OPEN ACCESS
International Journal of Advances in Artificial Intelligence and Machine Learning Vol 3 No 1 (2026) https://doi.org/10.58723/ijaaiml.v3i1.651

Check for updates

Research Article

Failure Mode Analysis of Machine Learning Models in Realistic Data Deployment Scenarios

Received: February 09, 2026	Revised: February 26, 2026	Accepted: March 25, 2026	Publish: March 31, 2026

Lau Meng Cheng*, Amel Zulfukar Hassan Adlan

Abstract:

Background: Machine learning models frequently demonstrate strong performance under controlled benchmark evaluations. However, such evaluations often fail to capture hidden vulnerabilities that emerge under realistic deployment conditions. In real-world environments, models are exposed to stressors such as label corruption, feature noise, distributional shifts, and operational constraints, including reduced computational precision and increased latency. These conditions can induce performance degradation and structural instability, highlighting the need for a systematic robustness evaluation framework that goes beyond conventional accuracy metrics.

Aims: This paper aims to introduce a formalized Failure Mode Analysis Protocol (FMAP) for evaluating machine learning model robustness under realistic operational stressors. The study reconceptualizes robustness evaluation as a distribution-based process, where model deployment itself generates a new distribution over time.

Methods: The proposed FMAP framework evaluates model behavior under progressively adverse conditions, including symmetric label corruption, additive feature noise, distributional shifts, and operational constraints such as reduced numerical precision and increased inference latency. Experiments were conducted across diverse tabular and image benchmark datasets using representative model architectures, including linear models, ensemble methods, margin-based models, and deep neural networks.

Result: The experiments reveal distinct robustness profiles across model architectures when exposed to escalating stress conditions. Operational constraints and compositional limitations were shown to induce measurable degradation patterns, including instability and output collapse under extreme stress. The findings demonstrate that model failure is not solely a function of predictive accuracy loss but is closely linked to operational constraints and evolving distributional conditions. The distribution-based evaluation framework effectively captures early-stage degradation and full failure transitions.

Conclusion: This study establishes a structured protocol for analyzing machine learning failure modes under realistic deployment scenarios. By framing robustness evaluation as a distribution-based process, the FMAP approach provides a systematic method for identifying operational risks and structural vulnerabilities.

Keywords: Data Science; Deployment Risks; Failure Analysis; Machine Learning Reliability; Model Robustness.

1. INTRODUCTION

Across industries like finance, cyber security, healthcare, industrial automation, and automated transportation, machine learning (ML) has become an integral part of the modern data-driven decision making process (Ige et al., 2025). ML's predictive capabilities and automation potential are the primary reasons for its increasing use (Ahmed, 2024). ML varables though are very difficult to identify, and deploying a model is very context dependent (Theng & Bhoyar, 2024). An ideal working environment for ML systems does not exist because most environments have unpredictable and dynamic users and limited resources (Truong et al., 2023). ML systems also have very hard to identify and hidden vulnerabilities which can lead to a high risk, underperforming system (Tripathi & Pandey, 2025). Operational environments can explain how benchmark datasets and controlled experimental protocols display such high performing benchmarks (Liao et al., 2022). As ML systems are applied to more and more safety and high risk environments, understanding, and mitigating those vulnerabilities is critical for the responsible development of artificial intelligence and the deployment of systems (Habbal et al., 2024).

Even with the advancements in technique optimization and model calibration, deployed systems continue to

demonstrate several unpredictable breakdown situations (Zhou et al., 2024). Under real-world scenarios, trained models will produce an inaccurate prediction with unstable decision edges and poorly calibrated confidence gauges when faced with corrupt labeling, missing feature sets, class imbalance, or unknown and ever-changing data distributions (Bauer et al., 2026). Specific factors contribute to the ability to not capture these breakdown situations. Such examples include missing feature sets and class imbalance towards previously unseen data distributions, which are rarely captured by traditional systems of evaluation like average loss or overall accuracy (Chen et al., 2024). As a result, models that appear to be high-performing and reliable during validation often turn out to be unfathomably deficient when applied in the real world, losing the high performance edge in real-world scenarios (Abdelkader & Csámer, 2025). This study looks to create a system to provide real-world data for the evaluation of ML models that encompasses the measurement, classification, and evaluation of failure modes due to the absence of such systems in existence (Wongkaew et al., 2024).

Prior work has analyzed robustness through adversarial learning, noise injection, and out-of-distribution (OOD) detection (Li et al., 2024). Adversarial studies show that small changes lead to misclassification, while data augmentation and regularization are shown to promote generalization under mild noise conditions (Mumuni & Mumuni, 2022). Other studies look at covariate shift and domain adaptation to address the loss of performance when the training and testing distributions are different (Ott et al., 2022). While these studies explore different types of perturbations, their analyses are often limited to laboratory controlled single perturbation settings (Monfort-lanzas et al., 2025). In addition, many studies shift their focus to performance retention rather than structurally characterizing failures, which makes it difficult to understand the breaking points of models in real-world operational systems (Faddi et al., 2025).

The second line of research looks at reliability through the lenses of uncertainty quantification, calibration, and interpretability of models (Salvi et al., 2025). Bayesian neural networks, ensemble methods, and calibration attempts to better align the predicted probabilities to the empirical risk (Ramesh et al., 2025). At the same time, explainable AI (XAI) approaches try to understand the models and identify anomalous behaviors of the predictions (Hassija et al., 2024). This illustrates how reliability is important beyond just accuracy. However, most studies still suffer from not addressing, or under-addressing, the stressors from the deployment of the deployed models, particularly the multi-factor, compounding effects of noise, distributional drift, and operational parameters constraints and limitations (Dong et al., 2024). Furthermore, the evaluation protocols lack, more often than not, a comprehensive set of standardized taxonomies of failures, horizontal degradation (as opposed to vertical) comparative analyses across families of models, and simulations addressing reproducibility (as opposed to those focusing on the models) that mimic deployment (Smith & Spencer, 2024).

Combining various studies demonstrates that although there has been extensive research on robustness, calibration, and domain adaptation, there is still an obvious lack of research on systematic failure mode analyses that take multiple varying deployment stressors into account in one unified evaluation framework (Huang et al., 2023). Particularly, there is a scarcity of research that aims to i) systematically simulate various deployment environments, ii) measure the varying degrees to which the functionality of different ML models is compromised, and iii) systematically interpret and categorize the breakdown behavior of models (Cabrera et al., 2023). This paper proposes a systematic failure mode analysis approach to machine learning reliability evaluation which entails realistic data deployment for research on the aforementioned gaps (An et al., 2024). The proposed framework is an integration of all the above-mentioned factors in an effort to diagnose the remaining hidden failure vulnerabilities in what are called the ‘standard’ benchmarks (Jung et al., 2022).

