

Revisiting Deep Neural Network Test Coverage from the Test Effectiveness Perspective

Journal:	<i>Journal of Software: Evolution and Process</i>
Manuscript ID	JSME-22-0232
Wiley - Manuscript type:	Research Article - Empirical
Date Submitted by the Author:	05-Sep-2022
Complete List of Authors:	Yan, Ming; Tianjin University Chen, Junjie; Tianjin University Cao, Xuejie; Tianjin University Wu, Zhuo; Tianjin University Kang, Yuning; Tianjin University Wang, Zan; Tianjin University
Editor Selection:	Guest Editor of the Special Issue
Keywords:	Deep Neural Network, Testing Criterion, Empirical Study

SCHOLARONE™
Manuscripts

RESEARCH ARTICLE - EMPIRICAL

Revisiting Deep Neural Network Test Coverage from the Test Effectiveness Perspective

Ming Yan¹ | Junjie Chen¹ | Xuejie Cao¹ | Zhuo Wu²
| Yuning Kang¹ | Zan Wang^{1,2}

¹College of Intelligence and Computing,
Tianjin University, Tianjin, Tianjin, 300350,
China
²School of new media and communication,
Tianjin University, Tianjin, Tianjin, 300072,
China

Correspondence
Junjie Chen, College of Intelligence and
Computing, Tianjin University, Tianjin,
Tianjin, 300350, China
Email: junjiechen@tju.edu.cn

Funding information
National Natural Science Foundation of
China, Grant/Award Number: 62002256

Many test coverage metrics have been proposed to measure the Deep Neural Network (DNN) testing effectiveness, including structural coverage and non-structural coverage. These test coverage metrics are proposed based on the fundamental assumption: they are correlated with test effectiveness. However, the fundamental assumption is still not validated sufficiently and reasonably, which brings question on the usefulness of DNN test coverage. This paper conducted a revisiting study on the existing DNN test coverage from the test effectiveness perspective, to effectively validate the fundamental assumption. Here, we carefully considered the diversity of subjects, three test effectiveness criteria, and both typical and state-of-the-art test coverage metrics. Different from all the existing studies that deliver negative conclusions on the usefulness of existing DNN test coverage, we identified some positive conclusions on their usefulness from the test effectiveness perspective. In particular, we found the complementary relationship between structural and non-structural coverage and identified the practical usage scenarios and promising research directions for these existing test coverage metrics.

KEYWORDS
Deep Neural Network, Testing Criterion, Empirical Study

1 | INTRODUCTION

Deep neural networks (DNNs) have been widely studied in recent years and have achieved great success in many domains, e.g., autonomous driving cars [1], face recognition [2], machine translation [3], medical diagnosis [4], and code analysis [5, 6]. However, like traditional software systems, DNN also contains bugs [7, 8, 9]. DNN bugs could lead to unexpected behaviors, even disasters in safety-critical domains. For example, an Uber autonomous driving car killed a pedestrian in Tempe, Arizona in 2018 [10]. Therefore, it is very critical to ensure the DNN quality.

DNN testing is one of the most effective methods to ensure the DNN quality. As with traditional software testing [11, 12, 13, 14, 15, 16, 17, 18], one important aspect in DNN testing is to measure test effectiveness. Indeed, many DNN test coverage metrics have been proposed to achieve this goal in recent years, e.g., neuron coverage measuring which DNN elements are covered during the testing process [7, 8] and surprise coverage measuring the difference of DNN behaviors between test inputs and training data [19]. All these test coverage metrics are proposed based on the same assumption: *the proposed test coverage metric is correlated with test effectiveness*.

However, this assumption is not sufficiently validated till now, which also brings question on the usefulness of DNN test coverage. Although some experiments have been conducted when these test coverage metrics were proposed, they ignored the influence of some reasonable factors such as test-set sizes. For example, in the existing experiments [8, 19], they first generated the same number of adversarial test inputs as the number of original test inputs (also called natural test inputs), and then compared the test coverage of natural test inputs and that of both natural and adversarial test inputs. They found that the latter is larger than the former, and thus concluded that the test coverage metrics are correlated with test effectiveness (measured by the generation of adversarial test inputs). However, the experiments did not control for the number of test inputs when comparing test coverage, and thus the increased coverage may be caused by the generation of adversarial test inputs or just increasing the number of test inputs.

Further, we conducted a small experiment to investigate whether adding natural test inputs can also increase test coverage by taking the model LeNet-5 based on MNIST and the test coverage KMNC (to be introduced in Section 2.1) as an example. We first randomly selected 5,000 test inputs from MNIST as the original test set and regarded the remaining 5,000 test inputs in MNIST as the newly added natural test inputs, and then generated 5,000 adversarial test inputs via C&W [20] (to be introduced in Section 3.4), a widely-used adversarial input generation method. We found that the KMNC values of the original test set, the original test set and added natural test inputs, and the original test set and adversarial test inputs are 61.53%, 68.77%, and 65.87%, respectively. The results demonstrate that adding natural test inputs can also increase test coverage, even the increment is larger than that achieved by the same number of adversarial test inputs. Hence, it is still unclear whether these test coverage metrics are correlated with test effectiveness by considering reasonable factors (e.g., controlling for the size of test sets).

Besides the experiments conducted in the work proposing the corresponding DNN test coverage metrics, an empirical study was also conducted to investigate the correlation between DNC [7] (the earliest DNN test coverage, to be introduced in Section 2.1) and test effectiveness through using DNC to guide test input generation [21]. They found increasing DNC is not a meaningful objective for generating DNN test inputs. However, this study only investigated the earliest DNN test coverage metric, which cannot validate the fundamental assumption sufficiently. Furthermore, there is another study investigating the usefulness of DNN test coverage from the perspective of model retraining quality (rather than the perspective of test effectiveness) [22], and they found DNN test coverage does not have monotonic relations with model retraining quality. Moreover, all these experiments and studies are conducted based on the subjects with limited diversity. To sum up, the fundamental assumption for DNN test coverage is still not validated sufficiently and reasonably. That is, *the usefulness of DNN test coverage is still unclear from the test effectiveness perspective*.

To answer the open question, in this work we conducted an extensive study to revisit DNN test coverage from the test effectiveness perspective. To sufficiently validate the fundamental assumption, we consider three test effectiveness criteria in DNN testing. 1) **Error-revealing capability**. Same as traditional software testing, revealing errors is also the most straightforward way of representing test effectiveness in DNN testing. Here, we used the ratio of error-revealing test inputs in test sets to reflect the error-revealing capability of test sets, where error-revealing test inputs refer to those making the DNN produce incorrect predictions. 2) **Test diversity**. In DNN testing, besides paying attention to the overall accuracy of DNNs, it is also important to keep a watchful eye on *each-class* accuracy for classification DNNs. Here, we used the number of covered classes to represent test diversity, since it requires to ensure the existence of test inputs for testing each class before measuring each-class accuracy. 3) Subsequently, we have the third criterion, i.e., **error-revealing capability per diversity unit**, which is represented by the ratio of error-revealing test inputs in test sets that cover only a small number of classes (i.e., we regard 10% of classes as a diversity unit in our study) rather than all classes (i.e., saturated diversity). It can help reflect the diverse error-revealing capability in DNN testing. In particular, to reasonably validate the fundamental assumption, we controlled the influence of the size of test sets when investigating the correlation between DNN test coverage and the three test effectiveness criteria.

In particular, we conducted a comprehensive study on DNN test coverage. In our study, we carefully considered the diversity of subjects (10 pairs of models and datasets in total), involving different tasks, domains, DNN types, and model accuracy. Besides the widely-used subjects in the existing studies [7, 8, 19, 22, 21, 23], we also added five subjects that have never been used to evaluate any of the studied test coverage metrics before. Besides, we not only studied all the DNN test coverage metrics used in the existing studies (except TKNP [8] since it is not a ratio but rather a number), but also added the state-of-the-art test coverage (i.e., IDC) [24].

Different from the existing studies [22, 21], which delivered negative conclusions on the usefulness of existing DNN test coverage, our conclusions are relatively positive in general because we identified the practical usage scenarios for these existing DNN test coverage by investigating them from the test effectiveness perspective more sufficiently and reasonably. Our study demonstrates that the studied structural coverage (i.e., DNC, TKNC, KMNC, and IDC) is complementary with the studied non-structural coverage (i.e., LSC and DSC), and they have different usage scenarios. Specifically, the studied non-structural coverage (i.e., LSC and DSC [19]) has positive correlation with error-revealing capability regardless of saturated diversity or one diversity unit, but has no obvious positive correlation with test diversity. Hence, they can be used to guide the generation of error-revealing test inputs but it is hard to guarantee the diversity of the generated test inputs guided by them. The studied structural coverage (i.e., DNC, TKNC, KMNC, and IDC) does not have positive correlation with error-revealing capability under saturated diversity, but has positive correlation under small test diversity. Hence, their usage scenario could be guiding the generation of error-revealing test inputs by taking a small number of classes of test inputs (rather than the whole test set with saturated diversity) as seed inputs¹. Also, they are positively correlated with test diversity, and thus the guidance of these structural metrics can also help improve the diversity of generated test inputs. More interesting findings and implications/suggestions can be found in Sections 4 and 5.

