POLITECNICO DI TORINO
Repository ISTITUZIONALE

M-EL2N: Data-Efficient Fault Injection Through Sample Difficulty Analysis

Original
M-EL2N: Data-Efficient Fault Injection Through Sample Difficulty Analysis / Bellarmino, N., Bosio, A., Al-Kaf, A., Benabdenbi, M., Leveugle, R., Cantoro, R.. - (2026), pp. 1-6. (27th Latin American Test Symposium, LATS 2026 Florianópolis (BRA) 17-20 March 2026) [10.1109/lats70329.2026.11480331].

Availability:
This version is available at: 11583/3016148 since: 2026-09-30T10:20:38Z

Publisher:
Institute of Electrical and Electronics Engineers

Published
DOI:10.1109/lats70329.2026.11480331

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

09 November 2026

M-EL2N: Data-Efficient Fault Injection through Sample Difficulty Analysis

Nicolò Bellarmino*, Alberto Bosio†, Ahmed Al-Kaf‡, Mounir Benabdenbi‡, Régis Leveugle ‡, and Riccardo Cantoro*

*CAD Research Group,
Politecnico di Torino,
Italy
{name.surname}@polito.it

† Centrale Lyon, INSA Lyon, CNRS
UCBL, CPE Lyon, INL,
UMR5270, Ecully, France.
alberto.bosio@ec-lyon.fr

‡Univ. Grenoble Alpes,
CNRS, Grenoble INP*,
TIMA, 38000 Grenoble, France
{name.surname}@univ-grenoble-alpes.fr

***Abstract*—Fault Injection (FI) is the standard technique for assessing the resilience of Convolutional Neural Networks (CNNs), but campaign time scales with the number of input stimuli applied per injected fault. Prior work has shown that selecting challenging inputs based on confidence metrics can significantly reduce this workload. In this paper, we propose Mirrored EL2N (M-EL2N), a novel sample-ranking metric that symmetrizes the original EL2N score by re-scaling the contribution of misclassified samples. This yields a unified notion of sample difficulty that prioritizes highly informative inputs regardless of prediction correctness, enabling faster exposure of critical faults, and requires no architectural knowledge. We evaluate Mirrored EL2N against established metrics (Cross-Entropy, DeepGini, EL2N) across multiple CNN architectures and datasets under a single-bit fault model in network weights, and we present initial results on a post–training quantized CNN. Experimental results show that Mirrored EL2N consistently achieves earlier detection of critical faults with up to three orders of magnitude reduction in the number of test samples required to reach 99% of the detected critical faults, while maintaining near-perfect detection curves (AUC $\approx$ 1). These results establish Mirrored EL2N as a practical and lightweight tool for data-efficient FI workload reduction in both floating-point and quantized CNNs.**

***Index Terms*—Fault Injection, Neural Networks, Reliability, Hardware Faults, Rational Sampling, Test Compaction, Fault Tolerance, Machine Learning.**

I. Introduction

Convolutional Neural Networks (CNNs) are increasingly deployed in safety-critical domains such as autonomous driving, industrial automation, and healthcare, where the reliability of inference is essential [1], [2], [3]. Transient or permanent hardware faults may alter the network's predictions and jeopardize system safety. Fault Injection (FI) is the de-facto methodology to assess CNN resilience under such conditions: by perturbing the hardware or model state and monitoring the resulting inference behavior, FI campaigns identify *critical faults*, i.e., faults that cause misclassification or unacceptable degradation [2]. Although effective, FI remains computationally demanding. A single campaign typically involves thousands of injected faults, each exercised over the entire test set, leading to hundreds of millions of fault–input evaluations. Such exhaustive exploration is often impractical. Recent works have shown that the FI workload can be substantially reduced by selecting a compact yet informative subset of input samples [4], [5]. Ranking inputs according to their classification difficulty—estimated through confidence-based metrics—has proved particularly effective in exposing fault-induced vulnerabilities early in the campaign.

