# Informed Repair of Deep Neural Networks

**Akshat Adsule**
University of California, Davis
aadsule@ucdavis.edu

**Darroll Saddi**
University of California, Davis
dwsaddi@ucdavis.edu

**Suyash Goel**
University of California, Davis
sngoel@ucdavis.edu

## Abstract

Deep neural networks (DNNs) have become increasingly prevalent in modern applications. DNNs are widely used in safety-critical settings such as air traffic control, healthcare, and self-driving vehicles. However, these networks are not infallible; they are susceptible to adversarial attacks, noisy input, and other corruptions that can have disastrous outcomes if not addressed. Recent approaches have introduced methods to provably repair DNNs over a defined edit set, which provide formal guarantees within the repaired region. However, these methods do not offer guidance on which internal components or edit sets to repair. To address these gaps, we present Informed Repair of DNNs (IRDNN), a framework for using heuristics to guide the editing process. IRDNN explores strategies such as activation-based repair, where the edit region is determined by where neuron activations are the highest over the repair set. Our goal is to enable more robust, generalizable, and predictable DNN repair through principled heuristic guidance.

## 1 Introduction

Recent research [1–3] has explored methods of repairing DNNs once a set of incorrect inputs is identified. These techniques typically aim to minimally adjust the network's parameters to correct its behavior on the specified repair set, often while providing formal guarantees on the outcome for those inputs or related input regions. These methods have the added benefits of formal guarantees, generalizability, scalability, architecture-preservation, and local repair. However, applying these repair methods in practice reveals significant ambiguities that can affect the quality, generalization, and efficiency of the repair process.

Take, for example, APRNN, the method proposed in [3] that offers provable, architecture-preserving repair over specified input regions (V-polytopes). APRNN achieves this guarantee, in part, by ensuring the network behaves linearly within the target region up to a chosen layer and then formulating the repair as a solvable linear problem. This involves modifying network weights starting primarily at that chosen layer and adjusting biases in subsequent layers.

While effective, this process introduces critical ambiguities in practice. One such ambiguity comes from selecting the starting layer for weight modifications. The paper itself does not prescribe how to choose this layer, and this choice can significantly impact the repair's effectiveness, efficiency, and generalization. Additionally, how a repair set is selected greatly influences the quality of the repair. In APRNN, the specific examples included in the repair set, which define the repair polytope, fundamentally determine the target behavior. The composition and scope of this set affect the repair outcome and how well the fix generalizes to similar, unseen inputs.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

This project aims to investigate and develop heuristics to guide the DNN repair process, specifically addressing the ambiguities highlighted above in methods like APRNN. By providing data-driven or structurally informed approaches for making these choices, we seek to enable more efficient, effective, and informed repairs of DNNs.

This work contributes to enhancing AI trustworthiness by making the crucial process of model repair, an essential step for restoring trust after failure, more principled and transparent. Current repair techniques often rely on arbitrary decisions, which can lead to unpredictable results. By introducing heuristics to guide key decisions, such as selecting the optimal network layer for modification or constructing meaningful repair sets, we aim to enable more systematic, reliable, and interpretable correction of model flaws. This ultimately increases confidence in the robustness and safety of AI systems by ensuring that necessary fixes are applied with clearer understanding and reduced risk.

## 2 Approach

We propose Informed Repair of DNNs (IRDNN), a framework for guiding the DNN repair process through two classes of heuristics: (1) for selecting the repair region within the network, and (2) for constructing the repair set from misclassified examples. These heuristics aim to make the repair process more targeted, effective, and generalizable.

Layer Selection Heuristics:

**Activation-Based** Choose the layer with the highest average activation magnitude or activation variance across the repair set, reflecting strong or unstable feature processing.

**Gradient/Sensitivity-Based** Select layers whose parameters exhibit the largest gradient norms with respect to the loss on the repair set, indicating high sensitivity to input perturbations.

**Feature-Similarity Based** Choose the layer where internal representations of repair set inputs are most similar, suggesting a semantically unified feature space relevant to the fix.

**Layer Type/Position** Use simple structural heuristics, such as always selecting the first fully connected layer after convolutional blocks, or the penultimate layer before classification.

**Change-Based (Adversarial Repairs)** Identify the layer where feature representations differ most between clean and adversarial inputs, indicating a breakdown in robustness.

**Neuron/Path Contribution Profiling** Run inference over the repair set (or defined input polytope) and record neuron activations across layers. Aggregate these to identify layers where neurons consistently fire or show strong contributions. Prioritize layers with high contribution density as effective intervention points.

Repair Set Analysis

**Diversity** Evaluating the diversity of the repair set, such as the number of unique classes or the distribution of inputs across the input space.

**Concentration** Analyzing the concentration of points defining the polytope, such as the number of points needed to define a convex hull or the dimensionality of the convex hull.

**Size** Considering the size of the repair set, including the number of points and the dimensionality of the input space.

