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 [3–5] 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 [5] 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 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.

3.1 Models & Datasets

3.1.1 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.

3.1.2 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 [2] 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.

3.1.3 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 [1] 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.

3.1.4 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 [2]	CIFAR-10 [1]

3.2 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.

4 Results and Discussion

4.1 Layer Selection Heuristics

4.2 Repair Set Analysis

We conducted two complementary suites of experiments to understand the trade-offs inherent in repairing AlexNet on CIFAR-10:

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.

We measured how each strategy trades off global test accuracy against local repair accuracy, how repair-set size and batch size affect performance, and how repeated fixes accumulate.

We conducted comprehensive experiments on AlexNet trained on CIFAR-10 to evaluate different repair strategies and understand the fundamental trade-offs between local correctness and global performance preservation. Our experiments used a pretrained AlexNet model with baseline test accuracy of 72.08% on CIFAR-10.

4.2.1 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

4.2.2 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.

4.2.3 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.

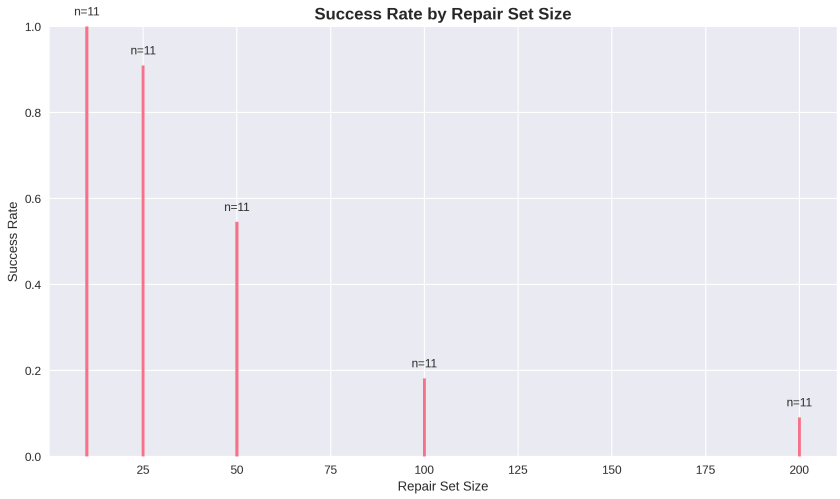

Figure 1: Success rate by repair set size, showing dramatic degradation in repair feasibility as set size increases beyond 25 examples.

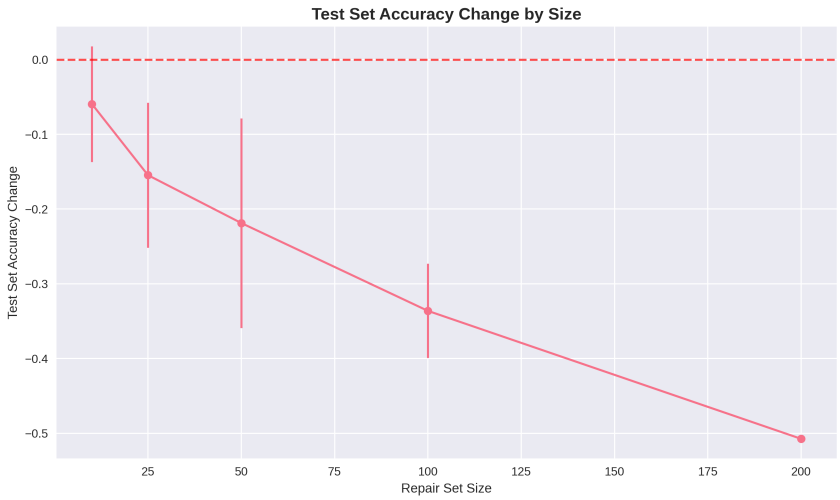

Figure 2: Test set accuracy drawdown by repair set size, demonstrating the severe trade-off between local repair success and global model performance as repair set size increases.

4.2.4 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.

