



# DeepSample: DNN sampling-based testing for operational accuracy assessment

Antonio Guerriero, Roberto Pietrantuono, Stefano Russo

DIETI, Università degli Studi di Napoli Federico II

Via Claudio 21, 80125 - Napoli, Italy

{antonio.guerriero,roberto.pietrantuono,stefano.russo}@unina.it

## ABSTRACT

Deep Neural Networks (DNN) are core components for classification and regression tasks of many software systems. Companies incur in high costs for testing DNN with datasets representative of the inputs expected in operation, as these need to be manually labelled. The challenge is to select a representative set of test inputs as small as possible to reduce the labelling cost, while sufficing to yield unbiased high-confidence estimates of the expected DNN accuracy. At the same time, testers are interested in exposing as many DNN mispredictions as possible to improve the DNN, ending up in the need for techniques pursuing a threefold aim: small dataset size, trustworthy estimates, mispredictions exposure.

This study presents DeepSample, a family of DNN testing techniques for cost-effective accuracy assessment based on probabilistic sampling. We investigate whether, to what extent, and under which conditions probabilistic sampling can help to tackle the outlined challenge. We implement five new sampling-based testing techniques, and perform a comprehensive comparison of such techniques and of three further state-of-the-art techniques for both DNN classification and regression tasks. Results serve as guidance for best use of sampling-based testing for faithful and high-confidence estimates of DNN accuracy in operation at low cost.

## CCS CONCEPTS

• **Software and its engineering** → **Software testing and debugging**.

## KEYWORDS

Software testing, Deep Neural Networks, Sampling

### ACM Reference Format:

Antonio Guerriero, Roberto Pietrantuono, Stefano Russo. 2024. DeepSample: DNN sampling-based testing for operational accuracy assessment. In *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE '24)*, April 14–20, 2024, Lisbon, Portugal. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3597503.3639584>

## 1 INTRODUCTION

A countless number of software systems today rely on Deep Neural Networks (DNN) predictions. Before release, engineers need to test

the DNN to estimate their accuracy (i.e., probability of not having mispredictions). This allows to establish a release criterion and to correct or tune the DNN until the criterion is met.

The reference scenario is the following: a DNN model meant to operate in a target context is trained with a *training dataset*. The goal of the tester is to select a small yet representative subset of (unlabelled) inputs from an *operational dataset*, to use as test cases to estimate the DNN accuracy [1]. Their manual labelling has a high cost. The challenge is to build a small test set able to provide an unbiased, high-confidence estimate of the DNN accuracy. At the same time, testers are interested in exposing DNN mispredictions, since they are input to DNN debugging and re-training [2]. The goal thus becomes threefold: build a *small dataset*, able to faithfully *estimate DNN accuracy*, and with a good ability to *expose mispredictions*.

Inspired by *operational testing*, a known practice in software reliability engineering [3–7], researchers proposed probabilistic sampling to test DNN. The basic scheme is simple random sampling (SRS). Li *et al.* proposed a sampling scheme aimed at minimizing cross-entropy between the selected tests and the operational dataset [1]. Guerriero *et al.* [2] leveraged adaptive sampling [8] to propose DeepEST, whose objective is to expose many DNN mispredictions while yielding good accuracy estimates. These techniques borrow basic concepts from sampling theory to derive algorithms working well for specific goals or contexts – for instance, CES and DeepEST outperform each other in their respective objectives (lower-variance estimate the former, better failure exposure the latter). However, better trade-offs can be achieved by exploiting advanced strategies from statistical sampling, e.g., by properly using the information available to drive the sampling process.

This work aims to give a high level view of sampling-based DNN testing to highlight what are the main knobs to tailor a technique according to the needs and improve performance, exploiting advanced sampling theory concepts besides the basic ones (e.g., auxiliary variables, unequal sampling, without-replacement schemes, stratification). To this aim:

- We propose DeepSample, a family of sampling-based DNN testing techniques differing from each other in the sampling strategy, in the auxiliary information used for sampling and for partitioning, and in the estimation process. The framework includes five new testing techniques, each implemented in three variants depending on the auxiliary information used to drive sampling.
- We present a comprehensive comparison of the new techniques and of three existing ones, SRS, CES, DeepEST, to evaluate their ability to assess DNN accuracy and select failing examples. The evaluation is conducted on classification and regression tasks,



This work is licensed under a Creative Commons Attribution International 4.0 License.  
*ICSE '24, April 14–20, 2024, Lisbon, Portugal*  
 © 2024 Copyright held by the owner/author(s).  
 ACM ISBN 979-8-4007-0217-4/24/04.  
<https://doi.org/10.1145/3597503.3639584>

under 5 testing budgets, 3 datasets, with 3 models per dataset for classification, and 1 dataset and 2 models for regression.<sup>1</sup>

The new algorithms turn out to outperform the existing ones in almost all the contexts. Overall, the results allow to draw guidelines for practitioners and researchers - on relevant factors like if and which auxiliary information to use and how to use it - for sampling-based DNN testing for high accuracy, high confidence estimates at low cost and with good mispredictions exposure ability.

## 2 RELATED WORK

Probabilistic sampling is used in *operational testing* (OT) to estimate the expected reliability of a software system after release. In OT, test suites are built by selecting or generating tests according to the expected *operational profile*, a probabilistic characterization of the expected usage. OT was central in Cleanroom software engineering [3–6] and in the Software Reliability Engineering Test process [7].

Over the years, researchers proposed better sampling strategies to improve estimates or lower their cost. Cai *et al.* [9–11] developed *Adaptive Testing*, still based on the operational profile, but with an adaptive selection of test cases from partitions. Adaptive Testing with Gradient Descent<sup>2</sup> [12] is one of the techniques considered in this study. Stratified sampling too has been used for reliability assessment [13, 14]. Later, Pietrantuono *et al.* [15, 16] stressed the use of unequal probability sampling to improve efficiency, formalizing several sampling schemes to this aim [17].

Li *et al.* [1] first proposed sampling for DNN operational accuracy assessment in the CES (Cross-Entropy Sampling) technique. Like OT, CES aims to select a small yet representative sample, by minimizing the cross-entropy between the selected and the operational dataset. A sample is expected to contain the same proportion of failing examples as in the operational dataset. Guerriero *et al.* [2] observed that the mere imitation of operational inputs may be inefficient, especially for accurate DNN, as much effort is wasted to label correctly classified inputs.

They propose DeepEST, exploiting an *adaptive sampling* algorithm for rare populations [8] to spot the more failing examples, hence spending effort to label examples useful for improvement besides assessment. The disproportional selection is balanced by an estimator that preserves unbiasedness.

A further technique is PACE (Practical accuracy estimation) [18], a heuristic method that uses clustering to partition tests into groups, and then uses adaptive random selection of test inputs representative of the clusters. Zhou *et al.* proposed DeepReduce [19], a two-stage heuristic method exploiting neuron coverage to select a subset of inputs, then using the Kullback-Leibler Divergence to drive the second-stage selection. These techniques are however not based on probabilistic sampling like those compared in this work, and they do not guarantee unbiasedness and convergence.

## 3 SAMPLING-BASED TESTING

### 3.1 Formulation

- $M$  is the DNN model under test;

- $D = \{d_1, \dots, d_N\}$  is the *operational dataset*, an arbitrarily large set of examples with unknown labels, which are possibly given as input to the model  $M$  in the operational phase. Its size is  $N = |D|$ ;
- $T \in D = \{t_1, \dots, t_n\}$  is the subset of examples to select from  $D$  and to be labelled. This set is used for estimating DNN accuracy, and can also be used to enlarge the training set and improve the DNN performance in new releases. Its size is  $n = |T| \ll N$ . When an example  $t_i$  is submitted to the DNN, a human oracle assigns the expected output to  $t_i$ , and then compares it with the actual output. In classification tasks, this gives a binary outcome  $z_i$  (whether actual and expected labels match or not). In regression tasks, the comparison gives an offset  $\delta_i$ , which is the absolute difference between the true ( $r_i$ ) and predicted ( $\hat{r}_i$ ) values – considering this a failure or not depends on the tolerable threshold. For our purposes, it suffices to focus on the value of  $\delta_i$ .
- $\theta = Pr(z_i = 1)$ , with  $i = 1, \dots, |D|$ , is, in classification tasks, the true failure probability on a randomly selected example from the entire operational dataset, and corresponds to the true (unknown) proportion  $\theta = \frac{1}{N} \sum_{i=1}^N z_i$ . Accuracy is defined as:  $\xi = 1 - \theta$ . In the case of regression, we look at the mean squared error between the true ( $r_i$ ) and predicted ( $\hat{r}_i$ ) value over the entire operational dataset:  $\Delta = \frac{1}{N} \sum_{i=1}^N \delta_i^2$ , and  $\xi = 1 - \Delta$ . Its estimate is  $\hat{\xi}$ .

Given a sample size budget  $n$ , the goal of DeepSample is to select a subset  $T$  able of giving an *unbiased* (i.e., such that  $\mathbb{E}[\hat{\xi}] = \xi$ ) estimate of  $\xi$  while maximizing the *efficiency* of the estimator (i.e., minimizing the variance of the estimate).<sup>3</sup> In addition, the set  $T$  is wanted to expose as many failing examples as possible.

### 3.2 Overview of DeepSample

DeepSample is a family of techniques leveraging prior knowledge available about the operational dataset, supposed to be correlated to the variable to estimate (namely, accuracy). Prior information is encoded in what are called *auxiliary variables* [20], here denoted as  $\chi$ ; for instance, the confidence value provided by classifiers when predicting a label can be assumed to be (negatively) correlated with the failure probability  $\theta$ . Clearly, accuracy and efficiency of estimates depend on the extent to which assumptions hold.

