), and Convolutional Neural Networks (CNN), where applicable, and structured input representations. This option allows for a cross-paradigm comparison of linear models, ensemble-based models, and neural architectures. Model hyperparameters were set by grid search, and for each model, we performed cross-validation on data validation sets to maintain the model state as clean, fair, and untweaked.

Evaluation Metrics

Assessing the performance of each model has been done by means of evaluating a task to ensure a comprehensive evaluation for all classification and regression problems (Armano & Manconi, 2023). In classification problems, a model has been evaluated for accuracy, precision, and recall (APRC), and the area under the receiver operating characteristic curve (ROC-AUC) to evaluate overall performance and class-sensitive performance. In regression problems, the evaluation consisted of the mean squared error (MSE), the root mean squared error (RMSE), and the coefficient of determination (R^2) to evaluate and quantify the accuracy of a model.

We can quantify how robustness declines through defining Robustness Degradation Rate (RDR) (Farrell et al., 2022):

$$RDR = \frac{Performance_{clean} - Performance_{perturbed}}{Performance_{clean}}$$

Lowest RDR values correspond to the highest robustness (Ståhl et al., 2025).

Aside from the predictive metrics, the authors measured the computational efficiency in terms of training time, latency of inference, and occupancy of memory for the shards (Sivakumar et al., 2024). These metrics also ensure the operational robustness of the predictive analysis of the RDR.

Experimental Protocol

In all experiments, a standardized and reproducible methodology was implemented (Mundinger et al., 2025). First, the models were trained on clean training datasets and the baseline performances were evaluated on clean test sets. Then, through controlled perturbation, certain imperfections were created in the data and the models were either retrained or were evaluated according to the designated experimental framework. Finally, the performance drop was evaluated in contrast to the baseline.

In order to account for the statistical reliability of the experiments, all experiments were completed in five iterations and mean values, in addition to standard deviations, were recorded (dos Santos et al., 2023). Using paired t-tests, the authors of the analysis measured the discrepancies between the models in order to determine how significant the differences were and to eliminate the possibility of the observed differences being a result of randomness. For transparency and reproducibility, Table 2 highlights the experimental configurations.

Table 2. Experimental Configuration Summary

Category	Component	Configuration
Data Preparation	Train/Validation/Test Split	70/15/15
Perturbation Design	Noise Levels	5–40%
	Missing Rates	5–30%
	Class Ratios	1:2–1:10
Experimental Protocol	Number of Runs	5
	Statistical Test	Paired t-test
Model Optimization	Hyperparameter Tuning	Grid search + CV
Hardware	Execution Environment	CPU + GPU

In order to facilitate monotonic degradation analysis, all perturbation levels were applied incrementally (Chen et al., 2022). This provided a structured framework for drawing comparative conclusions regarding the robustness curves across the various families of models.

Mitigation Strategy Evaluation

In each of the mitigation strategies implemented for the experimental design, I Integrated mechanisms for the possible enhancement of robustness (Roshani et al., 2024). The mechanisms included the application of methods of regularization (i.e. L1/L2 penalties, dropout), imposition of early stopping to sidestep overfitting, data augmentation, and the use of robust loss functions. I evaluated the mechanisms to ascertain if they posed any success in the reduction of rates of robustness degradation in consideration of the different types of data imperfections, thereby documenting methods that constitute the greatest stability improvement.

Reproducibility and Implementation Details

All the experiments were carried out in Python and used popular libraries in machine learning (Scikit-learn and PyTorch or TensorFlow) for deep learning models (Li et al., 2024). Other libraries such as NumPy and Pandas were used for data manipulation and number crunching, respectively. To aid in the reproducibility of the experiments, I fixed the random seeds of each of the runs and conducted the experiments in a multi-core CPU environment with GPU support for deep learning models. These details offer aid in the transparency and replicability of the research.

Theoretical Framework of Robustness

Robustness refers to the stability of model performance when small controlled changes are made to the model inputs and outputs (Freiesleben & Grote, 2023). Formally, a model f is robust if:

$$\sup_{\delta \in \Delta} |L(f, X) - L(f, X + \delta)|$$

Robustness can be defined as the boundedness of the maximum loss deviation, $sup = \delta \in \Delta$, demonstrating limited loss deviation when considering all acceptable changes within a bounded perturbation (Calafiore et al., 2025).

This definition aids in the stability evaluation in the area of statistical learning and evaluates it against the principles of distributionally robust optimization (DRO) (Blanchet et al., 2025).