To sum up, our work makes the following major contributions:

- We conducted a comprehensive study on DNN test coverage, which revisits the existing DNN test coverage metrics from the test effectiveness perspective.
- We obtained some positive conclusions on the usefulness of existing DNN test coverage (which is different from all the existing studies) and identified the practical usage scenarios and promising research directions for them.

¹In existing DNN fuzzing and adversarial input generation methods, new test inputs are generated based on the given set of test inputs (also called seed inputs).

- We optimized and integrated all the implementations of collecting these test coverage metrics, and the coverage collection efficiency has been largely improved after our optimization (e.g., 89.83%~95.54% improvements on LeNet-5 and MNIST). We have released our toolbox and all the experimental data on our project homepage [25].

2 | BACKGROUND

In DNN testing, many test coverage metrics have been proposed to measure test effectiveness [7, 8, 19]. In our study, we not only studied all the test coverage metrics considered in the existing studies [22, 21] (except TKNP [8] since it is not a ratio but rather a number) but also the recently-proposed test coverage (i.e., IDC [24]), in order to conduct a more comprehensive revisiting study. According to whether considering structural elements are covered during testing, we classify test coverage metrics into two categories: *structural coverage* and *non-structural coverage*. Here, we introduce these studied test coverage metrics in our work, and the details about other coverage will be presented in Section 7.

2.1 | Structural Coverage

Structural coverage considers which structural elements are covered during testing, mainly referring to various neuron coverage. Here, we studied six typical structural coverage metrics. The earliest one was proposed by Pei et al. [7], which is called **DNC** (DeepXplore's Neuron Coverage). DNC divides the state of a neuron into *activated* and *non-activated*, and measures the ratio of activated neurons in a DNN. If the output value of a neuron is larger than a pre-defined threshold t after executing a test input, DNC regards the neuron as an activated neuron.

Ma et al. [8] further proposed a set of neuron coverage metrics at several granularity levels, including TKNC (Top-K Neuron Coverage), KMNC (K-Multisection Neuron Coverage), NBC (Neuron Boundary Coverage), and SNAC (Strong Neuron Activation Coverage). **TKNC** is a layer-level metric, which considers top-k neurons (ranked in the descending order of their output values after executing a test input) in a layer to be covered. It measures the ratio of neurons that have once been top-k neurons in the corresponding layer. **KMNC** is a more fine-grained metric, which first obtains the range of its output values for each neuron from training data (denoted as $[low_i, high_i]$ for the neuron n_i) and then partitions the range into k sections. If the output value of a neuron falls into a section after executing a test input, KMNC regards the section as a covered section. KMNC measures the ratio of covered sections of all neurons in a DNN. The insight behind KMNC is that, different sections may represent different functional modules of a DNN and more fine-grained neuron coverage can increase discrimination of different test inputs. **NBC** also obtains the range of its output values for each neuron from training data like KMNC. The difference is that, NBC considers whether the corner-case region, i.e., $(-\infty, low_i)$ and $(high_i, +\infty)$, of a neuron is covered after executing a test input. **SNAC** is similar to NBC, but the difference is that SNAC only considers whether the upper corner-case region, i.e., $(high_i, +\infty)$, is covered after executing a test input.

Recently, Gerasimou et al. [24] proposed a novel test coverage metric **IDC** (Importance-Driven Coverage) that has never been investigated by the existing studies [22, 21]. Its key insight is that more important neurons have a stronger causal relationship with other neurons and can explain which high-level features contribute more to decision-making, and thus the importance of each neuron is measured by the widely-used relevance score [26]. Then, IDC partitions the space of each neuron's activity into a set of clusters via neuron-wise quantization, and measures the ratio of combinations of important neurons clusters covered by a test set.

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

2.2 | Non-Structural Coverage

Non-structural coverage does not consider whether structural elements of a DNN are covered during testing. The most typical non-structural coverage is surprise coverage [19]. It regards the difference of DNN behaviors between a test input and training data as surprise adequacy (SA) of the test input. Surprise coverage measures the range of SA values that a set of test inputs cover. Specifically, Kim et al. [19] proposed two kinds of SA: Likelihood-based Surprise Adequacy (LSA) and Distance-based Surprise Adequacy (DSA). LSA refers to the surprise of a test input with respect to the estimated density of each activation value in a set of activation traces observed over the neurons of a selected layer for training data via Kernel Density Estimation [27]. LSA only considers the training data whose label is the same as the predicted class of the test input. DSA is defined using the Euclidean distance between the activation trace observed over the neurons of a selected layer for a test input and a set of activation traces for training data. DSA needs find the closest neighbor of the test input that shares the same class, denoted as x_a , and also the closest neighbor of x_a in a different class.

According to LSA and DSA, the corresponding surprise coverage metrics are called **LSC**, and **DSC** respectively. Specifically, given the upper and lower bounds of LSA/DSA denoted as U and L , LSC/DSC first divides $[L,U]$ into n sections and then measures the ratio of covered sections. If the LSA/DSA value of a test input falls into a section, LSC/DSC regards the section as a covered section. In particular, a test input with a quite large SA value may be irrelevant to the problem domain (e.g., an image of an apple is irrelevant to the testing of traffic sign classifiers), and thus the parameter U can be used to filter out those irrelevant test inputs. Please note that DSC is not applicable to regression models [19].

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

3 | STUDY DESIGN

Our study aims to revisit the existing DNN test coverage metrics from the test effectiveness perspective, thus answering the open question: *are these existing metrics correlated with test effectiveness*. As presented in Section 1, we consider three test effectiveness criteria (i.e., error-revealing capability , test diversity, and error-revealing capability per diversity unit) in DNN testing to sufficiently validate the fundamental assumption. Specifically, we aim to answer it by addressing the following research questions (RQs):

- **RQ1:** Are the studied test coverage metrics correlated with the error-revealing capability of test sets (measured by the ratio of error-revealing test inputs in test sets)?
- **RQ2:** Are the studied test coverage metrics correlated with the test diversity of test sets (measured by the ratio of covered classes in test sets)?
- **RQ3:** Are the studied test coverage metrics correlated with error-revealing capability per diversity unit?

Please note that for a classification model, we treat covering all classes in a test set as *saturated diversity* and covering 10% of classes as *a diversity unit*. That is, RQ3 aims to investigate the correlation between test coverage and the error-revealing capability of test sets that covers 10% of classes. Actually, RQ1 investigates the correlation under saturated diversity and RQ3 is inspired by the conclusions from RQ1 and RQ2.

45

46

47

48

49

50

51

52

53

54

55

3.1 | Datasets and DNN Models

In our study, we used 10 pairs of datasets and DNN models as subjects in total. Table 1 shows the basic information on subjects, where Columns 4-10 represent the model size, the number of training instances, the number of original test

TABLE 1 Datasets and DNN models

ID	Dataset	Model	Size(KB)	#Train	#Test	Acc(%)	Task	Domain	Network	Used
1	MNIST	LeNet-5	1,300	60,000	10,000	98.72	classification	image	CNN	✓
2	Fashion-MNIST	LeNet-5	1,279	60,000	10,000	91.07	classification	image	CNN	✗
3	CIFAR-10	AlexNet	9,830	50,000	10,000	76.64	classification	image	CNN	✗
4	CIFAR-10	ResNet-20	3,507	50,000	10,000	91.21	classification	image	CNN	✓
5	CIFAR-10	VGG-16	19,639	50,000	10,000	87.41	classification	image	CNN	✓
6	CIFAR-100	ResNet-32	10,615	50,000	10,000	70.52	classification	image	CNN	✗
7	Driving	Dave-orig	8,306	101,396	5,614	90.33	regression	image	CNN	✓
8	Driving	Dave-dropout	13,139	101,396	5,614	91.82	regression	image	CNN	✓
9	Speech-Commands	DeepSpeech	6,734	51,776	6,471	94.53	classification	speech	RNN	✗
10	20-Newsgroups	TextCNN	38,974	11,314	7,532	77.68	classification	text	CNN	✗

We use 1 - MSE (Mean Square Error) as the accuracy of regression models. ✓ represents the subject has been used to evaluate at least one of the studied test coverage metrics before. ✗ represents the subject has never been used to evaluate any of the studied test coverage metrics before.

inputs, the model accuracy, the model task, the domain of the model, and the network type, respectively. Specifically, MNIST is a handwritten digit dataset [28], Fashion-MNIST is a MNIST-like dataset with greyscale images of Zalando's items [29]. CIFAR-10 and CIFAR-100 are 10-class and 100-class ubiquitous object datasets respectively [30], Driving is an Udacity-provided autonomous driving dataset [31], and Speech-Commands is a sequential dataset containing a set of one-second .wav audio files, each of which contains a single spoken English word [32]. 20-Newsgroups is a text dataset containing about 20,000 newsgroup documents, which has been widely used in natural language processing (NLP) tasks (e.g., text classification [33]).

