K-SNAC: Robust Neuron Coverage for OOD Generalization and Test Adequacy

Seungwon Woo*
Department of Artificial Intelligence
University of Seoul
Seoul, South Korea
wsw20047@uos.ac.kr

Hyunseo Shin*
Department of Artificial Intelligence
University of Seoul
Seoul, South Korea
hseo98@uos.ac.kr

Eunkyung Choi*
Department of Artificial Intelligence
University of Seoul
Seoul, South Korea
rmarud202@uos.ac.kr

Juheon Kang*,†
Department of Artificial Intelligence
University of Seoul
Seoul, South Korea
2715wngjs@uos.ac.kr

Wonseok Hwang†
Department of Artificial Intelligence
University of Seoul
Seoul, South Korea
wonseok.hwang@uos.ac.kr

***Abstract*—The rigorous testing of Deep Neural Networks (DNNs) is essential for their deployment in safety-critical applications. Neuron coverage criteria have been widely adopted for test adequacy, yet popular metrics such as Strong Neuron Activation Coverage (SNAC) suffer from a critical flaw: their reliance on a single maximum activation value makes them unstable, sensitive to the size of test set and outliers, and weakly correlated with robustness. To address this limitation, we propose K-SNAC, a distribution-aware coverage metric that replaces the unstable maximum threshold with a statistically grounded one derived from the mean and standard deviation of neuron pre-activations. We theoretically analyze the instability of SNAC and empirically demonstrate that K-SNAC achieves significantly stronger correlations with established robustness measures, while also maintaining low dependence on test set size. Furthermore, we validate the effectiveness of K-SNAC as an indicator of Out-of-Distribution (OOD) generalization, showing that it better aligns with existing OOD metrics than prior coverage criteria. Together, these results establish K-SNAC as a reliable and robust adequacy measure that bridges test coverage, robustness evaluation, and OOD generalization.**

***Index Terms*—Deep Learning, Neuron Coverage, Test Adequacy, Adversarial Robustness, OOD Generalization.**

I. Introduction

The remarkable success of Deep Neural Networks (DNNs) [1] has driven their deployment in safety-critical systems, such as autonomous driving [2] and medical diagnostics [3], making rigorous testing and verification of these components urgently important [4]. Prediction accuracy, while the primary performance indicator, reflects only average-case behavior and offers little insight into how and why models fail under safety-critical or adversarial conditions. In fact, models with similar accuracies may exhibit distinct robustness and failure patterns [5]. To bridge this gap, a growing body of research has drawn inspiration from traditional software coverage testing [6], proposing various neuron coverage criteria to assess the quality of test suites. The foundational works introduced Neuron Coverage (NC) metric [7], soon followed by more granular criteria such as k-Multisection Neuron Coverage (KMNC) and Neuron Boundary Coverage (NBC) [8].

Although previous studies have reported that increasing neuron coverages can paradoxically degrade testing quality [9], coverage itself should be viewed not as a target to optimize, but as a diagnostic measure reflecting the interaction between the model and a given test set. Rather than manipulating test data to artificially raise coverage, analyzing the distribution of different coverage criteria can reveal which activation regions are underexplored or which types of adversarial perturbations a model is particularly vulnerable to, for example, those inducing abnormal over activation or under activation in specific layers. Such insights can inform robustness oriented retraining or guide decisions on complementary defense strategies.

Beyond these foundational works, Caglar et al. [10] recently introduced CleanAI, a white-box testing framework that integrates coverage and dependability metrics to evaluate model quality and reliability. This demonstrates that, even in recent years, there remains active interest in using coverage metrics as efficient and lightweight measures of testing adequacy.

Despite their widespread use, however, the efficacy of existing coverage criteria has been questioned, with studies indicating a weak correlation between coverage scores and a test suite's ability to uncover defects [9, 11]. This issue is particularly pronounced for Strong Neuron Activation Coverage (SNAC), which defines its coverage boundary using a single maximum activation value from the training data. While SNAC represents a reasonable and intuitive attempt to formalize corner-case neuron activations, the metric has an important limitation: its reliance on a single maximum activation value makes the threshold for an individual neuron disproportionately sensitive to statistical outliers. A benchmark built upon such unstable, per-neuron thresholds are inherently unreliable, resulting in a weak and inconsistent correlation with

* These authors contributed equally to this work.
† Corresponding authors.