This project enhances AI trustworthiness by making the crucial process of model repair, itself a method for increasing trust after failures—more efficient and informed. Current repair techniques can involve arbitrary choices, leading to unpredictable outcomes. By developing heuristics to guide decisions within the repair process, such as selecting the optimal network layer for modification, we enable more systematic, reliable, and effective correction of identified flaws. This ultimately increases confidence in the robustness and safety of AI systems by ensuring that necessary fixes are applied more predictably and with a better understanding of their potential impact.

## 3 Related Work

Extensive work has been done on provably repairing deep neural networks given an edit set, with methods such as APRNN [3] and PRDNN [2]. These approaches guarantee that, if repair is successful,

the modified network will produce the correct output on the specified inputs or input regions. However, they do not provide guidance on how users should select internal repair targets, such as which layers, weights, or biases to modify, despite the fact that these decisions can significantly affect the repair's generalization, stability, and drawdown. Additionally, they offer no heuristics for edit set construction, even though the quality of the repair is highly dependent on the examples used to define the repair region (e.g., a V-polytope in APRNN). More recent approaches incorporate heuristic guidance into the repair process. VeRe [4] uses linear relaxation to estimate how changes to individual neuron activations affect the network's output, assigning a repair significance score to each neuron based on its potential to correct misclassified inputs. This score is used to guide iterative neuron-level repairs. INNER [5] takes an interpretability-driven approach: it identifies anomalous neurons whose attribution scores are high for corrupted inputs but normal for clean ones. It then repairs these neurons by reducing their influence on incorrect predictions while preserving their behavior on correctly classified inputs.

# 4 Experimental Setup

The experiments are designed to answer the primary research question: *How do different heuristics for layer selection (e.g., activation-based, gradient-based) and repair set analysis (e.g., diversity, concentration) influence the effectiveness, generalization, locality, and efficiency of DNN repair techniques like APRNN?* We aim to determine which heuristics provide the most significant improvements over arbitrary or naive selection strategies.

## 4.1 Models & Datasets

### 4.1.1 Selected Architectures

We aim to identify heuristics for a wide variety of model types and repair types. We choose models that are frequenty used in most machine learning tasks. These include:

**Multi-Level Perceptrons (MLPs)** MLPs are foundational neural networks with hidden layers, fully connected neurons, and non-linear activations used widely for classification and regression tasks.

**Convolutional Neural Networks (CNNs)** CNNs excel at processing grid-like data, especially images, by using convolutional layers to learn spatial feature hierarchies. CNNs are most used for vision tasks such as classification, object detection, and segmentation.

### 4.1.2 Selected Models

For experimentation purposes, we choose the following pre-trained models for each aforementioned model architecture. We picked models that are relativly well-established for their respective tasks. We also purposefully choose smaller models for ease of experimentation. However, we still expect our results to apply to larger and more complex models.

Our chosen models are:

**Clustering MLP** To simulate a classic use of MLPs, we implemented a simple supervised clustering problem, we can create a simple MLP model that is trained to classify points in a 2D space into two clusters.

**AlexNet [6]** AlexNet is a CNN architecture that achieved remarkable success in the ImageNet competition by using deep learning techniques, including ReLU activations and dropout for regularization.

### 4.1.3 Selected Datasets

We will use the following datasets for our experimentation that correspond to the selected pre-trained models.

**Custom Clusering Dataset** We will use the `make_moons` function from `sklearn.datasets` to generate a dataset, where an MLP will be insufficiently trained on this dataset to classify

the points into two clusters. By either increasing the complexity of the dataset (e.g. the number of points, noise), or by not training the MLP enough, we can apply our heuristics to see if any significant improvements can be made, or if there are any significant differences between the heuristics. By training an MLP on a supervised (known labels) clustering task, we can create a simple MLP model that is trained to classify points in a 2D space into two clusters.

**CIFAR-10 [7]** CIFAR-10 is a widely used image dataset consisting of 60,000 32x32 color images in 10 classes, with 6,000 images per class. CIFAR-10 is used by both AlexNet.

#### 4.1.4 Summary

The following table summarizes the models and datasets we use in our experimentation.

| Architecture | Model | Dataset |
|---|---|---|
| MLPs | Custom Clustering Model | Custom Clustering Dataset |
| CNNs | AlexNet [6] | CIFAR-10 [7] |

### 4.2 Evaluation Metrics

We will evaluate the heuristics based on the following metrics:

**Repair Success Rate** The percentage of inputs in the repair set for which the model produces the correct output after repair.

**Generalization** The model's performance on unseen inputs similar to those in the repair set, measured by accuracy or other relevant metrics.

**Locality** The extent to which the repair minimally affects the model's behavior on inputs outside the repair set, often quantified by changes in the model's predictions or parameter values.

**Efficiency** The computational cost of the repair process, including time and resources required.

**Scalability** The ability of the repair method to handle larger models or repair sets without significant degradation in performance or efficiency.