Among these metrics, the EL2N score measures the Euclidean distance between the predicted probability vector and the one-hot ground truth and exhibits strong correlation with fault activation. However, EL2N is inherently asymmetric: correctly and incorrectly classified samples occupy disjoint ranges, causing low-confidence misclassifications—often among the most informative for reliability—to be systematically underweighted. This limits the effectiveness of workload selection based solely on EL2N or similar confidence measures. To address this limitation, we introduce *Mirrored EL2N*, a symmetric reformulation of EL2N that rescales misclassified samples into the same numerical range as correctly classified ones, by reflecting misclassified samples within the EL2N range, correcting the asymmetry that suppresses low-confidence misclassifications. This simple transformation produces a unified measure of sample difficulty that preserves the computational efficiency of EL2N while enhancing the contribution of highly informative misclassifications. As a result, Mirrored EL2N more accurately captures fault sensitivity and enables more efficient workload selection. We evaluate Mirrored EL2N across multiple CNN architectures and datasets under a single-bit fault model in network weights, and we include initial experiments on a post-training quantized network to assess generality across numerical representations. Experimental results show that Mirrored EL2N consistently accelerates the detection of critical faults, achieving equivalent fault coverage with a fraction of the workload required by existing metrics. For example, on ResNet18, it reaches 99% of detected critical faults using as few as 2 images on MNIST and 6 images on CIFAR-10.

The remainder of the paper is organized as follows. Section II reviews the fundamentals of CNN reliability, fault injection, and existing workload-selection strategies. Section III presents the proposed Mirrored EL2N metric and the associ-

979-8-3195-4235-9/26/$31.00 ©2026 IEEE

ated fault-detection framework. Section IV describes the experimental setup, and Section V discusses the results. Finally, Section VI concludes the paper.

II. Background and Related Work

The integration of Deep Neural Networks (DNNs), and particularly Convolutional Neural Networks (CNNs), into safety-critical systems such as autonomous vehicles and industrial controllers demands rigorous reliability assessment. Even a single transient or permanent hardware fault (such as a bit-flip in memory or arithmetic units) can alter the network's output and potentially lead to hazardous decisions [2], [6], [7]. Ensuring dependable operation under such conditions is therefore a prerequisite for deployment [8].

Fault Injection (FI) is the primary methodology to evaluate the resilience of neural networks to hardware faults. By deliberately perturbing model parameters or hardware states and comparing faulty versus fault-free outputs, FI allows faults to be classified as masked, non-critical, or critical. This process provides quantitative insight into the network's vulnerability and the distribution of critical faults across layers.

However, the exhaustive nature of FI results in prohibitive computational effort. A typical campaign requires thousands of fault injections and inference over the entire test set, leading to millions of fault–input evaluations. Statistical Fault Injection (SFI) [9], [10] reduces this burden by sampling a representative subset of faults, maintaining statistical validity with fewer injections. Yet SFI addresses only the *fault* dimension of complexity. The second dimension—the selection of *input stimuli*, or workload—remains a dominant contributor to FI cost.

The workload strongly influences which faults are observed as critical [4], [5], [11], [12], [13]. Many faults remain masked across most inputs but become observable under specific activation patterns. Consequently, evaluating every injected fault across the full test set is often inefficient. Recent works have shown that selecting a compact yet informative subset of inputs can significantly accelerate FI without sacrificing coverage [4], [5]. These approaches rely on *sample difficulty metrics* derived from output probability vectors or internal activations to prioritize inputs likely to activate faulty behavior.

Several metrics have been proposed in this context. DeepGini measures output uncertainty and has been used for test-suite reduction. Cross-Entropy reflects classification loss. EL2N quantifies the Euclidean distance between the predicted probability vector and the one-hot ground truth and has been shown to correlate with sample difficulty and fault sensitivity [11], [14]. Although effective, these metrics implicitly treat correctly and incorrectly classified samples differently, often underweighting low-confidence misclassified inputs, despite these typically lying near decision boundaries and being disproportionately sensitive to internal perturbations. Moreover, most prior evaluations focus on floating-point models, leaving their behavior in quantized networks largely unexplored.

Building upon these foundations, this work introduces *Mirrored EL2N*, a symmetric reformulation of EL2N that rescales misclassified samples to align their values with those of correctly classified ones. This adjustment yields a unified and more expressive difficulty metric that better correlates with fault activation, enabling faster and more data-efficient workload selection for both floating-point and quantized CNNs.