2151-0849/25/$31.00 ©2025 IEEE
DOI 10.1109/ASEW67777.2025.00042

true model robustness [12]. In fact, we theoretically demonstrate that SNAC's reliance on such outliers allows its scores to be easily manipulated.

To address these limitations, we propose Strong Neuron Activation Coverage with K-sigma (K-SNAC), a distribution-aware coverage metric that replaces SNAC's outlier threshold with a statistically grounded criterion. By calculating each neuron's mean and standard deviation of pre-activation values from the training data, K-SNAC establishes a more stable baseline of normal behavior, thus enabling a probabilistic interpretation of coverage. This transition from an outlier driven benchmark to a distribution-based perspective allows K-SNAC to provide a more principled measure of test adequacy and a more interpretable basis for identifying anomalous activations. Our main contributions are as follows:

- We theoretically show that SNAC's maximum-value threshold is unstable and outlier-sensitive.
- We propose K-SNAC, a distribution-aware coverage using the mean and standard deviation of pre-activations.
- We demonstrate that K-SNAC has significantly stronger correlations with robustness metrics, including Loss Sensitivity [13], CLEVER scores (CL1, CL2, CLI) [14], and the global Lipschitz constant [15] than SNAC and other coverage criteria.
- We also show that K-SNAC consistently aligns more closely with NAC_ME [16], a widely used coverage-based OOD generalization metric.
- Finally, K-SNAC is far less sensitive to test set size, highlighting its stability as a practical adequacy measure.

From the perspective of explainable software engineering, coverage-based analysis offers a transparent, quantitative view of how a deep neural network utilizes its internal components during testing. By interpreting neuron activations as observable units of model behavior, our proposed metric (K-SNAC) enhances the explainability of testing outcomes—allowing developers to reason about which neurons, layers, or functional submodules contribute to robustness or failure. Consequently, K-SNAC bridges methodological advances in neural testing with the broader goal of building more interpretable and reliable AI systems.

II. Related Works

A. Limitations of Structural Coverage Metrics

Dong et al. [12] report that widely used structural coverage criteria such as NC, KMNC, NBC, and SNAC often exhibit weak and inconsistent correlation with a test suite's fault detection capability or robustness improvement, with this tendency being particularly pronounced in the cases of NBC and SNAC. Li et al. [17] argue that these criteria rely on arbitrary thresholds, which may not meaningfully represent the diversity of neuron behaviors. Yang et al. [9] provides updated empirical evidence that structural coverage still shows limited correlation with robustness and fault detection in modern architectures, reinforcing the need for alternative approaches.

B. Distribution-Aware Coverage Approaches

Kim et al. [18] propose Surprise Adequacy (SADL), which measures the novelty of a test input relative to the training distribution using activation trace analysis. Their method includes Likelihood-based Surprise Adequacy (LSA) and Distance-based Surprise Adequacy (DSA) to quantify "surprise" scores. Yuan et al. [19] revisit neuron coverage from a layer-wise perspective, aiming to account for coverage distribution across different network depths. Compared to these approaches, K-SNAC focuses on the statistical distribution of individual neuron pre-activations, using mean and standard deviation to define k-sigma thresholds.

C. OOD Detection and generalization

Lee et al. [20] demonstrate that many coverage-driven test generation methods produce out-of-distribution (OOD) inputs that can artificially inflate coverage metrics without revealing meaningful faults. Liu et al. [16] reframes neuron activation coverage as an OOD detection and generalization assessment tool, showing that coverage patterns can serve as lightweight indicators of distributional shifts. These works motivate the use of K-SNAC as a lightweight OOD detection heuristic by flagging inputs that cause statistically rare activations relative to the training distribution.

