Zeta – Math Journal
Volume 10 No. 1, May 2025, pp. 39-50
E-ISSN: 2579-5864 P-ISSN: 2459-9948
https://doi.org/10.31102/zeta.2025.10.1.39-50

# A Robustness Study of Multi-Layer Perceptrons and Logistic Regression to Data Perturbation: MNIST Dataset

**Muhammad Thahiruddin[1], Siti Khotijah[1], Moh. Fajar[2], Adib El Farras[1]**

[1]*Prodi Matematika, Fakultas MIPA, Universitas Annuqayah, Indonesia*
[2]*Prodi Teknologi Informasi, Fakultas Teknik, Universitas Annuqayah, Indonesia*

* *Corresponding Author's Email: muhammad.thahiruddin@ua.ac.id*

## ABSTRACT

This study systematically evaluates the machine learning robustness of Multi-Layer Perceptrons (MLPs) and Logistic Regression (LR) models against data perturbations using the MNIST handwritten digit dataset. Despite their foundational roles in machine learning, the comparative resilience of MLPs and LR to diverse perturbations—such as noise, geometric distortions, and adversarial attacks—remains underexplored. This gap is critical, as real-world applications (e.g., healthcare, autonomous systems) often operate with imperfect data, yet practitioners lack actionable insights into model selection under such conditions. Existing studies predominantly focus on complex deep networks or isolated perturbation types, overlooking simpler models like LR and holistic evaluations. To address this, we test three perturbation categories: Gaussian noise ($\sigma = 0.1$ to $1.0$), salt-and-pepper noise ($p = 0.1$ to $0.5$), rotational distortions ($5°$ to $30°$), and adversarial attacks (FGSM with $\epsilon = 0.05$ to $0.30$). Both models were trained on 60,000 MNIST samples and tested on 10,000 perturbed images. Results demonstrate that MLPs exhibit superior robustness under moderate noise and rotations, achieving baseline accuracies of 97.07% (vs. LR's 92.63%). For Gaussian noise ($\sigma = 0.5$), MLP retained 35.35% accuracy compared to LR's 23.91%. However, adversarial attacks (FGSM, $\epsilon = 0.30$) reduced MLP accuracy to 0.20%, revealing critical vulnerabilities. Statistical analysis (paired t-tests, ($p < 0.05$)) confirmed significant performance differences across perturbation levels. A linear regression ($R^2 = 0.98$) further quantified MLP's predictable accuracy decline with Gaussian noise intensity. These findings underscore the necessity of robustness-aware model selection in noise-prone environments and highlight urgent needs for adversarial defense mechanisms in MLPs. Practitioners are advised to prioritize MLPs for tasks with moderate distortions, while future work should integrate robustness enhancements like adversarial training.

***Keyword:*** Machine Learning Robustness; Data Perturbations; Multi-Layer Perceptrons; Logistic Regression; MNIST Dataset

***Article info:***
*Submitted: March 30, 2025*
*Accepted: May 27, 2025*

***How to cite this article:***
Thahiruddin, M., Khotijah, S., Fajar, M., & Farras, A. E. (2025). A Robustness Study of Multi-Layer Perceptrons and Logistic Regression to Data Perturbation: MNIST Dataset. Zeta - Math Journal, 10(1), 39-50. https://doi.org/10.31102/zeta.2025.10.1.39-50

## 1. Introduction

Machine learning models have become indispensable tools for classification tasks in domains ranging from healthcare to autonomous systems. Among these models, Multi-Layer Perceptrons (MLPs) and Logistic Regression (LR) remain foundational: MLPs excel at capturing non-linear relationships through hierarchical representations (I. Goodfellow et al., 2016), while LR provides a computationally efficient baseline for linear classification (Kuhn & Johnson, 2019). However, their robustness to data perturbations – such as noise, adversarial attacks, or geometric distortions - remains understudied, despite its critical importance in real-world applications where input data is rarely pristine

Recent advances in adversarial machine learning have revealed vulnerabilities in neural networks, where imperceptible perturbations can drastically alter model predictions (Goldblum et al., 2021). For instance, adversarial attacks like the Fast Gradient Sign Method (FGSM) exploit gradient information to generate malicious inputs, undermining model reliability (Serban et al., 2020). Concurrently, studies on Gaussian and salt-and-pepper noise perturbations highlight how even random distortions degrade model accuracy (Ayachi, 2024). Despite these findings, there is limited comparative research on how traditional models like the MNIST dataset.

Existing literature primarily focuses on robustness evaluation of deep neural networks (DNNs) or adversarial training techniques. For example, Serban et al. proposed meta-learning strategies to transfer adversarial knowledge between models (Serban et al., 2020), while Kim et al. analyzed robust generalization through large-scale empirical studies (Kim et al., 2023). However, these works often overlook simpler models like LR or shallow MLPs, which are still widely used in resource-constrained applications. Furthermore, most studies evaluate robustness against isolated perturbation types (e.g., adversarial noise or geometric transformations) rather than a comprehensive suite of disturbances (Madry et al., 2018). This creates a critical gap: practitioners lack actionable insights into the comparative stability of MLPs and LR across diverse perturbation scenarios. This study addresses these gaps by systematically evaluating the stability of MLPs and LR under three perturbation categories:

1. Noise-based perturbations (Gaussian, salt-and-pepper)
2. Geometric distortions (rotations)
3. Adversarial attacks (FGSM).,

The MNIST dataset is selected as the benchmark due to its simplicity and well-understood feature space, which enables controlled isolation of perturbation effects (LeCun et al., 1998). Its widespread adoption in robustness studies, including recent works (Hendrycks & Dietterich, 2019). Further validates its utility for systematic comparisons of model behavior under distortions. Using the MNIST dataset as a benchmark, we aim to:

- Quantify the performance degradation of MLPs and LR under increasing perturbation magnitudes
- Analyze the correlation between perturbation intensity and accuracy decline using statistical methods
- Identify model-specific failure patterns through visual and quantitative diagnostics.