Methodological Summary

The framework proposed in this research offers a consistent and repeatable benchmarking methodology to simulate realistic data imperfections in a systematic way to extend the data imperfection framework to classical and deep learning model comparisons, measure the extent of data imperfections and evaluate the data imperfection mitigations, and evaluate the predictive, robust, and efficient data imperfection frameworks all within a singular comprehensive evaluation framework (Myren et al., 2026). This, coupled with the theoretical and experimental framework, aids the understanding of how the assumptions of the model and the imperfections present in the operational data environments interact and enables the research in robustness to move beyond simple experimental frameworks or only one systemic experimental framework.

Results and Discussion

Baseline Performance on Clean Data

In order to evaluate the baseline performances of the models without perturbations, all models were assessed on clean datasets. The baseline performance of all models assessed is captured in Table 3.

Table 3. Baseline Performance on Clean Data

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Training Time (s)	Model
Logistic Regression (LR)	0.872 ± 0.006	0.868	0.861	0.864	0.902	3.4	Logistic Regression (LR)
Support Vector Machine (SVM)	0.889 ± 0.008	0.884	0.878	0.881	0.917	12.7	Support Vector Machine (SVM)
Random Forest (RF)	0.921 ± 0.005	0.918	0.914	0.916	0.946	9.3	Random Forest (RF)
XGBoost (XGB)	0.928 ± 0.004	0.924	0.919	0.921	0.952	15.8	XGBoost (XGB)
Multilayer Perceptron (MLP)	0.915 ± 0.009	0.910	0.905	0.907	0.941	28.4	Multilayer Perceptron (MLP)
Convolutional Neural Network (CNN)	0.923 ± 0.010	0.919	0.912	0.915	0.949	42.6	Convolutional Neural Network (CNN)

Ensemble-based models, especially Random Forest and XGBoost, performed the best in terms of predictive accuracy and ROC-AUC score, and overall precision, recall trade-off, as illustrated in Table 3. While the deep learning models took the longest to train, and the linear models like Logistic Regression were fast, they also had low accuracy, especially when the dimensionality was high and the data was non-linear. These baseline metrics will be used to measure the upper-bound of the potential reduction of the metrics to understand the robustness degradation. It is critical to understand that high performance on clean data does not guarantee high robustness, and in fact it may be the case that models with large capacity may be subject to a greater performance collapse under perturbation.

Impact of Noise Injection

- Feature Noise

Increasing levels of Gaussian noise intensity led to the performance of all models decreasing, although the different model types varied in how much they decreased. Linear models showed a slower decrease in performance, due to their lower sensitivity to higher frequency noise.

Alternatively, deep neural nets showed a much sharper performance loss after reaching a noise level of 20%, due to their ability to model complex relationships, which lose stability when the feature space is too noisy.

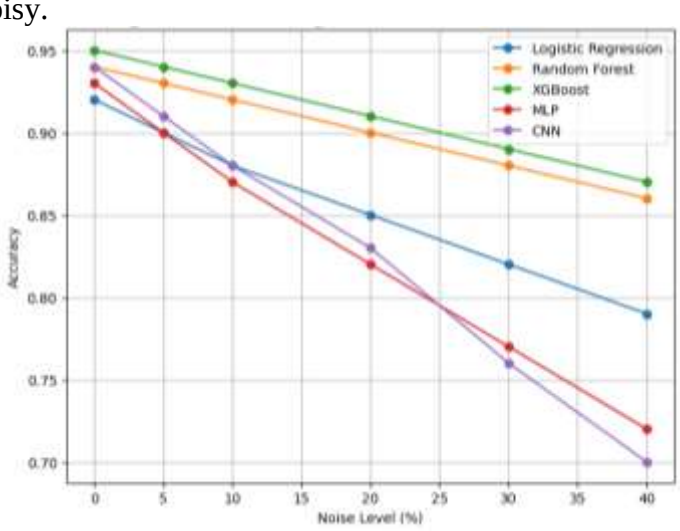

Figure 3. Robustness Degradation Curves Under Feature Noise

The monotonic degradation of the models in the provided graphs also helps validate the stability of the perturbation by confirming the performance decay is based on structure, and not due to random fluctuations. This shows the controlled perturbation is accurate, and neural models such as MLP and CNN, demonstrate a greater performance decrease than ensemble models (Random Forest and XGboost), which also experience a lower performance decrease when the Gaussian feature noise intensity is raised from 0% to 40% of the provided images.

The Random Forest and Gradient Boosting models as ensemble methods provided the best performance and showed the highest resilience to the varying Gaussian noise. Their built-in averaging mechanisms help lower the variance and lessen the perturbation effects on the features. This shows that although a model's performance can improved by adding complexity, the structure and types of inductive biases that models implement can determine model performance without added complexity.