## 5 Results and Discussion

We conducted comprehensive experiments to evaluate the effectiveness of different heuristic combinations across two distinct architectures and problem domains: Multi-Layer Perceptrons (MLPs) on a 2D clustering task and Convolutional Neural Networks (CNNs) on image classification. Our experimental design allows us to understand how heuristic performance varies with model complexity and problem characteristics.

### 5.1 MLP Heuristic Evaluation on 2D Classification

We first evaluate our heuristics on a controlled 2D binary classification problem using a simple MLP trained on the sklearn moons dataset. This controlled setting provides clear geometric intuition about decision boundaries and allows for direct visualization of repair effects.

#### 5.1.1 Experimental Setup for MLP Analysis

Our MLP experiments tested 9 edit heuristics against 10 set heuristics, creating 90 possible combinations. The baseline model achieved 93.67% accuracy, providing a clear reference point for measuring improvement or degradation. We successfully evaluated 71 out of 90 combinations, with 19 combinations failing due to optimization difficulties or constraint violations.

The tested edit heuristics included both traditional approaches (SingleLayer, FromLayer, ActivationBased, WeightsActivationBased) and novel methods (GradientBased, LayerImportance, NeuronPruning, Adversarial, Random). Set heuristics ranged from simple misclassification-based selection to sophisticated approaches like ConfidenceBased and BoundarySamples.

### 5.1.2 Key Findings from MLP Experiments

Our MLP experiments revealed several critical insights about heuristic effectiveness:

**Superior Performance of WeightsActivationBased:** The WeightsActivationBased edit heuristic emerged as the clear winner, appearing in the top-performing combination (WeightsActivationBased + ConfidenceBased) that achieved 97.33% accuracy—a substantial +3.66% improvement over the baseline. This heuristic consistently outperformed others across multiple set heuristics, demonstrating robust effectiveness.

**Importance of Targeted Set Selection:** The ConfidenceBased set heuristic proved highly effective, enabling the best overall performance when paired with appropriate edit heuristics. This suggests that focusing on low-confidence predictions provides a more principled approach to identifying repair targets than simple misclassification-based methods.

**Parameter Editing Intuition for Simple Models:** The success of parameter modification heuristics aligns with intuitive understanding of how simple neural networks function. For MLPs with clear decision boundaries, targeted weight adjustments can effectively shift these boundaries to correct misclassifications without significantly disrupting the overall learned representation. This is analogous to adjusting coefficients in a polynomial function to better fit data points.

**Failure Patterns and Robustness:** Several combinations failed entirely, particularly those involving the Adversarial edit heuristic (8 failures) and the FullDataset set heuristic (all 9 combinations failed). This suggests that some approaches may be fundamentally incompatible with certain repair formulations or may require different constraint parameterizations.

**Performance Distribution:** Successful repairs showed a wide range of outcomes, from significant improvements (+3.66%) to substantial degradations (-88.0%). This variability underscores the importance of principled heuristic selection rather than arbitrary choices.

## 5.2 CNN Analysis on Image Classification

We conducted complementary experiments on AlexNet trained on CIFAR-10 to understand how heuristic effectiveness scales to more complex architectures and higher-dimensional problems.

### 5.2.1 Deterministic vs. Stochastic Repair Strategies

**Deterministic ("full-set") repair** in which we solve a single repair problem over an entire edit set of varying size and homogeneity.

**Stochastic ("incremental") repair** in which we perform repeated, small-batch repair passes—carrying forward the model between passes—so as to trade batch size, number of passes, and homogeneity against overall fidelity.

Our CNN experiments used a pretrained AlexNet model with baseline test accuracy of 72.08% on CIFAR-10, allowing us to measure how each strategy trades off global test accuracy against local repair accuracy.

### 5.2.2 Repair Set Size Impact

Our experiments demonstrate that larger repair sets consistently lead to worse performance across all metrics. As repair set size increases from 10 to 200 examples, we observe dramatic degradation in success rates, increased accuracy drawdown, and longer repair times.

The data shows clear trends across all metrics:

- **Success Rate:** Drops from 100% for size 10 to just 9.1% for size 200
- **Accuracy Drawdown:** Increases from minimal impact ( -2.7%) for small sets to severe degradation (-50.8%) for large sets
- **Repair Time:** Scales linearly from 10 seconds to 50 seconds as set size increases