The DeepSample techniques are characterized by two dimensions: *i*) the **sampling algorithm**, and *ii*) the **auxiliary variable**.

The former specifies a *sampling scheme*, namely the sequence of steps required to select the tests  $t_i$ . The latter specifies what is the *auxiliary variable*  $\chi$ , if used by the sampling scheme (not all auxiliary variables can be used in all the schemes).

There are two ways of exploiting the auxiliary variables. The first is to partition the dataset into classes that are homogeneous with respect to the auxiliary variable (e.g., similar confidence), similarly to stratification in sampling theory [20]. If the variable is well correlated with the failure probability  $\theta$  (or  $\Delta$  for regression), partitions too should be homogeneous with respect to  $\theta$  (or  $\Delta$ ). This allows to wisely allocate the number of examples to draw from each partition with the aim to reduce the variance of the estimation. The second way is to let the sampling scheme select the examples proportionally to the auxiliary variable's value, so as to get the ones

<sup>1</sup>The replication package is at: <https://github.com/dessertlab/DeepSample.git>.

<sup>2</sup>At each step, the partition selected to draw the next test is the one that yields the greatest descent (i.e., negative gradient) of the variance of the reliability estimator.

<sup>3</sup>Minimizing the variance is equivalent to minimize the MSE since the estimators are required to be unbiased. Low variance (or MSE) implies maximizing the confidence.

with higher expected failure probability. A proper estimator is then needed to correct bias due to this *unequal* selection probability.

Techniques can be *with* or *without replacement*. The former ones (allowing an example to be selected more times) are associated with simpler estimators - a common choice in literature [1] [9] [10] [11] [12]; the latter ones are expected to give higher efficiency, though the gain in large populations can be marginal (with-replacement schemes will unlikely select twice the same example).

The **estimator** takes the result of submitting the selected sample  $T$  to the DNN  $M$  (denoted as  $z_i$  and  $\delta_i$  for classification and regression, respectively) and yields an unbiased estimate of  $\xi$  by counterbalancing the disproportional selection (cf. with Sec. 3.4).

### 3.3 Auxiliary variables

We consider three auxiliary variables for classification problems, and for regression as well. For classification, they are: Confidence, Distance-based Surprise Adequacy (DSA), and Likelihood-based Surprise Adequacy (LSA). We opted for these variables based on the literature [1, 2]. For regression, they are LSA, and two variables based on the reconstruction error of a simple autoencoder (SAE) and of a variational autoencoder (VAE), which have been demonstrated to be effective in detecting inputs likely to cause failure [21].

Confidence  $C_{d_i}$  of an input  $d_i$  is the maximum value in the probability vector obtained from the last layer's output of the DNN<sup>4</sup>; it is for classification problems only. DSA and LSA, defined by Kim *et al.* [22], exploit Activation Traces (AT), which are vectors of activation values of neurons belonging to a certain layer. DSA is defined as:  $DSA_{d_i} = \frac{\sigma_A}{\sigma_B}$ , where  $\sigma_A$  is the Euclidean distance between the ATs of the input  $d_i$  (whose predicted class is A) and its nearest neighbour belonging to the same class A,  $\sigma_B$  is the distance between the ATs of  $d_i$  and its nearest neighbour belonging to a different class B.<sup>5</sup> It makes sense for classification models only. LSA uses Kernel Density Estimation (KDE) [23] to estimate the probability density of each activation value, obtaining the surprise of a new input with respect to the estimated density. LSA is a measure of rareness computed as:  $LSA_{d_i} = -\log(\hat{f}(d_i))$ , where  $\hat{f}(d_i)$  is the KDE applied to the new input  $d_i$ . LSA is for both classification and regression.

For SAE/VAE-based variables, we leverage the reconstruction error  $\epsilon$ . We used the two best-performing autoencoders implemented by Stocco *et al.* [21], SAE (Simple Autoencoder) with a single hidden layer, and VAE (Variational Autoencoder). We consider autoencoders as single-image reconstructors, computing their outputs for all the operational examples, and then calculating the reconstruction error as:  $\epsilon_{d_i} = \frac{1}{WHC} \sum_{k=1, j=1, c=1}^{W, H, C} (d_i[c][k, j] - d'_i[c][k, j])^2$ , where  $d_i$  is the original image,  $d'_i$  is the reconstructed image,  $W$ ,  $H$ , and  $C$  are width, height, and channels respectively. The corresponding auxiliary variables are synthetically called SAE and VAE, meaning the  $\epsilon_{d_i}$  value obtained by SAE and VAE.

All variables are assumed to be correlated to accuracy: lower confidence, higher surprise (DSA, LSA), and higher reconstruction error (SAE, VAE) are expected to be related to higher failure probability. To have all positive variables (from which selection probabilities need to be derived), DSA and LSA for classification

**Table 1: Compared testing techniques**

Technique	SUPS	RHC-S	SSRS	GBS	2-UPS	SRS	CES	DeepEST
Partitioning	✗	✗	✓	✓	✓	✗	✗	✗
Unequal selection	✓	✓	✗	✗	✓	✗	✓	✓
Without replacement	✗	✓	✓	✗	✓	✗	✗	✓

are *min-max* normalized. For regression, as the min-max normalization affects the distribution of test data, we just shift the values:  $DSA_{d_i} = DSA_{d_i} + \lceil \min(DSA_{d_i}) \rceil$  (the same for *LSA*). All the above variables are denoted as  $\chi_i$  in the following, when there is no need to distinguish them. In the case of confidence,  $\chi_i = 1 - C_{d_i}$  since we assume that confidence is negatively correlated to the accuracy.

### 3.4 Testing techniques

The characteristics of the eight compared testing techniques are summarized in Table 1; their description follows.

#### 3.4.1 Without-partitioning techniques.

**Simple Random Sampling (SRS).** SRS with replacement, where all examples have the same probability to be selected, is the simplest and baseline technique [17][1]. For SRS; unbiased estimators of  $\theta$  (for classification) and  $\Delta$  (regression) are, respectively, the observed proportion and mean squared error over the subset of selected tests:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n z_i \quad (1) \quad \hat{\Delta} = \frac{1}{n} \sum_{i=1}^n \delta_i^2 \quad (2)$$

**Simple Unequal Probability Sampling (SUPS).** This scheme leverages auxiliary variables  $\chi$  for selecting the examples. The selection probability  $\pi_i$  for the  $i$ -th example  $t_i$  is obtained by normalizing the auxiliary variable  $\pi_i = \chi_i / \sum_{i=1}^N \chi_i$ ; this is known as probability-proportional-to-size (PPS) sampling [20]. The selection is with replacement. An unbiased estimator is the sample mean of the observed values re-scaled by the inverse of their selection probability  $\pi_i$  and by  $N$ , known as Hansen-Hurwitz estimator [24]:

$$\hat{\theta} = \frac{1}{nN} \sum_{i=1}^n \frac{z_i}{\pi_i} \quad (3) \quad \hat{\Delta} = \frac{1}{nN} \sum_{i=1}^n \frac{\delta_i^2}{\pi_i} \quad (4)$$

Note that this is a generalization of SRS, wherein the selection probability is  $\pi_i = 1/N$  for all the examples.

**RHC-Sampling (RHC-S).** This is another unequal probability selection scheme, but without replacement, and uses the Rao, Hartley, and Cochran (RHC) estimator [25]. The scheme is as follows:

- (1) Given the budget of  $n = |T|$  test cases, divide randomly the  $N = |D|$  units of the operational dataset into  $n$  groups, by selecting  $G_1$  inputs with SRS *without replacement* for the first group, then  $G_2$  inputs out of the remaining  $(N - G_1)$  for the second, and so on. This will lead to  $n$  groups of size  $G_1, \dots, G_n$  with  $\sum_{r=1}^n G_r = N$ . The group size is arbitrary, but we select  $G_1 = G_2 = \dots = G_n = N/n$ , as this minimizes the variance.
- (2) One test case is then drawn by taking an input  $t_i$  in each of these  $n$  groups independently and with a PPS sampling according to the above-defined  $\pi$  variable.
- (3) Denote with  $\pi_{i,r}$  the probability associated with the  $t_i$ -th unit in the  $r$ -th group, and with  $q_r = \sum_{i \in G_r} \pi_{i,r}$  the sum in the  $r$ -th group. The unbiased estimators are:

<sup>4</sup>In the case of binary classification with a single neuron, the confidence is the neuron output  $o$  when  $o \geq 0.5$  (e.g., class 1) and  $1 - o$  when  $o < 0.5$  (class 0).

<sup>5</sup>Note that the computation does not need the actual labels (but only predicted ones).

$$\hat{\theta} = \frac{1}{N} \sum_{r=1}^n \frac{z_r}{\pi_r/q_r} \quad (5) \quad \hat{\Delta} = \frac{1}{N} \sum_{r=1}^n \frac{\delta_r^2}{\pi_r/q_r}. \quad (6)$$

**Cross-entropy Sampling (CES).** Cross Entropy-based Sampling (CES) was proposed by Li *et al.* [1]. The CES algorithm builds the sample first selecting randomly an initial set of examples, and then selecting the remaining examples trying to minimize the average cross-entropy between the probability distribution of the  $m$ -dimensional representation of neurons output computed on the operational dataset and the selected images. The objective is to sample a set of examples as much as possible representative of the operational dataset, namely if it contains the same proportion of mispredictions as the operational dataset. For CES, the authors demonstrate that the estimator is the same as SRS (Eq. 1 and Eq. 2).