To more comprehensively study these test coverage metrics, we *carefully considered the diversity of subjects*, including 1) **different tasks**: normal classification, multi-label classification (DeepSpeech), and regression; 2) **different domains**: image domains, speech domains, and text domains; 3) **different DNN types**: CNN and RNN; and 4) **different model accuracy**: from 70.52% to 98.72%.

In our study, we not only selected five subjects that have been used to evaluate at least one of the studied test coverage metrics before without any subjective bias, but also used **five subjects that have never been used to evaluate any of the studied test coverage metrics before**, which is shown in the last column in Table 1.

3.2 | Implementations

Since the code for collecting DNC, LSC, DSC, and IDC is released by the existing work [7, 19, 24], we directly adapted the code to collect them in our study. As for TKNC, KMNC, NBC, and SNAC, we communicated with the corresponding authors and they shared the implementations with us. After carefully reviewing the implementations, we discussed with some authors about potential problems in code, and further optimized the code for KMNC, TKNC, NBC, SNAC, and IDC in order to improve their coverage collection efficiency. Table 2 shows the time spent on collecting these coverage metrics when executing 10,000 test inputs of the subject (ID: 1) before and after our optimization. Column *Original* presents the time spent using the original implementation without our optimization, while Column *Optimized* presents the time spent using our optimized implementation. The last column shows the efficiency improvement of our optimized implementation over the original one. We found that the coverage collection efficiency is largely improved, i.e., 89.83%~95.54% improvements. In particular, we checked that the coverage results are the same before and after our optimization.

TABLE 2 Coverage collection efficiency comparison between before and after our optimization for TKNC, KMNC, NBC, SNAC, and IDC on the subject (ID: 1) (seconds)

Coverage	Original	Optimized	$\uparrow rate(\%)$
TKNC	40.10	3.40	91.52%
KMNC	76.30	3.40	95.54%
NBC	59.10	3.40	94.25%
SNAC	57.20	3.30	94.23%
IDC	4,465.80	454.00	89.83%

To promote future research and practical usage, we have developed a toolbox that integrates the implementations of collecting the eight test coverage, each of which can be invoked by simply specifying the corresponding parameter value. Our toolbox and all the experimental data have been released on our project homepage [25]

For the parameters of these studied coverage metrics, we set them following the existing work [19, 7, 8, 24]. The t value of DNC is set to 0.5, the k value of TKNC is set to 10, the k value of KMNC is set to 1,000, and the n value of LSC and DSC is set to 1,000. LSC and DSC also have some parameters required to be specifically set for different subjects [19], i.e., the selected layer to calculate LSA/DSA, the upper bound U and lower bound L of LSC/DSC. However, there is no guide provided to help set them for different subjects. Hence, by communicating with the authors and referring to their implementation, for each subject we set the upper bound U and the lower bound L of LSC/DSC to the largest and smallest LSA/DSA values of all the test inputs respectively. We further investigated the influence of U on the performance of LSC/DSC in Section 4.1.2. Also, we set the selected layer required by LSA/DSA to the last-hidden layer of a model according to the existing work [34, 35], since the work has demonstrated that deeper layers tend to perform better in DNN testing.

We conducted our experiments on the Intel Xeon Silver 4214 machine with 128GB RAM, Ubuntu 18.04 LTS, and eight GTX 2080 Ti GPUs. We used the Python interpreter with version 3.6.13 and utilized the Anaconda to manage Python packages.

3.3 | Experimental Setup

In this section, we present the experimental setup to answer each RQ for each subject. Please note that we used all the subjects in RQ1, while used all the normal classification models (ID: 1~6) in both RQ2 and RQ3 since regression models do not have the concept of “class” and the number of 10% of classes of test inputs in the subjects (ID: 9 and ID: 10) is too small to support the setups of RQ2 and RQ3.

3.3.1 | RQ1’s Setup

RQ1 aims to investigate the correlation between test coverage and error-revealing capability (under saturated diversity). We used the ratio of error-revealing test inputs in a test set to measure the error-revealing capability of the test set. To answer RQ1, we constructed m test sets with various ratios of error-revealing test inputs. In particular, we controlled for the size of test sets, i.e., creating the test sets with the same size n . In practice, each test set should contain a number of passing test inputs (that make the DNN produce correct prediction) since passing test inputs are common for a well-trained DNN. Hence, we ensure that the percentage of error-revealing test inputs in each test set is less than or equal to $\alpha\%$. That is, when constructing a test set, we first randomly determined the percentage of

error-revealing test inputs $r\%$ from $[0, \alpha\%]$ (i.e., $0 \leq r\% \leq \alpha\%$) and randomly sampled $r\% * n$ error-revealing test inputs. Then, we randomly sampled the remaining $(1-r\%) * n$ passing test inputs from the whole set of natural test inputs to construct the test set. Here, we set $m = 500$, $n = 800$, $\alpha\% = 70\%$. The whole set of test inputs refers to all the natural test inputs provided by the original dataset and the generated adversarial test inputs. This is because the number of *natural* error-revealing test inputs tends to be small in the original set of test inputs, we then generated adversarial test inputs as supplementary to increase the number of error-revealing test inputs. We will introduce the adversarial test input generation methods used in our study in Section 3.4. Please note that we ensure that each test set for classification models has saturated diversity.

Further, we collected the studied eight DNN test coverage for each test set, and then measured the correlation between each test coverage and error-revealing capability (i.e., the ratio of error-revealing test inputs in test sets). In our study, we adopted the widely-used **Spearman correlation** method [36], which is used to assess how well the relationship between two variables can be described using a *monotonic* function. The sign of the Spearman correlation coefficient (denoted as c) represents the positive or negative correlation between two variables, $|c|$ represents the correlation degree, and the p value reported by the Spearman correlation method represents whether the correlation is statistically significant (at the level of 0.05). According to the existing work [11, 37], the correlation can be divided into four categories: weak correlation ($0 < |c| \leq 0.4$), moderate correlation ($0.4 < |c| \leq 0.7$), strong correlation ($0.7 < |c| \leq 0.9$), and extremely strong correlation ($0.9 < |c| \leq 1$).

3.3.2 | RQ2's Setup

RQ2 aims to investigate the correlation between test coverage and test diversity. We used the ratio of covered classes (among all classes involved in the classification model) in a test set to measure the test diversity of the test set. To answer RQ2, we constructed a number of test sets with various ratios of covered classes. In particular, we controlled for both the size of test sets (denoted as n) and the ratio of error-revealing test inputs in test sets (denoted as $r\%$). As presented before, we treat covering 10% of classes in a test set as a diversity unit. We first evenly divided the whole set of test inputs into 10 groups according to the ascending order of the class IDs. By taking CIFAR-100 (including 100 classes) as an example, we put all the natural test inputs with the class IDs 0 ~ 9 and the generated adversarial test inputs based on these natural test inputs into the first group. Then, from the first i ($i \in \{1, 2, \dots, 10\}$) groups, we constructed m test sets, each of which is formed by randomly sampling $r\% * n$ error-revealing test inputs and $(1-r\%) * n$ passing test inputs. Correspondingly, the test diversity of each test set constructed from the first i groups is i diversity units. Indeed, we ensure that each test set has the corresponding numbers of diversity units. In RQ2, we set $n = 800$, $r\% = 50\%$, $m = 100$. Next, we collected the eight test coverage for each test set, and then measured the correlation between each test coverage and test diversity via the Spearman correlation method.

3.3.3 | RQ3's Setup

By comprehensively considering RQ1 and RQ2, RQ3 subsequently investigates the correlation between test coverage and error-revealing capability per diversity unit. Actually, RQ1 considers all classes (i.e., saturated diversity) to investigate the correlation. RQ3 repeats the experiment of RQ1 under one diversity unit. We constructed m test sets ($m = 100$), and when constructing a test set, we sampled test inputs from *one* group of test inputs (defined in Section 3.3.2) instead of the whole set of test inputs in RQ1. The remaining settings are the same as RQ1. To reduce the bias from the selection of the target group, we repeated this experiment *five* times by randomly selecting five different groups as the target groups for test set construction respectively. For each time, we collected the eight test coverage

for each test set, and then measured the correlation between each test coverage and error-revealing capability per diversity unit via Spearman correlation.