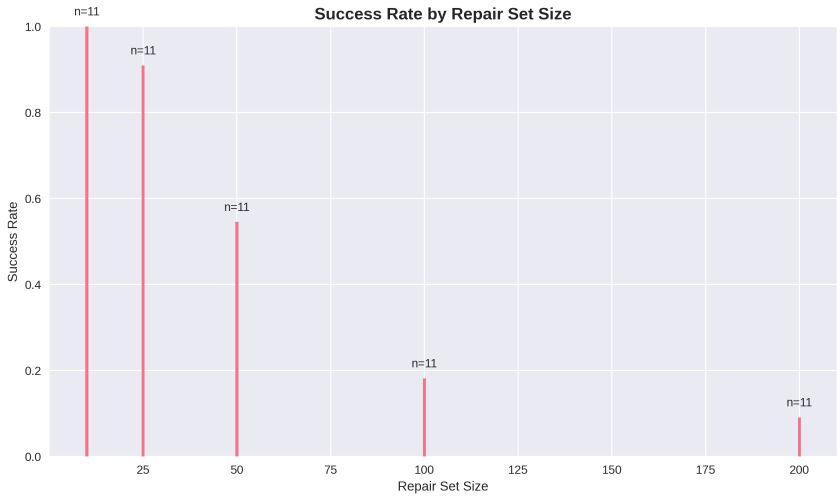


Figure 1: Success rate by repair set size, showing dramatic degradation in repair feasibility as set size increases beyond 25 examples.

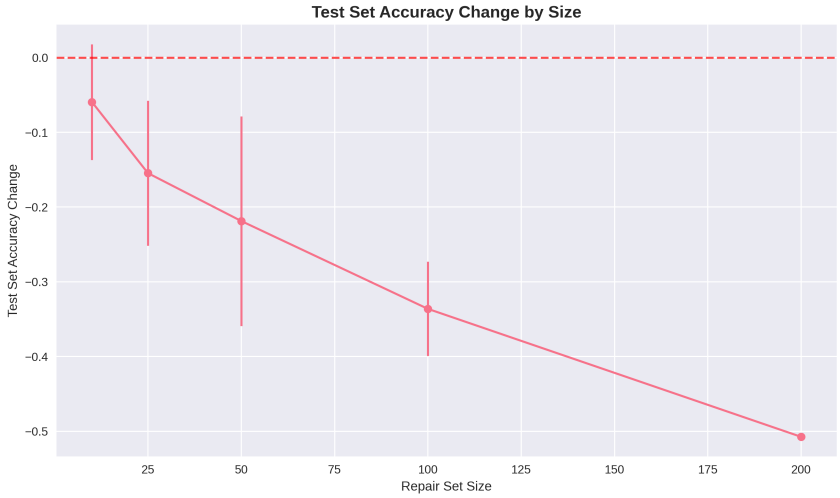


Figure 2: Test set accuracy drawdown by repair set size, demonstrating the severe trade-off between local repair success and global model performance as repair set size increases.

#### 5.2.3 Repair Set Homogeneity Effects

Repair set composition significantly impacts both success rates and performance degradation patterns. Class-homogeneous sets consistently outperform misclassified sets across all metrics.

#### 5.2.4 Stochastic Repair Efficiency for Class-Homogeneous Sets

Our stochastic experiments focused on batch sizes of 5 and 10, as batch size 20 resulted in too many infeasible repairs. For class-homogeneous sets, stochastic repair proves highly efficient, achieving rapid convergence to near-perfect repair set accuracy.

#### 5.2.5 Early Stopping to Reduce Drawdown

The key insight is that stochastic repair converges to near-perfect repair set accuracy within approximately 5 iterations. Stopping early can significantly reduce global accuracy drawdown while maintaining excellent local repair performance.

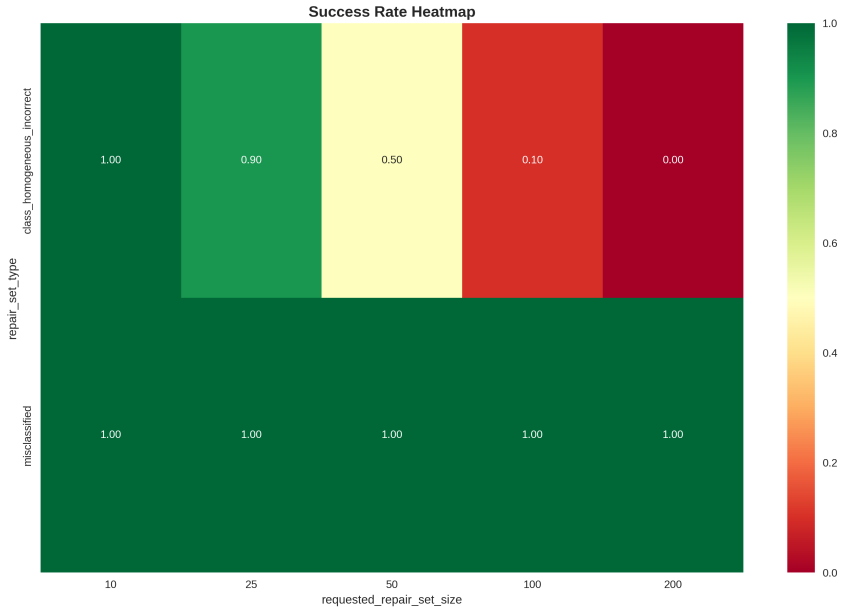


Figure 3: Success rate heatmap showing the interaction between repair set type and size, clearly illustrating the feasibility boundaries for different repair strategies.