III. Proposed Approach

Performing an exhaustive FI campaign on neural network parameters is computationally prohibitive due to the large dimensionality of the fault space. While Statistical Fault Injection (SFI) reduces the number of injected faults through random sampling, the choice of input stimuli, or *workload*, remains a dominant bottleneck. Using the full test set for each injected fault [2] is inefficient for two reasons: (1) many faults manifest as critical across multiple inputs, leading to redundant evaluations, and (2) most faults remain masked for the majority of the dataset. As a result, a smaller but well-chosen subset of inputs can provide equivalent fault coverage at a fraction of the computational cost.

Recent workload-selection strategies [4], [5] focus on *challenging samples*, which lie near decision boundaries and are therefore more sensitive to perturbations. These samples stress the network's internal representations and activate a disproportionately high number of critical faults. Among confidence-based difficulty metrics, the *EL2N* score [15] measures the Euclidean distance between the predicted probability vector and the one-hot ground truth. Although EL2N correlates with fault sensitivity, it is inherently asymmetric: correctly and incorrectly classified samples occupy distinct, poorly aligned regions of the metric space. This reduces its ability to provide a consistent notion of difficulty across the dataset and limits its effectiveness in guiding FI. To address this issue, we introduce the *Mirrored EL2N* metric.

A. Mirrored EL2N Metric

The Mirrored EL2N metric symmetrically re-scales the EL2N values of misclassified samples to align them with the range of correctly classified ones. This operation eliminates the bimodality observed in EL2N distributions and provides a unified measure of sample *criticality* that better reflects fault sensitivity.

Given ground-truth labels y, predicted probability vectors $\mathbf{p}_i$, and one-hot encodings $\mathbf{e}_{y_i}$, the original EL2N score is:

$$\mathrm{EL2N}_i = \left\| \mathbf{e}_{y_i} - \mathbf{p}_i \right\|_2 . \tag{1}$$

Let $M_{\text{mis}} = \max(\mathrm{EL2N}_{\text{mis}})$ denote the maximum EL2N value among misclassified samples. The Mirrored EL2N score is defined as:

$$\text{M-EL2N}_i = \begin{cases} \mathrm{EL2N}_i, & \text{if correctly classified,} \\ |\mathrm{EL2N}_i - M_{\text{mis}}|, & \text{if misclassified.} \end{cases} \tag{2}$$

Since $\mathrm{EL2N} \in [0, \sqrt{2}]$, the value M_{mis} typically approaches $\sqrt{2}$, causing high-confidence misclassifications to be mirrored toward low values and uncertain ones toward larger values. The resulting metric assigns consistent difficulty scores across

the entire dataset and produces a smooth, symmetric distribution that improves the stability of sample ranking.

The effect of this transformation is illustrated in Figure 1, which reports, for a representative network, the number of critical faults activated by each image as a function of its metric value. In the original EL2N space, correctly and incorrectly classified samples form two largely disjoint sub-distributions: misclassified inputs accumulate near the upper end of the range, while correctly classified ones occupy a narrower band of lower values. As a result, high-confidence misclassifications are numerically separated from confidently correct predictions, and the metric does not induce a smooth difficulty spectrum. After applying Mirrored EL2N, both correctly and incorrectly classified samples are redistributed over a common interval: high-confidence predictions (regardless of correctness) cluster at low scores, whereas uncertain samples populate the higher end of the range. This collapse of the two EL2N sub-distributions into a single, continuous manifold yields a more expressive ordering of samples and leads to a clearer monotonic trend between metric value and fault sensitivity.

B. Workload Selection Framework

The proposed workload-selection procedure consists of three lightweight steps:

1) **Metric computation:** compute Mirrored EL2N for each test image via a single forward pass.
2) **Ranking:** sort samples in descending order of their Mirrored EL2N score.
3) **Selection:** select the top n samples as the workload for the FI campaign.