Besides, we investigated the trend of the correlation between test coverage and error-revealing capability with the test diversity increasing from one diversity unit to saturated diversity. We additionally considered i ($i \in \{2, \dots, 9\}$) diversity units by constructing test sets from the first i groups of test inputs respectively. That is, we repeated the above one-diversity-unit (or saturated-diversity) experiment under i ($i \in \{2, \dots, 9\}$) diversity units respectively.

3.4 | Adversarial Test Input Generation Methods

In our study, we adopted four advanced adversarial test input generation methods to generate adversarial test inputs for general image datasets, i.e., MNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100. They are FGSM (Fast Gradient Sign Method) [38], C&W (Carlini&Wagner) [20], BIM (Basic Iterative Methods) [39], and JSMA (Jacobian-based Saliency Map Attack) [40]. These methods have been widely used in the existing work [8, 19]. For the dataset Driving, we used the two methods proposed by DeepXplore [7] (called *Patch* and *Light* in our work) to generate adversarial test inputs following most of the existing work using this dataset [19, 34, 35]. For the datasets Speech-Commands and 20-Newsgroups, all the above-mentioned methods cannot be applicable to them since they are not images, but audios and texts respectively. Therefore, we adopted the widely-used CTCM (CTC loss based Method) [41] and PWWS (Probability Weighted Word Saliency) [42] to generate adversarial test inputs for them, respectively.

4 | RESULTS AND ANALYSIS

4.1 | RQ1: Correlation between Test Coverage and Error-revealing Capability

4.1.1 | Results on Structural Coverage

Table 3 shows the Spearman correlation results between test coverage and error-revealing capability (under saturated diversity). From Columns 3-8, in general, the six structural coverage metrics have moderate to extremely strong positive correlation with the error-revealing capability of test sets (i.e., the ratio of error-revealing test inputs in test sets) *in few cases*, but have weak positive correlation, even no significant correlation and negative correlation, *in most cases*. For example, TKNC has weak positive correlation in 26.67% cases, no significant correlation in 36.67% cases, and negative correlation in 26.67% cases, with the ratio of error-revealing test inputs, but has moderate to extremely strong positive correlation in only 10% cases. This conclusion is similar to the one obtained in the existing study on DNC (i.e., increasing DNC is not a meaningful objective in DNN testing) [21]. That indicates when controlling for the size and saturated diversity of each test set, both passing test inputs and error-revealing test inputs have the similar capability to increase these structural coverage metrics, and sometimes the coverage-increasing capability of the former is stronger than that of the latter.

Through further observation, the results on different structural coverage have some differences. Among the cases with moderate to extremely strong positive correlation for ID:1~6, 55% of them occur on NBC and SNAC when adopting FGSM to generate adversarial test inputs. The reason is that when generating adversarial test inputs, C&W, BIM, and JSMA aim to minimize the difference between an original test input and an adversarial test input or perform a tiny perturbation in a single step while FGSM does not consider them. Hence, FGSM is more likely to generate the adversarial test inputs that have farther distance from the natural test inputs, further leading to irrelevant test inputs (that are meaningless in DNN testing). Those more different, even irrelevant test inputs, are much easier to

TABLE 3 RQ1: Spearman correlation between test coverage and the ratio of error-revealing test inputs.

ID	Adv.	DNC	NBC	SNAC	TKNC	KMNC	IDC	LSC	DSC
1	CW	-0.05	-0.3	-0.37	0.11	-0.94	0.08	0.97	0.98
	FGSM	0.22	0.51	0.3	0.2	-0.97	-0.53	0.97	0.98
	BIM	0.09	-0.25	-0.34	0.09	-0.94	-0.03	0.97	0.98
	JSMA	0.08	0.3	-0.32	0.2	-0.85	0.0	0.97	0.98
2	CW	0.08	-0.09	-0.14	-0.02	-0.82	0.32	0.97	0.95
	FGSM	0.4	0.59	0.57	0.03	-0.21	-0.02	0.97	0.97
	BIM	0.19	0.54	0.54	-0.0	-0.73	0.38	0.97	0.96
	JSMA	0.09	0.01	-0.14	0.01	-0.71	0.39	0.98	0.96
3	CW	0.01	-0.34	-0.34	0.02	-0.96	-0.66	0.95	0.75
	FGSM	0.25	0.8	0.79	0.08	-0.02	-0.37	0.94	0.92
	BIM	0.08	0.59	0.59	0.05	-0.92	-0.5	0.94	0.85
	JSMA	-0.04	-0.3	-0.31	0.04	-0.98	-0.77	0.94	0.81
4	CW	-0.03	-0.04	-0.05	0.0	-0.75	0.45	0.87	0.36
	FGSM	0.12	0.59	0.62	0.44	-0.54	0.24	0.8	0.59
	BIM	-0.05	-0.06	-0.04	-0.04	-0.75	0.37	0.86	0.4
	JSMA	-0.11	-0.11	-0.07	-0.17	-0.87	0.27	0.84	0.32
5	CW	-0.58	-0.22	-0.22	0.38	-0.95	-0.71	0.58	0.92
	FGSM	-0.63	0.47	0.47	0.32	-0.98	-0.77	0.18	0.95
	BIM	-0.64	0.42	0.42	0.33	-0.97	-0.77	0.24	0.94
	JSMA	-0.64	-0.27	-0.26	0.3	-0.97	-0.78	0.26	0.94
6	CW	-0.16	-0.13	-0.11	-0.33	-0.51	-0.19	0.51	-0.49
	FGSM	0.22	0.6	0.61	-0.13	0.15	-0.2	0.81	-0.26
	BIM	-0.11	-0.06	-0.04	-0.28	-0.49	-0.14	0.58	-0.5
	JSMA	-0.17	-0.15	-0.16	-0.5	-0.68	-0.05	0.54	-0.6
7	PATCH	0.65	0.77	0.77	0.7	0.81	0.71	0.49	—
	LIGHT	-0.76	0.86	0.86	-0.8	0.88	-0.4	0.87	—
8	PATCH	0.39	0.79	0.79	0.2	0.78	0.65	0.75	—
	LIGHT	0.8	-0.65	-0.65	0.73	0.08	-0.13	0.89	—
9	CTCM	-0.68	0.27	0.34	-0.91	-0.79	—	0.83	0.96
10	PWWS	-0.15	-0.2	-0.23	-0.12	0.8	0.84	0.96	0.21

The **bold** value represents that the p value is smaller than 0.05, indicating the correlation has statistical significance. The value marked with represents moderate positive correlation. The value marked with represents strong positive correlation. The value marked with represents extremely strong positive correlation. The cell marked with "—" represents the coverage is not applicable to the corresponding model.

cover corner-case regions of the output values for neurons, leading to stronger positive correlation for NBC and SNAC. However, such positive correlation may be meaningless due to irrelevant test inputs. That also indicates in the studies of DNN testing, it is not enough to only adopt FGSM to generate adversarial test inputs due to their different characteristics. Besides, almost all the cases for KMNC have negative correlation. We carefully analyzed the reason and found that with the ratio of error-revealing test inputs increasing, the sections closing to the boundary low_i and $high_i$ are covered more frequently, indicating that more test inputs focus on covering a small number of sections. This leads to the decrement of the overall KMNC coverage.

In general, the correlation on regression models is stronger than classification models. This is because the accuracy of regression models is measured based on MSE, which is more fine-grained, and thus is more sensitive to capture differences in test behaviors (such as neuron coverage differences). That suggests if we can design a more fine-grained accuracy measurement for classification models, the correlation between test coverage and error-revealing capability may be effectively improved.

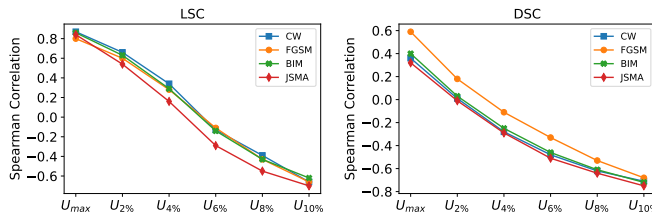


FIGURE 1 Spearman correlation under different settings of U for LSC and DSC on the subject (ID: 4).

4.1.2 | Results on Non-structural Coverage

We then analyzed the correlation between non-structural coverage and error-revealing capability from the last two columns in Table 3. Different from structural coverage, both LSC and DSC have moderate to extremely strong positive correlations in almost all the cases. The results demonstrate that measuring the difference of DNN behaviors between test inputs and training data is effective to measure the error-revealing capability of test sets, when controlling for the size and the saturated diversity of test sets. That is, the non-structural coverage outperforms the structural coverage in this scenario, and LSC slightly outperforms DSC.