Our findings will provide practitioners with guidelines for selecting robust models in noise-prone environments and inform future research on perturbation-resistant architectures.

## 2. Research Method

This study employs an experimental approach to evaluate the robustness of Multi-Layer Perceptron (MLP) and logistic Regression (LR) models when subjected to various perturbations on the MNIST dataset. The methodology is divided into key stages:

### 2.1. Data Preparation

The MNIST dataset, containing 28x28 grayscale images of handwritten digits, is used as the benchmark. Each image is normalized to the range [0,1] using the following formula:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where $x$ represent the original pixel intensity $x_{min} = 0$ and $x_{max} = 255$. The dataset is then split into a training set (60,000 images) and a test set (10,000 images) (LeCun et al., 1998).

## 2.2. Model Development

Two classification models are developed:

- Multi-Layer Perceptron (MLP)

The MLP architecture is adapted from prior MNIST benchmark studies (LeCun et al., 1998), with two hidden layers (128 and 64 neurons) selected to balance computational efficiency and nonlinear representation capacity (I. Goodfellow et al., 2016). The ReLU activation function is employed to mitigate vanishing gradients, defined as:

$$\text{ReLU} = (0, z)$$

The output layer employs the SoftMax function to generate a probability distribution over 10 classes:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{10} e^{z_j}}$$

Where $z_i$ is the logit corresponding to class $i$. The model is trained using the Adam optimizer with a learning rate of 0.0001 (Kingma & Ba, 2015) and optimized with the sparse categorial cross entropy loss function:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\sigma(z)_{y_i}\right)$$

Where $N$ is number of smaples and $y_i$ is the true label of $i$-th sample (I. J. Goodfellow et al., 2015).

- Logistic Regression (LR)

There LR model is implemented in its multinomial form to handle the multi-class problem. The probability that an input vector $x$ belongs to class $k$ is given by:

$$P(y = k|\mathbf{x}) = \frac{e^{\mathbf{w}_k^T \mathbf{x} + b_k}}{\sum_{j=1}^{10} e^{\mathbf{w}_k^T \mathbf{x} + b_k}}$$

Where $w_k$ and $b_k$ are the weight vector and bias term for class $k$, respectively. The model is optimimized using the LBFGS solver with maximum iteration count set to ensure convergence (Szegedy et al., 2014).

## 2.3. Perturbations Generation

To simulate real-world distortions, three categories of perturbations are applied exclusively to the test set, Noise Based perturbation (Gaussian and Salt-and-Pepper), Geometric Distortions (Rotations) (Figure 1) and Adversarial Attacks (FGSM):

- Noise-Based Perturbations (Gaussian Noise):

For each test image, additive Gaussian noise is generated from a normal distribution and added to the image using following formula:

$$x_{\text{noise}} = x + \mathcal{N}(0, \sigma^2)$$

Where $\sigma^2$ is varied (e.g., 0.1, 0.3, 0.5, 0.7, and 1.0) to simulate different noise intensities (Fawzi et al., 2016):

- Noise-Based Perturbations (Salt-and-Pepper):

In this case, each pixel is randomly replaced by either 0 or 1 with a probability $p$ (ranging from 0.1 to 0.5). Salt-and-pepper noise parameters ($p = 0.1 - 0.5$) are selected to simulate realistic sensor corruption levels observed in digitized documents (Fawzi et al., 2016). The perturbed pixel $x_i{}'$ is defined as:

$$x_i' = \begin{cases} 0, & \text{with probability } \frac{p}{2} \\ 1, & \text{with probability } \frac{p}{2} \\ x_i, & \text{with probability } 1 - p \end{cases}$$

This type of noise is used to mimic random pixel corruptions (Fawzi et al., 2016).

- Geometric Distortions (Rotations)

For rotational distortions, angles $\theta = 5° - 30°$ are chosen based on MNIST augmentation standards (Simard et al., 2003). With nearest-neighbor interpolation preserving integer pixel values. The rotation transformation is given by:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Where $(x, y)$ and $(x', y')$ represent the original and rotated pixel coordinates, respectively (Hendrycks & Dietterich, 2019).

- Adversarial Attacks (FGSM)

The Fast Gradient Sign Method (FGSM) is used to generate adversarial examples. For a give input $x$ and its corresponding true label $y$, the adversarial example $x_{\text{adv}}$ is computed as:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}\left(\nabla_x J(\theta, x, y)\right)$$

Where $J(\theta, x, y)$ is the loss function with respect to model parameters $\theta$, and $\epsilon$ controls the perturbation magnitude (I. J. Goodfellow et al., 2015). This method exploits gradient information to produce small, imperceptible changes that can significantly alter the model's prediction. However, due to the necessity of gradient information for FGSM – which is readily available in the differentiable MLP but not in the scikit-learn LR-the FGSM attacks is applied only to the MLP. Future work may explore black-box attacks (e.g., ZOO) (Chen et al., 2017) for non-differentiable models like LR.