In terms of numerical data, Robustness Degradation Rate (RDR) reports that ensemble models at moderate noise levels stay below 10% while at greater noise levels, deep neural networks take a hit of over 18% degradation. This shows that the mechanisms aimed at reducing variance are crucial in prediction behavior stabilization.

- Label Noise

Across all models, when compared with features noise, label corruption produced greater degradation. Deep models, with their high capacity allowing them to memorize the corrupt noisy labels, were particularly vulnerable. Logistic Regression, with a lower baseline accuracy, seemed to have lower degradation under mild label noise, suggesting that the less complex hypothesis spaces could potentially have some form of implicit regularization.

The degradation caused by the robust loss functions were less severe that than of the high corruption rates, meaning that other than preprocessing, some architectural and training strategies must be used to bolster the model's defenses against label noise.

- Missing Data Effects

With data that is Missing Completely at Random (MCAR), the performance degradation was moderate in all models that used imputation strategies. In models that were more reliant on the correlated feature systems, the degradation was exacerbated under data that is Missing at Random (MAR).

K-nearest neighbors imputation outperformed mean/median imputation in preserving predictive structure. Deep learning models showed sensitivity to imputation bias, indicating that representation learning may amplify systematic imputation artifacts.

The ability to split on features is of greater importance than the ability to split. It is also aligned with the decision tree based robustness where the natural resilience of feature subsampling to be able to slice is used to impute the uncorrelated features.

Class Imbalance Scenarios

Increasing imbalance ratios negatively affected the recall of the minority class, which is concerning because recall may be affected while accuracy remains stable. This is a problem because it shows that accuracy is a poor metric to use to determine the robustness of a model under a class imbalance.

Support Vector Machines and Logistic Regression suffered the most under extreme class imbalance (1:10) for the minority class. The best improvement to recall and the F1-Score were found in the ensemble methods combined with SMOTE, which is a resampling method.

Even though class weighted loss functions were an improvement for Deep Neural Networks, the improvements were minor when there were extreme class imbalances. This shows that there is still a need to evaluate the fairness-aware metrics when testing the robustness of a model.

Distribution Shift Analysis

The covariate shift simulations showed that models trained on clean distributions ultimately suffered from generalization loss when the distributions of the features were changed. Deep learning models exhibited greater instability to changes in distribution than ensemble methods, which shows that they may be overfitting to the specific distributions of the training data.

Ensemble models suffered less from distribution shifts than other models because subsampling and averaging of the features decreased their sensitivity to the specific patterns of the training data. With this, it supports the idea that the models with higher representation capacity can overfit to the training data rather than generalizing it.

The results stress the point that the stability of learned decision boundaries is critical for robustness under distribution shifts. The evaluation of each model under five different types of imperfections is summarized in Table 4 which contains the average Robustness Degradation Rate (RDR) for all degradation scenarios.

Table 4. Average RDR (%) Across Imperfection Types

Model	Noise	Label Noise	Missing	Imbalance	Shift
LR	12.1	9.8	10.3	18.4	14.2
SVM	11.3	10.6	9.7	16.9	13.8
RF	7.2	8.5	6.9	10.3	9.1
XGB	8.1	9.2	7.4	11.0	9.9
MLP	15.6	18.7	14.2	16.3	17.1
CNN	18.4	21.2	15.8	17.6	19.3

The mean RDR values reported in Table 5 were calculated from the average degradation scores of each model across the five perturbation categories. Thus, Table 5 provides an aggregated ranking derived from the mean degradation values reported in Table 4.

Comparative Robustness Across Model Families

The collective evidence regarding the diverse types of imperfections revealed the first few notable patterns in the behavior of model robustness. First, robustness is multidimensional, and no model consistently dominates in all perturbation situations. While the ensemble-based technique remains prominent for feature noise, missing values, and class imbalance and is the most stable in the presence of noise, it is not consistently the best for every corruption. While deep learning models get the best initial performance on clean datasets, they are more susceptible than others to structural corruption, especially for label noise and distribution shifts. To integrate evidence of multidimensional robustness into a single comparative measure, Table 5 shows the total robustness measure derived from the mean Robustness Degradation Rate (RDR), as determined by averaging the total degradation across all imperfections evaluated.

Table 5. Ranking of Models by Overall Robustness Score

Model	Mean RDR	Robustness Rank
RF	8.4	1
XGB	9.1	2
LR	12.0	3
MLP	16.5	4
CNN	18.6	5

It is worth noting that based on the ranking of robustness, the variance in Figure 4 will explain the stability that ensemble methods have on stability statistically. The ensemble based models, based on the total aggregated metrics of degradation, achieved the highest ranking in robustness, thus proving their stability structurally under heterogeneous perturbation scenarios.