**Deep neural networks Enhanced Sampler for operational Testing (DeepEST).** Guerriero *et al.* presented DeepEST [2], a technique for DNN operational testing with the twofold objective of accuracy estimation and accuracy improvement. DeepEST exploits adaptive sampling [8] to select a sample providing a close and efficient estimate *and*, at the same time, including a high number of failing examples. The original version of DeepEST works only for classification tasks. We hereafter extend it for regression too, defining the corresponding estimator. The auxiliary variable,  $\chi$ , is used by DeepEST to define a *weight*  $w_{i,j}$  between any pair of examples  $d_i$  and  $d_j$  of the operational dataset, used to explore the example space adaptively. The weight  $w_{i,j}$  is the value of  $\chi_{d_j}$  if  $\chi_{d_i}$  exceeds a threshold (i.e., it means that  $t_i$  is in an interesting cluster to explore), 0 otherwise. The thresholds are those of the original paper. The strategy acts as follows: the first input is selected via SRS, then a *weight-based sampling* (WBS) is used with probability  $r$  to sample the next example (or SRS with probability  $1-r$ ). The example  $d_i$  is selected at step  $k$  with probability  $q_{k,t_i}$ :

$$q_{k,i} = r \cdot \frac{\sum_{j \in s_k} w_{i,j}}{\sum_{h \in s_k, t_j \in s_k} w_{h,j}} + (1-r) \cdot \frac{1}{N - n_{s_k}} \quad (7)$$

where:

- $r$ : probability of using WBS;
- $s_k$ : current sample (all examples selected up to step  $k$ );
- $w_{i,j}$ : weight relating example  $d_j$  in  $s_k$  to example  $d_i$ ;
- $n_{s_k}$ : the size of the current sample  $s_k$ ;
- $N$ : the size of the operational dataset.

WBS selects an example  $d_i$  proportionally to the sum of weights  $w_{i,j}$  of already selected examples toward  $d_i$ . We compute the following step-by-step estimators to balance for the adaptive sampling:

$$\hat{\theta} = \frac{1}{n} (z_1 + \frac{1}{N} \sum_{k=2}^n \tilde{\theta}_k) \quad (8) \quad \hat{\Delta} = \frac{1}{n} (\delta_1^2 + \frac{1}{N} \sum_{k=2}^n \tilde{\Delta}_k) \quad (9)$$

where  $z_1$  and  $\delta_1^2$  are the estimates obtained at step  $k = 1$  (hence when  $n = 1$ ),  $\tilde{\theta}_k$  and  $\tilde{\Delta}_k$  are the Hansen-Hurwitz estimates at step  $k > 1$  for the total failures and for the mean-squared error:

$$\tilde{\theta}_k = \sum_{j \in s_k} z_j + \frac{z_i}{q_{k,i}} \quad (10) \quad \tilde{\Delta}_k = \sum_{j \in s_k} \frac{\delta_j^2}{k-1} + \frac{\delta_i^2/k}{q_{k,i}}. \quad (11)$$

The final estimators (Eq. 8, 9) are the sample mean of the step-by-step estimators. For regression, the  $k$ -th MSE estimate is  $\tilde{\Delta}_k$ .

**3.4.2 Partition-based techniques.** Partition-based techniques split the operational dataset into classes to improve sampling. In sampling theory, stratification splits the population to have a small expected intra-stratum variance of the variable to estimate  $\xi$  and a large inter-strata variance, so as to sample more from partitions with higher variance. Since the true variance of  $\xi$  is unknown, stratification can be done on an estimate of such variance (e.g., computed from a preliminary sample) [17]. However, this would require labeling a subset only just for the purpose of estimating the variance and then applying stratification. Another common solution, that we adopt, is to stratify based on auxiliary variables. Although risky (performance depends on the extent to which they are correlated to  $\xi$ ), this requires no prior knowledge about  $\xi$ . We used  $k$ -means clustering [26] on  $\chi$ , with  $k$  set to 10 after a preliminary tuning on 30 random samples from MNIST, with  $k = 6, 8, 10, 12$ .

**Stratified Simple Random Sampling (SSRS).** In this scheme, the number of examples to draw from each partition  $p$  is computed by the Neyman allocation [20] applied to  $\chi$ , namely proportionally to the standard deviation of the (normalized)  $\chi$  values for that partition, and to the size of the partition,  $N_p$ . Selection within the partition is without-replacement. The estimators are the weighted sum of the SRS estimates for partitions:

$$\hat{\theta} = \frac{1}{N} \left( \sum_{p=1}^P N_p \hat{\theta}_p \right) \quad (12) \quad \hat{\Delta} = \frac{1}{N} \left( \sum_{p=1}^P N_p \hat{\Delta}_p \right) \quad (13)$$

where  $\hat{\theta}_p$  and  $\hat{\Delta}_p$  are the within-partition SRS estimators (Section 3.4.1),  $P = k = 10$  is the number of partitions.

**Gradient-Based Sampling (GBS).** Unlike SSRS, this technique does not initially allocate a sample size for each stratum, but it decides step by step which partition the next example will be drawn from. Inspired by adaptive testing with gradient descent [12], at each step the partition is chosen so as to maximize the reduction of the variance  $Var(\hat{\xi})$  of the  $\xi$  estimator, by taking the partition with the largest negative gradient:  $-\partial Var(\hat{\xi})/\partial n_p$  (ties broken randomly),  $n_p$  being the number of examples selected from partition  $p$  up to the current step. The selection within the partition is then with replacement. The estimators are the same as SSRS (Eq. 12, 13). Note that the with-replacement SRS, used in GBS, and without-replacement SRS, used in SSRS, have the same mean estimators – they differ for the variance of these estimators.

**Two-stage Unequal Probability Sampling (2-UPS).** This technique implements a two-stage sampling scheme, where unequal probability sampling is adopted to select the partition (first stage), and SRS without replacement is adopted to select the example from the chosen partition (second stage). The selection probability for partition  $p$  is proportional to the sum of (normalized)  $\chi$  values (denoted as  $\pi_i$  as in SUPS and RHC-S) within that partition:

$$\psi_p = \frac{\sum_{i=1}^{N_p} \pi_i}{\sum_{p=1}^P \sum_{i=1}^{N_p} \pi_i}. \quad (14)$$

Clearly, selection of partitions is with replacement;  $Q_p$  is the number of times partition  $p$  is selected. The estimator for this technique is the average over  $n$  estimates:

$$\hat{\theta} = \frac{1}{Nn} \left( \sum_{p=1}^P \sum_{i=1}^{Q_p} \frac{z_i N_p}{\psi_p} \right) \quad (15) \quad \hat{\Delta} = \frac{1}{Nn} \left( \sum_{p=1}^P \sum_{i=1}^{Q_p} \frac{\delta_i^2 N_p}{\psi_p} \right) \quad (16)$$

Inner terms  $(z_i N_p)/(\psi_p)$  and  $(\delta_i^2 N_p)/(\psi_p)$  are Hansen-Hurwitz estimates for the total number of failures and squared errors in partition  $p$ , respectively. These estimates are summed up over all partitions and divided by the sample size  $n$  to get an average total estimate. The division by  $N$  gives  $\hat{\theta}$  and  $\hat{\Delta}$ .

## 4 EVALUATION

### 4.1 Research questions and metrics

**RQ1:** *How do the sampling techniques perform in assessing the operational accuracy of DNN models?*

- **RQ1.1:** *How do the techniques perform for classification?*
- **RQ1.2:** *How do the techniques perform for regression?*

Over  $R = 30$  repetitions, we measure the root mean squared error (RMSE) between the accuracy estimates  $\hat{\xi}$  and the true accuracy  $\xi$  computed on the operational datasets by labeling all the images:

$$RMSE = \sqrt{\frac{\sum_{r=1}^R (\xi - \hat{\xi})^2}{R}} \quad (17)$$

where  $\hat{\xi}$  for classification and regression is computed using  $\hat{\theta}$  and  $\hat{\Delta}$ , respectively. Lower RMSE means higher confidence in the estimate.

**RQ2:** *How do the sampling techniques perform in detecting failing examples?* An issue of some techniques like CES is that they, with reason, try to have in the sample the same proportion of failures as in the operational dataset, to faithfully estimate accuracy (what is called the *imitation bias* [2]); but in highly-accurate DNNs, this entails very few failures exposed, which requires engineers to run further tests to expose failures – an issue addressed by DeepEST [2]. Thus a desirable property is to expose a high number of failures, besides the ability to provide unbiased high-confidence estimates.

- **RQ2.1:** Classification task. *How many failures (namely, misclassifications) are exposed by the techniques?*
- **RQ2.2:** Regression task. *How many examples with an inaccurate prediction are selected by the techniques?* Since in regression we have continuous outputs, we measure the number of examples having a difference between true and predicted output (i.e., the offset:  $\delta_i = |r - \hat{r}_i|$ ) greater than or equal to a given value  $y$ :  $N_{\delta \geq y}$  with  $y$  ranging from  $0^\circ$  to  $25^\circ$ , with a step of  $2.5^\circ$ .<sup>6</sup>

**RQ3:** *How does the budgeted sample size affect performance?* The sample size is directly related to the cost of labelling, as it determines the number of examples to be manually labelled.

- **RQ3.1:** *How does the size affect the accuracy estimate?*
- **RQ3.2:** *How does the size affect the failing examples detection?*

To answer RQ1 and RQ2, we consider a budget size of 200, as in [1, 2]; the total runs are 6,600 [11 models  $\times$  30 repetitions  $\times$  (6 techniques  $\times$  3 auxiliary variables + the 2 techniques CES and SRS not using auxiliary variables)]. For RQ3, with 5 sample size values (50, 100, 200, 400, 800), there are additional  $6,600 \times 4 = 26,400$  runs, for a total of 33,000 runs.