This work departs from traditional robustness studies which focus on resistance to perturbations (Giacobazzi et al., 2024). Instead, it proposes the conceptualization of the failure of deployments as a phenomenon of a defined structure with identifiable degradation pathways and defined tipping points (Qiu et al., 2023).

This work makes three main contributions. First, it creates a deployment-oriented evaluation framework that demonstrates realistic stressors such as label noise, feature corruption, distribution shift, and operational constraints. Second, it creates a taxonomy of failure modes in a structured way in order to classify breakdown behaviors by triggering conditions, observable signatures, and diagnostic signs. Third, it performs a cross-model family analysis using a multi-metric approach to reveal degradation trends empirically, which go unnoticed in traditional benchmarking practices.

In the remainder of this paper, Section 2 provides the proposed framework for failure mode analysis and describes the methodologies for simulating deployment stress. Section 3 describes the experimental design and evaluation methodology. Section 4 provides empirical data and analysis, focusing on comparative degradation. In Section 5, I will focus on the implications, limitations, and threats concerning the validity of this study. In the final section, I will conclude this study and pursue a discussion on possible research avenues.

2. MATERIAL AND METHOD

In this section, explain how the research was conducted. The main points of this section are: (1) type of research; (2) sample or object of research (who is the object or sample of research and tell what kind of sample method was used); (3) time of research (tell when, where, and how long the research was conducted); (4) research

procedures; (5) data collection techniques; (6) research instruments, (7) analysis plan (explain the statistical tests and comparisons made; ordinary statistical methods should be used without comment; advanced or unusual methods may require literature citations); (8) scope and/or limitations of the research used.

An experimental simulation-based design was developed for this study to analyze the failure modes of machine learning algorithms in the presence of realistic deployment hurdles. The goal is to understand the behavior of trained models under conditions they have not been trained on. The entire workflow has been segmented for the purpose of this study into four primary components: data collection and the establishment of a baseline, model training and preprocessing, simulation of deployment stresses in a controlled environment, and evaluation of robustness across multiple dimensions. This systematic approach ensures transparency and reproducibility, while enabling the comparison of models. The study framework is designed to highlight failure behaviors running in a realistic environment while controlling training simulation failures.

The outlined process is the overall workflow of the proposed Failure Mode Analysis Protocol (FMAP) and its components have been planned in a structured manner, as illustrated in the accompanying diagram (Figure 1).

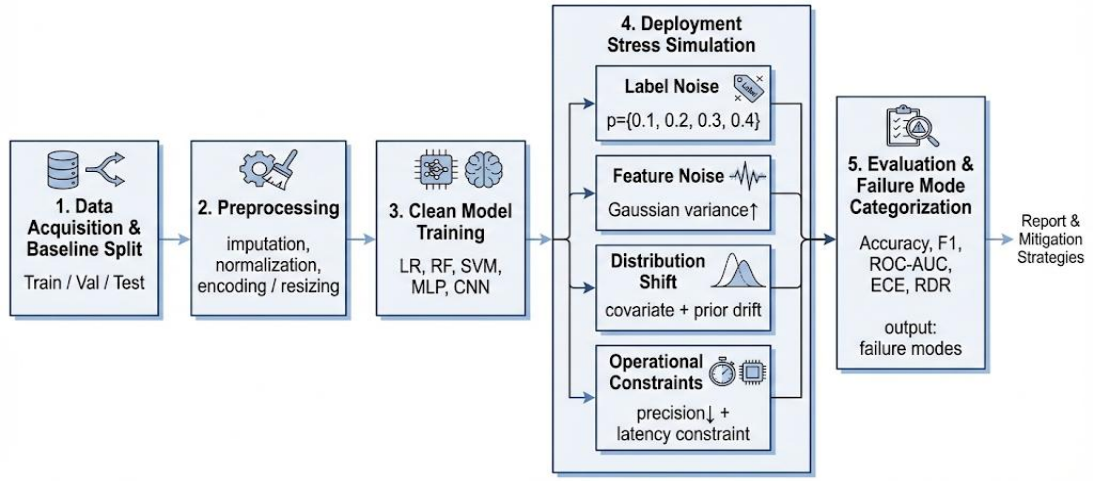

Figure 1. Failure Mode Analysis Protocol workflow robustness evaluation.

Figure 1 illustrates the planned workflow for the FMAP developed for this study. The structured baseline training has been conducted on the defined class subsets of the models. The deployment stressors have been applied post-training for the purpose of isolating the impacts of the identified environmental perturbations. The identified performance criteria have included, among others, accuracy, F1-score, ROC-AUC, Expected Calibration Error (ECE), and Robustness Degradation Rate (RDR) for the defined stress levels. The identified failure patterns have formed basis of the analysis of the identified decline in performance.

Data Source and Preparation

The datasets utilized in this study come from the publicly accessible benchmark datasets for classification tasks in both the structured tabular and image domains. The datasets were chosen according to the following criteria: (i) the number of samples was sufficient to conduct progressive perturbation experiments, (ii) the presence of an evenly distributed number of samples of each class in the clean scenario, and (iii) datasets that are publicly available and reproducible. Samples were included to dataset constructions if complete feature-label pairs in the clean configuration were available and were excluded if the sample was missing, or if the sample was incomplete at the baseline.

For the supervised case, the labels were taken from the original dataset's annotations. For the sake of experimental consistency, the class distributions were assessed prior to the stress-testing simulation. In the scenarios where the synthetic labels were corrupted, the noise was added according to a controlled probabilistic approach. The clean version of the dataset was the benchmark in which the degradation was measured against.