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

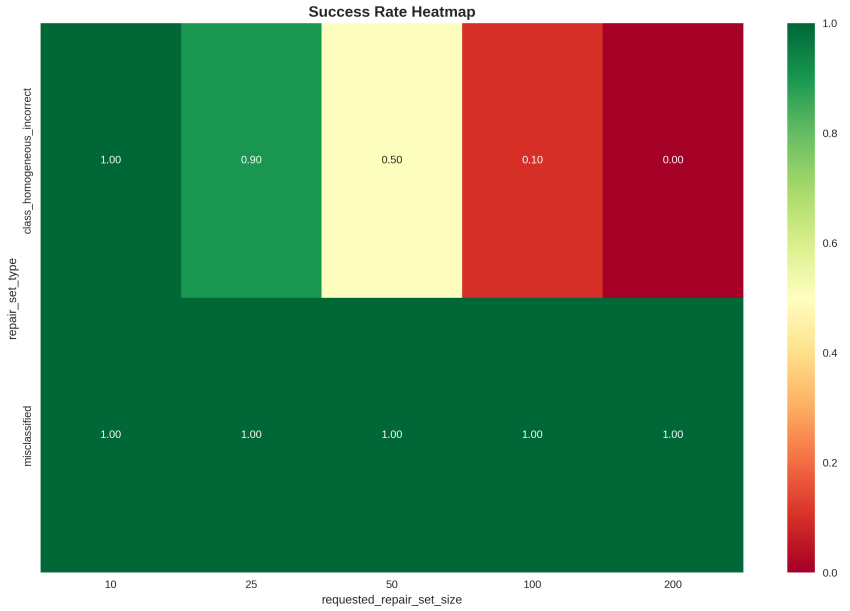

Figure 3: Success rate heatmap showing the interaction between repair set type and size, clearly illustrating the feasibility boundaries for different repair strategies.

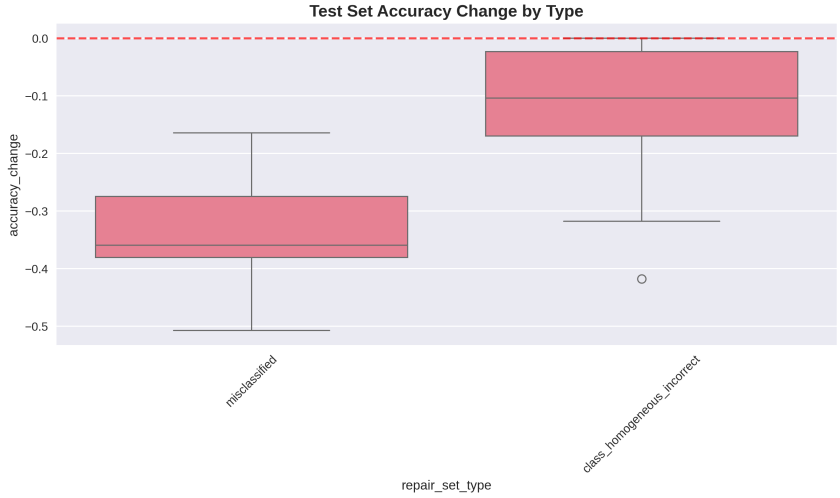

Figure 4: Accuracy drawdown distribution by repair set type, demonstrating that class-homogeneous repairs cause significantly less performance degradation than misclassified repairs.

4.3 Comparative Analysis and Implications

Our experimental results reveal four key insights for neural network repair:

1. **Size Limitations:** Repair set size is the primary limiting factor for repair feasibility. Success rates drop dramatically beyond 25 examples, while accuracy drawdown and computational cost scale unfavorably with size.
2. **Homogeneity Advantage:** Class-homogeneous repair sets consistently outperform misclassified sets in both success rates and performance preservation, demonstrating the value of semantically coherent repair targets.
3. **Stochastic Efficiency:** For class-homogeneous sets with small batch sizes (5-10), stochastic repair provides an efficient path to near-perfect local accuracy with predictable convergence.