### 4.2 Subjects

The evaluation is on 11 DNN models on popular datasets (Table 2). For classification we consider 3 models for each of the following

**Table 2: List of experimental subjects**

Model	Dataset	Layers	Parameters	Accuracy
A	MNIST	7	6,237	90.3%
B		6	97,114	94.8%
C		8	545,546	93.3%
D	CIFAR10	13	1,084,234	71.5%
E		10	258,762	79.0%
F		12	550,570	65.1%
G	CIFAR100	16	15,047,588	66.3%
H		9	564,484	57.4%
I		13	1,465,220	58.8%
DO	Udacity	13	2,116,983	0.904
DD		15	3,276,225	0.918

3 datasets: MNIST [27], CIFAR10 and CIFAR100 [28]. MNIST has 70,000 entries; CIFAR10 has 60,000 entries; both have 10 classes. CIFAR100 also has 60,000 entries, with 100 classes. For regression, we consider 2 models for the Udacity dataset<sup>7</sup> (101,396 entries for training, 5,614 for test) for steering angle prediction in Autonomous Driving Systems: *Dave\_orig* (DO) and *Dave\_dropout* (DD) [1][29].

Recht *et al.* [30] showed that if the accuracy is computed on previously unseen data, it is actually smaller than the claimed one by a value ranging from 3% to 15% on CIFAR10 and from 11% to 14% on ImageNet. Therefore, for a more realistic accuracy, each DNN is trained “from scratch” by separating training, verification, and operational sets, as in [31]. The verification set is the set used to evaluate the DNN. The operational set contains unlabelled images.

The three datasets are split as follows. For MNIST, 7,000 images are for training and 2,500 for verification; the remaining 60,500 entries are the operational dataset (*big size*). All models trained with this configuration achieve an accuracy greater than 90%. For CIFAR10, we use 24,000 images for training and 2,500 for verification; the remaining 33,500 entries are the operational dataset (*medium size*). For CIFAR100, 40,000 entries are for training and 5,000 for verification; thus, the operational dataset has 15,000 images (*small size*). The operational datasets are chosen to have MNIST (big) almost double than CIFAR10 (medium) and four times CIFAR100 (small). The greater training set sizes for CIFAR10 and CIFAR100 are due to the higher complexity of the images, to pursue an acceptable accuracy. For regression models, we use as operational dataset the entire test dataset, as all its examples are unseen during training.

## 5 RESULTS

### 5.1 RQ1: operational accuracy assessment

**5.1.1 RQ1.1: Classification.** To check if techniques have pairwise a statistically significant difference, we run the Friedman test [32] on all subjects/auxiliary variable pairs. The  $p$ -value is lower than  $\alpha = 0.05$  in all cases, hence the null hypothesis of no difference among techniques is always rejected. For pairwise comparison, we run the non-parametric *post hoc* Dunn test [33] with the Holm adjustment. The results are in Figure 1, where gray squares mean *no significant difference* for the pair, white (black) squares mean the technique on the row is statistically better (worse) than the one on the column. All exact  $p$ -values are in the replication package.<sup>1</sup>

On MNIST, DeepEST and 2-UPS significantly differ from the other techniques (which perform similarly). We show three examples in Figures 2a-2c. The first is on Model A (top-left box in Fig. 1)

<sup>6</sup>The output of the DNN for regression is a steering angle degree; a difference greater than  $25^\circ$  is unrealistic, and never occurred in our experiments.

<sup>7</sup><https://github.com/udacity/self-driving-car>.

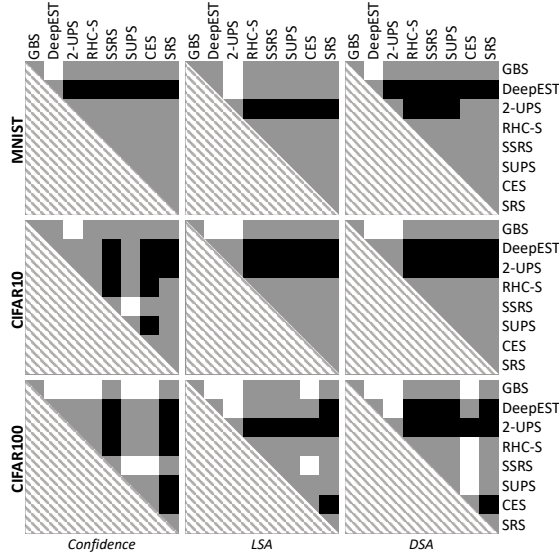


Figure 1: RQ1.1: Dunn test on the classification task

with *confidence* as auxiliary variable. Here, the RMSE of 2-UPS is by far the worst; however, it is affected by few outliers due to the inability of the estimator to balance, within the given budget, the examples whose auxiliary information is incoherent with the result (e.g., failures with high confidence). If we take the root square of the *median* of squared errors, called RMedSE, we see 2-UPS is in line with the others. This causes 2-UPS to go unreported as significantly different by the Dunn test (non-parametric, hence robust to outliers). DeepEST, instead, shows to be significantly worse.

The second example (Fig. 2b) is on Model B with LSA (top-middle box in Fig. 1). 2-UPS performs worse than the others, the second being DeepEST although the difference is not detected by the Dunn test. The third example (Fig. 2c) is on Model C with DSA (top-right box in Fig. 1). In this case, both DeepEST and 2-UPS perform worse. In the second and third examples, the values of the RMSE and RMedSE for 2-UPS are close (no outliers); this is attributable to the higher representativeness of LSA and DSA, which were more robust than *confidence* to misclassification on inputs closer to training set.

On CIFAR10 with *confidence*, the outliers in 2-UPS are even more pronounced (Fig. 2d). DeepEST and 2-UPS again give the worst estimates. The other algorithms are similar (Fig. 2e).

On CIFAR100 with *confidence*, GBS, SSRS and SRS differ significantly from the other techniques. Consider Fig. 2f (Model I). GBS, SSRS and SRS exhibit the best values. Outliers in 2-UPS are confirmed; they are more frequent, especially on low-accurate models (the behaviour is more evident with CIFAR10 and CIFAR100, less accurate for MNIST). On the other hand, it is worth to stress that not all the algorithms relying on the auxiliary variable suffer from unstable results; RHC-S and SUPS are more stable. With LSA (Fig. 2g) and DSA (Fig. 2h), the previous results are confirmed; after DeepEST and 2-UPS, CES turned out to be the third worst one.

**5.1.2 RQ1.2: Regression.** The Friedman test gives a  $p$ -value lower than  $\alpha = 0.05$  in all the cases, except for DO with the SAE auxiliary variable. Figure 3 shows the results of the Dunn test for pairwise comparison. When using LSA, DeepEST and 2-UPS are significantly

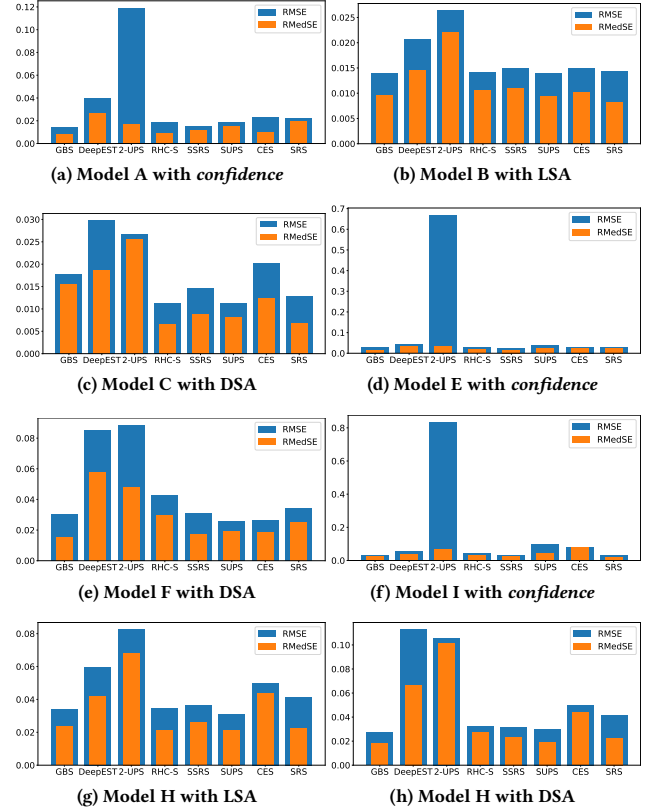


Figure 2: RQ1.1: Examples

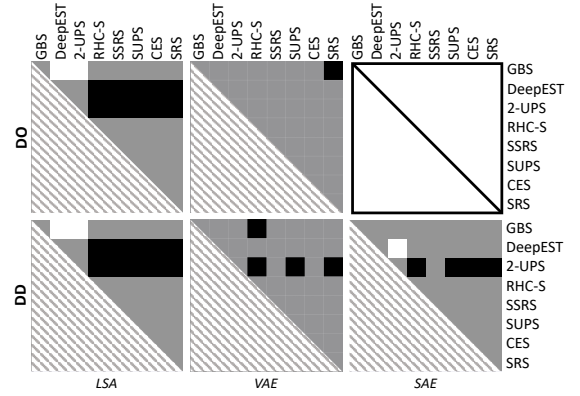


Figure 3: RQ1.2: Dunn test on the regression task

worse. Figures 4 and 5 confirm their higher values of RMSE and RMedSE for both DO and DD models.

With the VAE auxiliary variable, GBS differs from SRS, but it is almost equivalent to the other algorithms in the DO model. Figure 4 confirms that GBS has higher RMSE than the others. For the DD model, 2-UPS differs from SUPS and RHC-S. 2-UPS has the highest RMSE values, while RHC-S has the lowest ones (Figure 5).