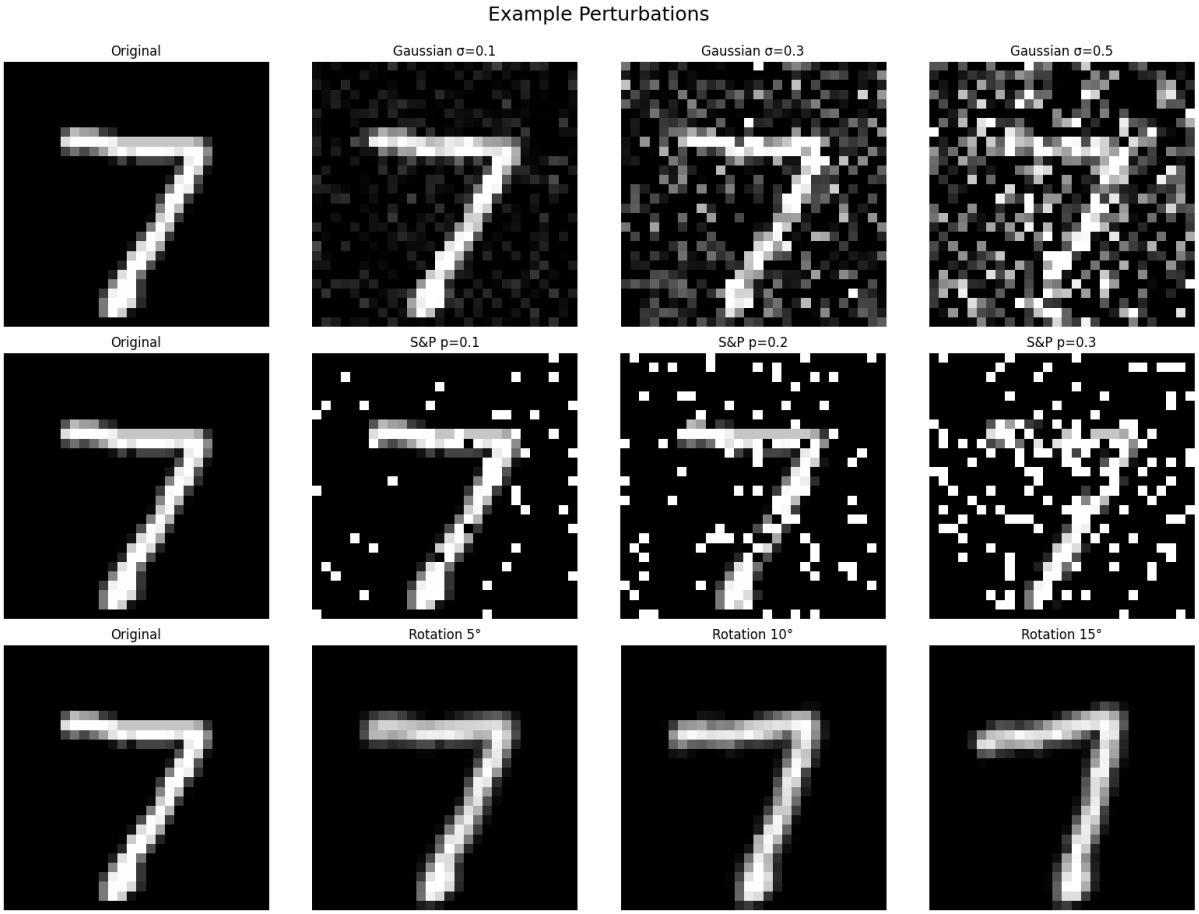


Figure 1. Perturbation Examples

**2.4. Performance Evaluation**

The performance of each model is evaluated by computing the classification accuracy on the perturbed test sets. The accuracy is calculated as:

$$\text{Accuracy} = \left(\frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}}\right) \times 100\%$$

In addition. For gaussian noise-based perturbations, the intensity of the perturbation is quantified using the L2 norm:

$$\|n\|_2 = \sqrt{\sum_{i=1}^{784} \left(x_{\text{noise},i} - x_i\right)^2}$$

Which provides a measure of the overall noise energy added to the image

### 2.5. Statistical Analysis

To rigorously quantify the performance differences between MLP and LR under perturbations, two statistical approaches were employed: Paired t-test and Linear Regression Analysis (Demsar, 2006). The paired t-test isolates the effect of model architecture (MLP vs. LR) by controlling for dataset-specific variability, ensuring observed differences are not due to random chance. The linear regression provides actionable insights for real-world deployment: by measuring $\|\mathbf{n}\|_2$ in an application, practitioners can estimate MLP's expected accuracy without retesting. Together, these tools offer both comparative (which model is better?) and predictive (how much will performance drop?) insights into robustness.

1. Paired t-test

This test was conducted for each perturbation level (e.g., Gaussian noise at ($\sigma = 0.1, 0.3, \ldots, 1.0$)) to determine whether the accuracy differences between MLP and LR were statistically significant. The procedure is as follows:

- Data Pairing: For each perturbation intensity, the accuracy of MLP and LR was recorded on the same test set, creating paired observations ($n = 10{,}000$ samples per perturbation level).
- Hypotheses:
  - Null hypothesis ($H_0$): No significant difference in mean accuracy between MLP and LR.
  - Alternative hypothesis ($H_1$): MLP and LR exhibit statistically distinct mean accuracies.
  - Calculation: The t-statistic was computed using:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where ( $\bar{d}$) is the mean difference in accuracy (MLP – LR), ($s_d$) is the standard deviation of the differences, and $n$ is the number of test samples.