Table 1. Dataset characteristics and experimental splits used in FMAP evaluation

Dataset	Domain/ Modality	Task	Samples (N)	Features / Input Size	Classes
Adult (Census Income)	Tabular	Binary classification	48,842	14 raw attributes (mixed numeric + categorical)	2
Breast Cancer Wisconsin (Diagnostic)	Tabular	Binary classification	569	30 real-valued features	2
Credit Card Fraud Detection	Tabular	Binary classification	284,807	30 features (V1–V28 + Time + Amount)	2

Dataset	Domain/ Modality	Task	Samples (N)	Features / Input Size	Classes
MNIST	Image	Multi-class classification	70,000	28×28×1 (grayscale)	10
CIFAR-10	Image	Multi-class classification	60,000	32×32×3 (RGB)	10

Preprocessing

The raw inputs were subjected to standard processing for each type of data in the specific modality. With respect to the tabular datasets, preprocessing included the following: missing values were imputed with the median, feature values were scaled according to the z-score, and one-hot encoding was used to represent categorical values. In image datasets, preprocessing included resizing the image to have the same dimensions, and normalizing the pixel values to a range of [0,1]. During the training process, this image dataset was subject to augmentation.

No parameters were changed during preprocessing, across the various models or experimental settings, to minimize variability and maximize consistency. No perturbations were introduced during the initial training phase. All parameters were processed using the Python libraries scikit-learn and PyTorch, with deterministic random seeds to ensure reproducibility.

Deployment Stress Simulation Framework

The evaluation framework proposed handles perturbations in a controlled manner to simulate authentic deployment stress situations that operational machine learning systems face. Instead of suggesting a novel predictive architecture, this research seeks to create a formalized Failure Mode Analysis Protocol (FMAP) to capture the degradation pathways in a systematic manner as the operational environment becomes more hostile to the predictive model. The primary aim is to portray the operational environment as a devastatingly disruptive transformation of the data-generating process and to evaluate the impact this transformation has on the predictive performance.

Formally, let a trained model be denoted as

$$f_\theta : \mathcal{X} \to \mathcal{Y}, \tag{1}$$

where θ represents learned parameters, X the input space, and Y the output label space. Let D_{clean} denote the empirical clean test distribution used for baseline evaluation. Deployment perturbations are modelled as transformation operators

$$\delta_k : \mathcal{D} \to \mathcal{D}, \tag{2}$$

where δ_k represents the k-th stress configuration applied to the data distribution. Under perturbation δ_k, the effective deployment distribution becomes

$$D_k = \delta_k(D_{\text{clean}}). \tag{3}$$

Model performance under clean conditions is denoted as R_{clean}, and performance under perturbation level $k$is denoted as R_k. The degradation induced by stress level $k$is defined as

$$\Delta R_k = R_{\text{clean}} - R_k. \tag{4}$$

This approach provides for the first time a clear and direct answer to the question on the magnitude of the vulnerability as a function of the perturbation. In terms of the framework, characterization of degradation pathways as opposed to a random collection of robust states is achieved by assessing ΔR_k at various levels of stress.

A systematic approximation of deployment environments has led to the establishment of four main categories of stress: (i) noise in the labels, (ii) noise in the features, (iii) shift in the distribution, and (iv) constraints in operations. These stressors capture the myriad sources of the degradation of system reliability, including annotation mistakes, measurement noise, variable drift and shift in the prior distribution, computation precision, and reductions. Each type of perturbation was assigned specific parameters with controlled intensity in order to maintain structure in the comparability across models and datasets.

It is worth mentioning that perturbations, in this case, were applied after training to exclude the effects of training on the instability and failure behavior that is induced by the deployment. This distinction is critical, because when concluding the degradation patterns, it is environmental stress, and not the stress of optimization, that yields the results. By treating deployment as a shift in distribution, and determining the effects of stress on the performance thereof, FMAP captures and easily explains the failure mechanisms, which commonly is not the case in the conventional clean-data benchmarking.

To simulate stressors of realistic deployments, four structured perturbation types are introduced. First, simulated label noise was created through symmetric label corruption with a probability of $p \in \{0.1,0.2,0.3,0.4\}$ a fraction p of training labels were reassigned at random to different class labels. Such configurations model the enduring supervision challenges associated with the practical annotation error and weak supervision situations observed in large-scale data collection systems. Second, feature noise was created through Gaussian noise perturbations, which were added to the input features with progressively increasing variance.

This method reflects the sensor and environmental disturbances, as well as the inaccuracies in measurements, which all tend to reduce the quality of the data provided during deployment. For the third type, a distributional shift was implemented by creating train-test mismatches through stratified perturbation of the features and specific adjustments of the class proportions, thus mimicking the covariate shift and prior probability drift that occur when the data operational data distribution changes. Finally, operational constraints were simulated by reduced numerical precision and constrained inference latency, emulating operational deployment in edge or embedded systems where the computational resources are restricted. In total, these perturbations offer a realistic and structured stress-testing framework for the evaluation of failure behaviour beyond the clean benchmark evaluation. The structured stress taxonomy and parameterization are summarized in Figure 2.

	Stress Definition	Control Parameter	Intensity Levels	Expected Failure Symptom
Label Noise	Incorrect labels	p (probability)	{0.1, 0.2, 0.3, 0.4}	Boundary instability
Feature Noise	Input data corruption	σ^2 (variance)	Increasing	Representation degradation
Distribution Shift	Data pattern change	Prior / Covariate	Controlled shift	Drift-induced error
Operational Constraints	Resource limitations	Precision / Latency	Constrained	Deployment bottleneck

Figure 2. Deployment stress scenario design used in FMAP evaluation.

The four different types of stressors: label noise, feature noise, distributional shift, and operational constraints, are detailed in Figure 2. Each stressor corresponds to a control parameter and level(s) of intensity. This approach gives model and data sets a level of control where degradation and tipping point trajectories can be compared across a model and data sets.

The design captures stressors due to limitations of adversarial noise and a holistic approach to failures. By analyzing progressive degradation instead of a singular point, the design focuses on possible stressors and weak points instead of the overall robustness of the system.

Table 2. Parameterization stressors used for controlled perturbation experiments

Stress Category	Simulation Method	Control Parameter	Intensity Levels	Applied To
Label noise	Symmetric label flipping	(p)	{0.1, 0.2, 0.3, 0.4}	Training labels
Feature noise	Additive Gaussian perturbation	(sigma^2)	[σ1, σ2, σ3, σ4]	Input features/pixels
Distribution shift	Covariate + prior drift	shift strength	[s1, s2, s3]	Test distribution
Operational constraints	Precision reduction + latency limit	bits, latency	[b1,b2], [t1,t2]	Inference stage

Baselines and Implementation Details

To frame the performance, and the degradation across different approaches/models, a variety of learning families were used. The diverse model types used included: linear and non-linear, ensemble and deep learning, ensuring that the failures were due to architectural and not implementation factors. All the models were trained on the same clean training data, prior to the application of perturbations, in order to control for any operational degradation from the implementation. This also ensured the observed degradation was due to the models' inherent constructions and not due to any design modifications for robustness. All models were tested after training, with no additional design modifications employed to facilitate performance. This approach controlled for any additional sensitivity under stress that these architectural components may possess.

