

Benchmarking Generative AI Models for Deep Learning Test Input Generation

Maryam

University of Udine
Udine, Italy
maryam@spes.uniud.it

Matteo Biagiola

Università della Svizzera italiana
Lugano, Switzerland
matteo.biagiola@usi.ch

Andrea Stocco

Technical University of Munich, fortiss
Munich, Germany
andrea.stocco@tum.de

Vincenzo Riccio

University of Udine
Udine, Italy
vincenzo.riccio@uniud.it

Abstract—Test Input Generators (TIGs) are crucial to assess the ability of Deep Learning (DL) image classifiers to provide correct predictions for inputs beyond their training and test sets. Recent advancements in Generative AI (GenAI) models have made them a powerful tool for creating and manipulating synthetic images, although these advancements also imply increased complexity and resource demands for training.

In this work, we benchmark and combine different GenAI models with TIGs, assessing their effectiveness, efficiency, and quality of the generated test images, in terms of domain validity and label preservation. We conduct an empirical study involving three different GenAI architectures (VAEs, GANs, Diffusion Models), five classification tasks of increasing complexity, and 364 human evaluations. Our results show that simpler architectures, such as VAEs, are sufficient for less complex datasets like MNIST. However, when dealing with feature-rich datasets, such as ImageNet, more sophisticated architectures like Diffusion Models achieve superior performance by generating a higher number of valid, misclassification-inducing inputs.

Index Terms—Software Testing, Generative AI, Deep Learning

I. INTRODUCTION

Deep Learning (DL) has reshaped several fields, including image processing, where DL image classifiers often outperform traditional vision methods and even human experts in accuracy and efficiency [1]. This advancement has enabled greater automation, especially in life- and safety-critical areas such as healthcare and autonomous driving [2]–[5].

Ensuring the quality of DL image classifiers remains crucial, as it is difficult to assess their ability to generalize to unseen data. In fact, their training and test sets may not fully capture the range of real-world scenarios they will encounter after deployment [6], [7]. A significant challenge for software testers is generating test images that accurately reflect real-world operating conditions and trigger *misclassifications*, i.e., unexpected behaviors where predicted labels deviate from the expected ones. Thus, researchers have proposed test input generators (TIGs), aimed at automatically generating synthetic images to assess the quality of DL classifiers [8]–[11].

Most TIGs apply small perturbations to inputs to maintain their expected label, addressing the oracle problem [12]–[14],

This work was supported by the project SOP CUP N.H73C22000890001, PNRR M4 C2 I1.3 “SEcurity and RIghts in the Cyberspace (SERICS)” PE0000014 PE7, funded by Next-Generation EU, and the Bavarian Ministry of Economic Affairs, Regional Development and Energy.

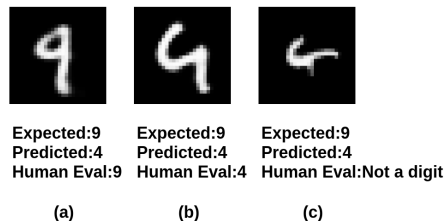


Fig. 1. Misclassification-inducing test images for handwritten digit classifiers: (a) valid and label-preserving, (b) valid but not label-preserving, (c) invalid.

as the correct output is unknown for synthetic data. However, such perturbations can lead to invalid inputs (i.e., outside the classification task’s domain) or fail to preserve the label of the original image (i.e., no longer belong to the same class) [11], [15]. Existing TIGs often overlook input validity and assume that ground-truth labels remain preserved [11]. Figure 1 shows three misclassified inputs for an handwritten digit classifier: input (a) is valid and the expected label (i.e., 9) matches with the human assessment; (b) is valid but fails to preserve the expected label; and (c) is invalid, because it does not belong to the domain of handwritten digits.

The first propositions of TIGs primarily focused on manipulating raw inputs (i.e., pixel perturbations [16]–[19]) or parametrized semantic representations (e.g., control points in vector graphics [7], [20], [21] or parameters in simulators [22], [23]). However, these approaches are restricted to modifying initial images with known ground-truth labels, limiting exploration to regions near the original inputs and leaving significant portions of the input space untested.