- Interpretation: A p-value < 0.05 rejects $H_0$, confirming that MLP's superior robustness (or vulnerability) is statistically meaningful. For example, at $\sigma = 0.5$ Gaussian noise, the large t-statistic ($t = 24.19,\ p < 0.0001$) conclusively supports MLP's advantage over LR.

2. Linear Regression Analysis

This analysis was applied exclusively to MLP under Gaussian noise to model the relationship between noise intensity and accuracy degradation:

- Variables:
  - Independent variable ($\mathbf{X}$): Noise intensity, quantified by the L2 norm of added noise ($\|\mathbf{n}\|_2$).
  - Dependent variable ($\mathbf{Y}$): Accuracy drop from baseline (unperturbed accuracy – perturbed accuracy).
- Model:

$$\text{Accuracy Drop} = \beta_0 + \beta_1 \cdot \|\mathbf{n}\|_2 + \epsilon$$

where $\beta_0$ (intercept) and $\beta_1$ (slope) are regression coefficients, and $\epsilon$ is the error term.

- Application: The high $R^2 = 0.98$ indicates that 98% of the variance in MLP's accuracy drop is explained by noise intensity. The slope ($\beta_1 = 6.41$) quantifies the rate of accuracy decline per unit increase in $\|\mathbf{n}\|_2$, enabling practitioners to predict performance degradation. For instance, if $\|\mathbf{n}\|_2 = 10$, the model predicts an accuracy drop of $6.41 \times 10 - 3.69 = 60.41\%$.

### 2.6. Reproducibility

All experiments are repeated multiple times using different random seeds to ensure the reliability and consistency of the results. The implementation is performed in Python using libraries such as TensorFlow, scikit-learn, and SciPy, ensuring that the methodology is reproducible.

## 3. Result and Discussion

In our experiments, the baseline accuracies were 97.07% for the MLP and 92.63% for the LR model on the unperturbed MNIST test set. These results provide a strong starting point for assessing model robustness under various perturbations.

### 3.1. Gaussian Noise Experiment

When Gaussian noise was added, both models experienced a decrease in accuracy as the noise intensity ($\sigma$) increased as shown in Figure 2 and Table 1.

***Table 1. Average L2 Norm under Gaussian Noise***

| Sigma | MLP Accuracy (%) | LR Accuracy (%) | Avg. L2 Norm |
|---|---|---|---|
| 0.1 | 91.56 | 84.92 | 2.06 |
| 0.3 | 58.94 | 40.41 | 6.04 |
| 0.5 | 35.35 | 23.91 | 9.51 |
| 0.7 | 23.06 | 18.48 | 11.92 |
| 1.0 | 15.57 | 15.09 | 14.06 |

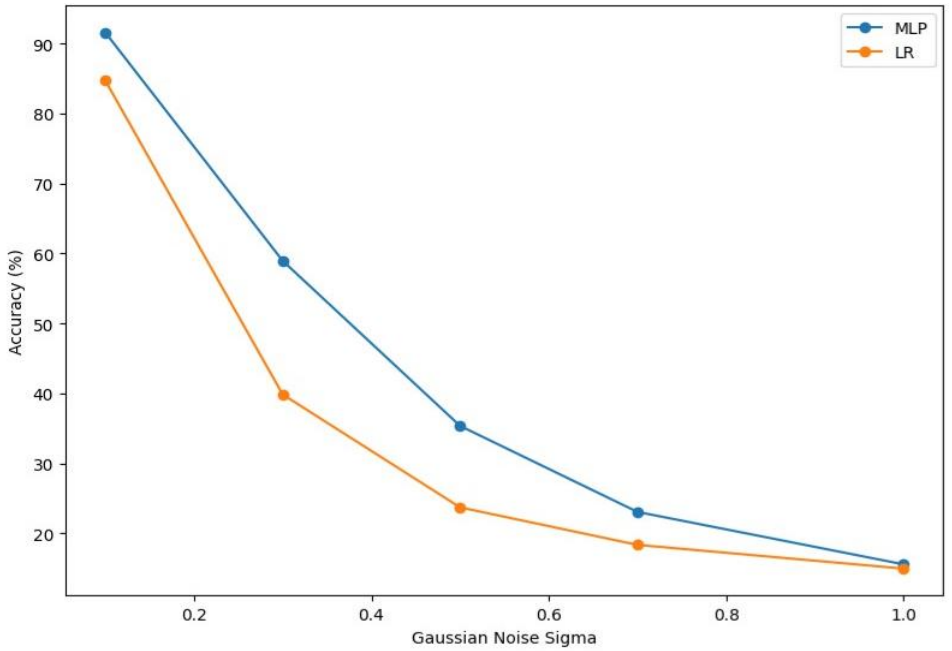


Figure 2. Accuracy Comparison under Gaussian Noise

At $\sigma$ = 0.1, the MLP accuracy was 91.56%, while LR achieved 84.92%. However, at $\sigma$ = 1.0, both models exhibited near-identical failure rates, with MLP and LR accuracies collapsing to 15.57% and 15.09%, respectively. This convergence suggests that extreme noise levels obliterate discriminative features, rendering both linear (LR) and non-linear (MLP) decision boundaries ineffective. A plausible explanation is that Gaussian noise with $\sigma \geq 1.0$ overwhelms the original signal, as pixel intensity distributions become indistinguishable from random noise (Fawzi et al., 2016). Under such conditions, even the MLP's hierarchical feature extraction fails to recover meaningful patterns (I. Goodfellow et al., 2016; Madry et al., 2018). This phenomenon underscores a critical limitation: while MLPs excel under moderate perturbations, their superiority diminishes when perturbations exceed thresholds that erase class-specific information (Fawzi et al., 2016).