With SAE, the Friedman test did not detect any difference for DO, while, for DD, 2-UPS is still the worst technique, although it is closer to GBS and SSRS than the previous cases.



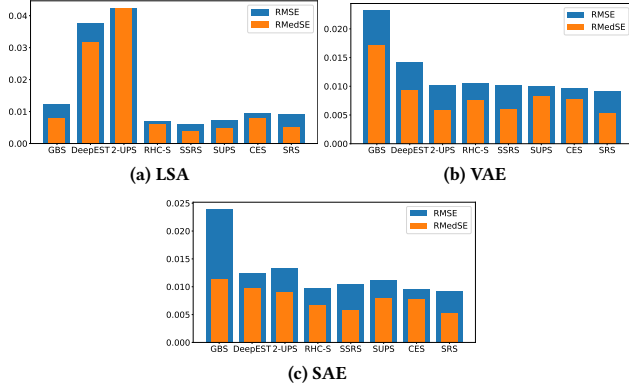


Figure 4: RQ1.2: DO model - Bar charts

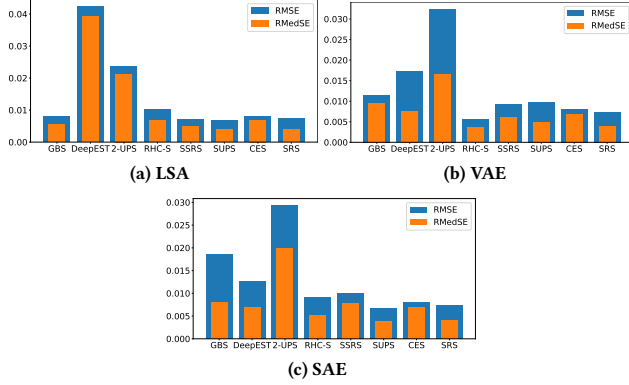


Figure 5: RQ1.2: DD model - Bar charts

As for RMedSE%, the worst values are always with the LSA variable: for DO, 2-UPS and DeepEST show the worst values (**4.2%** and **3.2%**, respectively); SSRS has the best value (**0.4%**); for DD, DeepEST and 2-UPS show **3.9%** and **2.1%**, respectively, against SUPS with **0.4%**. The worst case with autoencoders is for DD with the SAE variable, with 2-UPS (**2.0%**), while the best one is SRS (**0.4%**). These results are in line with the classification ones.

Overall, the techniques are all equivalently effective in assessing the operational accuracy of DNN models for classification and regression tasks, except for DeepEST and 2-UPS. While DeepEST was expected to show worse results (its primary objective is on failure exposure), 2-UPS shows many outliers since it is strongly affected by auxiliary variable representativeness.

## 5.2 RQ2: failing examples detection

For RQ2, we treat classification and regression differently. For the former, we count the number of misclassifications. For the latter, we count the number of examples whose offset  $\delta$  (predicted vs actual value) is greater than a threshold  $y$ , with  $y \in [0^\circ, 2.5^\circ, 5^\circ, \dots, 25^\circ]$  – the higher the difference, the more severe the misprediction.

**5.2.1 RQ2.1: Classification.** Table 3 reports the number of misclassifications broken down by dataset and auxiliary variable – the best mean values are in bold. DeepEST exposes more failures than the others in 7 out of 9 times. 2-UPS has the highest value only for

Table 3: RQ2.1: Number of exposed failures (classification)
























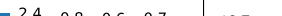
$\chi$	Technique	MNIST		CIFAR10		CIFAR100	
		mean	std	mean	std	mean	std
confidence	GBS	28.2	7.2	68.4	8.0	86.2	10.6
	DeepEST	<b>80.5</b>	10.3	108.7	9.4	136.4	6.6
	2-UPS	69.5	17.4	<b>108.9</b>	11.7	141.5	6.2
	RHC-S	70.6	16.5	106.0	12.8	142.3	6.4
	SSRS	38.4	10.2	78.7	13.5	109.2	6.2
	SUPS	69.8	16.9	106.9	12.3	<b>143.6</b>	5.5
	CES	15.6	5.8	55.8	12.6	70.4	7.5
	SRS	14.6	5.3	57.3	13.1	78.0	12.8
LSA	GBS	21.0	5.2	58.2	10.6	84.6	7.4
	DeepEST	<b>35.9</b>	10.7	<b>69.2</b>	10.4	<b>119.9</b>	12.4
	2-UPS	25.0	7.5	61.8	11.3	110.5	20.1
	RHC-S	25.5	7.1	62.9	11.4	110.3	19.3
	SSRS	27.3	6.4	60.1	10.7	93.2	10.7
	SUPS	25.4	6.8	63.5	11.2	110.2	21.7
	CES	15.6	5.8	55.8	12.6	70.4	7.5
	SRS	14.6	5.3	57.3	13.1	78.0	12.8
DSA	GBS	20.3	6.0	63.6	11.0	88.5	7.5
	DeepEST	<b>73.0</b>	17.9	<b>102.4</b>	8.9	<b>136.7</b>	4.9
	2-UPS	25.2	7.2	65.2	12.0	96.3	8.7
	RHC-S	23.9	7.0	65.5	13.8	96.9	9.6
	SSRS	21.7	5.6	58.3	12.0	82.4	6.4
	SUPS	24.7	6.9	65.7	12.1	97.3	10.7
	CES	15.6	5.8	55.8	12.6	70.4	7.5
	SRS	14.6	5.3	57.3	13.1	78.0	12.8

CIFAR10 with confidence, and SUPS for CIFAR100 with confidence. 2-UPS, SUPS, and RHC-S almost equivalently follow DeepEST.

These results counterbalance the DeepEST and 2-UPS results on the estimates, which were worse than the others (RQ1.1). DeepEST assumes that failures belong to a rare population, and is conceived to spot them. The greater ability to find misclassifications causes a greater variability of the estimates, and more budget is needed to converge. A similar problem is observed for 2-UPS. We hypothesize that partitioning combined with unequal sampling (both based on the auxiliary variable  $\chi$ ) can push toward failing examples, but the estimator needs more time to converge. Unlike DeepEST, 2-UPS showed many spikes in the accuracy estimation; the estimator generates spikes every time a failure is detected with “misleading” values of  $\chi$ , namely misclassified examples with values of  $\chi$  that would indicate a correct classification. For instance, failures with high *confidence*, or with low LSA/DSA. SUPS and RHC-S seem very good compromises between the two – more details in the final discussion. GBS, CES, and SRS detect fewer failures. For SRS and CES this is likely because the former does not use any auxiliary variable, the latter uses cross-entropy, not supposed to be related to failures. GBS and SSRS both use  $\chi$  only for partitioning; but GBS detects fewer failures likely because the algorithm is thought to minimize the variance of the estimate.

**5.2.2 RQ2.2: Regression.** Tables 4 and 5 report the histograms of the offset, starting from  $12.5^\circ$  to  $25^\circ$ . Compared to the classification case, the differences here are less pronounced. Looking at the sum of the bins, we notice that CES and SRS select less examples with higher offset with respect to the others under the LSA case, while all

**Table 4: RQ2.2: Average number of failures whose offset is  $12.5^\circ \leq \delta < 15^\circ$ ,  $15^\circ \leq \delta < 17.5^\circ$ ,  $\dots 22.5^\circ \leq \delta < 25^\circ$ . DO model**

Technique	LSA							VAE							SAE						
GBS							10.2							23.1							16.6
DeepEST							28.9							19.5							19.2
2-UPS							31.0							19.6							18.2
RHC-S							32.8							18.7							18.7
SSRS							31.4							19.1							17.1
SUPS							32.7							18.1							19.1
CES							18.9							18.9							18.9
SRS							17.7							17.7							17.7

**Table 5: RQ2.2: Average number of failures whose offset is  $12.5^\circ \leq \delta < 15^\circ$ ,  $15^\circ \leq \delta < 17.5^\circ$ ,  $\dots 22.5^\circ \leq \delta < 25^\circ$ . DD model**