III. Methodology

A. Strong Neuron Activation Coverage (SNAC)

Strong Neuron Activation Coverage (SNAC) is a structural coverage criterion designed to capture corner-case behaviors of Deep Neural Networks (DNNs). For each neuron n in the network, SNAC first determines the maximum pre-activation value across the entire training dataset. A test input x is considered to *cover* neuron n under SNAC if the neuron's pre-activation value $a_n(x)$ exceeds this maximum:

$$\mathrm{COVER}_{\mathrm{SNAC}}(n, x) = \mathbb{I}\left[a_n(x) > \max_{x \in \mathcal{D}_{train}} a_n(x)\right] \quad (1)$$

where $\mathbb{I}[\cdot]$ is the indicator function. The overall SNAC score is computed as the ratio of neurons covered by at least one test input to the total number of neurons in the network.

Although SNAC is intuitive and easy to compute, its use of a single maximum value per neuron as a threshold makes it highly sensitive to outliers. This can lead to instability and weak, inconsistent correlations with model robustness and fault detection performance.

B. Instability of SNAC's Maximum-Value Threshold

We provide a theoretical motivation demonstrating why SNAC's maximum-value-based threshold is an unstable and unreliable standard.

a) Probability Analysis under Identical Distributions: Let us assume that the training set $\mathcal{D}_{train}$ and the test set $\mathcal{D}_{test}$ are drawn independently and identically (i.i.d.) from the same underlying distribution $\mathcal{P}$. The SNAC coverage criterion for a neuron n is defined as the maximum pre-activation T_n^{SNAC}, observed in the training set:

$$T_n^{\text{SNAC}} = \max_{x \in \mathcal{D}_{train}} a_n(x), \tag{2}$$

where $a_n(x)$ denotes the pre-activation of neuron n for input x.

Under the i.i.d. assumption, principles from order statistics show that the probability of a test input exceeding this threshold can be approximated as:

$$\mathbb{P}\big[\exists x \in \mathcal{D}_{test} : a_n(x) > T_n^{\text{SNAC}}\big] \;\approx\; \frac{|\mathcal{D}_{test}|}{|\mathcal{D}_{train}| + |\mathcal{D}_{test}|}. \tag{3}$$

This ratio emerges because, in the combined pool of $|\mathcal{D}_{train}| + |\mathcal{D}_{test}|$ samples, the global maximum is equally likely to originate from any given sample. Consequently, in the common scenario where $|\mathcal{D}_{train}| \gg |\mathcal{D}_{test}|$, the probability of coverage for any individual neuron becomes vanishingly small. Since this represents the probability of covering any individual neuron, the expected value of the overall SNAC score—which by definition is the fraction of covered neurons—is also equivalent to this small probability. This statistical property is the fundamental reason why SNAC scores are typically very low, particularly in the common scenario where the training set is much larger than the test set ($|\mathcal{D}_{train}| \gg |\mathcal{D}_{test}|$) and the test data originates from the same distribution.

b) Manipulating SNAC via Small Perturbations: Consider a training set $\mathcal{D}_{train}$ and let $x^* \in \mathcal{D}_{train}$ be the input that determines T_n^{SNAC} for neuron n. We construct an adversarial variant $\tilde{x}^*$ using the Fast Gradient Sign Method (FGSM) [21], but with the attack objective specifically designed to *increase the pre-activation of neuron* n:

$$\tilde{x}^* = x^* + \epsilon \cdot \operatorname{sign}(\nabla_x a_n(x)), \tag{4}$$

where ϵ is a small perturbation bound in the input space. For sufficiently small ϵ, $\tilde{x}^*$ remains statistically similar to x^*, yet is crafted to exceed T_n^{SNAC} with high probability by explicitly maximizing the activation $a_n(x)$.

By selectively injecting such perturbed samples into the training set at a ratio $\rho \in [0, 1]$, we can directly control the SNAC score:

- $\rho = 0$: No perturbed samples; coverage probability ≈ 0.
- $\rho \to 1$: Coverage probability approaches its maximum possible value.