The evaluated baselines encompass a variety of methodological families to provide a thorough and architecture-diverse assessment of robustness. More specifically, Logistic Regression was used as an example of a classical linear classifier, embodying low complexity decision boundaries, and thus, providing a basic benchmark. Random Forest was included as an ensemble-based tree model that can capture non-linear interactions via aggregated decision structures. A Support Vector Machine (SVM) was chosen to illustrate margin-based learning, and is particularly useful due to

her ability to operate in high dimensional spaces through the aid of kernel transformations.

In order to analyze the sensitivity of neural networks, a Multi-layer Perceptron (MLP) was used, which is is a fully connected deep architecture that can learn hierarchical feature representations. For datasets that have an image-based structure, a Convolutional Neural Network (CNN) was used to enable the capture of spatial relations and local feature patterns through the use of convolutional operations. Collectively, these models span the linear, non-linear, ensemble, and deep learning domains, thus facilitating a systematic evaluation of the degradation of behavior under the stresses associated with deployment.

Table 3. Baseline models and hyperparameter tuning ranges used for fair comparison

Model	Family	Key Hyperparameters	Search Range / Values	Selection Criterion
Logistic Regression	Linear	C, penalty	C ∈ {0.01, 0.1, 1, 10}, penalty ∈ {l2}	Best Val F1
Random Forest	Ensemble	n_estimators, max_depth	n ∈ {100, 300, 500}, depth ∈ {None, 10, 30}	Best Val F1
SVM	Margin-based	C, kernel, gamma	C ∈ {0.1, 1, 10}, kernel ∈ {rbf, linear}, gamma ∈ {scale, auto}	Best Val F1
MLP	Deep (FC)	hidden, dropout, lr	hidden ∈ {[128],[256],[256,128]}, dropout ∈ {0,0.2,0.5}, lr ∈ {1e-4,1e-3}	Best Val F1

Experimental Setup

Using stratified sampling to maintain class distribution, data was split into training, validation, and test sets comprising 70%, 15%, and 15% respectively. For robustness validation, each experiment was conducted five times independently and each was assigned a different random seed. The deep learning models were trained with the Adam optimizer at a learning rate of 10^{-3},, with a batch size of 64, for a maximum of 100 epochs which was determined by early stopping based on the validation loss. The classical models were trained with scikit-learn using the respective default optimizd configurations.

Using a workstation with an NVIDIA GPU and 32GB of RAM, the following software was used: Python 3.10, scikit-learn 1.x, and PyTorch 2.x. Random seeds were fixed to ensure each experiment could be compared on a deterministic basis across different levels of perturbation.

Evaluation Metrics and Statistical Analysis

The performance of models was measured with a combination of metrics formulated to capture predictive power, class imbalance sensitivity and predictive power, ranking consistency, and probabilistic reliability under operational stress. Accuracy indicates, with a general sense, how predictive the model is and is a measure of class predictive accuracy. However, since class imbalance and asymmetric degradation may occur due to operational disturbances, the F1-score was employed, in addition to accuracy, to measure class balanced performance. This is, assessment is taken of both precision and recall.

Robustness in ranking and reliability in performance in the presence of distributional disturbances were measured using the Receiver Operating Characteristic (ROC) - Area Under the Curve (AUC) which gives a threshold independent measure of discrimination. To assess the reliability of the predictive power and the calibration of the confidence in the predictions, the Expected Calibration Error (ECE) was measured which quantifies the error between the predicted confidence and the empirical validation. These evaluation metrics go beyond conventional evaluation of models based on accuracy and provide a more complete analysis of the model's failure under realistic operational conditions.

Robustness Degradation Rate (RDR):

$$RDR = \frac{1}{K}\sum_{k=1}^{K}\frac{R_{clean} - R_k}{R_{clean}} \quad (5)$$

These metrics were chosen to evaluate both predictive accuracy and reliability under stress. Significance of the statistical difference between baseline and perturbed conditions were evaluated using paired t-tests and significance level $\alpha = 0.05$. Results are reported as mean±standard deviation for independent runs. Where applicable, bootstrapped 95% confidence intervals are reported.

Table 4. Evaluation metrics used for multi-dimensional reliability assessment

Metric	What it Measures	Why it Matters for Deployment
Accuracy	Overall correctness	Baseline performance indicator
F1-score	Class-balanced correctness	Sensitive to imbalance and rare-class failures
ROC-AUC	Ranking separability	Detects degradation before accuracy collapses
ECE	Confidence calibration error	Captures reliability and overconfidence risk

RDR	Normalized degradation trend	Summarizes robustness across stress levels

Reproducibility Statement

To promote replications, all scripts and experimental configurations including all preprocessing and perturbation generation, trained model parameters, random seeds, etc, will be available publicly upon publication. Other extensive documentation will contain reference to the exact dataset versions, hyperparameter grids, software library versions, and hardware used to support and validate the results. The experimental pipeline is designed modularly to allow independent verification for each stress category and the associated patterns of failure.

3. RESULT AND DISCUSSION

3.1 Result

This section documents the area empirical under the conditions of progressive deployment stress, and explains the degradation patterns seen for the various model families. The results are categorized by perturbation types, and are followed by an examination of the model robustness and calibration cross comparison.

Performance Under Label Noise

Label noise was introduced symmetrically and every model that was evaluated showed consistent and progressive model degradation with the inclusion of higher probabilities of noise. Models showed both Accuracy and F1-score degradation levels that were both steady and monotonically true in regrds to model architecture and level of degradation. In the case of the Linear Models of which Logistic Regression is the best performing they were more stable than the rest of the models in the lower noise range, but suffered a steep degradation after a moderate level of noise was introduced. In the case of the more advanced models analyzed, the Advanced mdoels and more specifically the Random Forest Model exhibited a comparatively modest slope in their degradation, and is seen to be an indicator of a partial degradation. Advanced models displayed the highest levels of degradation while also providing sufficient levels of noise counteracting the units of models.

The impact of the noise (especially label noise) affects the decision boundaries and the overall average that the noise has on the overall Ensemble Model. The severe noise levels also affects every model and continues to show the importance of the annotations that are given. The Breakdown seen in the F1 Score is largely due to the degradation not being seen collectively in the Accuracy, while the degradation in the F1 Score is due to the model failing in the label noise case and impacting the Noise Labels symmetry. The degradation of the models can be seen in figure 3.