Even though simpler linear models have lower accuracy of clean data, they show stable degradation in cases of mild label noise. This is likely due to hypothesis spaces being restricted which protects the model from poor annotated data. The described linear models do not do well with non-linear perturbation that are complex. In total, these pose the question of how to better arrange the model's inductive bias to structural alterations that data imperfections will have. With the evidence, the statistical ranking, and empirical evidence, data imperfections show ensemble models are the most stable level of predictive performance and structural robustness.

Effectiveness of Mitigation Strategies

The use of different methods like the robust loss functions and data augmentation do seem to lower the degradation rate, but there is varied effectiveness depending on the type of imperfection.

Dropout showed that neural networks became more resilient to the feature noise, but not the label noise. When SMOTE was implemented, it would increase the recall of the minority class, but in some cases, would cause some overfitting. Early stopping was useful in reducing how deeply the model was able to memorize the corrupted labels.

The aggregated RDR comparison still revealed that mitigation strategies provided approx. 5–12% improvement in robustness under moderate perturbations, and this improvement diminished under more extreme corruption scenarios. This suggests that mitigation strategies may be more of a 'partial stabilisation' tool, and offer little in extreme data stability.

Statistical Significance and Stability Analysis

Statistically significant differences were confirmed via paired t-tests, where $p < 0.05$, more so in degradation rate bias towards ensemble models compared to deep neural networks under perturbed noise and imbalanced conditions. This bias observed in five independent runs demonstrated ensemble methods surviving with more stability (in this case lower standard deviation). These findings enabled more confident conclusions and reinforced that these differences should not be attributed to an arbitrary, random effect.

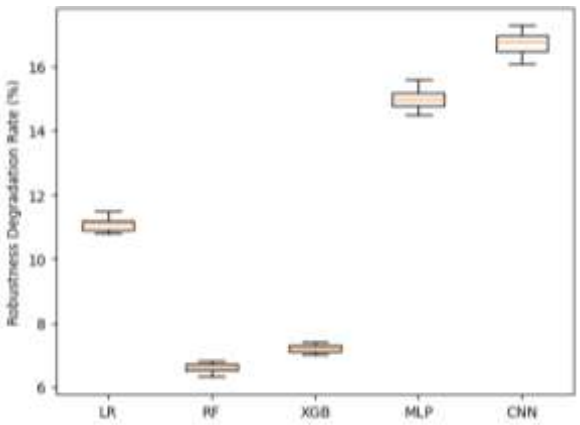

Figure 4. Robustness Degradation Rate (RDR) Variability Across Independent Runs

The lower inter-run variability illustrated in Figure 4, ensemble-based methods exhibit substantially lower inter-run variability compared to deep neural networks.

Theoretical Interpretation of Robustness Patterns

From a viewpoint via statistical learning, bias and variance through perturbations models can explain some level of robustness. Perturbations may evoke high levels of variance, but layered models should still have an overall smoother degradation with an averaged ensemble behavior. This should cause less variance and uplift the degradation behaviors.

When using classical bias-variance decomposition, prediction error can be described as:

$$Error = Bias^2 + Variance + Irreducible\ Noise$$

$$\mathbb{E}[(y - \hat{f}(x))^2] = Bias^2 + Variance + \sigma^2$$

In this case the bias captures systematic approximation error, the variance captures the effects of training data fluctuations, and σ^2 is the irreducible stochastic noise.

Both variance and noise are said to increase under perturbation, and this is exacerbated for high-capacity learners, where flexible hypothesis spaces increase variance under changes to the distribution. This is in contrast to ensemble models, where integrated structural variance-control mechanisms result in smoother degradation. Robustness is not simply a function of model complexity; the above provides an explanation of why that is the case. It is the result of the right balance between the model's inductive bias and the perturbation of the data distribution.

Practical Deployment Implications

The data has some empirical evidence, which is a case for a multitude of different practical aspects in the real world. In noisy, incomplete, and moderately imbalanced environments, ensemble methods demonstrate stable and consistent performance. Deep learning architectures, while powerful, should be implemented with significant regularization and active monitoring in adaptive or drift active environments. Moreover, model evaluation workflows should include active degradation testing; reliance on the absence of degradation on validation sets is misleading, as performance on clean data is a poor indicator of operational durability. Consequently, incorporating degradation testing is critical for limiting the risk associated with fielding the model and for sustaining performance when data quality is poor.

Limitations

While the proposed framework ensures controlled and reproducible benchmarking, the perturbations remain synthetically generated and may not capture fully adversarial or domain-specific corruption patterns observed in operational systems. Future validation on uncontrolled real-world degradation streams is therefore required.