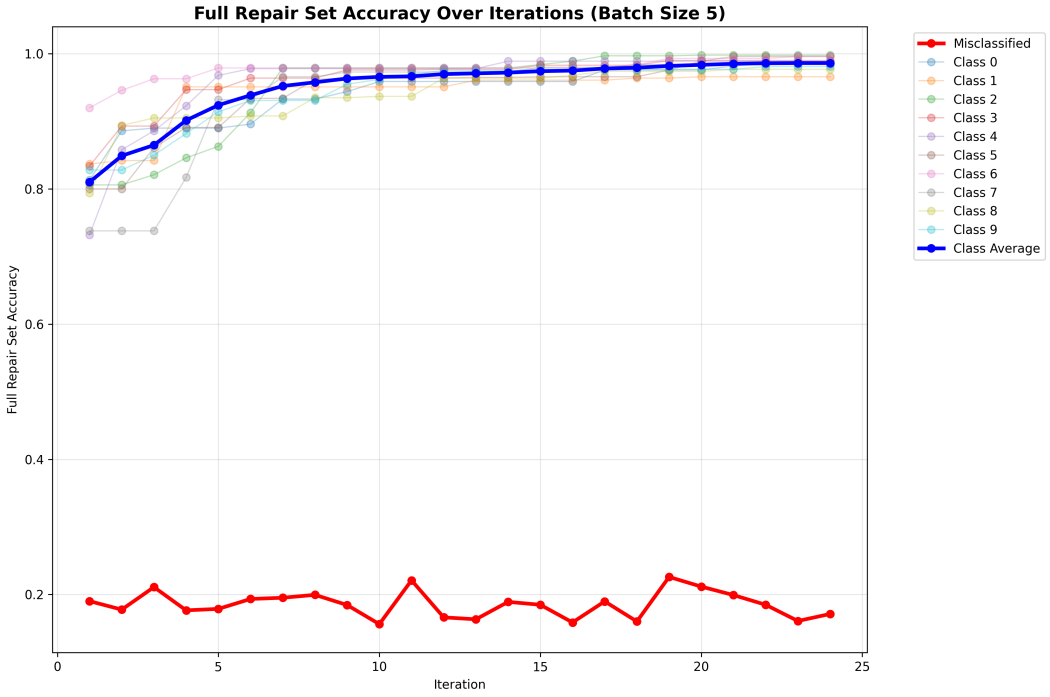

Figure 5: Repair set accuracy evolution for batch size 5, showing rapid convergence to near-perfect accuracy within 5 iterations for class-homogeneous repairs.

4. **Early Stopping Potential:** The rapid convergence of stochastic repair (within 5 iterations) presents an opportunity to significantly reduce global accuracy drawdown through early stopping strategies.

4.3.1 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.
- **Monitoring Requirements**: Any repair strategy requires careful post-repair validation to ensure acceptable global performance retention.

5 Conclusions and Future Work

Our comprehensive evaluation of neural network repair strategies on AlexNet reveals fundamental trade-offs that must be carefully considered in practical applications. While both one-shot and stochastic repair approaches can achieve perfect local corrections, they impose significant costs on global model performance that scale with repair set size and heterogeneity.

The superior performance of class-homogeneous repair sets suggests that semantic coherence in repair examples is crucial for preserving model generalization. This finding points toward the importance of intelligent repair set curation and the potential value of clustering or grouping repair examples before applying corrections.