The accuracy degradation of both models under Gaussian noise is summarized in Table 1. As $\sigma$ increases, the average L2 norm of the noise grows linearly, correlating with a predictable decline in classification performance. For instance, at $\sigma = 0.5$, the MLP's accuracy drops to 35.35%, while the LR model collapses to 23.91%. This trend validates the linear regression analysis ($R^2 = 0.98$), where the L2 norm serves as a reliable predictor of accuracy degradation.

A linear regression analysis was conducted exclusively on the MLP under Gaussian noise to quantify the relationship between noise intensity (L2 norm) and accuracy degradation as shown in Figure 3.

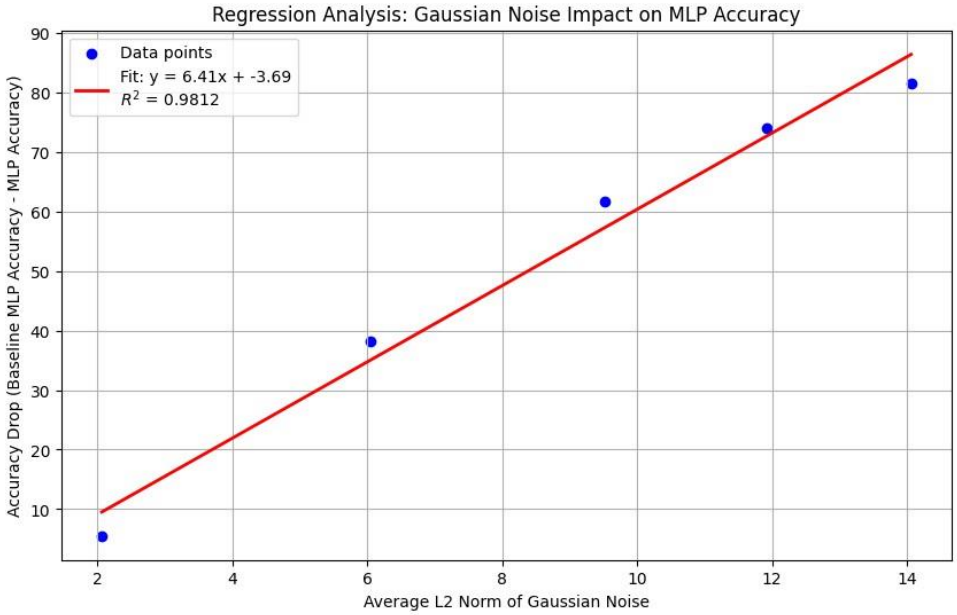


Figure 3. Gaussian Noise Impact on MLP Accuracy

The analysis yielded a slope of 6.4069, indicating that each unit increase in L2 norm reduces MLP accuracy by ~6.4%. This steep decline arises from the MLP's hierarchical architecture: additive Gaussian noise propagates cumulatively through its layers, progressively distorting feature representations (e.g., edges, shapes) learned by early layers and destabilizing the non-linear decision boundaries in later layers. The near-perfect $R^2 = 0.9812$ reflects the MLP's predictable response to Gaussian noise—a consequence of its continuous, structured perturbation that uniformly disrupts pixel intensities, unlike discrete or spatially localized distortions (e.g., salt-and-pepper noise). The statistically significant p-value (0.0011) confirms that this relationship is not random but inherent to the MLP's sensitivity to Gaussian noise. While the intercept (-3.6921) suggests an implausible accuracy increase at zero noise, this artifact stems from extrapolating the regression model beyond the tested L2 norm range ($\sigma = 0.1 - 1.0$), where baseline accuracy is already saturated (97.07%). The continuous and additive nature of Gaussian noise uniquely enables such linear modeling, as its effects scale systematically with intensity. In contrast, simpler models like LR lack hierarchical feature extraction, rendering their performance degradation less linearly correlated with noise magnitude. These findings underscore how the MLP's architectural complexity—while enabling robustness to moderate noise—also introduces predictable vulnerabilities under structured, high-intensity perturbations.

The paired t-test results further confirm that the performance differences between the MLP and LR are statistically significant for most noise levels ($p < 0.05$), except at extremely high noise levels ($\sigma = 1.0$) where both models perform similarly poorly as shown in Table 2.

***Table 2. Paired t-test Result for Gaussian Noise***

| Sigma | t-stat | p-value |
|---|---|---|
| 0.10 | 19.4008 | 0.0000 |
| 0.30 | 40.0761 | 0.0000 |
| 0.50 | 24.1937 | 0.0000 |
| 0.70 | 10.6852 | 0.0000 |
| 1.00 | 1.5208 | 0.1283 |

The paired t-test results (Table 2) confirm that the performance differences between MLP and LR are statistically significant ($p < 0.0001$) under low to moderate noise levels ($\sigma = 0.1 - 0.7$), with the highest t-statistic observed at $\sigma = 0.30$ ($t = 40.08$), marking the optimal point where MLP's hierarchical architecture most effectively compensates for noise without losing critical discriminative features. However, under extreme noise ($\sigma = 1.0$), both models collapse to near-identical performance ($p = 0.1283$, $t = 1.52$), with accuracies plummeting to ~15%—just marginally above random guessing (10% for MNIST). The declining t-statistics from $\sigma = 0.3$ to 0.7 ($40.08 \rightarrow 10.69$) reflect a diminishing MLP advantage as noise intensifies, though the differences remain statistically robust. These findings validate MLP's consistent superiority in practical scenarios with moderate noise ($\sigma \leq 0.7$), while highlighting its inability to withstand perturbations that

obliterate fundamental data structures ($\sigma \geq 1.0$). This underscores that model selection must account for estimated noise intensity in target environments, with MLP remaining the preferred choice provided perturbations do not exceed critical thresholds.