C. K-SNAC: A Distribution-Aware Coverage Metric

To address SNAC's instability, we propose **Strong Neuron Activation Coverage with K-sigma (K-SNAC)**, a distribution-aware variant that replaces the outlier-sensitive maximum threshold with a statistically grounded one.

For each neuron n, K-SNAC calculates the mean μ_n and standard deviation σ_n of its pre-activation values across the training set. The coverage threshold is then defined as:

$$T_n^{\text{K-SNAC}} = \mu_n + k \cdot \sigma_n, \tag{5}$$

where k is a parameter that allows practitioners to set the coverage threshold based on statistical confidence levels (e.g., 95%, 99%), enabling a multi-level analysis. A neuron n is considered covered for an input x if:

$$\text{COVER}_{\text{K-SNAC}}(n, x) = \mathbb{I}\big[a_n(x) > T_n^{\text{K-SNAC}}\big]. \tag{6}$$

The K-SNAC score is the fraction of neurons covered by at least one test input under these k-sigma thresholds.

By grounding the threshold in the statistical distribution of training activations, K-SNAC maintains a lightweight computation by adopting an incremental algorithm: for each neuron n, we store the cumulative sum of activations (denoted as P_n) and the cumulative squared sum of activations (denoted as Q_n) across the training data. Using these two quantities, the mean and variance are computed as

$$\mu_n = \frac{P_n}{|\mathcal{D}_{\text{train}}|}, \qquad \sigma_n^2 = \frac{Q_n}{|\mathcal{D}_{\text{train}}|} - \mu_n^2. \tag{7}$$

From a statistical perspective, the $k\sigma$ threshold corresponds to a specific upper-tail quantile of the activation distribution. Under the assumption of an approximately Gaussian distribution, k can be directly mapped to common confidence levels: for example, 95% corresponds to $k \approx 1.645$, and 99% to $k \approx 2.326$. This mapping allows practitioners to set k with an explicit probabilistic interpretation, ensuring that coverage reflects statistically rare events while remaining robust to outliers and resistant to adversarial manipulation.

IV. Experiments

A. Models and Datasets

We perform our evaluation using two families of architectures: the LeNet family (LeNet-1, LeNet-4, and LeNet-5) and the ResNet family (ResNet-18, ResNet-34, and ResNet-50). The LeNet models are trained on MNIST, while each ResNet architecture is trained separately on MNIST, CIFAR-10 and CIFAR-100, providing a broader spectrum of architectures and data domains.

For each architecture–dataset pair, we first train a baseline model on the original clean dataset. In addition, we construct two adversarial training sets by perturbing the data using FGSM and PGD [22], respectively. Each adversarial dataset is further combined with the original clean data to create "mixed" training sets that include both clean and adversarial examples. This results in three variants (clean, FGSM-trained, PGD-trained) for each architecture–dataset pair. Overall, we obtain 9 models from the LeNet family (trained on MNIST), 27 models from the ResNet family (trained separately on MNIST, CIFAR-10 and CIFAR-100) yielding a total of 36 trained models.

B. Evaluation Protocol

For each of the 36 trained models, we evaluate coverage, robustness, and OOD generalization metrics on test subsets of increasing size. For the nine baseline (clean) models, we sample 1,000 to 10,000 images in 1,000 increments, yielding 10 evaluation points per model ($12 \times 10 = 120$). For 24 adversarially trained models, we evaluate on mixed test sets formed by concatenating clean and adversarial examples, from 1,000 to 20,000 in 1,000 increments, yielding 20 points per model ($24 \times 20 = 480$). In total, this procedure yields **600** evaluation points across all models.

For each evaluation point, we compute a comprehensive set of metrics:

- **Coverage criteria:** NC, KMNC, NBC, TKNC, TKNP, SNAC, and our proposal K-SNAC95 and K-SNAC99.
- **Robustness metrics:** Loss Sensitivity, CLEVER-based scores (CL1, CL2, CLI), and the global Lipschitz Constant.
- **OOD generalization metric:** NAC_ME

C. Correlation Analysis Protocol

Following Dong et al. [12], we adopt correlation analysis as our primary evaluation protocol. For each coverage criterion, we compute Pearson correlations with the defined robustness and OOD generalization metrics. Since metric scales (e.g., Lipschitz constants) differ substantially across model families, we calculate correlations separately within each family and then report averaged results.