But the remarkable thing for LSC and DSC is that some parameters in them are required to be specially set for different subjects. These parameters may affect the performance of LSC/DSC, but there is no guide to help set them for different subjects. Here, we conducted an experiment to investigate the influence of U on both LSC and DSC since this parameter is very important, which is used to filter out irrelevant test inputs. We studied six settings of U : U_{max} (the maximum value of LSA/DSA), which is the setting used in our study and means that we did not filter out any test inputs (even though some are irrelevant), and $U_{2\%} \sim U_{10\%}$ with the step size of 2% (the 2%~10% quantile values of all the ranked LSA/DSA values in their descending order), which simulates different degrees of filtering out irrelevant test inputs. Figure 1 shows the results, where the x-axis represents different settings of U while the y-axis represents the Spearman correlation coefficient between LSC/DSC and error-revealing capability. From this figure, the correlation decreases monotonically as the upper bound U becomes smaller, demonstrating the significant influence of U . Therefore, providing a guide for setting the parameters of LSC and DSC or designing an automatic method of adaptively setting the parameters for different subjects is an important future work for improving LSC and DSC.

Answer to RQ1: In general, the six structural coverage metrics do not have moderate to extremely strong positive correlation while the two non-structural coverage metrics have, when controlling for the size and saturated diversity of test sets, showing the superiority of non-structural coverage in this scenario. However, setting parameters of LSC and DSC properly is a challenge for their practical use.

4.2 | RQ2: Correlation between Test Coverage and Test Diversity

For each coverage on each subject, we plot the correlation between test coverage and test diversity, where the x-axis represents the ratio of covered classes (representing test diversity) and the y-axis represents the achieved coverage. Different colors represent the results of different adversarial test input generation methods. We found that the conclusions across different subjects are similar and due to space limit, we show the results of CIFAR-10 and ResNet-20 (ID: 4) as the representative in Figure 2. All the results can be found on our project homepage [25]. In particular, we calculated the Spearman correlation between test coverage and test diversity by using the medium of each box as the achieved coverage under the corresponding degree of test diversity for calculation, as shown in

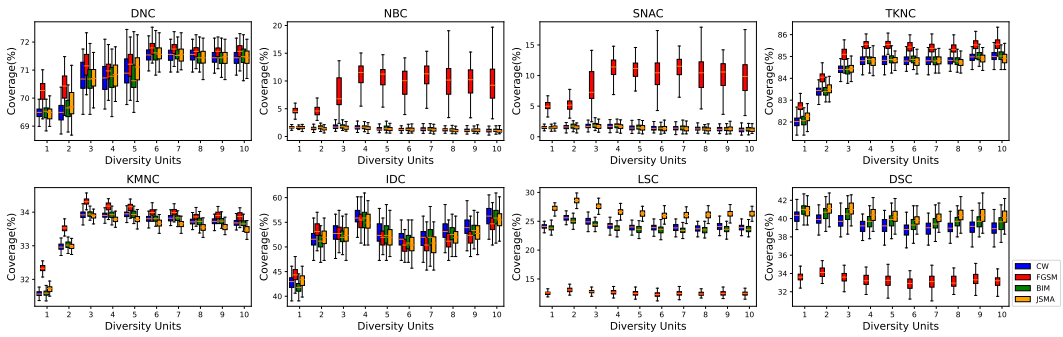


FIGURE 2 Correlation between test coverage and test diversity on the subject (ID: 4).

TABLE 4 RQ2: Spearman correlation between test coverage and test diversity.

ID	Adv.	DNC	NBC	SNAC	TKNC	KMNC	IDC	LSC	DSC
1	CW	0.99	0.75	0.81	0.99	0.62	0.94	0.01	0.43
	FGSM	0.99	0.66	0.83	0.98	0.77	0.99	0.25	0.83
	BIM	0.99	0.15	0.66	0.98	0.62	0.95	0.02	0.54
	JSMA	0.99	-0.41	0.76	0.99	0.62	0.95	-0.12	0.43
2	CW	0.93	0.2	0.14	0.96	0.88	0.87	-0.15	-0.15
	FGSM	0.96	0.7	0.7	0.96	0.84	0.87	0.2	0.82
	BIM	0.99	0.41	0.41	0.98	0.85	0.88	-0.15	-0.38
	JSMA	0.96	0.49	0.08	0.98	0.85	0.88	0.2	0.42
3	CW	0.9	-0.75	-0.75	1.0	0.33	0.83	-0.36	0.14
	FGSM	0.65	-0.03	-0.03	0.99	0.03	0.84	-0.27	-0.26
	BIM	0.75	-0.82	-0.82	1.0	0.22	0.87	-0.27	-0.15
	JSMA	0.89	-0.71	-0.71	1.0	0.07	0.84	-0.26	0.5
4	CW	0.75	-0.9	-0.88	0.85	0.04	0.7	-0.75	-0.78
	FGSM	0.76	0.45	0.45	0.72	0.03	0.12	-0.7	-0.64
	BIM	0.79	-0.87	-0.82	0.81	0.02	0.7	-0.62	-0.72
	JSMA	0.79	-0.93	-0.88	0.82	0.01	0.45	-0.61	-0.67
5	CW	0.98	-0.93	-0.93	0.89	0.36	0.15	-0.61	-0.52
	FGSM	0.98	-0.13	-0.13	0.85	0.35	0.05	-0.46	0.32
	BIM	0.98	-0.43	-0.43	0.85	0.66	0.13	-0.59	0.92
	JSMA	0.98	-0.94	-0.95	0.92	0.25	0.14	-0.83	-0.2
6	CW	0.98	0.53	0.56	0.95	0.98	0.6	-0.18	-0.37
	FGSM	0.99	0.43	0.45	0.96	0.99	0.6	0.05	-0.68
	BIM	0.98	0.81	0.82	0.98	0.98	0.65	-0.2	-0.29
	JSMA	1.0	0.53	0.53	0.98	0.96	0.71	0.13	-0.26

Table 4. According to these tables and figures, we obtained the following major conclusions.

First, regarding the four structural coverage metrics (i.e., DNC, TKNC, KMNC, and IDC), with the test diversity increasing, the achieved coverage increases (especially when the test diversity is small). That indicates when controlling for the size and the ratio of error-revealing capability of test sets, they show positive correlation (especially for DNC and TKNC). That is, different neurons are responsible for the prediction of different classes. When covering more classes, more neurons could be activated, leading to the increment of these structural coverage metrics. When the test diversity reaches a certain extent, the increment is starting to flatten. The possible reason lies in the limitation of the test-set size. When the number of covered classes becomes large in a test set, the number of test inputs for each class becomes small accordingly, which may cause the responsible neurons for each class are not sufficiently activated. That is, the coverage increment brought by increasing test diversity may be counteracted by the coverage decrement caused by decreasing the number of test inputs for each class.

TABLE 5 RQ3: Spearman correlation between test coverage and error-revealing capability per diversity unit (1st group).

ID	Adv.	DNC	NBC	SNAC	TKNC	KMNC	IDC	LSC	DSC
1	CW	0.72	-0.43	-0.43	0.8	0.79	0.77	0.98	0.99
	FGSM	0.76	0.46	0.32	0.86	0.53	0.06	0.97	0.97
	BIM	0.8	-0.47	-0.47	0.9	0.72	0.77	0.97	0.97
	JSMA	0.85	-0.46	-0.46	0.89	0.96	0.68	0.99	0.99
2	CW	0.86	0.39	0.04	0.79	0.71	0.44	0.87	0.94
	FGSM	0.8	0.43	0.23	0.84	0.75	0.42	0.79	0.88
	BIM	0.77	0.23	0.04	0.73	0.43	0.36	0.81	0.94
	JSMA	0.84	0.54	0.32	0.8	0.8	0.5	0.82	0.95
3	CW	0.94	-0.6	-0.65	0.98	0.96	0.61	0.95	0.5
	FGSM	0.95	0.87	0.86	0.98	0.99	0.67	0.88	0.94
	BIM	0.95	0.8	0.8	0.98	0.97	0.73	0.94	0.89
	JSMA	0.96	-0.57	-0.69	0.99	0.96	0.47	0.92	0.67
4	CW	0.63	-0.19	-0.09	0.85	0.32	0.96	0.84	0.73
	FGSM	0.63	0.8	0.85	0.89	0.93	0.94	0.83	0.62
	BIM	0.38	-0.06	0.13	0.85	0.6	0.94	0.82	0.62
	JSMA	0.41	-0.22	-0.09	0.84	0.45	0.94	0.86	0.51
5	CW	0.84	-0.56	-0.54	0.89	-0.72	-0.33	0.55	0.95
	FGSM	0.73	0.5	0.49	0.91	-0.53	-0.48	-0.04	0.95
	BIM	0.68	0.63	0.64	0.89	-0.85	-0.48	0.04	0.95
	JSMA	0.72	-0.5	-0.5	0.92	-0.77	-0.37	0.22	0.95
6	CW	0.67	-0.08	-0.05	0.12	-0.53	0.5	0.19	-0.85
	FGSM	0.76	0.64	0.63	0.68	0.66	0.65	0.83	-0.63
	BIM	0.69	-0.24	-0.21	0.15	-0.42	0.67	0.26	-0.65
	JSMA	0.71	0.14	0.14	0.01	-0.37	0.6	0.57	-0.78

Second, regarding the remaining two structural coverage metrics (i.e., NBC and SNAC), in most cases there is no obvious positive correlation between the two metrics and test diversity. The correlation tends to be stronger when using FGSM to generate adversarial test inputs. The reason is that it is difficult to improve both NBC and SNAC due to their inherent “corner-case” characteristics, and thus regardless of test diversity, both NBC and SNAC stay small values stably. Regarding the results on FGSM, as presented before, the adversarial test inputs generated by FGSM are more likely to cover corner-case regions of the output values for neurons, causing that both NBC and SNAC may break away from small values.