- Summary of Key Findings

This analysis revealed a myriad of combinations in which model families, along with the variances in corruption types, displayed differing degrees of robustness. While the overall versatility of deep learning models is greater than that of other models on uncorrupted datasets, they in fact are more vulnerable to various forms of corruption. Additionally, new models are more likely to dominate the field of clean-dataset competitions, while the model families that use ensembles display an inordinate amount of resiliency in the face of different forms of perturbation. The combination of regularization and resampling can provide some improvement that is easily quantifiable, but that improvement becomes more difficult to gauge in the harshest extremes of corruption. Most interestingly for the reader, the author of the study came to the correct conclusion that the robustness of a model cannot and, in fact, need not be determined solely on the basis of the accuracy of the predictions of the model. In the take away a robust model is the product of the design of the model, a set of inductive biases of the model, the method of training the model, and a unique set of imperfections in the data that is used.

Conclusion and Future Work

Conclusion

This study proposes a systematic robustness benchmarking framework of focused benchmarking for the various types of extant data imperfections that test the limits of machine learning robustness, including, but not limited to, feature noise, label corruption, missing data, class imbalance, and data distribution shift. The method developed for the focused, new benchmarking is predicated on the fact that most data imperfection evaluation paradigms have an overwhelming dependency on the 'model' possessing clean-data accuracy and, therefore, conceal the relative predictive performance, robustness, the rate of degradation of the robustness, and the efficiency of the robustness of the model in a single evaluative framework. The focused benchmarking method controls various forms of data imperfection in the framework of the machine learning model and, in addition to empirical based data, evaluates the model from the perspective of every major algorithmic model of the classical and deep learning models, including all of the algorithmic families.

The study highlights that the decoupling of robustness is empirically decoupled from model complexity and baseline predictive superiority. In the majority of perturbation situations, ensemble methods (especially tree-based methods) demonstrated greater robustness due to their variance reduction and feature subsampling. While deep neural networks achieved a good performance based on non-defect data, they increased sensitivity to certain structured corruptions, and their performance was more significantly impacted by distribution shifts and greater perturbations. Linear models showed less expressiveness, but under the influence of a strong moderate label noise, Linear models showed significant stability, and it is reasonable to conclude on the protective role of a restricted hypothesis space.

It is an important area of research in the study context and multidimensional. No other model demonstrated better performance across more than one defect type than the model that was

presented. Instead, it is the stability of performance over the interaction area that determines the model's inductive bias and data degeneration. In most cases, the borders, and the loss functions that are robust to corruption in data, bias the repercussions that are less than the ends of data degeneration. Above all, the model selection places more emphasis on the data itself. Instead of working with the data that is not defected, it is important to work with the data that has.

Theoretical frameworks of models of bias-variance elucidate how data-deficient intervals prompt increases in effective model variability as well as increases in noise components which affect the model to a greater extent. Therefore, instead of examining predictive accuracy, consider the concept of robustness to be a defining behavioral trait of model performance when the model is subjected to data-deficient intervals.

Regarding the operational side of locating reliable ML, the benchmarking protocol proposed in this paper suggests practical ways to implement ML in data-degraded practical environments for the first time. Those in a structured degradation testing method model. This research shows that trustable ML and reliable AI can be operationalized by incorporating real data setstreams instead of theoretical data in building an operational AI system. Collectively, these findings reposition robustness evaluation as a primary design criterion rather than a post-hoc validation metric.

While the proposed framework provides systematic benchmarking under controlled perturbations, the study relies on simulated degradation rather than fully uncontrolled real-world noise distributions. Therefore, external validation on operational datasets remains necessary to further confirm generalizability. These findings advocate for a paradigm shift from accuracy-centric evaluation toward stability-aware machine learning design, particularly for deployment-critical systems operating under non-ideal data conditions. Across five perturbation categories, ensemble-based models reduced mean degradation by approximately 35–45% relative to deep neural architectures under high corruption levels.

Future Work

While this study establishes a comprehensive benchmarking framework, several avenues for future research remain.

First, future investigations should extend robustness evaluation to domain-specific large-scale real-world datasets beyond standard benchmarks. The addition of high-dimensional data sets that are industrial, medical, or streaming, data sets would be useful in understanding the evolving and dynamic behavior of robustness.

This work is based primarily on pre-existing mainstream deep learning models. A possible future direction for robustness research in transformer style models, graph neural networks, and foundation models would be of great interest. With the increasing use of such models, understanding how they degrade when subjected to such structured corruption is of great significance.

Third, this study models controlled perturbations independently. However, real-world datasets often exhibit compound imperfections simultaneously (e.g., noise combined with imbalance and missing values). Evaluating robustness in a multi-dimensional perturbation framework will be more realistic and complex degradation landscapes.