Technique	LSA							VAE							SAE						
GBS	<div><div>3.7</div><div>1.9</div><div>1.4</div><div>0.8</div><div>0.8</div><div>0.8</div></div>						9.4	<div><div>8.3</div><div>4.9</div><div>1.7</div><div>2.0</div><div>0.9</div><div>1.9</div></div>						19.7	<div><div>5.7</div><div>4.9</div><div>1.7</div><div>0.9</div><div>0.6</div><div>1.0</div></div>						14.7
DeepEST	<div><div>8.8</div><div>6.9</div><div>4.2</div><div>2.8</div><div>1.2</div><div>2.2</div></div>						26.1	<div><div>4.7</div><div>3.0</div><div>1.9</div><div>1.1</div><div>0.5</div><div>0.8</div></div>						12.0	<div><div>4.8</div><div>2.9</div><div>2.1</div><div>1.1</div><div>0.5</div><div>0.7</div></div>						12.1
2-UPS	<div><div>10.5</div><div>7.9</div><div>5.9</div><div>3.0</div><div>1.5</div><div>2.3</div></div>						31.1	<div><div>4.2</div><div>3.0</div><div>1.7</div><div>1.0</div><div>0.4</div><div>0.7</div></div>						11.0	<div><div>4.8</div><div>3.1</div><div>1.7</div><div>1.1</div><div>0.6</div><div>0.8</div></div>						12.0
RHC-S	<div><div>11.1</div><div>7.2</div><div>4.7</div><div>2.7</div><div>1.9</div><div>2.3</div></div>						29.9	<div><div>4.7</div><div>2.5</div><div>1.3</div><div>0.8</div><div>0.5</div><div>0.9</div></div>						10.6	<div><div>5.3</div><div>2.5</div><div>1.9</div><div>0.9</div><div>0.5</div><div>0.6</div></div>						11.8
SSRS	<div><div>9.1</div><div>5.9</div><div>4.3</div><div>3.3</div><div>2.1</div><div>2.5</div></div>						27.3	<div><div>5.7</div><div>2.7</div><div>1.8</div><div>0.9</div><div>0.6</div><div>0.7</div></div>						12.4	<div><div>5.0</div><div>3.5</div><div>2.9</div><div>1.5</div><div>0.9</div><div>1.6</div></div>						15.4
SUPS	<div><div>11.1</div><div>8.1</div><div>5.2</div><div>2.6</div><div>2.1</div><div>2.8</div></div>						31.9	<div><div>4.6</div><div>3.1</div><div>1.7</div><div>1.2</div><div>0.7</div><div>0.7</div></div>						12.1	<div><div>4.4</div><div>2.9</div><div>1.5</div><div>1.0</div><div>0.7</div><div>0.8</div></div>						11.4
CES	<div><div>3.9</div><div>2.6</div><div>1.5</div><div>1.1</div><div>0.8</div><div>0.5</div></div>						10.3	<div><div>3.9</div><div>2.6</div><div>1.5</div><div>1.1</div><div>0.8</div><div>0.5</div></div>						10.3	<div><div>3.9</div><div>2.6</div><div>1.5</div><div>1.1</div><div>0.8</div><div>0.5</div></div>						10.3
SRS	<div><div>4.6</div><div>2.3</div><div>2.1</div><div>0.3</div><div>0.4</div><div>0.1</div></div>						9.8	<div><div>4.6</div><div>2.3</div><div>2.1</div><div>0.3</div><div>0.4</div><div>0.1</div></div>						9.8	<div><div>4.6</div><div>2.3</div><div>2.1</div><div>0.3</div><div>0.4</div><div>0.1</div></div>						9.8

the techniques are roughly equivalent with VAE and SAE<sup>8</sup>. GBS has similar poor performance, but it performs much better when used with VAE (consistently with the more unstable RMSE (Fig. 4). SUPS is the best one with LSA. The good performance of partitioning-based techniques (which achieve or even outperform DeepEST) is attributable to a better effect of partitioning when applied to regression compared to classification (since the auxiliary variable, used for partitioning, and the offset are more correlated).

### 5.3 RQ3: efficiency analysis

**5.3.1 RQ3.1. Accuracy assessment.** We synthesize in Tables 6 and 7 the results for classification and regression. Besides the RMSE value at each point,<sup>9</sup> we are interested in figuring out if the techniques smoothly converge as the sample size increase. First, we report for each dataset, technique, auxiliary variable, and model, how many times the minimum RMSE is reached under the given sample size. For instance, 3/3/3 of GBS for sample size 800 in MNIST, means

<sup>8</sup>Note that the histograms for CES and SRS are the same along the three columns of the Table since they do not use LSA/VAE/SAE.

<sup>9</sup>The full set of graphs for each dataset-auxiliary variable-model combination over the sample size are in the replication package.

that the minimum RMSE was reached for all the 3 models used with MNIST, using respectively *confidence*/LSA/DSA as auxiliary variable. This is marked as green, and is the expected behaviour. When this is not true for at least one case, we mark it as red, and correspondingly mark as yellow those cells in the same row (with sample size smaller than 800) where the minimum was reached.

There are many cases where the minimum is not achieved with the largest sample size (red cells). For instance, the instability of 2-UPS makes it even reach the best values with a sample size 50 (MNIST and CIFAR100) and sample size 100 (MNIST and CIFAR10). CES with CIFAR100 has the same convergence problem, while it is stable in MNIST and CIFAR10. In remaining red cases, the minimum is at 400. SRS is the most stable technique, for independence from auxiliary variables. GBS and DeepEST are stable for 2 of 3 datasets; in the bad case, they converge at size 400. For regression, performance is better; GBS is more unstable, while the others converge at 800 with few exceptions at 400 and one (CES) at 200.

Tables 6 and 7 report also in how many cases the RMSE with budget 50 is smaller than that with budget 800 (red cells). We call this *inversions*, denoting convergence problems. There are 5 such cases: 3 for 2-UPS (2 with MNIST and 1 with CIFAR100), 1 for SUPS and 1



**Table 6: RQ3.1: RMSE sensitivity analysis (conf./LSA/DSA)**

	Technique	Minimum RMSE					Inversions 50>800
		50	100	200	400	800	
MNIST	GBS	0/0/0	0/0/0	0/0/0	0/0/0	3/3/3	3/3/3
	DeepEST	0/0/0	0/0/0	0/0/0	0/0/1	3/3/2	3/3/3
	2-UPS	1/0/0	1/0/0	0/0/0	0/1/0	1/2/3	1/3/3
	RHC-S	0/0/0	0/0/0	0/0/0	0/1/1	3/2/2	3/3/3
	SSRS	0/0/0	0/0/0	0/0/0	0/0/1	3/3/2	3/3/3
	SUPS	0/0/0	0/0/0	1/0/0	0/0/0	2/3/3	3/3/3
	CES	0	0	0	0	3	3
	SRS	0	0	0	0	3	3
CIFAR10	GBS	0/0/0	0/0/0	0/0/0	0/0/0	3/3/3	3/3/3
	DeepEST	0/0/0	0/0/0	0/0/0	0/0/0	3/3/3	3/3/3
	2-UPS	0/0/0	0/0/1	0/0/0	1/0/0	2/3/2	3/3/3
	RHC-S	0/0/0	0/0/0	0/0/0	1/0/0	2/3/3	3/3/3
	SSRS	0/0/0	0/0/0	0/0/0	0/0/0	3/3/3	3/3/3
	SUPS	0/0/0	0/0/0	0/0/0	1/0/0	2/3/3	3/3/3
	CES	0	0	0	0	3	3
	SRS	0	0	0	0	3	3
CIFAR100	GBS	0/0/0	0/0/0	0/0/0	0/0/1	3/3/2	3/3/3
	DeepEST	0/0/0	0/0/0	0/0/0	0/0/0	3/3/3	3/3/3
	2-UPS	0/0/1	0/0/0	0/0/0	0/0/0	3/3/2	3/3/2
	RHC-S	0/0/0	0/0/0	0/0/0	0/1/0	3/2/3	3/3/3
	SSRS	0/0/0	0/0/0	0/0/0	0/0/0	3/3/3	3/3/3
	SUPS	0/0/0	0/0/0	0/0/0	2/1/0	1/2/3	2/3/3
	CES	0	1	1	0	1	2
	SRS	0	0	0	0	3	3

**Table 7: RQ3.1: RMSE sensitivity analysis (LSA/VAE/SAE)**

	Technique	Minimum RMSE					Inversions 50>800
		50	100	200	400	800	
Udacity	GBS	0/0/0	0/0/0	0/2/0	0/0/1	2/0/1	2/2/2
	DeepEST	0/0/0	0/0/0	0/0/0	0/0/0	2/2/2	2/2/2
	2-UPS	0/0/0	0/0/0	0/0/0	1/1/0	1/1/2	2/2/2
	RHC-S	0/0/0	0/0/0	0/0/0	0/0/0	2/2/2	2/2/2
	SSRS	0/0/0	0/0/0	0/0/0	0/0/0	2/2/2	2/2/2
	SUPS	0/0/0	0/0/0	0/0/0	1/0/0	1/2/2	2/2/2
	CES	0	0	1	0	1	2
	SRS	0	0	0	0	2	2

for CES (both with CIFAR100). Inversions never occur for regression. SRS is still the most stable technique for both classification and regression. Again, 2-UPS is the most affected one.

**5.3.2 RQ3.2: Failing examples detection.** Results for this RQ are in Tables 8 and 9. For regression, we consider as failures all the predictions with an error on the steering angle greater than 12.5°. The Table reports the mean (over the sample sizes) of the minimum and maximum number of failures, and the ratio between the number of failures detected with sizes 800 and 50 ( $F_{800/50}$ ).

We observe there is no *inversion*: failures constantly increase with the budget size – they roughly double as the sample size doubles for both classification and regression (detailed values are in the replication package). The expectation in this case is fully matched by all techniques. Looking at  $F_{800/50}$ , all the results of RQ2 are confirmed for all budget sizes.

## 6 DISCUSSION

We analyze the results with respect to the main impacting factors, to provide guidance to both practitioners (to select the technique best fitting the needs) and researchers (to design new techniques).