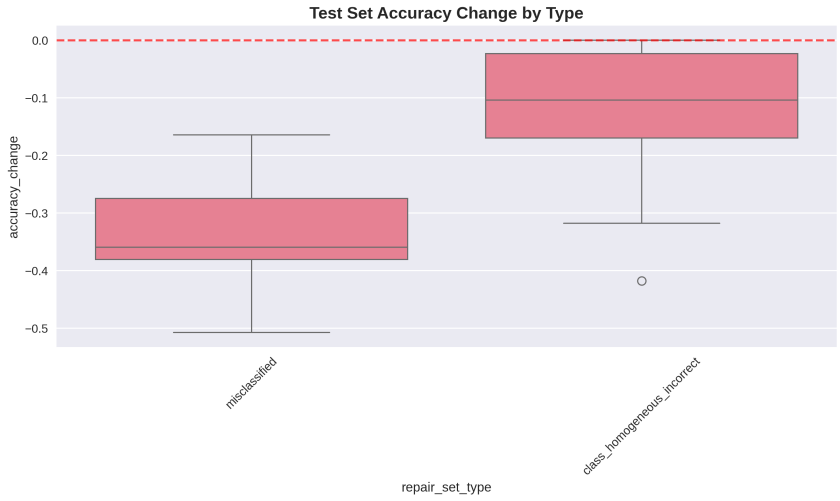


Figure 4: Accuracy drawdown distribution by repair set type, demonstrating that class-homogeneous repairs cause significantly less performance degradation than misclassified repairs.

Key findings from our analysis show that:

- **Rapid Convergence:** Class-homogeneous repairs reach 99% repair set accuracy within 5 iterations
- **Early Stopping Benefit:** Stopping at iteration 5 preserves 60-65% test accuracy vs. 45-29% when running to completion
- **Batch Size Efficiency:** Both batch sizes 5 and 10 show similar convergence patterns, with batch 5 being slightly more stable

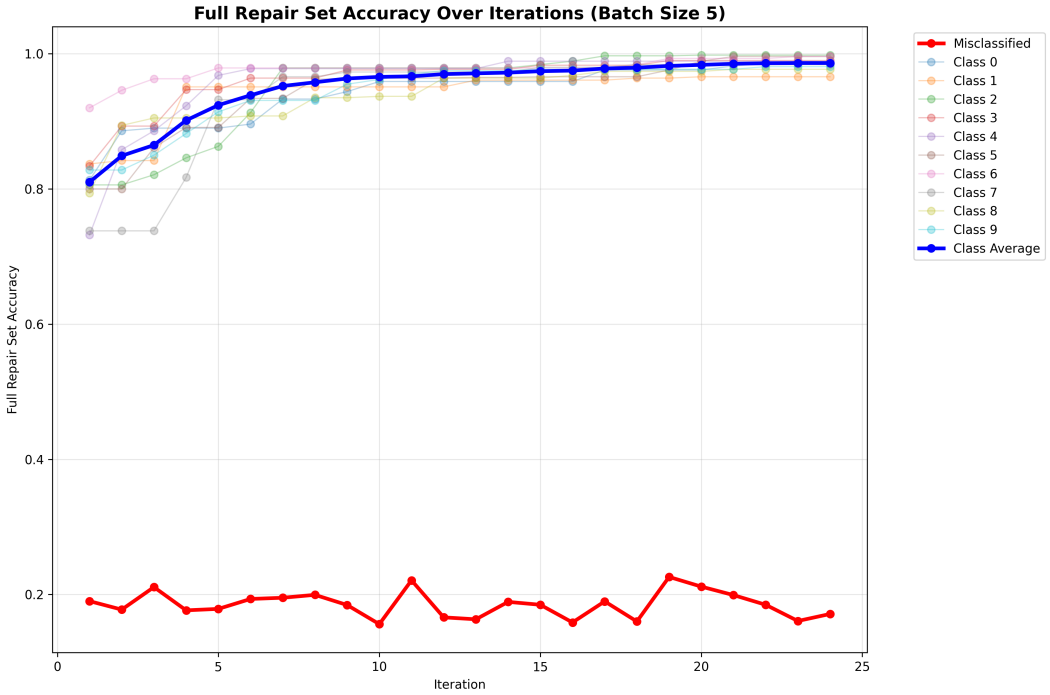


Figure 5: Repair set accuracy evolution for batch size 5, showing rapid convergence to near-perfect accuracy within 5 iterations for class-homogeneous repairs.

### 5.3 Cross-Architecture Analysis and Implications

Our comprehensive evaluation across both simple MLP models and complex CNN architectures reveals fundamental insights about heuristic effectiveness and the nature of neural network repair.