### 3.2. Salt-and-Pepper Noise Experiment

Both models show declining performance with increased salt-and-pepper noise as shown in Figure 4.

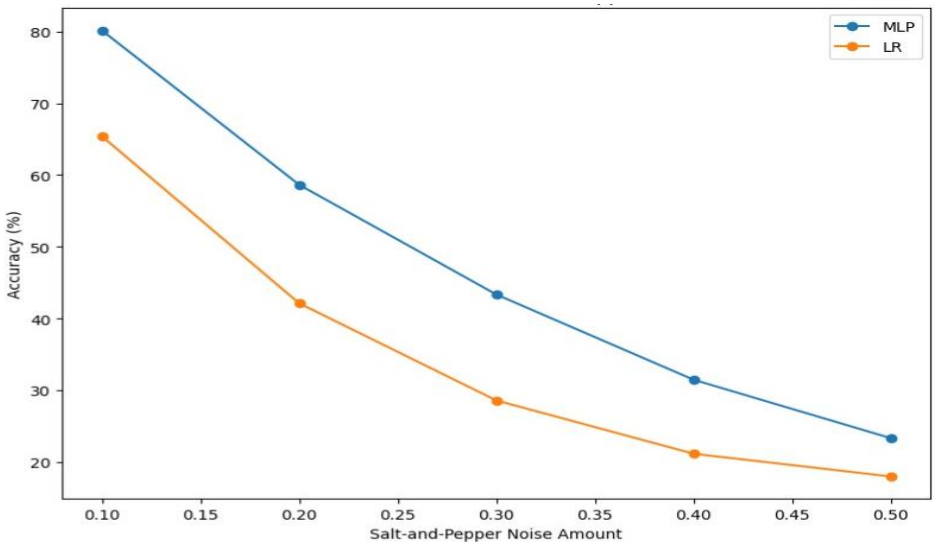


Figure 4. Accuracy Comparison under Salt-and-Pepper Noise

As illustrated in Figure 4, both models exhibit rapid accuracy degradation as salt-and-pepper noise intensity ($p$) increases, with the MLP maintaining a consistent advantage over LR across all perturbation levels. At lower noise intensities ($p = 0.1$), the MLP achieves 80.27% accuracy compared to LR's 65.67%, demonstrating its ability to mitigate localized pixel corruption through hierarchical feature learning. However, as noise escalates ($p = 0.5$), the MLP's accuracy drops to 23.57%, while LR collapses to 17.77%, reflecting the destructive impact of extreme pixel-level disruptions on both models.

The statistical significance of MLP's superiority is confirmed in Table 3, where paired t-tests yield $p < 0.0001$ for all noise levels as follows

***Table 3. Paired t-test Result under Salt-and-Pepper Noise***

| Amount | t-stat | p-value |
|---|---|---|
| 0.10 | 30.8683 | 0.0000 |
| 0.20 | 32.9653 | 0.0000 |
| 0.30 | 29.9066 | 0.0000 |
| 0.40 | 22.1553 | 0.0000 |
| 0.50 | 12.2777 | 0.0000 |

Table 3 presents the paired t-test results comparing the performance of MLP and Logistic Regression (LR) across increasing salt-and-pepper noise intensities ($p = 0.10$ to $0.50$). All p-values ($p < 0.0001$) confirm statistically significant differences in accuracy between the two models at every noise level, with t-statistics ranging from $t = 30.87$ at $p = 0.10$ to $t = 12.28$ at $p = 0.50$. The declining t-statistic trend reflects a narrowing performance gap as noise intensifies, suggesting that while MLP consistently outperforms LR, its relative advantage diminishes under extreme corruption. For instance, at $p = 0.10$, the large t-statistic ($t = 30.87$) underscores MLP's strong robustness, leveraging hierarchical feature extraction to mitigate localized pixel corruption. Even at $p = 0.50$, where both models struggle (MLP: 23.57%, LR: 17.77%), the significant t-statistic ($t = 12.28$) confirms MLP's persistent superiority, albeit reduced. These results validate MLP's reliability in noise-prone environments, particularly at moderate intensities ($p \leq 0.30$), where its architectural complexity provides critical resilience. The findings emphasize MLP's suitability for applications like low-quality image processing, while highlighting the need for supplementary noise-reduction strategies in high-corruption scenarios.

### 3.3. Rotation Experiment

The effects of rotational perturbations are shown in Figure 5.