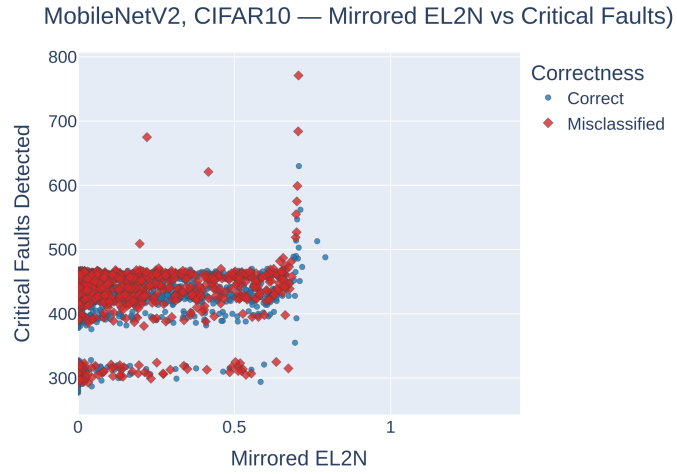
This process incurs negligible computational overhead, and the resulting ranked list can be reused across multiple FI runs, fault models, or hyperparameter settings. By reducing the workload size without altering the statistical assumptions behind SFI, Mirrored EL2N improves efficiency while preserving confidence and error bounds on the estimated fault coverage.

C. Fault Detection Analysis

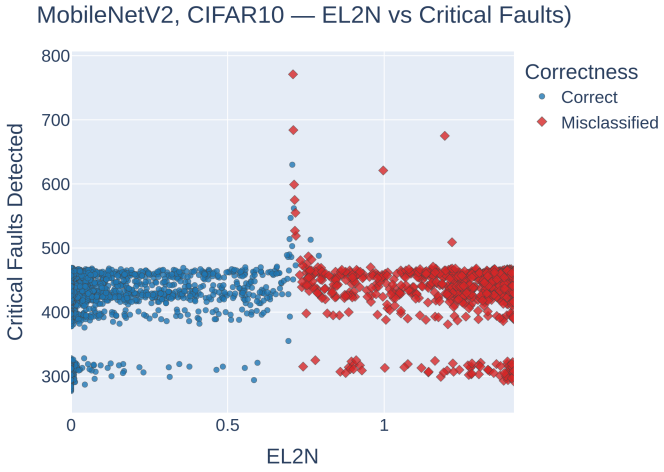
To evaluate workload efficiency, we interpret FI as a *fault-detection process*. For each injected fault, the outputs of the fault-free and faulty networks are compared across the workload, and the fault is labeled according to the most severe deviation observed. Let $C(n)$ denote the number of distinct critical faults detected using the top n ranked images. By progressively increasing n, we obtain a *fault-detection curve* $C(n)$ describing the rate at which critical faults are exposed.

A steeper initial slope indicates a more efficient workload, as critical faults are identified using fewer inputs. Empirically, Mirrored EL2N yields faster convergence than alternative metrics, concentrating the majority of critical-fault activations within a small prefix of the ranked list. This behavior results in substantial reductions in FI time with no loss in coverage.

In practice, $C(n)$ often saturates early, revealing the redundancy of many test inputs. We extract such plateau points



(a) MEL2N



(b) EL2N

Fig. 1: Metric values vs. number of critical faults activated by each image (MobileNet, CIFAR10). Mirrored EL2N provides a more uniform distribution of sensitivity, effectively capturing both easy and hard samples.

a posteriori from the detection curves to estimate workload compactness and quantify the achievable time savings.

IV. Experimental Setup

A. Networks and Datasets

The proposed Mirrored EL2N metric was evaluated across a diverse set of convolutional architectures and datasets to ensure robustness and generality. We considered models representative of different design paradigms: a Residual Network (ResNet-18 [16]), a LeNet-5 [17], a VGG architecture (VGG-11) [18], and a gated hybrid-connectivity CNN optimized for dense classification (HCGNet-A2 [19]). These models cover different architectural properties that influence fault propagation, such as residual connections (ResNet-18), shallow convolutional pipelines (LeNet-5), deep sequential architectures (VGG-11), and hybrid multi-branch connectivity (HCGNet-A2). Experiments were conducted on the MNIST [20],

CIFAR-10 and CIFAR-100 [21] datasets, each comprising 10,000 test images. MNIST provides a low-complexity workload, whereas CIFAR-10 and CIFAR-100 introduce growing levels of visual variability and classification difficulty, allowing us to evaluate the metric under progressively more challenging conditions.

To assess generality across numerical representations, we also evaluated a post-training uniformly quantized version of the LeNet network (weights and activations) [22]. As quantized CNNs increasingly dominate edge deployments, understanding whether difficulty-based workload selection remains effective under low-precision arithmetic is also important.