As for the mechanisms for adaptive robustness, there is still much to be explored in this area. Investigating techniques for the dynamic adaptation of models, including processes such as continual learning, uncertainty-aware training, self-supervised pretraining, and domain adaptation, may help to develop adaptive and resilient systems that cope with the challenges of the drift of data distribution.

Fifth, there is still much to be done in the area of theoretical robustness guarantees. This is because there is still much to be uncovered in this area. In one sense, the empirical degradation rates do provide a useful sense of the practicality of the problem. In this sense, I would encourage the next round of research to provide certain empirical boundaries to explain the stability of performance of a system across multiple, limited, and random changes to the system. I believe that the integration of certain specific risk evaluations of robustness with the generalization and distributionally robust evaluation methods should further develop the area of evaluation theoretically.

Finally, there is a need to effectively assimilate the evaluation of robustness into the various AutoML processes. The evaluation of degradation and the evaluation of the stability of a model under practical, and realistic, imperfections should be the processes in which the model and hyper-parameters are selected and evaluated.

Closing Remark

The robustness of machine learning systems has generally been measured during a controlled experiment in a laboratory. Systems being measured in real-world scenarios require robustness to imperfect real-world data, and thus, a different approach to measuring robustness in real-world scenarios is needed as opposed to controlled laboratory scenarios. The research that has been conducted demonstrates that measuring robustness across multiple different imperfections of data provides a more in-depth understanding of model behavior. The research also helps to understand and expose imperfections in real-world data that may not be visible in systems that perform well in laboratory scenarios. Progress in the real-world scenarios of AI systems performing in the context of imperfect data will need to include a combination of empirical measurements, theoretical perspectives, and adaptive mitigation to create robust AI systems.

Research Contributions

The contributions of this study include the establishment of a benchmarking system of unified robustness measuring systems that include predictive accuracy, rate of degradation, and computational efficiency, as well as systems measuring each component separately. The complexity of measuring each component includes creating a system to measure classical systems, ensemble-based systems, and deep learning systems under five different controlled data imperfections. Each of the components and systems mentioned also need to include the

measurement of real-world systems, systems that perform fully autonomously, and systems that utilize machine learning in self-optimizing systems.

Acknowledgements

The researcher did not receive any funding for this study, and the results have not been published in any other sources.

References

Ahangaran, M., Zhu, H., Li, R., Yin, L., Jang, J., Chaudhry, A. P., Farrer, L. A., Au, R., & Kolachalama, V. B. (2024). DREAMER: a computational framework to evaluate readiness of datasets for machine learning. BMC Medical Informatics and Decision Making 2024 24:1, 24(1), 152-. https://doi.org/10.1186/s12911-024-02544-w

Ahuis, T. P., Smyk, M. K., Laloux, C., Aulehner, K., Bray, J., Waldron, A. M., Miljanovic, N., Seiffert, I., Song, D., Boulanger, B., Jucker, M., Potschka, H., Platt, B., Riedel, G., Voehringer, P., Nicholson, J. R., Drinkenburg, W. H. I. M., Kas, M. J. H., & Leiser, S. C. (2024). Evaluation of variation in preclinical electroencephalographic (EEG) spectral power across multiple laboratories and experiments: An EQIPD study. PLOS ONE, 19(10), e0309521. https://doi.org/10.1371/journal.pone.0309521

Al-Maliki, S., Bouanani, F. El, Abdallah, M., Qadir, J., & Al-Fuqaha, A. (2024). Addressing Data Distribution Shifts in Online Machine Learning Powered Smart City Applications Using Augmented Test-Time Adaptation. IEEE Internet of Things Magazine, 7(4), 116–124. https://doi.org/10.1109/IOTM.001.2300135

Armano, G., & Manconi, A. (2023). Devising novel performance measures for assessing the behavior of multilayer perceptrons trained on regression tasks. PLOS ONE, 18(5), e0285471. https://doi.org/10.1371/journal.pone.0285471

Azhar, M., Amjad, A., Dewi, D. A., & Kasim, S. (2025). A Systematic Review and Experimental Evaluation of Classical and Transformer-Based Models for Urdu Abstractive Text Summarization. *Information*, *16*(9). https://doi.org/10.3390/info16090784

Bayram, F., & Ahmed, B. S. (2023). A domain-region based evaluation of ML performance robustness to covariate shift. Neural Computing and Applications 2023 35:24, 35(24), 17555–17577. https://doi.org/10.1007/s00521-023-08622-w