#### 5.3.1 Architecture-Dependent Heuristic Performance

**Model Complexity and Repair Sensitivity:** Our MLP experiments demonstrate that simpler models with clear geometric decision boundaries respond well to targeted parameter modifications. The WeightsActivationBased heuristic's success (achieving +3.66% improvement) illustrates how activation-guided weight adjustments can effectively reshape decision boundaries without catastrophic interference. This mirrors the intuitive process of adjusting polynomial coefficients to better fit data points—small, targeted changes to critical parameters can yield significant improvements in model behavior.

**Scaling Challenges in Complex Models:** In contrast, our CNN experiments on CIFAR-10 reveal the increased difficulty of repair in high-dimensional parameter spaces. The dramatic performance degradation observed in AlexNet repairs (up to -50.8

**Set Heuristic Effectiveness Patterns:** The ConfidenceBased set heuristic's superior performance in MLP experiments (+3.66% with WeightsActivationBased) demonstrates the value of principled sample selection. Low-confidence predictions often correspond to points near decision boundaries, making them ideal targets for local repairs that preserve global model behavior.

#### 5.3.2 Fundamental Trade-offs in Neural Network Repair

Our experiments reveal several critical trade-offs that practitioners must navigate:

1. **Model Complexity vs. Repair Feasibility:** Simpler models (MLPs) show higher repair success rates and more predictable outcomes, while complex models (CNNs) exhibit greater sensitivity to parameter modifications and higher failure rates.

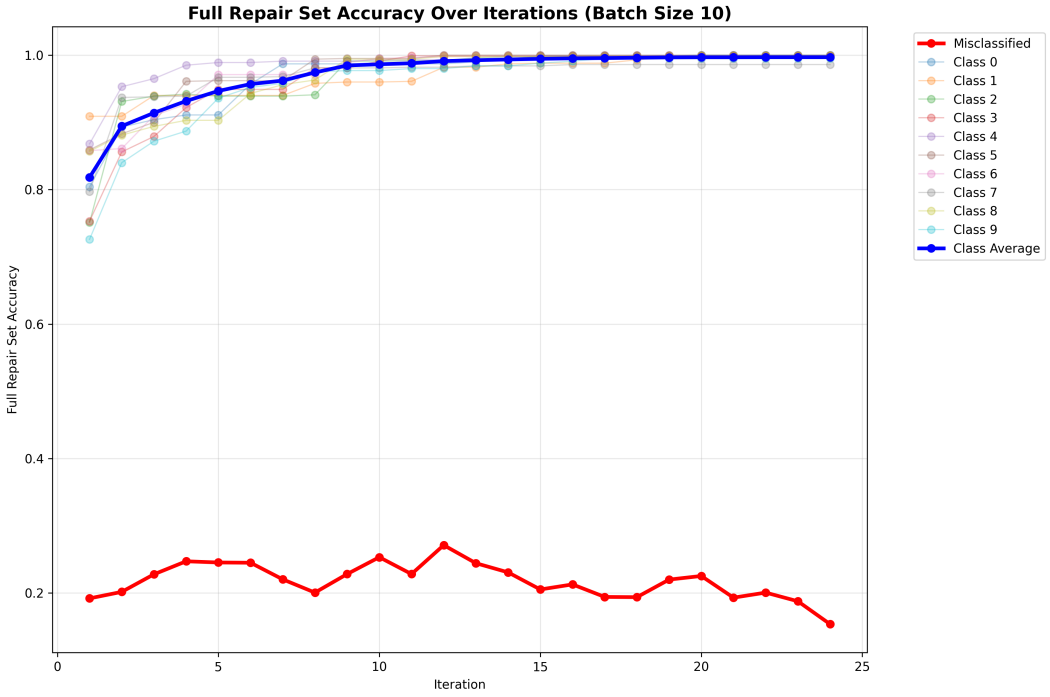


Figure 6: Repair set accuracy evolution for batch size 10, demonstrating similar rapid convergence patterns to batch size 5.

2. **Repair Set Size vs. Global Performance:** Both architectures demonstrate that larger repair sets lead to worse global performance, but the degradation is more severe in complex models. CNN experiments show success rates dropping from 100% (size 10) to 9.1% (size 200), while MLP repairs maintain higher success rates across different set sizes.
3. **Local Accuracy vs. Generalization:** The rapid convergence observed in CNN stochastic repair (99% repair set accuracy within 5 iterations) comes at the cost of significant global accuracy degradation, highlighting the fundamental tension between local correctness and generalization.
4. **Heuristic Sophistication vs. Robustness:** Advanced heuristics (Adversarial, Gradient-Based) showed higher failure rates in our MLP experiments, suggesting that simpler, more robust approaches may be preferable for practical deployment.