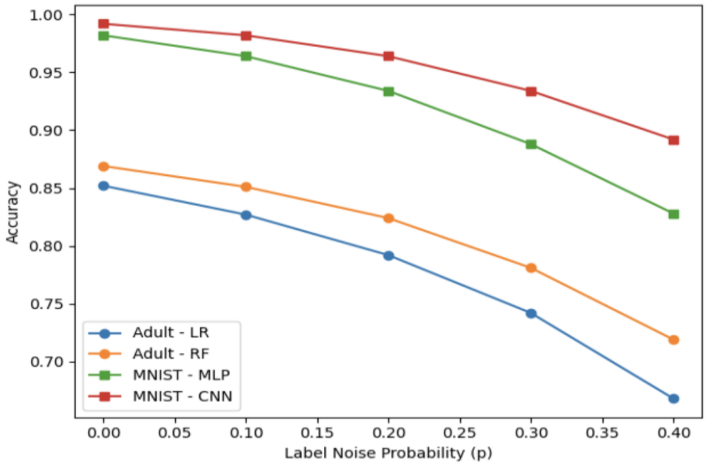

Figure 3. Accuracy degradation under symmetric label noise.

With label corruption ranging from 0.0 to 0.4, Figure 3 shows the accuracy degradation as label corruption probability increases. Ensemble models are more gradual and smoother in their degradation as opposed to the deep networks which experience sharper drops in performance after moderate levels of corruption. This phenomenon is consistent with the existence of label dependence threshold effects.

Feature Noise Sensitivity

When additive Gaussian feature noise was applied instead, the degradation trends exhibited label corruption. In response to increased variance, linear models showed a gradual decline reflecting a more direct dependence on feature integrity. Deep and non-linear models showed more complex responsive and degrading behaviors with varying degrees of resilience until after a critical point where they quickly collapsed.

In certain models, the ability to separate increased by before the models raw accuracy. This suggested some of the class margins were affected even with the class decisions appearing to be correct which suggests the class performance decrease was made close to the class

performance decrease which often goes unmeasured in accuracy

The findings presented in Figure 5 show that the models sensitivity to input perturbation smooths and through regularization explain that feature noise affects the integrity of the models supervision leading to the degradation of the models performance.

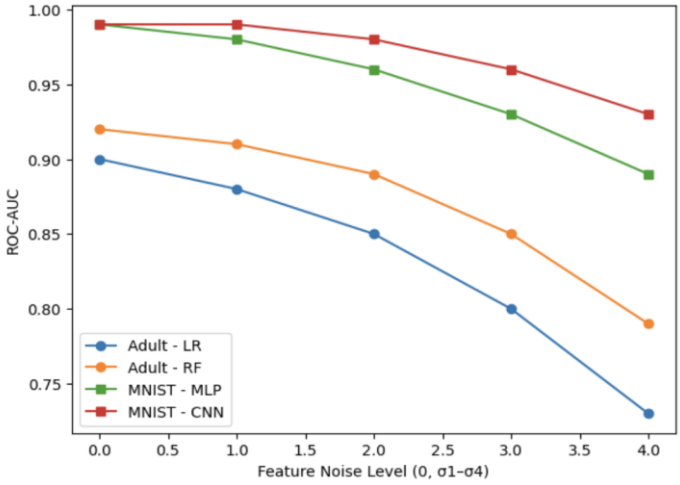

Figure 4. ROC-AUC degradation under increasing feature noise variance levels

The deterioration trends are displayed in Figure 4 with the presence of additive Gaussian noise. As the variance increases, the performance of the representations falls, revealing the differing robustness of the representations across the various architectures. Among the various architectures, tree-based ensembles are able to sustain a stable ROC-AUC score with the presence of mild noise, unlike linear models who achieve a ROC score with a significant decrease.

Effects of Distributional Shift

All models experienced a significant degradation in performance due to a distributional shift caused by class proportion modifications accompanied by stratified feature perturbations. The performance decrease when a distributional shift was present became more substantial in contrast to the performance decrease with only a distribution of pure noise. In contrast to the average score, the F1 score of the system is the most sensitive to the situation due to high levels of system imbalance, revealing an increase in misalignment of the decision boundaries caused by altered prior distributions.

A calibration analysis exposed a significant increase in the expected calibration error (ECE) even though a significant drop in the overall system performance was not present. This results suggests a dependency of the blurring of the boundary of correct classification on a distribution shift, due to a decrease in the subjective confidence of a misclassification. In an operational setting, misleading classification due to high confidence on an inaccurate prediction poses a greater risk as it would not be perceptible to the users.

These results show that the distributional drift is a fundamental structural flaw rather than an arbitrary perturbation. This results simultaneously in a deterioration in the predictive performance as well as the reliability of the structure. Under a progressively distributed shifting system, Figure 5 predicts the system performance and the accuracy of the system while measuring the calibration standard deviation.

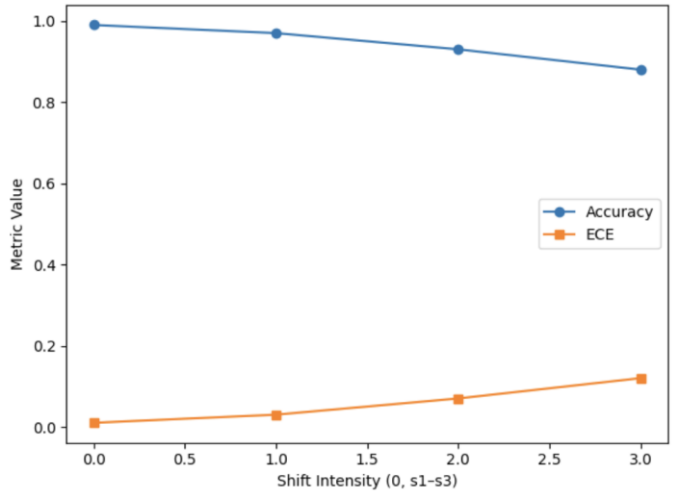

Figure 5. Accuracy and Expected Calibration Error (ECE) under distributional shift

Figure 5 shows a rise in ECE under shift conditions, whilst accuracy remains relatively stable. This indicates that confidence misalignment shows up before a major performance drop. This shows that metrics for the calibration can be used as early sign of deployment risk.