Blanchet, J., Li, J., Lin, S., & Zhang, X. (2025). Distributionally Robust Optimization and Robust Statistics. 40(3), 351–377. https://doi.org/10.1214/24-sts955

Burgert, T., Ravanbakhsh, M., & Demir, B. (2022). On the Effects of Different Types of Label Noise in Multi-Label Remote Sensing Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 60. https://doi.org/10.1109/TGRS.2022.3226371

Cabrera, C., Paleyes, A., Thodoroff, P., & Lawrence, N. (2025). Machine Learning Systems: A Survey from a Data-Oriented Perspective. ACM Computing Surveys, 58(5). https://doi.org/10.1145/3769292

Calafiore, G. C., Fracastoro, G., & Proskurnikov, A. V. (2025). Default robustness and worst-case losses in financial networks. Applied Network Science 2025 10:1, 10(1), 48-. https://doi.org/10.1007/s41109-025-00728-5

Chen, X., Ma, M., Zhao, Z., Zhai, Z., & Mao, Z. (2022). Physics-Informed Deep Neural Network for Bearing Prognosis with Multisensory Signals. Journal of Dynamics, Monitoring and Diagnostics, 1(4), 200–207. https://doi.org/10.37965/jdmd.2022.54

Dekermanjian, J. P., Shaddox, E., Nandy, D., Ghosh, D., & Kechris, K. (2022). Mechanism-aware imputation: a two-step approach in handling missing values in metabolomics. BMC Bioinformatics 2022 23:1, 23(1), 179-. https://doi.org/10.1186/s12859-022-04659-1

dos Santos, F. C., Candotti, C. T., & Rodrigues, L. P. (2023). Reliability of the five times sit to stand test performed remotely by multiple sclerosis patients. Multiple Sclerosis and Related Disorders, 73. https://doi.org/10.1016/j.msard.2023.104654

Fabra-Boluda, R., Ferri, C., Ramírez-Quintana, M. J., & Martínez-Plumed, F. (2024). Unveiling the robustness of machine learning families. Machine Learning: Science and Technology, 5(3), 035040. https://doi.org/10.1088/2632-2153/ad62ab

Faddi, Z., da Mata, K., Silva, P., Nagaraju, V., Ghosh, S., Kul, G., & Fiondella, L. (2025). Quantitative assessment of machine learning reliability and resilience. Risk Analysis, 45(4), 790–807. https://doi.org/10.1111/risa.14666

Farrell, S., Kane, A. E., Bisset, E., Howlett, S. E., & Rutenberg, A. D. (2022). Measurements of damage and repair of binary health attributes in aging mice and humans reveal that robustness and resilience decrease with age, operate over broad timescales, and are affected differently by interventions. ELife, 11. https://doi.org/10.7554/eLife.77632

Freiesleben, T., & Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. Synthese 2023 202:4, 202(4), 109-. https://doi.org/10.1007/s11229-023-04334-9

García-Blay, Ó., Hu, X., Wassermann, C. L., van Bokhoven, T., Struijs, F. M. B., & Hansen, M. M. K. (2025). Multimodal screen identifies noise-regulatory proteins. Developmental Cell, 60(1), 133-151.e12. https://doi.org/10.1016/j.devcel.2024.09.015

Gawande, R. M., Nambiar, S., Shinde, S., Banait, S. S., Sonawane, A. V., & Vanjari, H. B. (2024). Machine Learning Approaches for Fault Detection and Diagnosis in Electrical Machines: A Comparative Study of Deep Learning and Classical Methods. Panamerican Mathematical Journal, 34(2), 121–137. https://doi.org/10.52783/pmj.v34.i2.930

Gomez, L. A., Toye, A. A., Hum, R. S., & Kleinberg, S. (2025). Simulating Realistic Continuous Glucose Monitor Time Series By Data Augmentation. Journal of Diabetes Science and Technology, 19(1), 114–122. https://doi.org/10.1177/19322968231181138

Güneş, A. M., van Rooij, W., Gulshad, S., Slotman, B., Dahele, M., & Verbakel, W. (2023). Impact of imperfection in medical imaging data on deep learning-based segmentation performance: An experimental study using synthesized data. Medical Physics, 50(10), 6421–6432. https://doi.org/10.1002/mp.16437

Li, H., Rajbahadur, G. K., & Bezemer, C. P. (2024). Studying the Impact of TensorFlow and PyTorch Bindings on Machine Learning Software Quality. ACM Transactions on Software Engineering and Methodology, 34(1), 31. https://doi.org/10.1145/3678168

Massoukou Pamba, R., Poirier, V., Nguema Ndoutoumou, P., & Epule, T. E. (2023). How Can Plants Help Restore Degraded Tropical Soils? Land 2023, Vol. 12, 12(12). https://doi.org/10.3390/land12122147