B. Fault Model

Faults were injected at the bit level into the network parameters following a Statistical Fault Injection (SFI) process [9], [10]. Each injection flips a single randomly selected bit in the weight tensor under a stuck-at fault model and the single-fault hypothesis. A total of about $N_f = 16{,}640$ faults were injected to ensure statistical significance with 99% confidence and an estimation error below 1%, following [10]. In the quantized LeNet-5, previously analyzed in [22], 16,046 faults were injected to preserve the same confidence guarantees as for the floating-point models A fault is labeled as *critical* if it causes at least one misclassification across the workload; distinct critical faults refer to unique bit locations that trigger incorrect predictions under at least one input. Results of the SFI campaign are listed in Table I.

C. Evaluation Procedure

For each injected fault, the outputs of the faulty and fault-free networks were compared across a selected workload derived from the ranking of test samples according to each metric. The cumulative number of distinct critical faults detected as a function of the workload size defines the *fault-detection curve*, which quantifies how efficiently each workload selection strategy exposes critical faults.

Mirrored EL2N was compared against established difficulty-based metrics, namely EL2N, Cross-Entropy, and DeepGini [4], [5], [15], [23], [24]. All experiments were implemented in Python using the PyTorch framework and executed on a workstation equipped with an Intel® Core™ i9-9900K CPU, 32 GB of RAM, and an NVIDIA® RTX 2080 Ti GPU.

D. Detection-Curve Metrics

For each $(\text{network}, \text{metric})$ pair, a cumulative detection curve $C(n)$ was generated. $C(n)$ represents the cumulative number of unique critical faults detected when evaluating the first n samples in the ranked workload. From each curve, we extracted:

1) the number of samples required to reach 90% and 99% of total detected critical faults,
2) the area under the normalized curve (AUC), summarizing detection efficiency, and
3) the relative speed-up in workload size compared to the random baseline at 99% coverage.

TABLE I: Accuracy metrics, fault-type distributions, and critical faults for an SFI process.

Model	Clean Acc	Faulty Acc	Not Critical Inferences	Critical Inferences	Critical Faults
CIFAR10					
ResNet18	93.52%	91.24%	97.52%	2.46%	826
LeNet	79.36%	79.07%	99.64%	0.34%	1158
VGG11	90.10%	88.15%	97.82%	2.18%	819
MNIST					
ResNet18	99.63%	97.08%	97.45%	2.55%	581
LeNet	99.25%	99.09%	99.85%	0.15%	484
LeNet (int-8)	98.05%	98.05%	99.99%	0.01%	3554
VGG11	99.68%	97.36%	97.67%	2.33%	661
CIFAR100					
HCGNet	82.41%	81.36%	98.69%	1.31%	2756

V. Results

A. Fault-Detection Curves and Metric Behavior

As discussed in Section III, the mirroring transformation produces a continuous difficulty spectrum that enhances early fault detection. Figure 2 shows the cumulative fault-detection curves $C(n)$ for ResNet18 on CIFAR-10. The vertical dashed line marks the point where 90% and 99% of all detected critical faults are reached. Mirrored EL2N exhibits the steepest initial slope and the earliest saturation, consistently revealing most critical faults using only a small fraction of the test set. Similar trends were observed across all evaluated architectures and datasets, confirming that high-ranked samples concentrate the majority of fault activations.

B. Efficiency of Workload Selection Metrics

Table II reports the number of samples needed to reach 90% and 99% of detected critical faults, the AUC, and the speed-up over random selection. On MNIST, where most inputs are confidently classified, all difficulty-based metrics achieve near-perfect AUC, but Mirrored EL2N provides substantial gains for deeper models. For ResNet18, it reaches 99% coverage with only 2 samples, yielding a speed-up exceeding $4{,}000\times$.

On CIFAR-10 and CIFAR-100, the advantage becomes more pronounced: Mirrored EL2N achieves 99% coverage using less than 0.5% of the test set for ResNet18 and VGG11, and reduces the required workload by $4.6\times$ on HCGNet for CIFAR-100. Across all cases, it outperforms EL2N, DeepGini, and Cross-Entropy in both early detection and final workload size.