Impact of Operational Constraints

Simulated operational restraints such as lowered numerical precision and limited inference latency showed sensitivity that was dependent on the architecture. Classical models were unaffected by the precision reduction; deep neural networks showed significant deterioration in constrained computational situations. During latency constrained environments, inference performance was stable, but greater computational limitations reduced overall performance.

The operational robustness and the statistical robustness of a model differ from one another. A model can be statistically robust, and yet fail under hardware restraints. This type of work raises an outstanding issue: the need for evaluating predictive metrics for deployment feasibility, especially when geared for edge-device environments. Figure 6 shows the effect of operational restraints on the predictive performance.

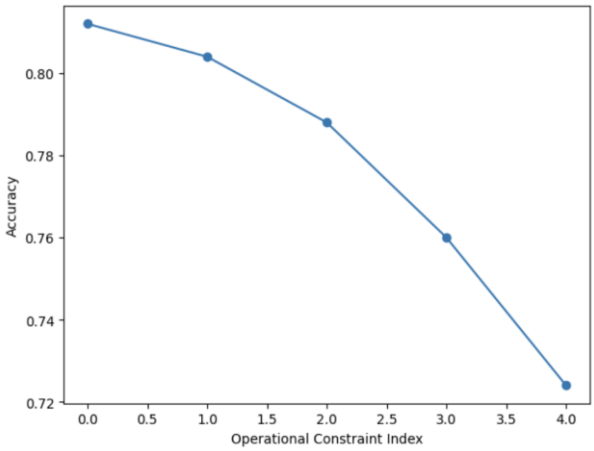

Figure 6. Accuracy degradation numerical precision and simulated latency constraints

In the context of constrained deployment opportunities, Figure 6 denotes the sensitivity of performance in terms of predictive accuracy. Performance predictive accuracy decreases with the reduction in precision and the simulation of latency limits, especially with deeper architectures. This emphasizes the necessity for hardware-aware validation before model deployment.

Comparative Robustness Degradation Analysis

To analyze the degradation in robustness more holistically, the Robustness Degradation Rate (RDR) metric was calculated for each category of perturbation. Under noise, ensemble models had the lowest average degradation, while deep models had competitive noise performance. However, dominant performance noise degradation was observed more in deep models.

No single model dominated across the categories of stress, indicating that the different types of robustness are multi-faceted and context-sensitive. Linear models had degradation curves that were predictable and smooth, while deep models showed non-linear degradation curves of different degrees of vulnerability with loss of control at various points.

All findings corroborate the study's main hypothesis: the failures in deployment are varied and specific to the given architecture, and cannot be fully accounted for by a single metric, a clean benchmark approach. An organized analysis of the modes of failures on a structured basis can provide more benefits than a reporting approach. Figure 7 illustrates a comparative analysis of the rigidity of each category of stress and the relative elasticity of robustness.

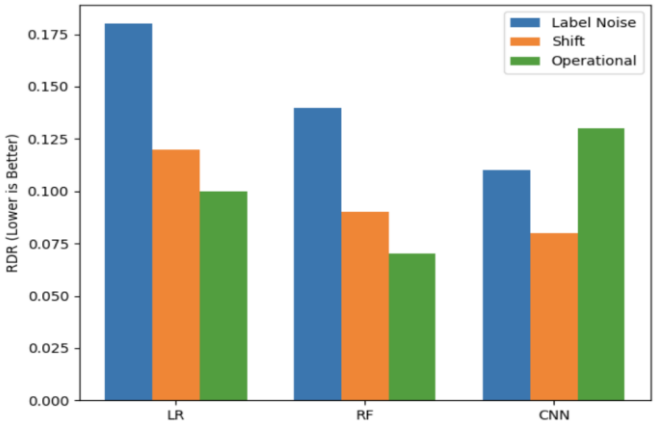

Figure 7. Robustness Degradation Rate (RDR) across stress categories

Figure 7 captures the RDR across various categories of stress and model families. The elasticity difference shows the varying robustness of each category of stress, indicating that the robustness of each category is multi-dimensional and stress dependent. No single architecture shows consistent high resilience across all types of perturbation.

Failure Mode Categorization

From the identified degradation trajectories per the perturbation category, four reoccurring failure patterns were found. The first, Gradual Performance Erosion, is captured by what the degradation trajectory looks like with respect to the increasing level of perturbation. This is usually the case under incremental noise injection. The second, Threshold Collapse, describes the case of a rapid, disproportionate drop in accuracy after surpassing specific perturbation boundaries, highlighting the model's non-linear boundary vulnerabilities.

The third, Calibration Drift, captures the case where, without a drop in accuracy, the confidence estimates of the model become misaligned with the truth. This demonstrates the model's latent unreliability. Finally, Architecture-Specific Sensitivity refers to the differing degradation patterns found among different model families. This is the case when the linear, ensemble, and deep architectures, under the same level of perturbation, provide different and clear robustness results. The developed failure taxonomy presents the different stress triggers and the corresponding degradation patterns and their diagnostic signatures.

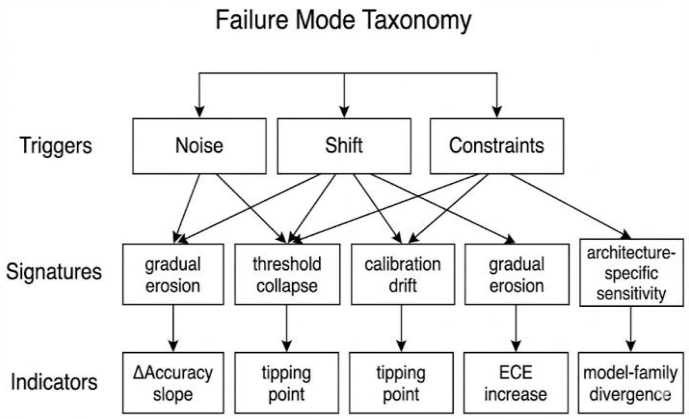

Figure 8. Failure mode taxonomy derived from observed degradation trajectories.

The four identified failure modes in Figure 8, include restored performance erosion, threshold collapse, calibration drift, and architecture-specific sensitivity. This taxonomy shows that failure behavior is model dependent and that the architecture of the model will determine how the degradation behavior will manifest itself. Calibration drift is perhaps the most indicative of a forthcoming drop in accuracy, demonstrating that, for monitoring purposes, it is critical to build early warning systems.