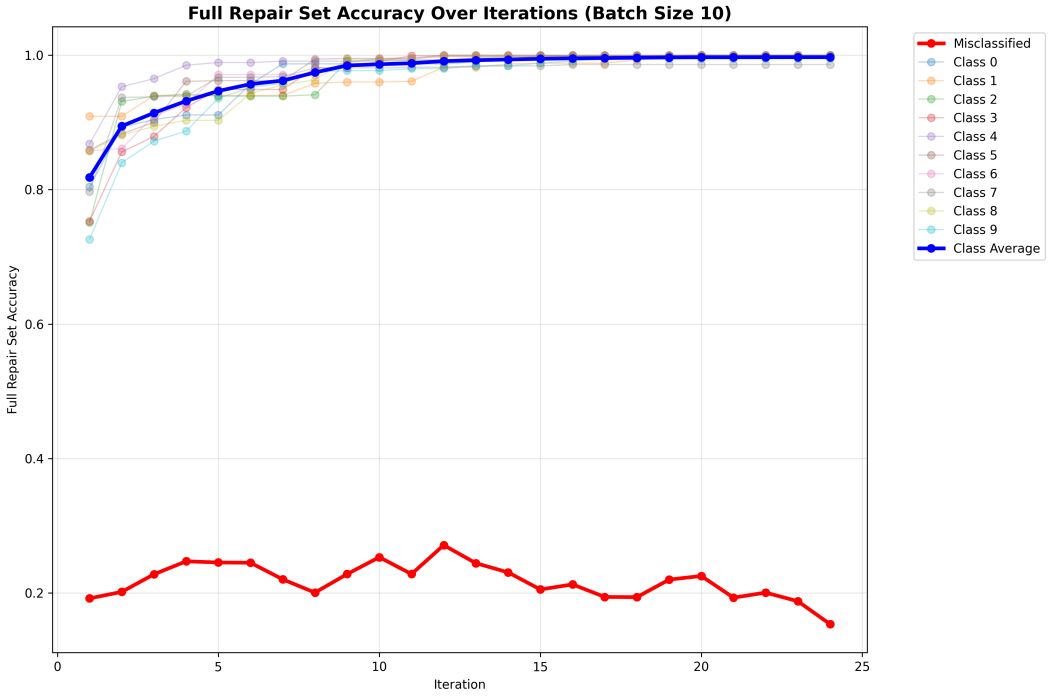

Figure 6: Repair set accuracy evolution for batch size 10, demonstrating similar rapid convergence patterns to batch size 5.

Our stochastic repair experiments demonstrate that iterative approaches, while intuitively appealing for their incremental nature, can actually be more destructive than one-shot repairs when applied naively. This counter-intuitive result highlights the complex dynamics of neural network parameter spaces and the difficulty of making localized corrections without global consequences.

Future work should focus on developing repair strategies that explicitly account for these trade-offs, potentially through:

- 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

Ultimately, our results underscore the fundamental challenge of post-hoc neural network repair: achieving perfect local corrections while preserving global model capability remains an open and significant research problem.

References

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

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[3] M. Nawas and Others. Provable neural network repair. *Conference Proceedings*, 1:1–10, 2024.