To assess robustness under low-precision arithmetic, we evaluated a post–training 8-bit quantized LeNet5. Mirrored EL2N preserved its behavior, reaching 90% and 99% detection with 59 and 211 samples (AUC = 0.997), corresponding to a $46\times$ speed-up over random. Although our quantized experiments are limited to a single architecture, the preserved performance suggests that Mirrored EL2N is robust to the altered activation statistics introduced by low-precision arithmetic that alters the numerical landscape of the model. It is worth noting that, although Mirrored EL2N consistently delivers the best early detection (90% coverage) and the highest AUC, we observe a few configurations in which DeepGini

TABLE II: Efficiency of metric-based workload selection. For each network and metric, we report the number of images required to reach 90% and 99% of the detected critical faults, the AUC of the detection curve, and the relative speed-up with respect to the random baseline at 99% coverage.

Network	Metric	@90%	@99%	AUC	Speed-up@99%
MNIST					
LeNet5	Random	2419	8870	0.928	1.00×
	Cross-Entropy	664	8311	0.957	1.07×
	DeepGini	956	8311	0.957	1.07×
	EL2N	682	8838	0.958	1.00×
	Mirrored EL2N	**969**	**8837**	**0.957**	**1.00×**
LeNet5 (Int8)	Random	8530	9746	0.649	1.00×
	Cross-Entropy	221	258	0.983	37.78×
	DeepGini	110	162	0.995	60.16×
	EL2N	215	248	0.982	39.30×
	Mirrored EL2N	**59**	**211**	**0.997**	**46.19×**
ResNet18	Random	3318	8567	0.926	1.00×
	Cross-Entropy	35	36	0.999	238.0×
	DeepGini	4	5	0.9999	1713.4×
	EL2N	35	36	0.999	238.0×
	Mirrored EL2N	**2**	**2**	**0.99999**	**4283.5×**
VGG11	Random	5429	7874	0.899	1.00×
	Cross-Entropy	34	34	0.9992	231.6×
	DeepGini	12	19	0.9996	414.4×
	EL2N	33	34	0.9993	231.6×
	Mirrored EL2N	**3**	**20**	**0.9999**	**393.7×**
CIFAR-10					
LeNet5	Random	7999	9780	0.748	1.00×
	Cross-Entropy	2425	2491	0.903	3.93×
	DeepGini	1920	2262	0.944	4.32×
	EL2N	2212	2224	0.899	4.40×
	Mirrored EL2N	**387**	**2635**	**0.979**	**3.71×**
ResNet18	Random	6213	9313	0.853	1.00×
	Cross-Entropy	652	664	0.977	14.0×
	DeepGini	58	72	0.998	129.3×
	EL2N	649	655	0.974	14.2×
	Mirrored EL2N	**6**	**46**	**0.9996**	**202.5×**
VGG11	Random	5573	9342	0.853	1.00×
	Cross-Entropy	995	997	0.969	9.37×
	DeepGini	67	124	0.998	75.3×
	EL2N	990	995	0.964	9.39×
	Mirrored EL2N	**11**	**276**	**0.999**	**33.8×**
CIFAR-100					
HCGNet	Random	6188	9208	0.757	1.00×
	Cross-Entropy	2026	2026	0.902	4.54×
	DeepGini	1497	1497	0.947	6.15×
	EL2N	1878	1878	0.884	4.90×
	Mirrored EL2N	**200**	**1979**	**0.983**	**4.65×**

achieves 99% coverage with fewer samples. DeepGini heavily prioritizes low-confidence predictions, which occasionally correspond to outlier stimuli that activate the last few remaining critical faults. In contrast, Mirrored EL2N produces a more balanced ranking and achieves faster convergence overall. These complementary behaviors explain why Mirrored EL2N excels in AUC and early detection while remaining competitive at 99% coverage.

C. Aggregated Performance Across Networks

Table III summarizes the mean AUC and speed-up across all evaluated configurations. Mirrored EL2N achieves near-perfect detection curves (AUC ≈ 1) and the highest overall efficiency, requiring on average less than 0.2% of the samples needed by random selection to reach 99% coverage. DeepGini provides competitive early convergence but exhibits greater variability, while EL2N and Cross-Entropy remain less selective. These results confirm the robustness and architecture-agnostic nature of the proposed metric.

TABLE III: Average AUC and relative speed-up (with respect to random sampling) across the evaluated networks and datasets. Speed-up is computed at 99% of detected critical faults.