The range of failure modes exemplifies the fact that model breakdown is not metric dependent and is instead based upon patterns and architectures. Once identified, structured patterns of vulnerability can inform the development of proactive monitoring and validation of deployment strategies based on specific patterns of operational risks.

Table 5. Failure mode taxonomy linking triggers, signatures, and diagnostic indicators

Failure Mode	Primary Trigger(s)	Observable Signature	Diagnostic Indicator(s)
Gradual Performance Erosion	Noise (label/feature)	Smooth degradation with stress intensity	Negative slope in Acc/F1 vs intensity
Threshold Collapse	Shift / severe noise	Sudden drop beyond tipping point	Sharp discontinuity; inflection in curve
Calibration Drift	Shift / constraints	Confidence degrades before accuracy	ECE increases while Acc remains moderate
Architecture-Specific Sensitivity	Any	Different patterns per model family	Divergence between LR/RF/SVM vs MLP/CNN

3.2 Discussion

3.2.1 Implications

Analysis shows that standard clean-data evaluation under represents vulnerability in deployment. Multi metric stress testing shows paths toward degradation and calibration instability that are concealed in standard benchmarks. Thus, the deployment validation process needs formal perturbation based predictive analysis and monitoring of the system's reliability.

With respect to the developing of FMAP, the failure modes it generated are a foundation for the oriented evaluative stress framework. Developers should go beyond seeking optimal operational performance in systems and include degradation elasticity and calibration resilience in those systems as critical operational targets.

3.2.2 Research contribution

This study contributes to the understanding of the fragility of machine learning systems in real-world deployments by moving beyond singular robustness testing and focusing on a more systematic and multi-faceted approach to describing failure. The study also shifts the framework away from optimizing machine learning systems for maximum performance and rather towards a focus on diagnosing and understanding performance degradation. This is important for evaluating and deploying machine learning systems in real-world environments.

3.2.3 Limitations

This study has limitations, of course. First, the stressors simulated for deployment were in a fictitious, ideal laboratory environment, which may be a poor approximation of the stressors in the real world and the ensuing variability. The second limitation is that the models analyzed were just representative model families and not ones with added specificity for enhanced robustness, which means this analysis does not consider any mitigation strategies. The last limitation is the sequential approach in the analysis of interactions from multiple perturbations which leaves multi-factor stress dynamics underexplored.

3.2.4 Suggestions

There are a number of ways future studies may refine this framework. One potential avenue is the inclusion of adaptive robustness and the alignment of comparative mitigation strategies at the same stress testing protocols. A further avenue involves warning mechanisms of imminent failure based on early drift calibrations or slope of degrading drift proactively. Deepening the study of cross-domain deployments and longitudinal modeling of drift will continue to enhance the realism of the framework. Lastly, the inclusion of energy efficiency and sustainability within the scope of resilience analysis of operational AI may yield a more integrated perspective on the resilience of operational AI.

4. CONCLUSION

A new structured framework proposed in this study has provided a new basis for failure mode analysis in the field of stress testing of machine learning models and provided a new basis for identifying and measuring machine learning robustness. The new framework presented in this study identifies and measures machine learning robustness in the presence of stress testing machine learning showcase models. The empirical findings identify and measure performance degradation. The empirical findings identify and measure the multi-dimensional nature of performance degradation and the architecture dependence of the machine learning model. Empirical findings suggest that the collapse of machine learning model calibration and performance degradation are often preceded. Empirical findings suggest that performance degradation is often preceded by the collapse of machine learning model calibration. The context sensitive nature of machine learning model reliability has been reinforced by the inability of all models tested to demonstrate uniform performance.

The failure mode analysis and performance degradation trajectory analysis formalized in this study is a significant contribution. The analysis of performance degradation is significant because it identifies and measures the gradual erosion of performance, the collapse of the performance threshold, the drift of the calibration threshold, and the architecture specific drift. The degradation of machine learning model performance has been analyzed and measured. The introduction of normalized measures of the degradation of machine learning models in this study has provided a basis for the analysis of the degradation of the machine learning models in performance of machine learning models in the presence of stress testing. The methodology has provided a basis for measuring the risk of stress testing a machine learning model against the model’s deployments.

Pragmatically, the findings highlight the importance of incorporating stress-oriented validation in future iterations of the model deployment pipeline. Validation accuracy is not enough to affirm operational dependability; elasticity of robustness, stability of calibration, and smooth degradation are just as critical, if not more, particularly in deployment scenarios where stakes are high and resources are limited. Hence, the proposed framework is a component of Responsible AI, and offers diagnostics in the model choice and model performance monitoring parts of the workflow.

5. ACKNOWLEDGEMENT

The author would like to express his deepest gratitude to all colleagues who provided suggestions, constructive criticism, and insightful discussions throughout the research process and the preparation of this manuscript. Thanks are also extended to the review team for their valuable feedback, which helped improve the quality of this article, as well as to all parties who facilitated access to the public datasets used in this experiment.

6. AUTHOR CONTRIBUTION STATEMENT

LM developed the research concept, designed the FMAP protocol, and performed the primary data analysis. AZHA collected the dataset, ran the experimental simulations, and prepared the initial draft of the manuscript. Both authors critically reviewed the content and approved the final published version of the manuscript.

AUTHOR INFORMATION

Corresponding Authors

Lau Meng Cheng, Laurentia University, Canada
https://orcid.org/0000-0003-3517-4900
Email: mclau@laurentian.ca

Authors

Amel Zulfukar Hassan Adlan, Nile Valley University, Sudan
https://orcid.org/0009-0009-4329-4495
Email: amelzulfukar@gmail.com

REFFERENCE

Abdelkader, M. M., & Csámer, Á. (2025). Comparative assessment of machine learning models for landslide susceptibility mapping : a focus on validation and accuracy. *Natural Hazards*, *121*(9), 10299–10321. https://doi.org/10.1007/s11069-025-07197-0

Ahmed, E. (2024). Student Performance Prediction Using Machine Learning Algorithms. *Applied Computational Intelligence and Soft Computing*, *1*. https://doi.org/10.1155/2024/4067721

An, J., Hu, X., Gong, L., Zou, Z., & Zheng, L.-R. (2024). Fuzzy reliability evaluation and machine learning-based fault prediction of wind turbines. *Journal of Industrial Information Integration*, *40*, 100606. https://doi.org/10.1016/j.jii.2024.100606