**Table 8: RQ3.2: Failures sensitivity analysis (conf./LSA/DSA)**

	Technique	$mean(min)$	$F_{800/50}$	$mean(max)$
MNIST	GBS	5.3/5.8/4.5	26.3/15.7/21.9	136.5/91.9/95.9
	DeepEST	19.3/8.0/17.9	16.7/17.7/16.5	321.6/140.0/295.9
	2-UPS	17.4/5.8/5.9	15.9/17.1/16.9	277.2/100.5/98.0
	RHC-S	17.1/6.7/6.2	16.1/15.1/16.0	274.2/101.8/100.3
	SSRS	10.0/6.9/5.5	15.5/15.8/16.6	153.3/107.4/90.9
	SUPS	17.7/6.2/5.8	16.0/16.4/17.3	282.6/102.6/100.9
	CES	3.8	16.1	61.5
	SRS	3.8	15.9	58.6
CIFAR10	GBS	15.4/14.9/14.4	17.9/15.9/18.3	272.5/238.7/261.1
	DeepEST	26.9/17.1/25.3	16.1/16.2/16.2	432.7/277.7/408.6
	2-UPS	26.7/15.8/16.8	16.1/16.1/15.8	431.8/252.7/264.2
	RHC-S	27.2/15.9/16.1	15.7/16.0/16.3	427.7/252.6/262.4
	SSRS	19.7/15.3/14.4	16.0/15.7/16.0	315.1/239.4/231.1
	SUPS	26.7/15.1/16.4	16.2/16.6/16.2	433.8/250.6/263.3
	CES	14.1	15.2	216.5
	SRS	13.8	16.4	227.6
CIFAR100	GBS	21.5/21.3/21.7	15.9/15.9/16.5	341.0/339.2/357.8
	DeepEST	33.7/29.8/34.9	16.2/16.0/11.3	546.6/475.3/393.3
	2-UPS	35.7/27.2/24.5	15.9/16.0/15.7	565.9/434.7/385.8
	RHC-S	35.3/27.6/24.1	15.9/15.7/15.9	560.9/432.6/383.3
	SSRS	28.2/22.7/20.8	15.5/16.5/15.9	436.5/374.5/329.8
	SUPS	35.2/27.4/23.8	16.2/16.2/16.4	571.2/441.1/389.7
	CES	19.7	13.8	270.5
	SRS	20.8	15.2	315.0

**Table 9: RQ3.2: Failures sensitivity analysis (LSA/VAE/SAE)**

	Technique	$mean(min)$	$F_{800/50}$	$mean(max)$
Udacity	GBS	4.2/3.7/4.1	3.0/34.1/14.9	12.4/122.8/57.8
	DeepEST	6.8/3.6/3.6	16.2/16.6/18.0	109.3/60.3/61.2
	2-UPS	7.4/3.8/3.4	16.6/15.5/18.9	123.1/59.1/61.2
	RHC-S	8.1/3.5/3.7	13.8/17.3/16.4	112.1/60.8/61.5
	SSRS	7.5/4.2/4.2	15.6/15.1/15.2	117.5/64.2/64.9
	SUPS	8.1/3.7/3.8	15.9/16.5/16.2	127.3/60.6/61.1
	CES	3.1	20.9	65.3
	SRS	3.3	17.9	58.3

The performance of a sampling technique depends on the tester's *objective* and on the application *context*.

As for the *objective* (set in the problem formulation, Sec. 3.1), while a tester is always interested in an *unbiased assessment* of the DNN accuracy, s/he can specifically focus on:

- ① *High confidence (i.e., low variance)*, e.g., as criterion to release a DNN, or to choose which DNN to deploy among various alternatives – a high-confidence estimate is usually required in critical domains. This can be achieved by reducing the RMSE or RMedSE: in the former case, one looks for high-confidence estimate even in presence of outliers; in the latter case, one neglects the negative effect of outliers.
- ② *High failure exposure ability*, e.g., when the tester needs to assess and improve the DNN accuracy efficiently, and the high-confidence requirements can be relaxed (e.g., in non-critical domains). The simultaneous assessment and improvement can help during subsequent re-training/fine-tuning iterations to efficiently track progress in the achieved accuracy.

**Table 10: Top 3 techniques across configurations (two factors) - (T: trade-off, nf: number of failures)**

(a) Task						
	Classification			Regression		
	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>
RMSE	GBS	SSRS	SUPS	CES	SSRS	RHC-S
RMedSE	SSRS	GBS	SRS	SSRS	SUPS	RHC-S
Failures	DeepEST	SUPS	RHC-S	SSRS	SUPS	GBS
T <sub>RMSE-nf</sub>	SUPS	RHC-S	DeepEST	SSRS	SUPS	RHC-S
T <sub>RMedSE-nf</sub>	SUPS	RHC-S	DeepEST	SSRS	SUPS	RHC-S

(b) Sample size						
	Small/Medium (50, 100, 200)			Large (400, 800)		
	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>
RMSE	SSRS	GBS	SUPS	GBS	SSRS	SRS
RMedSE	SSRS	SRS	SUPS	SSRS	GBS	SRS
Failures	DeepEST	RHC-S	SUPS	SUPS	DeepEST	2-UPS
T <sub>RMSE-nf</sub>	RHC-S	SUPS	SSRS	SUPS	RHC-S	GBS
T <sub>RMedSE-nf</sub>	RHC-S	SUPS	SSRS	SUPS	RHC-S	SSRS

(c) Dataset/model accuracy						
	Low (CIFAR10, CIFAR100)			High (MNIST, Udcity)		
	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>
RMSE	GBS	SSRS	SRS	SSRS	RHC-S	SUPS
RMedSE	SSRS	GBS	SRS	SSRS	RHC-S	SUPS
Failures	RHC-S	2-UPS	DeepEST	DeepEST	SUPS	RHC-S
T <sub>RMSE-nf</sub>	SUPS	RHC-S	GBS	SUPS	SSRS	RHC-S
T <sub>RMedSE-nf</sub>	SUPS	RHC-S	GBS/2-UPS	SUPS	SSRS	RHC-S

(d) Auxiliary variable					
	RMSE	RMedSE	Failures	T <sub>RMSE-nf</sub>	T <sub>RMedSE-nf</sub>
conf.	1 <sup>st</sup> : SSRS	SSRS	SUPS	SUPS	SUPS
	2 <sup>nd</sup> : GBS	GBS	2-UPS	RHC-S	RHC-S
	3 <sup>rd</sup> : SRS	SRS	RHC-S	2-UPS	2-UPS
LSA classification	GBS	SSRS	DeepEST	RHC-S	RHC-S
	SUPS	GBS	RHC-S	SUPS	SUPS
	RHC-S	SRS	SUPS	DeepEST	DeepEST
DSA	SUPS	SUPS	DeepEST	SUPS	SUPS
	GBS	GBS	2-UPS	RHC-S	RHC-S
	SSRS	SSRS	SUPS	DeepEST	DeepEST
LSA regression	SSRS	SSRS	SSRS	SSRS	SSRS
	RHC-S	SUPS	SUPS	RHC-S	SUPS
	SUPS	RHC-S	2-UPS	SUPS	RHC-S
VAE	CES	SRS	GBS	SSRS	SSRS
	SRS	SSRS	SSRS	CES	SUPS
	RHC-S	RHC-S	SUPS	SUPS	GBS
SAE	SRS	SSRS	GBS	CES	SSRS
	CES	RHC-S	DeepEST	SRS	RHC-S
	SSRS	CES	RHC-S	RHC-S	SUPS

- ③ *A trade-off between confidence in the accuracy estimate and number of exposed failures*, e.g., when a good confidence estimate is used to monitor the accuracy of a DNN and engineers want to use the exposed failing examples in the re-training actions (these may be triggered only when the accuracy drops under a certain threshold) [34].

As for the *context*, following our experimental design, the factors that we identified as potentially impacting are: the **task** (classification or regression), the **sample size** (hence the budget available), the **dataset**<sup>10</sup>, and the **auxiliary variable**, if available, for sampling.

Table 10 reports a two-way analysis of the ranking performance of the techniques. On the row, we list the objective. On the column,

<sup>10</sup>Datasets and models are considered together; the average accuracy of the models on the datasets capture three distinct cases of low, medium and high accuracy (Tab. 2)

**Table 11: Number of best-performing occurrences out of 270 (classification) and 60 (regression) configurations**

	Classification				Regression		
aux.	RMSE	RMedSE	Failures	aux.	RMSE	RMedSE	Failures
conf.	73	82	243	LSA	35	30	52
LSA	85	82	8	VAE	12	17	7
DSA	112	106	19	SAE	13	13	1

we break down the results by the impacting factor. For each combination (e.g. RMSE with Classification, Table 10a), we count the number of times a technique was among the top-3 ones, and report the best 3 techniques according to this count.

A practitioner should consider the combination reflecting more his/her needs and context. For instance, one might want a high-confidence robust-to-outlier assessment (row 1), with a medium (200) labelling effort (Table 10b); or (s)he might not want to use LSA or DSA, which are more expensive to compute, preferring the use of confidence (Table 10d).<sup>11</sup> Since exploring any  $n$ -way combinations could be of interest too (e.g., small RMSE *and* small sample size *and* high-accuracy dataset), we release a notebook in our replication package<sup>1</sup> to specify the factors of interest and query the results.

Besides combination-specific findings easily inferable from the Tables, some interesting patterns are hereafter highlighted:

- In high-confidence assessment ① (RMSE, RMedSE), SSRS is among the best three techniques in 22 out of 24 combinations, followed by GBS and SRS (12/24), SUPS (11/24) and RHC-S (10/24). The existing techniques CES and DeepEST are never in the top 3. Surprisingly, SRS appears often, especially for large sample size, and for low-accuracy models;
- In high failure exposure ②, SUPS stands out 10 out of 12 times, followed by DeepEST (8/12) and RHC-S (7/12);
- For good trade-offs ③ ( $T_{RMSE-nf}$ ,  $T_{RMedSE-nf}$ ), SUPS and RHC-S appear almost always (23/24 and 22/24, respectively). The others are far less common (SSRS 12/24, DeepEST 6/24).

It is worth to note that the new algorithms proposed (GBS, 2-UPS, RHC-S, SSRS, SUPS) appear among the best three in the vast majority of cases. The following specific considerations can be drawn.

SSRS is particularly good for high-confidence estimates; SUPS (and to a lesser extent RHC-S) outperforms the others for high failure exposure, where it even defeats DeepEST that is specifically conceived for that task via adaptive sampling.

SUPS and RHC-S give the best trade-offs. This indicates that they perform generally well for all the objectives.

The distinguishing feature of the new techniques is that they exploit the auxiliary variable for just partitioning (SSRS, GBS) and/or for inputs selection (RHC-S, SUPS, 2-UPS). This in essence allows to direct the sampling toward higher-variance areas of the population, reducing the estimator variance and exposing more failures.