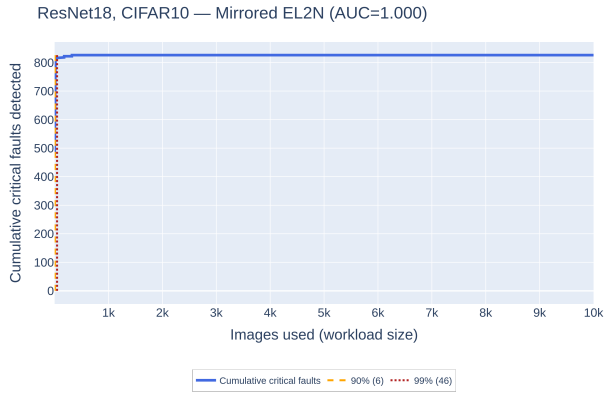
Metric	Mean AUC	Mean Speed-up @99%
Random (5× avg)	0.807	1.0×
Cross-Entropy	0.964	64×
DeepGini	0.982	274×
EL2N	0.960	65×
Mirrored EL2N	**0.990**	**557×**

D. Interpretation

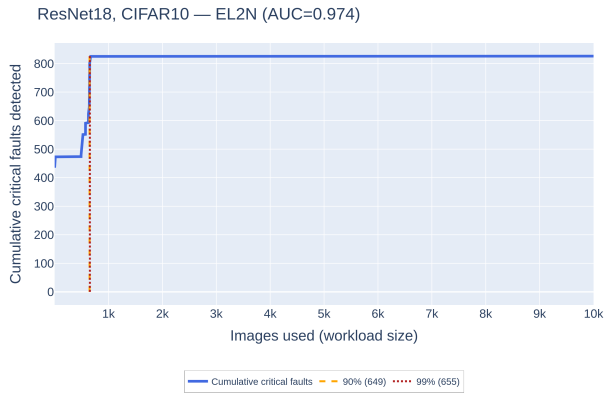
The superiority of Mirrored EL2N over EL2N can be traced back to its symmetric treatment of correctly and incorrectly classified samples. By mirroring the EL2N values of misclassified inputs, the metric maps both groups onto a single difficulty scale, in which high metric values correspond to samples close to the decision boundary, regardless of correctness. These inputs are precisely those most likely to activate and propagate fault effects. The resulting ranking improves the early fault-detection rate, increases the area under the detection curve, and enables substantial reductions in the required workload size. These results confirm that Mirrored EL2N provides a data-efficient and architecture-agnostic strategy for guiding fault injection campaigns. In practice, our results indicate that selecting the top 0.1–1% samples ranked by Mirrored EL2N is sufficient to expose the vast majority of critical faults across architectures.

VI. Conclusion

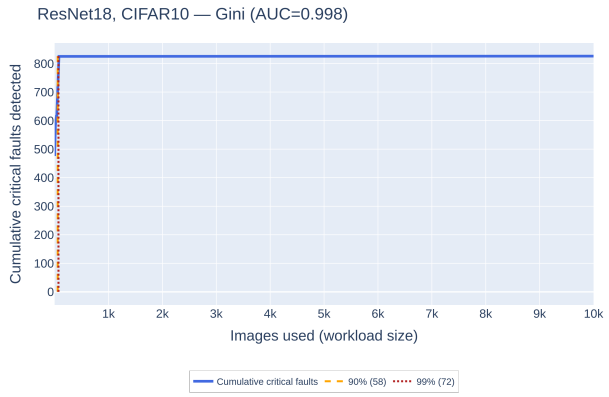
This work presented a data-efficient methodology for optimizing fault-injection campaigns in CNNs through sample-difficulty metrics. We introduced Mirrored EL2N, a symmetric reformulation of EL2N that aligns the difficulty scale of correct and incorrect samples and improves the prediction of fault sensitivity. Across multiple architectures, datasets, and a quantized model, Mirrored EL2N consistently accelerates the detection of critical faults, achieving speed-ups of up to three orders of magnitude over random selection while preserving near-perfect detection curves. The results confirm that fault activation is strongly concentrated in a small subset of difficult inputs and that exploiting this structure enables substantial reductions in workload size without compromising coverage. Future work will investigate how these findings can support more adaptive and automated fault-injection methodologies for safety-critical neural systems.