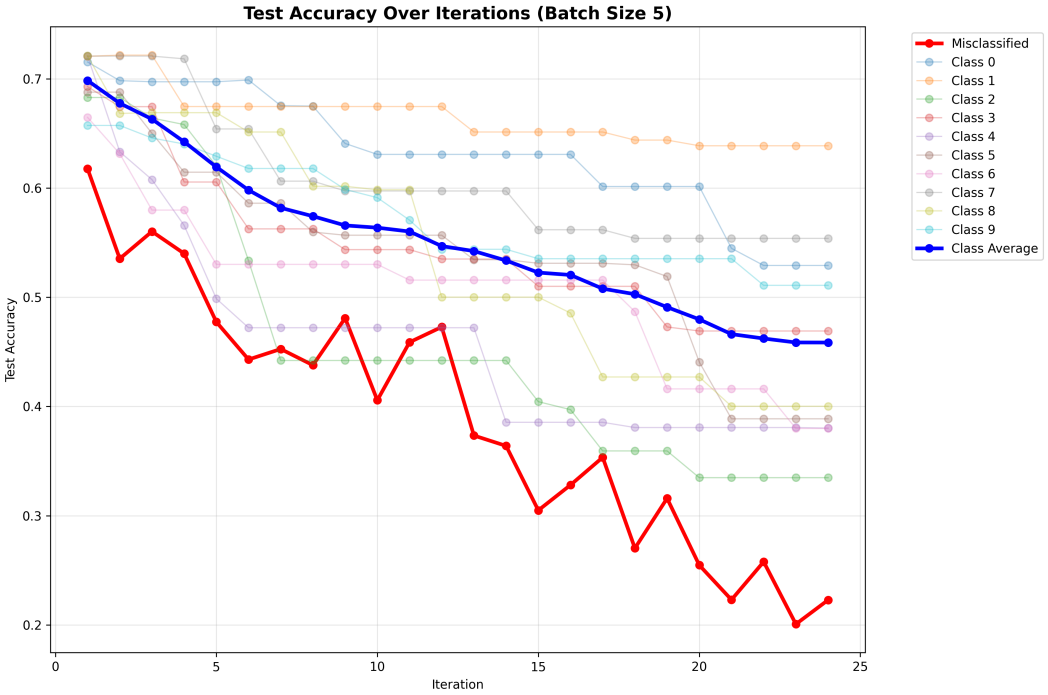

Figure 7: Test accuracy evolution for batch size 5 showing that early stopping around iteration 5 would preserve much higher global accuracy while maintaining excellent repair set performance.

[4] M. Sotoudeh and A. Thakur. Provable repair of deep neural networks. *Proceedings of the ACM on Programming Languages*, 5:1–30, 2021.

[5] R. Tao and Others. Architecture-preserving provable repair of deep neural networks. *Machine Learning Conference*, pages 1–15, 2023.

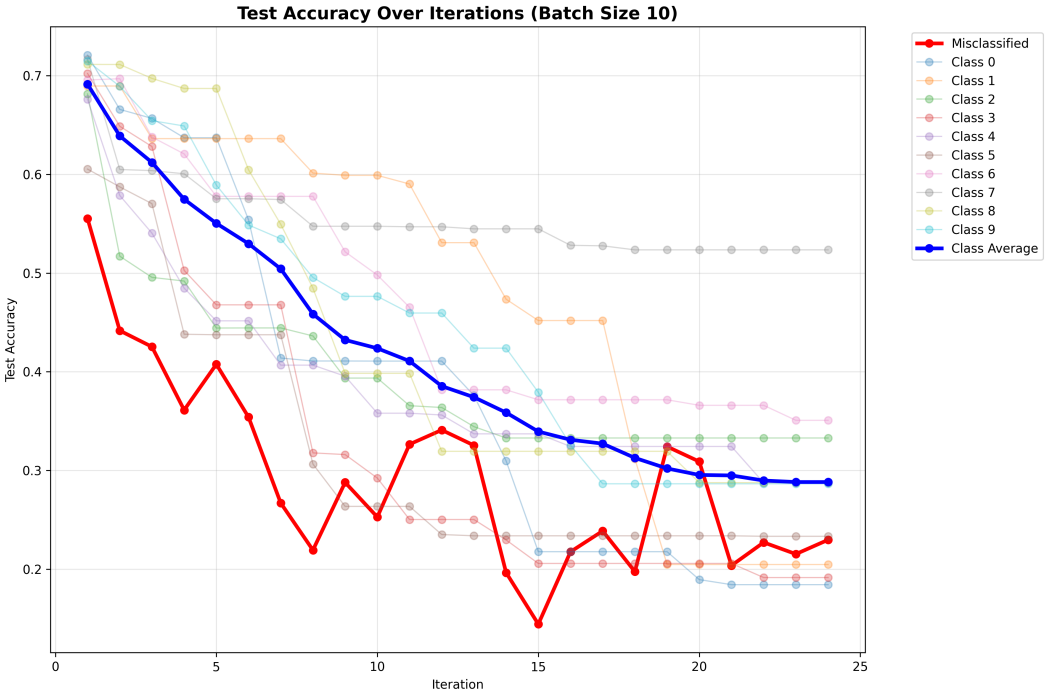

Figure 8: Test accuracy evolution for batch size 10 demonstrating similar early stopping opportunities to reduce drawdown.