V. Results and Analysis

We expect K-SNAC to exhibit stronger correlations (i.e., correlation coefficients closer to ± 1) with robustness-related metrics—such as CLEVER-based scores, the Lipschitz constant, and Loss Sensitivity—as well as with the OOD generalization measure NAC_ME, while showing only weak correlations (i.e., values closer to 0) with Test Size.

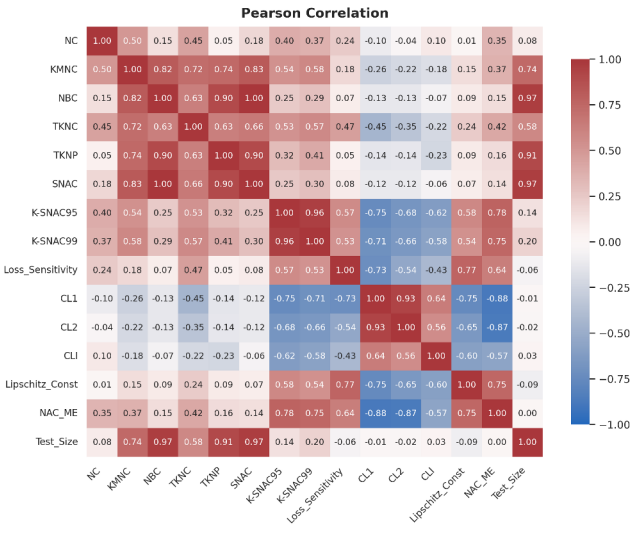

Fig. 1. Overall results with various metrics.

Our overall results with all metrics are summarized in Fig. 1. A key observation is the strong dependence of most traditional coverage criteria on test set size. In particular, SNAC and NBC exhibit correlations of 0.97 with test size, indicating that their reported coverage values are almost entirely determined by the number of test samples rather than by meaningful model behavior. In contrast, our proposed metric shows only weak correlations (0.14, 0.20) with test size, demonstrating that it provides a more stable adequacy measure independent of dataset size.

Beyond dataset size, K-SNAC also shows more consistent relationships with robustness and OOD generalization metrics. SNAC, for example, shows correlations close to zero with robustness indicators, whereas K-SNAC achieves strong negative correlations (-0.75, -0.68, -0.62, -0.71, -0.66, -0.58) with CLEVER-based scores, meaning that higher K-SNAC coverage corresponds to greater adversarial robustness. A positive correlation (0.58, 0.54) is also observed with the Lipschitz constant, indicating alignment with global robustness bounds. Moreover, K-SNAC exhibits high correlations with NAC_ME (0.78, 0.75) and loss sensitivity (0.57, 0.53), suggesting that it captures both OOD generalization and gradient-based stability.

Among the baselines, only TKNC shows a modest positive correlation with loss sensitivity, while others exhibit virtually no relationship. However, these effects are minor compared to the strong and consistent correlations observed for K-SNAC. Overall, the key takeaway from Fig. 1 is that K-SNAC uniquely maintains stability with respect to dataset size while aligning closely with multiple robustness and OOD generalization indicators, as reflected in the dark blue and red regions of the correlation map.

VI. Limitation and Future Work

While K-SNAC shows clear advantages in terms of robustness, OOD generalization, and reduced dependence on test size, our study remains limited in scope. The current design of K-SNAC relies on a Gaussian assumption and a fixed $k\sigma$ threshold, leaving open questions about how to define outliers in activation distributions. In practice, however, the Gaussianity of neuron activations is only approximate. The effective theory of deep learning [23] indicates that deviations from Gaussianity increase with network depth and diminish with width, suggesting that non-Gaussian models may yield biased thresholds or unstable coverage estimates. The choice of k may also depend on test size, yet there is no theoretical guideline for setting it.