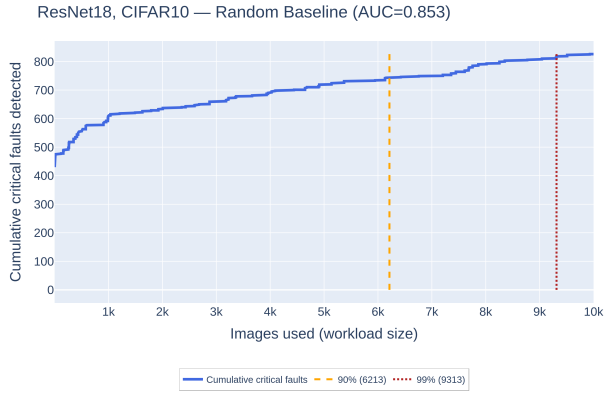
(a) MEL2N



(b) EL2N



(c) Gini-Index



(d) Random Sampling

Fig. 2: Cumulative fault-detection curves $C(n)$ for different selection metrics (ResNet18, CIFAR10).Mirrored EL2N achieves earlier convergence and the steepest initial growth, indicating superior FI efficiency.

ACKNOWLEDGEMENT

This work has been funded by the ANR France 2030 AdaptING project "ANR-23-PEIA-0009".

REFERENCES

[1] G. D. Natale et al., *Cross-Layer Reliability of Computing Systems*. Jan. 2020.

[2] A. Ruospo et al., "A survey on deep learning resilience assessment methodologies," *Computer*, 2023.

[3] N. Bellarmino et al., "Covid-19 detection from exhaled breath," *Scientific Reports*, 2024.

[4] N. Bellarmino et al., "Opra: Optimizing resiliency assessment for deep neural networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2025.

[5] N. Bellarmino et al., "Dear-cnn: Data-efficient assessment of resiliency in convolutional neural networks," in *2025 IEEE 28th International Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS)*, 2025.

[6] M. H. Ahmadilivani et al., "Special session: Approximation and fault resiliency of dnn accelerators," in *2023 IEEE 41st VLSI Test Symposium (VTS)*, 2023.

[7] A. Bosio et al., "A reliability analysis of a deep neural network," in *2019 IEEE Latin American Test Symposium (LATS)*, 2019.

[8] C. Qian et al., "A survey of bit-flip attacks on deep neural network and corresponding defense methods," *Electronics*, 2023.

[9] A. Ruospo et al., "Assessing convolutional neural networks reliability through statistical fault injections," in *2023 Design, Automation and Test in Europe Conference I& Exhibition (DATE)*, 2023.

[10] R. Leveugle et al., "Statistical fault injection: Quantified error and confidence," in *2009 Design, Automation & Test in Europe Conference & Exhibition*, 2009.

[11] M. Barbareschi et al., "Investigating the resilience source of classification systems for approximate computing techniques," *IEEE Transactions on Emerging Topics in Computing*, 2024.

[12] F. Meng et al., "Exploring image selection for self-testing in neural network accelerators," in *2022 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2022.

[13] F. Meng et al., "A self-test framework for detecting fault-induced accuracy drop in neural network accelerators," in *2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2021.

[14] S. A. El-Sayed et al., "Compact functional testing for neuromorphic computing circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023.

[15] M. Paul et al., "Deep learning on a data diet: Finding important examples early in training," in *Advances in Neural Information Processing Systems*, M. Ranzato et al., Eds., Curran Associates, Inc., 2021.

[16] K. He et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[17] Y. LeCun et al., "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.

[18] K. Simonyan et al., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] C. Yang et al., *Gated convolutional networks with hybrid connectivity for image classification*, https://github.com/winycg/HCGNet, 2020.

[20] Y. LeCun et al., *The mnist database of handwritten digits*, http://yann.lecun.com/exdb/mnist/, Accessed on April 11, 2024, 1998.

[21] A. Krizhevsky, *Cifar-10 dataset*, https://www.cs.toronto.edu/~kriz/cifar.html, Accessed on April 11, 2024, 2009.

[22] R. Leveugle et al., "Edge AI Trustworthiness: Revisiting Bit-Flip Impacts and Criticality Conditions," *Electronics*, 2025.

[23] Y. Feng et al., *Deepgini: Prioritizing massive tests to enhance the robustness of deep neural networks*, 2020.

[24] A. Mao et al., *Cross-entropy loss functions: Theoretical analysis and applications*, 2023.