Third, regarding the two non-structural coverage metrics, the achieved coverage does not increase with the test diversity increasing, indicating no obvious positive correlation between them. This is because these metrics measure the distance between a test input and the training data with the same label. Since training data tend to be sufficient, the distributions of the training data for different classes are stable, causing that the distance cannot be affected obviously by different classes of test inputs.

Answer to RQ2: DNC, TKNC, KMNC, and IDC have positive correlation with test diversity especially under small test diversity, demonstrating that the four structural coverage metrics are able to guide the improvement of test diversity, while the remaining four coverage metrics have no obvious positive correlation with test diversity.

4.3 | RQ3: Correlation between Test Coverage and Error-revealing Capability per Diversity Unit

According to RQ1 and RQ2, we found that some of test coverage metrics are positively correlated with test diversity, but are not positively correlated with error-revealing capability under saturated diversity. One subsequent question

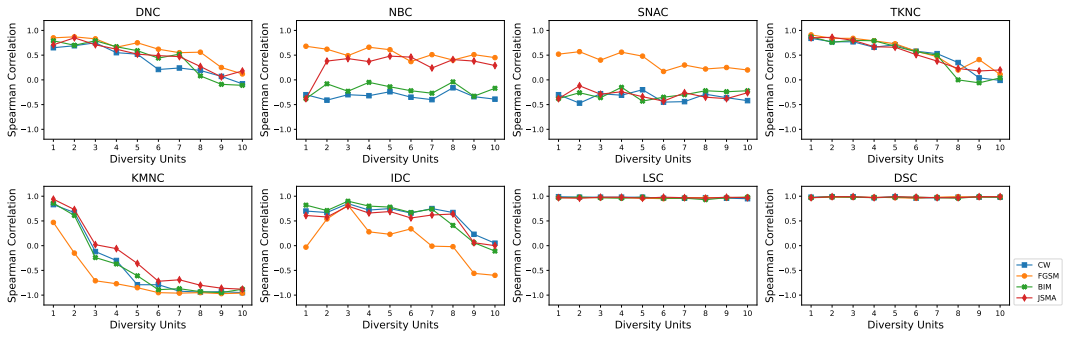


FIGURE 3 Trend of correlation between coverage and error-revealing capability with test diversity increasing on subject (ID: 1)

is that when test diversity is not saturated, what about the correlation between test coverage and error-revealing capability?

To answer this question, for each coverage on each subject, we measured the correlation between test coverage and error-revealing capability per diversity unit. Although we repeated the experiment five times by selecting different groups to construct the test sets with one diversity unit (presented in Section 3.3.3), we only show the results on the first group in Table 5 due to the space limit. Indeed, the conclusions are consistent across different groups, which can be found on our project homepage [25]. From Table 5, we found that the four structural coverage metrics (DNC, TKNC, KMNC, and IDC) that are positively correlated with test diversity, also have positive correlation with error-revealing capability under small test diversity (i.e., one diversity unit). However, as shown in RQ1, they have no correlation or negative correlation with error-revealing capability under saturated diversity. The reason is that although the test inputs in a test set concentrate on only 10% of classes, the error-revealing ones among them may predict the classes out of the 10% of classes, causing that they are more likely to share similar traces in the DNN with the other classes of test inputs. That is, under such a small degree of test diversity, increasing the ratio of error-revealing test inputs may cover different neurons with those covered by the 10% of classes of test inputs.

Furthermore, we investigated the correlation trend with the degree of test diversity increasing from one diversity unit to saturated diversity (i.e., 10 diversity units) with the step size of a diversity unit. Due to the space limit, we show the trend results by taking MNIST and LeNet-5 (ID: 1) as the representative in Figure 3. Indeed, there are almost consistent conclusions across different subjects, and we put all the results on our project homepage [25]. In Figure 3, the x-axis represents the degree of test diversity and the y-axis represents the Spearman correlation coefficient. From Figure 3, with the degree of test diversity increasing, the correlation between test coverage and error-revealing capability becomes weaker for the four structural coverage metrics (DNC, TKNC, KMNC, and IDC). This is as expected, since when the degree of test diversity is large, more error-revealing test inputs cannot bring more coverage of different neurons that are covered by other classes. Therefore, the four structural coverage metrics are not useless to guide the generation of test inputs, but the proper usage scenario for them is not found before. In actual, they can be effective to guide to generate error-revealing test inputs when using a set of test inputs with small test diversity as seed inputs.

Regarding the remaining four test coverage metrics (i.e., NBC, SNAC, LSC, and DSC), with the degree of test diversity increasing, the correlation between test coverage and error-revealing capability is stable. That is, their correlation between test coverage and error-revealing capability is not largely affected by various degrees of test diversity, which is as expected based on the conclusions obtained in RQ1 and RQ2. Specifically, the correlation is stably strong for LSC and DSC, but is stably weak for NBC and SNAC.

Answer to RQ3: DNC, TKNC, KMNC, and IDC are positively correlated with error-revealing capability under small test diversity, and the correlation becomes weaker with the degree of test diversity increasing, indicating that they are useful to guide the generation of error-revealing test inputs by taking the test inputs with small test diversity as seed inputs.

5 | IMPLICATIONS AND SUGGESTIONS

Based on the conclusions from our extensive study, we summarize a series of implications from both academia and practice perspectives:

Implications from the academia perspective: (1) Some empirical studies have been conducted to investigate the usefulness of existing DNN test coverage from some perspectives, and they delivered negative conclusions that may drive future research towards the direction away from the existing coverage. Different from them, our work investigates the usefulness of these coverage metrics from another perspective (i.e., test effectiveness measured by different criteria), which identifies the practical usage scenarios for the existing coverage. That is, our study can inspire future research that continue around the *existing* coverage but in their proper scenarios. In recent years, many studies have been conducted to overturn a (small) area of existing techniques/measurements, opening another direction for the area. They are appealing, but revisiting these existing techniques/measurements from different perspectives and identifying the proper usage scenarios for them are also inspiring for the development of the area.

(2) The usefulness of some structural coverage metrics can be manifested by incorporating test diversity. In our study, we propose a coarse-grained measurement (i.e., the ratio of covered classes among all classes involved in a DNN) for test diversity, but such a coarse-grained measurement can indeed inspire the usefulness of these coverage metrics. In the future, it is promising to design more fine-grained measurements for test diversity in order to better utilize the existing structural coverage (such as measuring test-input diversity under each class based on input features). Moreover, it is also important to incorporate the concept of test diversity in regression tasks.

(3) The correlation between test coverage and error-revealing capability is much stronger on regression models than classification models due to the fine-grained accuracy measurement for the former. Such a fine-grained measurement is more sensitive to capture the differences in test behaviors (such as neuron coverage differences). Therefore, decomposing the accuracy measurement for classification models to be more fine-grained may be a promising direction to improve the usefulness of these coverage metrics (especially structural coverage).

Implications from the practice perspective: (1) The usage scenario of the non-structural coverage (i.e., LSC and DSC) is using it to guide the generation of error-revealing test inputs regardless of taking the whole test set or a small number of classes of test inputs as seed inputs. But the remarkable thing is how to properly set the parameters of LSC and DSC for different subjects, which can be studied carefully in the future. Moreover, it is hard to guarantee the diversity of the generated test inputs guided by LSC and DSC.

(2) The usage scenario of the structural coverage (i.e., DNC, TKNC, KMNC, and IDC) is using it to guide the generation of error-revealing test inputs by taking a small number of classes of test inputs as seed inputs. In particular, the guidance of these structural coverage can also help improve the diversity of the generated test inputs. However, they are not capable of guiding to generate error-revealing test inputs by taking the whole test set (with saturated diversity) as seed inputs.

(3) In practice, we can first use the four structural coverage metrics (i.e., DNC, TKNC, KMNC, and IDC) to guide the generation of *diverse* error-revealing test inputs by taking each class of test inputs as seed inputs respectively. When the generated error-revealing test inputs based on each class of test inputs are diverse enough, we can then

adopt the non-structural coverage (i.e., LSC and DSC) to guide to generate *more* error-revealing test inputs by taking diverse test inputs as seed inputs, which can complement with the four structural coverage metrics.