Moreover, our evaluation primarily focuses on correlation-based evidence rather than end-to-end improvements. In particular, K-SNAC has not yet demonstrated direct performance gains in fault discovery, coverage-guided test generation, or OOD detection rates. Future work will aim to establish a more principled framework for threshold selection that accounts for data size, non-Gaussian behavior, and the statistical meaning of outliers, as well as to integrate K-SNAC into downstream testing pipelines to verify its practical impact on end-to-end robustness and generalization.

ACKNOWLEDGEMENT

We are grateful to Prof. Sangki Ko, who ran the course "Trustworthy AI" and gave us the motivation for this paper. We also thank the members of IDA.L Lab. (Hongseok Oh and Kyeongjun Cho) for their support.

Juheon Kang was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2025-25433878).

Wonseok Hwang was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-23524855).

REFERENCES

[1] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[2] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[4] J. Zhang and J. Li, "Testing and verification of neural-network-based safety-critical control software: A systematic literature review," *Information and Software Technology*, vol. 123, p. 106296, 2020.

[5] D. Stutz, M. Hein, and B. Schiele, "Disentangling adversarial robustness and generalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6976–6987.

[6] M. R. Lyu, J. Horgan, and S. London, "A coverage analysis tool for the effectiveness of software testing," *IEEE transactions on reliability*, vol. 43, no. 4, pp. 527–535, 2002.

[7] K. Pei, Y. Cao, J. Yang, and S. Jana, "Deepxplore: Automated whitebox testing of deep learning systems," in *proceedings of the 26th Symposium on Operating Systems Principles*, 2017, pp. 1–18.

[8] L. Ma, F. Juefei-Xu, F. Zhang, J. Sun, M. Xue, B. Li, C. Chen, T. Su, L. Li, Y. Liu *et al.*, "Deepgauge: Multi-granularity testing criteria for deep learning systems," in *Proceedings of the 33rd ACM/IEEE international conference on automated software engineering*, 2018, pp. 120–131.

[9] Z. Yang, J. Shi, M. H. Asyrofi, and D. Lo, "Revisiting neuron coverage metrics and quality of deep neural networks," in *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2022, pp. 408–419.

[10] O. Caglar, C. Baglum, and U. Yayan, "Cleanai: Deep neural network model quality evaluation tool," *SoftwareX*, vol. 29, p. 102015, 2025.

[11] F. Harel-Canada, L. Wang, M. A. Gulzar, Q. Gu, and M. Kim, "Is neuron coverage a meaningful measure for testing deep neural networks?" in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, pp. 851–862.

[12] Y. Dong, P. Zhang, J. Wang, S. Liu, J. Sun, J. Hao, X. Wang, L. Wang, J. S. Dong, and D. Ting, "There is limited correlation between coverage and robustness for deep neural networks," *arXiv preprint arXiv:1911.05904*, 2019.

[13] J. Z. Zhu, D. Hwang, and A. Sadjadpour, "Loss sensitivity calculation and analysis," in *2003 IEEE Power Engineering Society General Meeting (IEEE Cat. No. 03CH37491)*, vol. 2. IEEE, 2003, pp. 962–967.

[14] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel, "Evaluating the robustness of neural networks: An extreme value theory approach," *arXiv preprint arXiv:1801.10578*, 2018.

[15] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *International conference on machine learning*. PMLR, 2017, pp. 854–863.

[16] Y. Liu, C. X. Tian, H. Li, L. Ma, and S. Wang, "Neuron activation coverage: Rethinking out-of-distribution detection and generalization," *arXiv preprint arXiv:2306.02879*, 2023.

[17] Z. Li, X. Ma, C. Xu, and C. Cao, "Structural coverage criteria for neural networks could be misleading," in *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*. IEEE, 2019, pp. 89–92.

[18] J. Kim, R. Feldt, and S. Yoo, "Guiding deep learning system testing using surprise adequacy," in *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 2019, pp. 1039–1049.

[19] Y. Yuan, Q. Pang, and S. Wang, "Revisiting neuron coverage for dnn testing: A layer-wise and distribution-aware criterion," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 1200–1212.

[20] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[23] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge, UK: Cambridge University Press, 2021.