In the perspective of a researcher devising a new technique, attention has to be paid to these aspects: auxiliary variable (if and which one to use), partitioning, and replacement scheme (Tab. 1).

<sup>11</sup>These results have to be read with the pairwise statistical test results, as the best 3 techniques could be negligibly different (in which case one can be chosen arbitrarily).

**Auxiliary variable.** The performance of auxiliary variables is useful not only for selecting a technique, but also to design new ones. The results in Table 10.d highlight that the only techniques not using auxiliary variables (SRS and CES) are rarely among the top-3 ones, especially for the failure exposure ability ( $A_f$ ).

Table 11 reports how many times each auxiliary variable yields the best RMSE and RMedSE, and the number of failures. For classification, DSA and *confidence* are the best variables for RMSE/RMedSE ① and number of failures ②, respectively. It is important to highlight that *confidence* is cheaper to collect, as it comes with the output of the classification. For regression, LSA shows the best results ①②③. The variables derived by SAE/VAE perform poorly.

**Partitioning.** Partitioning based on auxiliary variables is particularly beneficial for good accuracy estimates ①; SSRS and GBS are the best ones for this aim. The benefit of partitioning is lower when the aim is to expose failures ②③; performance is better when partitioning with LSA, especially for regression, as it is better correlated to (in)accuracy.

**Replacement.** We found no remarkable advantage of without-replacement sampling; for instance, SUPS (with replacement) works well in all scenarios. This is likely due to the negligible sample size compared to the operational dataset, hence sampling with replacement is unlikely to pick the same example twice.

## 7 THREATS TO VALIDITY

As for the selection of the experimental subjects, we have considered publicly available DNNs [31]; we have however re-trained them from scratch to have realistic accuracy and avoid the mentioned inflated accuracy issue described in [30].

The choice of the sample size affects the results. We ran a sensitivity analysis with five (from 50 to 800) values of the sample size. Different values could yield different results.

The evaluation does not include an extensive analysis of partitioning. We ran  $k$ -means, with  $k = 10$  partitions, after a preliminary tuning on 30 random samples from MNIST and  $k = 6, 8, 10, 12$ . Extending the tuning of  $k$  to all cases would improve performance.

Despite extensive code inspection, the presence of defects in the algorithms cannot be excluded.

External validity is undermined by the number of models and datasets; we considered state-of-the-art DNNs and widely-used datasets. The replicability of the experiments mitigates this threat.

## 8 CONCLUSIONS

We presented DeepSample, a framework encompassing a set of sampling-based techniques for DNN operational accuracy assessment. We implemented techniques with and without partitioning, with and without replacement, with and without auxiliary variables to drive the selection, and we empirically evaluated them in terms of accuracy estimation and number of failures, on both classification and regression problems.

The findings pertaining to the individual techniques, as well as to the key factors impacting the sampling algorithms, serve: *i*) as guidance for testers to select the technique depending on the needs and on the auxiliary information available to expedite sampling, and *ii*) for researchers to devise new techniques.

We conclude that the tester's objective and the application context are crucial in selecting a sampling technique. Techniques yielding high-confidence estimates (such as SSRS) are well suited to check the DNN against a release criterion, or for choosing among different DNNs. Techniques with high failure exposure ability (such as SUPS and DeepEST) are well suited for the simultaneous DNN accuracy assessment and improvement in iterative life cycle models. Techniques exhibiting a good trade-off between high-confidence estimates and high failure exposure (such as SUPS and RHC-S) are appropriate for cost-effective assessment and retraining.

In devising new techniques, the use of auxiliary variables and partitioning is strongly encouraged, as they have been shown to be beneficial for both accuracy estimation and failure exposure – LSA was the best choice for regression, while *confidence* (for failures exposure) and DSA (for accuracy estimation) were the best ones for classification.

## 9 DATA AVAILABILITY

All results and the artefacts for replication are available at: <https://github.com/dessertlab/DeepSample.git>.

## ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 871342 "uDEVOPS".

## REFERENCES

- [1] Z. Li, X. Ma, C. Xu, C. Cao, J. Xu, and J. Lü. Boosting Operational DNN Testing Efficiency through Conditioning. In *Proc. 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, pages 499–509. ACM, 2019.
- [2] A. Guerriero, R. Pietrantuono, and S. Russo. Operation is the Hardest Teacher: Estimating DNN Accuracy Looking for Mispredictions. In *Proceedings of the IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 348–358. IEEE, 2021.
- [3] P. A. Currit, M. Dyer, and H. D. Mills. Certifying the reliability of software. *IEEE Transactions on Software Engineering*, SE-12(1):3–11, 1986.
- [4] H. D. Mills, M. Dyer, and R. C. Linger. Cleanroom software engineering. *IEEE Software*, 4(55):19–24, 1987.
- [5] R. C. Linger and H. D. Mills. A case study in cleanroom software engineering: the IBM COBOL Structuring Facility. In *12th International Computer Software and Applications Conference (COMPSAC)*, pages 10–17. IEEE, 1988.
- [6] R. H. Cobb and H. D. Mills. Engineering software under statistical quality control. *IEEE Software*, 7(6):45–54, 1990.
- [7] J. D. Musa. Software reliability-engineered testing. *Computer*, 29(11):61–68, 1996.
- [8] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):pp. 663–685, 1952.
- [9] J. Lv, B. Yin, and K. Cai. Estimating confidence interval of software reliability with adaptive testing strategy. *Journal of Systems and Software*, 97:192–206, 2014.
- [10] K. Cai, C. Jiang, H. Hu, and C. Bai. An experimental study of adaptive testing for software reliability assessment. *Journal of Systems and Software*, 81(8):1406–1429, 2008.
- [11] K. Cai, Y. Li, and K. Liu. Optimal and adaptive testing for software reliability assessment. *Information and Software Technology*, 46(15):989–1000, 2004.
- [12] J. Lv, B. Yin, and K. Cai. On the asymptotic behavior of adaptive testing strategy for software reliability assessment. *IEEE Transactions on Software Engineering*, 40(4):396–412, 2014.
- [13] A. Podgurski, W. Masri, Y. McCleese, F.G. Wolff, and C. Yang. Estimation of software reliability by stratified sampling, 1999.
- [14] F.b.N. Omri. Weighted statistical white-box testing with proportional-optimal stratification. In *Proc. 19th International Doctoral Symposium on Components and Architecture, WCOP'14*, pages 19–24. ACM, 2014.
- [15] D. Cotroneo, R. Pietrantuono, and S. Russo. RELAI Testing: A Technique to Assess and Improve Software Reliability. *IEEE Transactions on Software Engineering*, 42(5):452–475, 2016.

- [16] R. Pietrantuono and S. Russo. Probabilistic sampling-based testing for accelerated reliability assessment. In *IEEE International Conference on Software Quality, Reliability and Security (QRS)*, pages 35–46. IEEE, 2018.
- [17] R. Pietrantuono and S. Russo. On adaptive sampling-based testing for software reliability assessment. In *27th International Symposium on Software Reliability Engineering, ISSRE*, pages 1–11. IEEE, 2016.
- [18] J. Chen, Z. Wu, Z. Wang, H. You, L. Zhang, and M. Yan. Practical accuracy estimation for efficient deep neural network testing. *ACM Trans. Softw. Eng. Methodol.*, 29(4), oct 2020.
- [19] J. Zhou, F. Li, J. Dong, H. Zhang, and D. Hao. Cost-effective testing of a deep learning model through input reduction. In *2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE)*, pages 289–300, 2020.
- [20] S. L. Lohr. *Sampling: Design and Analysis*. Duxbury Press, 2nd edition, 2009.
- [21] A. Stocco, M. Weiss, M. Calzana, and P. Tonella. Misbehaviour prediction for autonomous driving systems. In *Proc. of the IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pages 359–371. ACM, 2020.
- [22] J. Kim, R. Feldt, and S. Yoo. Guiding Deep Learning System Testing Using Surprise Adequacy. In *Proceedings of the 41st International Conference on Software Engineering (ICSE)*, pages 1039–1049. IEEE, 2019.
- [23] M. P. Wand and M. C. Jones. *Kernel smoothing*. CRC press, 1994.
- [24] H. H. Morris and N. H. William. On the Theory of Sampling from Finite Populations. *The Annals of Mathematical Statistics*, 14(4):333–362, 1943.
- [25] J.N.K. Rao, H.O. Hartley, and W.G. Cochran. On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2):482–491, 1962.
- [26] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.
- [27] Y. LeCun and C. Cortes. MNIST handwritten digit database. <http://yann.lecun.com/exdb/mnist/>, 2010.
- [28] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, 2009.
- [29] K. Pei, Y. Cao, J. Yang, and S. Jana. DeepXplore: Automated Whitebox Testing of Deep Learning Systems. *Communications of the ACM*, 62(11):137–145, 2019.
- [30] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet Classifiers Generalize to ImageNet? In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of Machine Learning Research (PMLR)*, volume 97, pages 5389–5400, 2019.
- [31] A. Guerriero, M. R. Lyu, R. Pietrantuono, and S. Russo. Assessing operational accuracy of cnn-based image classifiers using an oracle surrogate. *Intelligent Systems with Applications*, 17:200172, 2023.
- [32] R. L. Iman and J. M. Davenport. Approximations of the critical region of the fbietkan statistic. *Communications in Statistics - Theory and Methods*, 9(6):571–595, 1980.
- [33] A. Dinno. Nonparametric pairwise multiple comparisons in independent groups using dunn's test. *The Stata Journal*, 15(1):292–300, 2015.
- [34] A. Guerriero, R. Pietrantuono, and S. Russo. Iterative assessment and improvement of dnn operational accuracy. In *2023 IEEE/ACM 45th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*, pages 43–48. IEEE, 2023.