6 | THREATS TO VALIDITY

Our study has five kinds of threats to *external validity*. (1) *Subjects*: Our used subjects may not represent other subjects. To reduce this kind of threat, we carefully considered the diversity of subjects and used some new subjects that have never been used to evaluate these studied DNN test coverage metrics. (2) *Created test sets*: We controlled for the size of created test sets, i.e., 800, which may not represent other sizes. Actually, we have conducted a small experiment to investigate the influence of different test-set sizes (i.e., 100, 500, 800, and 1000) and obtained almost consistent conclusions. We put the experimental results on our project homepage [25]. (3) *Studied test coverage*: Our studied test coverage may not represent other coverage. To reduce this kind of threat, we considered all the DNN test coverage studied in the existing studies [22, 21] (except TKNP [8]) and additionally studied the most recent test coverage. In the future, we will study more coverage metrics, such as t-way combination coverage [43] and state coverage [44]. (4) *Adversarial test input generation methods*: We adopted the most widely-used FGSM, BIM, JSMA, C&W for general image-domain DNN models, the widely-used DeepXplore-Patch and DeepXplore-Light for Driving following the existing work [34, 19], CTCM for Speech-Commands and PWWS for 20-Newsgroups, which may not represent other methods. However, this kind of threat may be not serious since the conclusions obtained from BIM, JSMA, and C&W are almost consistent. (5) *Used errors*: We regard a test input predicted wrongly as an error-revealing test input, but do not distinguish whether they trigger the same bugs. This kind of threat exists in all existing studies to evaluate DNN coverage since it is challenging to distinguish them due to involving huge manual effort. Actually, the number of error-revealing test inputs is also important in DNN testing since identifying them can help improve the accuracy of the DNN model [45].

Our study also has a kind of threat to *construct validity*: *diversity unit partition method*. In RQ2 and RQ3, we treat covering 10% of classes in a test set as a diversity unit. For a subject, we evenly divided all classes into 10 groups according to the *default* order of the class IDs. Also, we constructed test sets with different numbers of diversity units according to the *default* order of the group IDs. However, the method of diversity unit partition may be not a serious kind of threat. This is because we investigated the influence of different groups when studying the correlation between test coverage and error-revealing capability per diversity unit in RQ3, and found that the conclusions are consistent.

7 | RELATED WORK

DNN Testing. There is a lot of research on DNN testing [46, 47, 48, 49, 50, 51, 52, 53]. We classify them into: measuring test effectiveness, generating test inputs, and improving test efficiency.

Regarding measuring test effectiveness, we have introduced six structural coverage metrics and two non-structural coverage metrics in Section 2. Besides, Ma et al. [43] proposed t-way combination coverage for DNN testing by borrowing the idea of traditional combinatorial testing. Du et al. [44] proposed state coverage for RNN-based stateful DNN testing. Ma et al. [54] proposed a mutation-based metric by designing a set of mutation operators for DNN source programs and models. Sun et al. [48] proposed MC/DC-inspired neuron coverage, which is tailored to the distinct features of a DNN. It instantiates the decision change and condition change in a DNN by defining the *sign change* and *value change* of neurons, which show the degree of the change of neuron activation values under two different test inputs. Same as the existing studies on DNN coverage [21, 22], we did not investigate MC/DC-inspired neuron coverage [51] in our study, since it is limited in scalability (hard to apply to large subjects) as demonstrated by

the existing work [8, 55].

Besides the two empirical studies on DNN test coverage discussed in Section 1, there are some other studies [56, 57, 23] on evaluating DNN test coverage. Dong et al. [57] conducted a study to investigate the correlation between test coverage and model robustness and found the correlation is very limited. Li et al. [56] conducted a preliminary study on structural coverage using only two datasets, and found that adversary-oriented search may make more contributions to the error-revealing capability than structural coverage. Yang et al. [23] extended the existing study [22], further confirming the negative conclusions on the existing test coverage from the perspective of model retraining quality. They also claimed that the defense strategies against coverage-driven methods are worth further investigating. Different from them, our study is to systematically revisit DNN test coverage from the test effectiveness perspective. In particular, we designed our study systematically, i.e., considering both structural and non-structural coverage (including the state-of-the-art one), three test effectiveness criteria, and more comprehensive subjects. Also, we did not deliver only negative conclusions like existing studies, but identified the practical usage scenarios for the existing coverage. Furthermore, there are some metrics to measure DNN robustness [58, 59, 60, 61]. Since they are different from the metrics measuring test effectiveness (the target of our work), we do not discuss them.

Regarding generating test inputs, many approaches have been proposed. For example, Guo et al. [62] proposed *DLFuzz*, the first differential testing framework for DNNs by maximizing neuron coverage. Xie et al. [63] proposed *Deephunter*, a coverage-guided fuzz testing framework for DNNs. Sun et al. [64] proposed to generate test inputs for DNNs through concolic testing. Regarding improving test efficiency, some test input selection and prioritization approaches have been proposed for DNNs [65, 34, 66, 35]. For example, Li et al. [34] proposed to select test inputs by minimizing the cross entropy between the selected test inputs and the whole test set, so as to save labeling costs. Feng et al. [66] proposed *DeepGini* to prioritize test inputs by measuring the purity of test inputs, so as to label error-revealing test inputs earlier. Different from them, our work aims to investigate the correlation between DNN coverage and test effectiveness through a *systematic study*.

Empirical Studies on Traditional Test Coverage. In traditional software testing, there are some studies to investigate the correlation between traditional test coverage and test effectiveness. For example, Namin and Andrews [67] conducted a study to explore the relationship among the test-suite size, structural coverage, and test effectiveness based on C/C++ programs, and found that coverage is sometimes correlated with test effectiveness when controlling for the test-suite size. Inozemtseva and Holmes [11] conducted a study to investigate the above relationship based on Java programs, and found that the correlation between test coverage and test effectiveness is low to moderate when controlling for the test-suite size. Zhang and Mesbah [12] conducted a study to investigate the correlation between assertions and test effectiveness based on Java programs, and found that both assertion numbers and assertion coverage are strongly correlated with test effectiveness.

Different from them, our study is to investigate the correlation between test coverage and test effectiveness in *DNN testing*. Traditional software testing and DNN testing have different test coverage and their subjects have totally different characteristics.

references

- [1] Chen C, Seff A, Kornhauser A, Xiao J. Deepdriving: Learning affordance for direct perception in autonomous driving. In: ICCV; 2015. p. 2722–2730.
- [2] Sun Y, Chen Y, Wang X, Tang X. Deep Learning Face Representation by Joint Identification-Verification. In: NeurIPS; 2014. p. 1988–1996.

- [3] Cheng Y. Semi-supervised learning for neural machine translation. In: Joint Training for Neural Machine Translation; 2019.p. 25–40.
- [4] Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 2016;375(13):1216.
- [5] Gu X, Zhang H, Kim S. Deep code search. In: ICSE; 2018. p. 933–944.
- [6] Gu X, Zhang H, Zhang D, Kim S. Deep API learning. In: FSE; 2016. p. 631–642.
- [7] Pei K, Cao Y, Yang J, Jana S. Deepxplore: Automated whitebox testing of deep learning systems. In: proceedings of the 26th Symposium on Operating Systems Principles; 2017. p. 1–18.
- [8] Ma L, Juefei-Xu F, Zhang F, Sun J, Xue M, Li B, et al. Deepgauge: Multi-granularity testing criteria for deep learning systems. In: ASE; 2018. p. 120–131.
- [9] Odena A, Olsson C, Andersen D, Goodfellow I. Tensorfuzz: Debugging neural networks with coverage-guided fuzzing. In: ICML; 2019. p. 4901–4911.
- [10] News; Accessed: 2022. https://www.vice.com/en_us/article/9kga85/uber-is-giving-up-on-self-driving-cars-in-california-after-deadly-crash.
- [11] Inozemtseva L, Holmes R. Coverage is not strongly correlated with test suite effectiveness. In: ICSE; 2014. p. 435–445.
- [12] Zhang Y, Mesbah A. Assertions are strongly correlated with test suite effectiveness. In: FSE; 2015. p. 214–224.
- [13] Chekam TT, Papadakis M, Traon YL, Harman M. An empirical study on mutation, statement and branch coverage fault revelation that avoids the unreliable clean program assumption. In: ICSE; 2017. p. 597–608.
- [14] Morrison GC, Inggs CP, Visser WC. Automated coverage calculation and test case generation. In: SAICSIT ACM; 2012. p. 84–93.
- [15] Hilton M, Bell J, Marinov D. A large-scale study of test coverage evolution. In: ASE; 2018. p. 53–63.
- [16] Gligoric M, Groce A, Zhang C, Sharma R, Alipour MA, Marinov D. Comparing non-adequate test suites using coverage criteria. In: ISSTA; 2013. p. 302–313.
- [17] Gay G, Rajan A, Staats M, Whalen MW, Heimdahl MPE. The Effect of Program and Model Structure on the Effectiveness of MC/DC Test Adequacy Coverage. *TOSEM* 2016;25(3):25:1–25:34.
- [18] Gligoric M, Groce A, Zhang C, Sharma R, Alipour MA, Marinov D. Guidelines for Coverage-Based Comparisons of Non-Adequate Test Suites. *TOSEM* 2015;24(4):22:1–22:33.
- [19] Kim J, Feldt R, Yoo S. Guiding deep learning system testing using surprise adequacy. In: ICSE; 2019. p. 1039–1049.
- [20] Carlini N, Wagner DA. Towards Evaluating the Robustness of Neural Networks. In: S&P; 2017. p. 39–57.
- [21] Harel-Canada F, Wang L, Gulzar MA, Gu Q, Kim M. 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. p. 851–862.
- [22] Yan S, Tao G, Liu X, Zhai J, Ma S, Xu L, et al. Correlations between deep neural network model coverage criteria and model quality. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering; 2020. p. 775–787.
- [23] Yang Z, Shi J, Asyofi MH, Lo D. Revisiting Neuron Coverage Metrics and Quality of Deep Neural Networks. *arXiv preprint arXiv:220100191* 2022;.