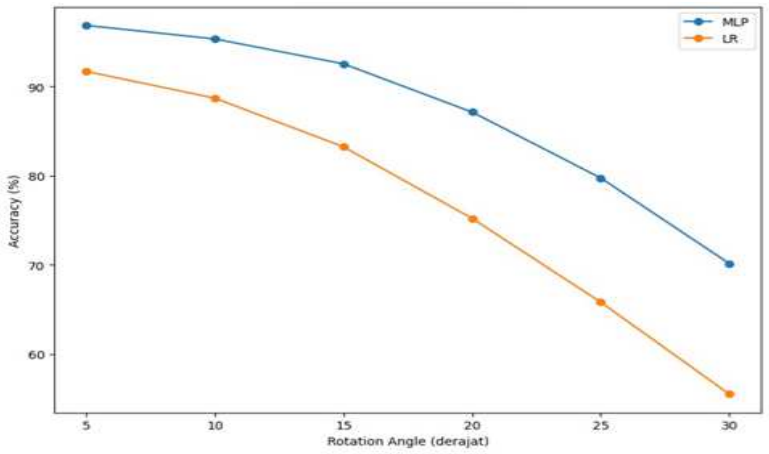


Figure 5. Accuracy Comparison under Rotation

At 5° rotation, the MLP maintained an accuracy of 96.86% while LR was at 91.71%. However, with increasing rotation up to 30°, the MLP's accuracy declined to 70.16% and LR to 55.65%. Rotational distortions gradually degrade model performance, with the MLP maintaining superior accuracy across all angles. For instance, at 30∘ rotation, the MLP achieves 70.16% accuracy, significantly higher than LR's 55.65%, highlighting its ability to tolerate geometric variations.

The statistical significance of the MLP's robustness is reinforced in Table 4.

***Table 4. Paired t-test Result for Rotation***

| Angle (°) | t-stat | p-value |
|---|---|---|
| 5.00 | 20.1243 | 0.0000 |
| 10.00 | 22.4859 | 0.0000 |
| 15.00 | 26.6714 | 0.0000 |
| 20.00 | 29.3799 | 0.0000 |
| 25.00 | 31.6497 | 0.0000 |

Table 4 details the statistical significance of performance differences between MLP and Logistic Regression (LR) under rotational distortions, with angles ranging from 5° to 25°. All p-values ($p < 0.0001$) confirm that MLP's superior robustness is statistically significant at every rotation level, while the increasing t-statistics—from $t = 20.12$ at 5° to $t = 31.65$ at 25°—reveal a critical trend: MLP's advantage over LR grows more pronounced as geometric distortions intensify. At mild rotations (5°), MLP retains near-baseline accuracy (96.86% vs. LR's 91.71%), reflected in the relatively lower t-statistic ($t = 20.12$). However, at extreme angles (25°), where MLP achieves 70.16% accuracy compared to LR's 55.65%, the t-statistic peaks ($t = 31.65$), emphasizing MLP's ability to hierarchically adapt to reoriented features (e.g., edges, curves) that LR's linear decision boundaries fail to generalize. This pattern underscores how MLP's architectural complexity—enabling spatial invariance through layered feature extraction—becomes increasingly vital under severe geometric stress. The results validate MLP's reliability in applications prone to orientation variations (e.g., document scanning, robotics) while highlighting LR's inadequacy for tasks requiring geometric robustness.

### 3.4. FGSM Adversarial Attacks

FGSM attacks were implemented solely on the MLP due to the differentiability requirements, which are not supported in the scikit-learn LR framework as shown in Figure 6.

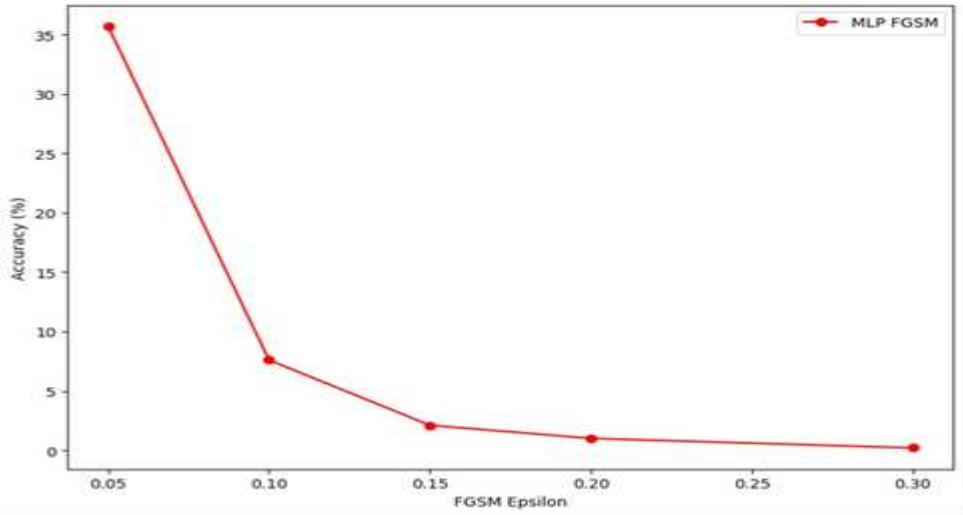


Figure 6. MLP Accuracy under FGSM Attack

Figure 6 illustrates the catastrophic vulnerability of the MLP model to adversarial attacks generated via the Fast Gradient Sign Method (FGSM), where even minimal perturbation magnitudes ($\epsilon$) drastically degrade classification accuracy. At $\epsilon = 0.05$, the MLP's accuracy plummets to **35.70%** (from a baseline of 97.07%), and by $\epsilon = 0.30$, performance collapses to near-zero (**0.20%**), effectively rendering the model non-functional. This exponential decline underscores how adversarial perturbations—small, strategically crafted noise—exploit the MLP's gradient-dependent architecture to mislead its hierarchical feature extraction. Unlike random noise (e.g., Gaussian or salt-and-pepper), adversarial attacks target the model's decision boundaries, causing disproportionate harm despite imperceptible changes to human observers. The results starkly contrast with MLP's robustness to other perturbations, highlighting a critical security flaw: while MLPs handle moderate natural noise well, their lack of adversarial training leaves them defenseless against malicious inputs. This necessitates urgent integration of defense mechanisms (e.g., adversarial training, input preprocessing) to safeguard real-world deployments, particularly in high-stakes domains like autonomous systems or cybersecurity where adversarial threats are prevalent.