#### 5.3.3 Implications for Repair Strategy Selection

- Repair set curation is critical—semantic coherence significantly improves outcomes
- Size limitations exist for all architectures, though thresholds vary
- Confidence-based sample selection outperforms naive misclassification targeting
- One-shot repairs generally preserve global performance better than iterative approaches

#### 5.3.4 Practical Implications

These findings have important implications for practical DNN repair deployment:

- **Repair Set Curation**: Careful selection and grouping of repair examples by semantic similarity can substantially improve repair outcomes.
- **Size Limitations**: There exist practical upper bounds on repair set size beyond which the cure becomes worse than the disease.
- **Strategy Selection**: One-shot repair is generally preferable to iterative approaches for maintaining global model performance.

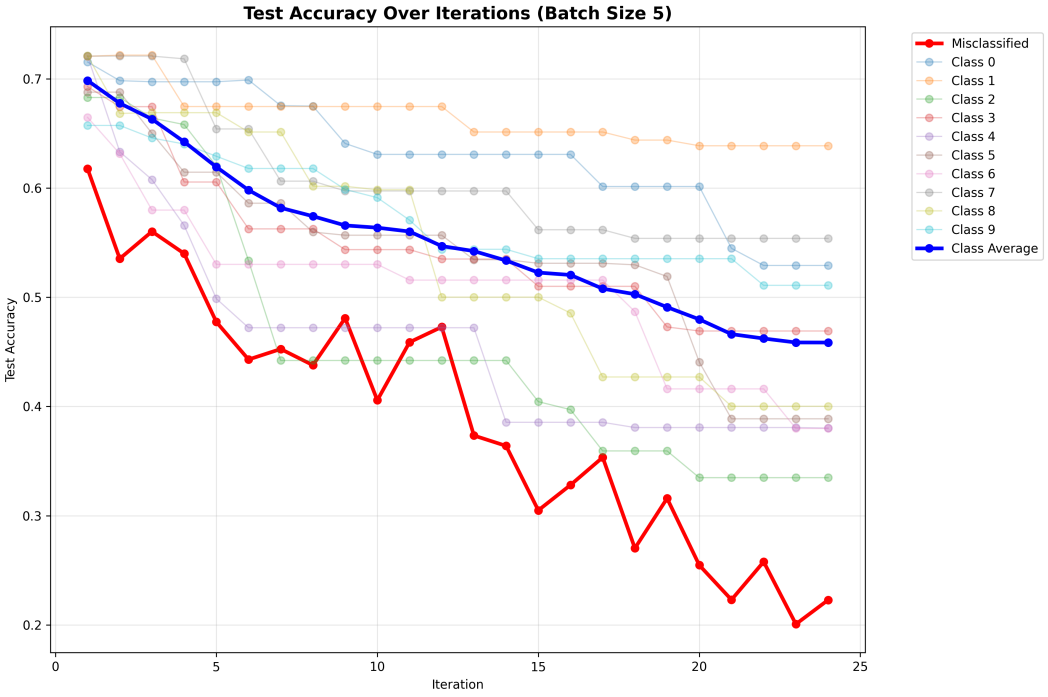


Figure 7: Test accuracy evolution for batch size 5 showing that early stopping around iteration 5 would preserve much higher global accuracy while maintaining excellent repair set performance.

- **Monitoring Requirements**: Any repair strategy requires careful post-repair validation to ensure acceptable global performance retention.

This geometric intuition suggests that repair effectiveness is fundamentally linked to the interpretability and locality of a model's decision-making process.

## 6 Conclusions and Future Work

Our comprehensive evaluation of informed neural network repair strategies across multiple architectures reveals fundamental insights about the nature of neural network modification and the effectiveness of principled heuristic guidance. Through systematic evaluation of 90 heuristic combinations on MLPs and extensive analysis of repair strategies on CNNs, we have established both the promise and limitations of guided repair approaches.

**Key Contributions and Findings:**

Our experimental results demonstrate that intelligent heuristic selection can significantly improve repair outcomes. The WeightsActivationBased + ConfidenceBased combination achieved a remarkable +3.66% improvement over baseline MLP performance, showcasing the potential for substantial model enhancement through principled parameter modification. This success illustrates a key insight: for models with interpretable decision boundaries, targeted parameter adjustments function analogously to coefficient optimization in polynomial fitting, enabling precise control over model behavior.

The effectiveness of confidence-based set selection across both architectures validates our hypothesis that principled sample curation outperforms naive approaches. By targeting low-confidence predictions—often corresponding to decision boundary regions—we can achieve more effective repairs with less risk of global performance degradation.