- [24] Gerasimou S, Eniser HF, Sen A, Cakan A. Importance-driven deep learning system testing. In: 2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE) IEEE; 2020. p. 702–713.
- [25] Homepage; Accessed: 2022. <https://github.com/anonymousprojs/DL-Coverage-Study>.
- [26] Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 2015;10(7):e0130140.
- [27] Wand MP, Jones MC. Kernel smoothing. Chapman and Hall/CRC; 1994.
- [28] MNIST; Accessed: 2022. <http://yann.lecun.com/exdb/mnist/>.
- [29] Xiao H, Rasul K, Vollgraf R, Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms; 2017.
- [30] CIFAR-10; Accessed: 2022. <http://www.cs.toronto.edu/~kriz/cifar.html>.
- [31] Driving; Accessed: 2022. <https://udacity.com/self-driving-car>.
- [32] Speech-Commands; Accessed: 2022. https://github.com/bjtommychen/Keras_DeepSpeech2_SpeechRecognition.
- [33] 20-Newsgroups; Accessed: 2022. <http://qwone.com/~jason/20Newsgroups/>.
- [34] Li Z, Ma X, Xu C, Cao C, Xu J, Lü J. Boosting Operational DNN Testing Efficiency Through Conditioning. In: FSE; 2019. p. 499–509.
- [35] Chen J, Wu Z, Wang Z, You H, Zhang L, Yan M. Practical Accuracy Estimation for Efficient Deep Neural Network Testing. TOSEM 2020; To appear.
- [36] Myers L, Sirois MJ. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences* 2004;12.
- [37] Chen J, Bai Y, Hao D, Zhang L, Zhang L, Xie B. How do assertions impact coverage-based test-suite reduction? In: ICST; 2017. p. 418–423.
- [38] Goodfellow IJ, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. In: ICLR; 2015. .
- [39] Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. In: ICLR; 2017. .
- [40] Papernot N, McDaniel PD, Jha S, Fredrikson M, Celik ZB, Swami A. The Limitations of Deep Learning in Adversarial Settings. In: S&P; 2016. p. 372–387.
- [41] Carlini N, Wagner DA. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. In: S&P Workshops; 2018. p. 1–7.
- [42] Ren S, Deng Y, He K, Che W. Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. In: ACL (1) Association for Computational Linguistics; 2019. p. 1085–1097.
- [43] Ma L, Juefei-Xu F, Xue M, Li B, Li L, Liu Y, et al. DeepCT: Tomographic Combinatorial Testing for Deep Learning Systems. In: SANER; 2019. p. 614–618.
- [44] Du X, Xie X, Li Y, Ma L, Liu Y, Zhao J. DeepStellar: model-based quantitative analysis of stateful deep learning systems. In: ESEC/SIGSOFT FSE; 2019. p. 477–487.
- [45] Ma S, Liu Y, Lee W, Zhang X, Grama A. MODE: automated neural network model debugging via state differential analysis and input selection. In: FSE; 2018. p. 175–186.
- [46] Sun Y, Huang X, Kroening D. Testing deep neural networks. *arXiv preprint arXiv:180304792* 2018;.

- [47] Zhang JM, Harman M, Ma L, Liu Y. Machine learning testing: Survey, landscapes and horizons. TSE 2020;To a.
- [48] Sun Y, Huang X, Kroening D, Sharp J, Hill M, Ashmore R. Structural Test Coverage Criteria for Deep Neural Networks. TECS 2019;18(5s):94:1–94:23.
- [49] Zhang F, Chowdhury SP, Christakis M. DeepSearch: Simple and Effective Blackbox Fuzzing of Deep Neural Networks. CoRR 2019;abs/1910.06296.
- [50] Tian Y, Pei K, Jana S, Ray B. DeepTest: automated testing of deep-neural-network-driven autonomous cars. In: ICSE; 2018. p. 303–314.
- [51] Sun Y, Huang X, Kroening D, Sharp J, Hill M, Ashmore R. DeepConcolic: testing and debugging deep neural networks. In: ICSE; 2019. p. 111–114.
- [52] Gerasimou S, Eniser HF, Sen A, Cakan A. Importance-Driven Deep Learning System Testing. In: ICSE; 2020. To appear.
- [53] Lee S, Cha S, Lee D, Oh H. Effective white-box testing of deep neural networks with adaptive neuron-selection strategy. In: ISSTA; 2020. p. 165–176.
- [54] Ma L, Zhang F, Sun J, Xue M, Li B, Juefei-Xu F, et al. Deepmutation: Mutation testing of deep learning systems. In: ISSRE; 2018. p. 100–111.
- [55] Zhou Z, Dou W, Liu J, Zhang C, Wei J, Ye D. DeepCon: Contribution Coverage Testing for Deep Learning Systems. In: 2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER) IEEE; 2021. p. 189–200.
- [56] Li Z, Ma X, Xu C, Cao C. Structural coverage criteria for neural networks could be misleading. In: ICSE (NIER); 2019. p. 89–92.
- [57] Dong Y, Zhang P, Wang J, Liu S, Sun J, Hao J, et al. An Empirical Study on Correlation between Coverage and Robustness for Deep Neural Networks. In: 2020 25th International Conference on Engineering of Complex Computer Systems (ICECCS) IEEE; 2020. p. 73–82.
- [58] Bastani O, Ioannou Y, Lampropoulos L, Vytiniotis D, Nori A, Criminisi A. Measuring neural net robustness with constraints. In: NeurIPS; 2016. p. 2613–2621.
- [59] Jha S, Raj S, Fernandes SL, Jha SK, Jha S, Jalaian B, et al. Attribution-Based Confidence Metric For Deep Neural Networks. In: NeurIPS; 2019. p. 11826–11837.
- [60] Gopinath D, Katz G, Pasareanu CS, Barrett CW. DeepSafe: A Data-Driven Approach for Assessing Robustness of Neural Networks. In: ATVA, vol. 11138; 2018. p. 3–19.
- [61] Katz G, Barrett CW, Dill DL, Julian K, Kochenderfer MJ. Towards Proving the Adversarial Robustness of Deep Neural Networks. In: FVAV@iFM, vol. 257 of EPTCS; 2017. p. 19–26.
- [62] Guo J, Jiang Y, Zhao Y, Chen Q, Sun J. Dlfuzz: Differential fuzzing testing of deep learning systems. In: FSE; 2018. p. 739–743.
- [63] Xie X, Ma L, Juefei-Xu F, Xue M, Chen H, Liu Y, et al. DeepHunter: a coverage-guided fuzz testing framework for deep neural networks. In: ISSTA; 2019. p. 146–157.
- [64] Sun Y, Wu M, Ruan W, Huang X, Kwiatkowska M, Kroening D. Concolic testing for deep neural networks. In: ASE; 2018. p. 109–119.
- [65] Ma W, Papadakis M, Tsakmalis A, Cordy M, Traon YL. Test Selection for Deep Learning Systems. CoRR 2019;abs/1904.13195.
- [66] Feng Y, Shi Q, Gao X, Wan J, Fang C, Chen Z. DeepGini: prioritizing massive tests to enhance the robustness of deep neural networks. In: ISSTA; 2020. p. 177–188.
- [67] Namin AS, Andrews JH. The influence of size and coverage on test suite effectiveness. In: ISSTA; 2009. p. 57–68.