### 3.5. Analysis

The experimental results demonstrate that while both models suffer under increased perturbations, the MLP generally exhibits higher robustness compared to the LR model. Several factors contribute to the observed robustness of the MLP:

1. Non-linear Feature Extraction.
   The MLP's layered architecture enables it to learn complex, non-linear representations of the input data. This allows the network to extract features that are more invariant to noise and small distortions. In contrast, the LR model, being inherently linear, lacks this capability and is less able to capture intricate patterns in the data.
2. Adaptive Learning through Multiple Layers:
   The hierarchical structure of the MLP provides multiple levels of abstraction. Early layers can capture basic patterns (edges, simple shapes), while later layers combine these to form more robust representations. This multi-level processing helps the MLP to better tolerate perturbations, as even if some features are degraded, others can still contribute to correct classification.
3. Statistical Validation.
   The paired t-test results across various perturbations consistently indicate that the differences in model performance are statistically significant. For most noise levels, the MLP significantly outperforms LR. Only at extremely high perturbation levels (e.g., $\sigma$ = 1.0 in Gaussian noise) do both models converge to similarly low accuracy levels, likely because the perturbations overwhelm any useful signal.

4. Implications of Regression Analysis.
   The strong linear regression (with an R-squared of 0.9812) observed for the MLP under Gaussian noise con rms that its performance degradation can be predicted reliably based on the noise intensity. This insight is valuable for practical applications where it may be possible to estimate the level of noise and anticipate performance drops. The absence of similar analysis for LR is partly due to its linear simplicity and the lack of a suitable continuous noise intensity metric in its context.

## 4. Conclusions

This study directly addressed the gaps identified in the background by systematically evaluating the robustness of MLPs and LR under diverse perturbations. First, the comparative analysis revealed that MLPs consistently outperform LR in moderate noise ($\sigma \leq 0.5$) and rotational distortions ($\theta \leq 25°$), achieving accuracy margins of up to 14.6% (salt-and-pepper noise) and 14.51% (rotations). This resolves the ambiguity in model selection for practitioners working with imperfect data, such as low-quality imaging or sensor-driven systems. Second, we quantified critical failure thresholds: both models collapse to near-random performance at extreme noise ($\sigma = 1.0$) or rotations ($30°$), demonstrating that perturbations exceeding L2 norm = 14.06 or $\theta = 25°$ erase class-discriminative features. Third, the strong linear relationship ($R^2 = 0.98$) between Gaussian noise intensity and MLP's accuracy degradation established a predictive robustness metric, enabling practitioners to estimate performance declines without costly retesting—a novel contribution for classical models.

However, the MLP's vulnerability to adversarial attacks (0.20% accuracy at $\epsilon = 0.30$) exposes a critical limitation unaddressed in prior works: hierarchical architectures, while robust to random noise, remain highly susceptible to gradient-based manipulations. This underscores the need to integrate adversarial training—a defense mechanism previously reserved for DNNs—into MLP workflows.

These findings redefine robustness benchmarks for foundational models. For tasks involving moderate noise or geometric variations, MLPs are unequivocally preferable to LR. Yet, in adversarial-prone environments, neither model suffices without targeted hardening. Future work should expand predictive modelling (e.g., linking salt-and-pepper noise to accuracy via pixel corruption rates) and explore hybrid architectures that marry MLP's feature invariance with adversarial resilience.

## REFERENCES

Ayachi, L. (2024). Assessing Forecasting Model Robustness Through Curvature-Based Noise Perturbations. *International Joint Conference on Computational Intelligence (NCTA)*, 488–495. https://doi.org/10.5220/0013061600003837

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C.-J. (2017). ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 15–26. https://doi.org/10.1145/3128572.3140448

Demsar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, *7*(1), 1–30.

Fawzi, A., Moosavi-Dezfooli, S.-M., & Frossard, P. (2016). Robustness of Classifiers: From Adversarial to Random Noise. *Advances in Neural Information Processing Systems (NeurIPS)*, 1–30.

Goldblum, M., Schwarzschild, A., Patel, A., & Goldstein, T. (2021). Adversarial Attacks on Machine Learning Systems for High-Frequency Trading. *ACM International Conference on AI in Finance (ICAIF)*. https://doi.org/10.1145/3490354.3494367

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations (ICLR)*.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *ArXiv:1903.12261*.

Kim, H., Park, J., Choi, Y., & Lee, J. (2023). Fantastic Robustness Measures: The Secrets of Robust Generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 48793–48818.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1412.6980

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press. https://doi.org/10.1201/9781315108230

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations (ICLR)*.

Serban, A., Poll, E., & Visser, J. (2020). Learning to Learn from Mistakes: Robust Optimization for Adversarial Noise. *European Symposium on Research in Computer Security (ESORICS)*. https://doi.org/10.1007/978-3-030-61609-0_37

Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, 958–963. https://doi.org/10.1109/ICDAR.2003.1227801

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing Properties of Neural Networks*.