Our cross-architectural analysis reveals a fundamental scaling challenge: repair complexity and failure rates increase dramatically with model sophistication. While simple MLPs demonstrate robust improvements with appropriate heuristics, complex CNNs exhibit severe sensitivity to repair

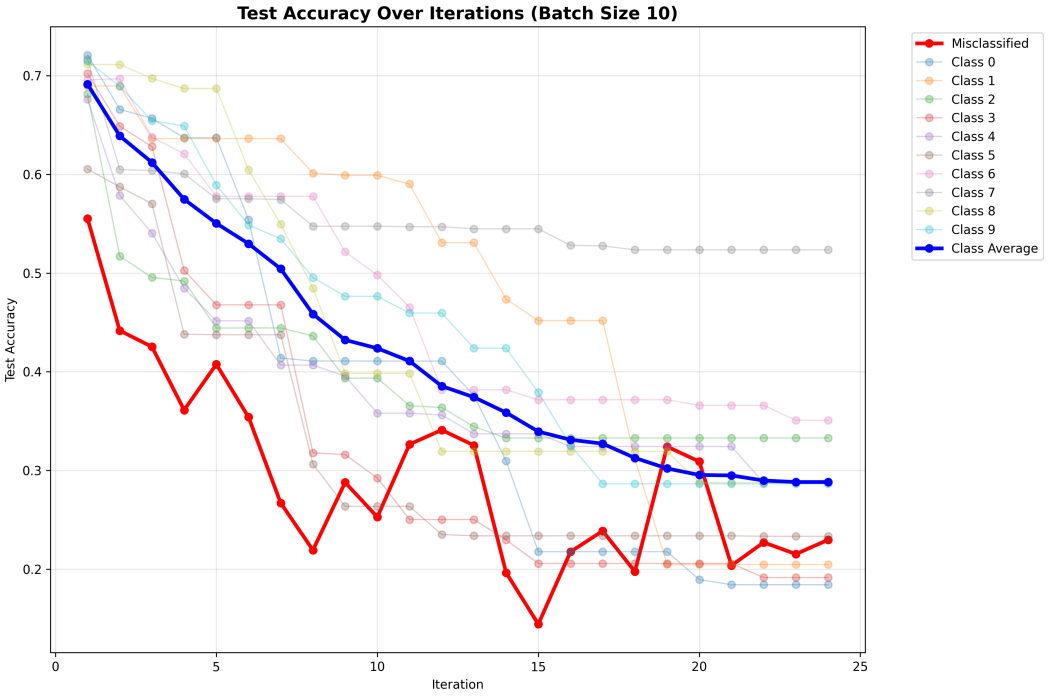


Figure 8: Test accuracy evolution for batch size 10 demonstrating similar early stopping opportunities to reduce drawdown.

interventions, with accuracy drawdowns reaching -50.8% for large repair sets. This scaling behavior suggests that repair strategies must be carefully tailored to model complexity.

**Future Research Directions:**

Our work opens several promising avenues for future investigation:

- Regularization techniques that preserve global decision boundaries while enabling local corrections
- Sophisticated layer selection heuristics that minimize repair impact on unrelated model functionality
- Adaptive repair strategies that adjust their approach based on repair set characteristics
- Multi-objective optimization frameworks that balance local repair success against global performance preservation

**Broader Impact:**

This work contributes to the broader goal of trustworthy AI by making neural network repair more systematic and predictable. By providing principled guidance for repair strategy selection, we help move the field away from ad-hoc approaches toward evidence-based practices. The geometric insights we provide also contribute to our theoretical understanding of neural network parameter spaces and their modification.

Ultimately, our findings demonstrate that while neural network repair remains a challenging problem, informed heuristic selection can significantly improve outcomes. The path forward requires continued research into architecture-specific strategies, better theoretical understanding of parameter modification effects, and the development of more sophisticated tools for predicting and controlling repair outcomes. Through such advances, we can work toward making neural network repair a reliable tool for enhancing AI system trustworthiness and safety.

## References


[1] Stephanie Nawas, Zhe Tao, and Aditya V. Thakur. Provable repair of vision transformers. In Guy Avni, Mirco Giacobbe, Taylor T. Johnson, Guy Katz, Anna Lukina, Nina Narodytska, and Christian Schilling, editors, *AI Verification*, pages 156–178. Springer Nature Switzerland, 2024.

[2] Matthew Sotoudeh and Aditya V. Thakur. Provable repair of deep neural networks. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, PLDI 2021, pages 588–603. Association for Computing Machinery, 2021.

[3] Zhe Tao, Stephanie Nawas, Jacqueline Mitchell, and Aditya V. Thakur. Architecture-preserving provable repair of deep neural networks. *Proceedings of the ACM on Programming Languages*, 7:124:443–124:467, 2023.

[4] J. Ma, P. Yang, J. Wang, Y. Sun, C.-C. Huang, and Z. Wang. VeRe: Verification guided synthesis for repairing deep neural networks. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13, February 2024.

[5] Z. Chen, J. Zhou, Y. Sun, J. Wang, Q. Xuan, and X. Yang. Interpretability based neural network repair. *33rd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA) 2024*, 35:908–919, September 2024.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[7] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.