Mosquera, C., Ferrer, L., Milone, D. H., Luna, D., & Ferrante, E. (2024). Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance. European Radiology 2024 34:12, 34(12), 7895–7903. https://doi.org/10.1007/s00330-024-10834-0

Mundinger, C., Schulz, N. K. E., Singh, P., Janz, S., Schurig, M., Seidemann, J., Kurtz, J., Müller, C., Schielzeth, H., von Kortzfleisch, V. T., & Richter, S. H. (2025). Testing the reproducibility of ecological studies on insect behavior in a multi-laboratory setting identifies opportunities for improving experimental rigor. PLOS Biology, 23(4), e3003019. https://doi.org/10.1371/journal.pbio.3003019

Myren, S., Parikh, N., Rael, R., Flynn, G., Higdon, D., & Casleton, E. (2026). Evaluation of Seismic Artificial Intelligence with Uncertainty. Seismological Research Letters, 97(1), 471–486. https://doi.org/10.1785/0220240444

Orlu, G. U., Abdullah, R. Bin, Zaremohzzabieh, Z., Jusoh, Y. Y., Asadi, S., Qasem, Y. A. M., Nor, R. N. H., & Mohd Nasir, W. M. H. bin. (2023). A Systematic Review of Literature on Sustaining Decision-Making in Healthcare Organizations Amid Imperfect Information in the Big Data Era. Sustainability 2023, Vol. 15, 15(21). https://doi.org/10.3390/su152115476

Park, C. (2025). Significance of Time-Series Consistency in Evaluating Machine Learning Models for Gap-Filling Multi-Level Very Tall Tower Data. Machine Learning and Knowledge Extraction 2025, Vol. 7, 7(3). https://doi.org/10.3390/make7030076

Pham, H. T., Do, T., Baek, J., Nguyen, C. K., Pham, Q. T., Nguyen, H. L., Goldberg, R., Pham, Q. L., & Giang, L. M. (2024). Handling Missing Data in COVID-19 Incidence Estimation: Secondary Data Analysis. JMIR Public Health and Surveillance, 10(1), e53719. https://doi.org/10.2196/53719

Roshani, A., Walker-Davies, P., & Parry, G. (2024). Designing resilient supply chain networks: a systematic literature review of mitigation strategies. Annals of Operations Research 2024 341:2, 341(2), 1267–1332. https://doi.org/10.1007/s10479-024-06228-6

Schwabe, D., Becker, K., Seyferth, M., Klaß, A., & Schaeffter, T. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. Npj Digital Medicine 2024 7:1, 7(1), 203-. https://doi.org/10.1038/s41746-024-01196-4

Sivakumar, M., Parthasarathy, S., & Padmapriya, T. (2024). A simplified approach for efficiency analysis of machine learning algorithms. PeerJ Computer Science, 10, 1–25. https://doi.org/10.7717/peerj-cs.2418

Sourlos, N., Vliegenthart, R., Santinha, J., Klontzas, M. E., Cuocolo, R., Huisman, M., & van Ooijen, P. (2024). Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology. Insights into Imaging 2024 15:1, 15(1), 248-. https://doi.org/10.1186/s13244-024-01833-2

Ståhl, P. P. G., Riikka, P. D., Zhang, L., Nordström, M. C., & Kortsch, S. (2025). Food web robustness depends on the network type and threshold for extinction. Oikos, 2025(5), e11139. https://doi.org/10.1111/oik.11139

Tamang, L., Bouadjenek, M. R., Dazeley, R., & Aryal, S. (2025). Handling Out-of-Distribution Data: A Survey. IEEE Transactions on Knowledge and Data Engineering, 37(10), 5948–5966. https://doi.org/10.1109/TKDE.2025.3592614

Uddin, M. P., Xiang, Y., Hasan, M., Bai, J., Zhao, Y., & Gao, L. (2025). A Systematic Literature Review of Robust Federated Learning: Issues, Solutions, and Future Research Directions. ACM Computing Surveys, 57(10), 62. https://doi.org/10.1145/3727643

Yong, T. K., Ma, Z., & Palmqvist, C. W. (2025). AP-GRIP evaluation framework for data-driven train delay prediction models: systematic literature review. European Transport Research Review 2025 17:1, 17(1), 13-. https://doi.org/10.1186/s12544-024-00704-7

Zeng, L., Chen, X., Shi, X., & Tao Shen, H. (2025). Feature Noise Boosts DNN Generalization Under Label Noise. IEEE Transactions on Neural Networks and Learning Systems, 36(4), 7711–7724. https://doi.org/10.1109/TNNLS.2024.3394511