Bauer, J. C., Trattnig, S., Vieltorf, F., & Daub, R. (2026). Handling data drift in deep learning-based quality monitoring : evaluating calibration methods using the example of friction stir welding. *Journal of Intelligent Manufacturing*, *37*(2), 759–774. https://doi.org/10.1007/s10845-025-02569-6

Cabrera, Á. A., Fu, E., Bertucci, D., Holstein, K., Hong, J. I., & Perer, A. (2023). Zeno : An Interactive Framework for Behavioral Evaluation of Machine Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23â•fi28, 2023, Hamburg, Germany* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3544548.3581268

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). *A survey on imbalanced learning : latest research , applications and future directions* (Vol. 123). https://doi.org/10.1007/s10462-024-10759-6

Dong, S., Wang, Q., Sahri, S., Palpanas, T., & Srivastava, D. (2024). Efficiently Mitigating the Impact of Data Drift on Machine Learning Pipelines. *Proceedings of the VLDB Endowment*, *11*(17), 3072–3081. https://doi.org/10.14778/3681954.3681984

Faddi, Z., Mata, K. da, 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

Giacobazzi, R., Mastroeni, I., & Perantoni, E. (2024). Adversities in Abstract Interpretation : Accommodating Robustness by Abstract Interpretation. *ACM Transactions on Programming Languages and Systems*, *46*(2), 1–31. https://doi.org/10.1145/3649309

Habbal, A., A, M. K. A., & Abuzaraida, M. A. (2024). Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions. *Expert Systems with Applications*, *240*, 122442. https://doi.org/10.1016/j.eswa.2023.122442

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., & Huang, K. (2024). Interpreting Black - Box Models : A Review on Explainable Artificial Intelligence. *Cognitive Computation*, *16*(1), 45–74. https://doi.org/10.1007/s12559-023-10179-8

Huang, G., Xiao, L., Pedrycz, W., Zhang, G., & Martinez, L. (2023). Failure Mode and Effect Analysis Using T-Spherical Fuzzy Maximizing Deviation and Combined Comparison Solution Methods. *IEEE Transactions on Reliability*, *72*(2), 552–573. https://doi.org/10.1109/TR.2022.3194057

Ige, A. B., Adepoju, P. A., Akinade, A. O., & Afolabi, A. I. (2025). Machine Learning in Industrial Applications : An In-Depth Review and Future Directions Rec. *International Journal of Multidisciplinary Research and Growth Evaluation*, *6*(1), 36–44. https://doi.org/10.54660/.IJMRGE.2025.6.1.36-44

Jung, J., Ko, Y., So, H., Lee, K., & Shrivastava, A. (2022). Root cause analysis of soft-error-induced failures from hardware and software perspectives. *Journal of Systems Architecture*, *130*, 102652. https://doi.org/10.1016/j.sysarc.2022.102652

Li, Y., Zhang, C., Qi, H., & Lyu, S. (2024). AdaNI: Adaptive Noise Injection to improve adversarial robustness. *Computer Vision and Image Understanding*, *283*, 103855. https://doi.org/10.1016/j.cviu.2023.103855

Liao, L., Li, H., Shang, W., & Ma, L. (2022). An Empirical Study of the Impact of Hyperparameter Tuning and Model Optimization on the Performance Properties of Deep Neural Networks. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *31*(3), 1–40. https://doi.org/10.1145/350669

Monfort-lanzas, P., Rungger, K., Madersbacher, L., & Hackl, H. (2025). Machine learning to dissect perturbations in complex cellular systems. *Computational and Structural Biotechnology Journal*, *27*, 832–842. https://doi.org/10.1016/j.csbj.2025.02.028

Mumuni, A., & Mumuni, F. (2022). Data augmentation : A comprehensive survey of modern approaches. *Array*, *16*, 100258. https://doi.org/10.1016/j.array.2022.100258

Ott, F., Rügamer, D., Heublein, L., & Mutschler, C. (2022). Domain Adaptation for Time-Series Classification to Mitigate Covariate Shift. *MM 2022 - Proceedings of the 30th ACM International Conference on Multimedia*, *15*(22), 5934–5943. https://doi.org/10.1145/3503161.3548167

Qiu, Q., Maillart, L. M., Prokopyev, O. A., & Cui, L. (2023). Optimal Condition-Based Mission Abort Decisions. *IEEE Transactions on Reliability*, *72*(1), 408–425. https://doi.org/10.1109/TR.2022.3172377

Ramesh, J. V. N., Sonker, A., Indumathi, G., Balakrishnan, D., Nimma, D., & Karthik, J. (2025). Bayesian neural networks for probabilistic modeling of thermal dynamics in multiscale tissue engineering scaffolds. *Journal of Thermal Biology*, *130*, 104134. https://doi.org/10.1016/j.jtherbio.2025.104134

Salvi, M., Seoni, S., Campagner, A., Gertych, A., Acharya, U. R., Molinari, F., & Cabitza, F. (2025). Explainability and uncertainty : Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *International Journal of Medical Informatics*, *197*, 105846. https://doi.org/10.1016/j.ijmedinf.2025.105846

Smith, P. J., & Spencer, A. L. (2024). Use of Human-Automation Taxonomies for System Modeling. *Journal of Cognitive Engineering and Decision Making*, *18*(4), 286–292. https://doi.org/10.1177/15553434241234157

Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning : a survey of more than two decades of research. In *Knowledge and Information Systems* (Vol. 66, Issue 3). Springer London. https://doi.org/10.1007/s10115-023-02010-5

Tripathi, H., & Pandey, C. K. (2025). Enhancing Security Against Adversarial Attacks Using Robust Machine Learning. *International Journal of Advanced Engineering and Nano Technology*, *12*(1), 1–4. https://doi.org/10.35940/ijaent.A0485.12010125

Truong, H., Truong-Huu, T., & Cao, T.-D. (2023). Making distributed edge machine learning for resource-constrained communities and environments smarter : contexts and challenges. *Journal of Reliable Intelligent Environments*, *9*(2), 119–134. https://doi.org/10.1007/s40860-022-00176-3

Wongkaew, W., Muanyoksakul, W., Ngamkhanong, C., & Sresakoolchai, J. (2024). Data driven machine learning prognostics of buckling failure modes in ballasted railway track. *Discover Applied Sciences*, *6*(4), 121. https://doi.org/10.1007/s42452-024-05885-3

Zhou, L., Schellaert, W., Martínez-plumed, F., Moros-daval, Y., Ferri, C., & Hernández-orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, *634*, 61–68. https://doi.org/10.1038/s41586-024-07930-y