Researchers have recently started leveraging the creativity of distribution-aware Generative AI (GenAI) models [15], [24]–[27], which learn the input data distribution in the form of a *latent space*, i.e., a compressed low-dimensional representation capturing the key features of the problem domain [28]. GenAI-based TIGs manipulate inputs in the latent space, where small changes can yield significant image variations (e.g., style, pose, color), before converting latent representations back into pixel space. In this way, GenAI models can generate unique images, blending features from learned patterns in ways not present in the original training data.

Existing studies have demonstrated the effectiveness of GenAI-based TIGs [11], [24], [25], [29]. However, they are

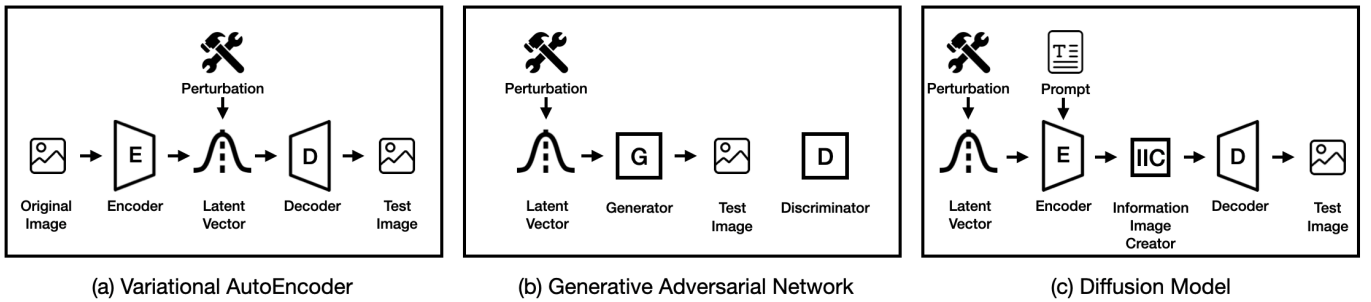


Fig. 2. Summary of the GenAI models considered in this paper and the process by which TIGs perturb their latent vectors.

often influenced by confounding factors such as differences in testing algorithms and the absence of standardized training, experimental setups, and hyperparameter tuning. Moreover, they have not yet incorporated recent innovations like diffusion models, making it difficult to fairly compare the test generation capabilities across different GenAI models.

In our study, we fill this gap by providing the most comprehensive cross-evaluation of GenAI TIGs, benchmarking three key GenAI models, i.e., VAEs, GANs, and diffusion models, across popular datasets of increasing complexity. We provide a standardized experimental framework, including search-based optimization for latent space manipulation, and systematically evaluate these models in terms of their ability to generate valid, label-preserving, and misclassification-inducing inputs, considering efficiency and effectiveness. Our results show that GenAI TIGs can produce valid, label-preserving inputs, although the degree of success varies. Simpler models, such as VAEs and GANs, perform well on less complex datasets like MNIST and SVHN. Diffusion models excel in complex tasks like CIFAR-10 and ImageNet, generating up to 80% more valid label-preserving inputs. Our key contributions are:

- A large-scale empirical study comparing three GenAI architectures (VAEs, GANs, and diffusion models) across five classification tasks, along with a validity assessment involving 364 human assessors.
- A search-based test generation framework that integrates different GenAI models, enabling automated testing of DL classifiers through latent space manipulation.

To encourage open research, our test generation framework and experimental data are available [30].

II. BACKGROUND

In the following, we describe the three main GenAI architectures considered in this paper (Figure 2).

A. Variational Autoencoders (VAEs)

VAEs, introduced by Kingma et al. [31], consist of two neural networks: an encoder and a decoder (Figure 2 (a)). The encoder learns to map images to lower-dimensional representations (i.e., latent space vectors sampled from a predefined distribution). The decoder learns to reconstruct images from the latent vectors. This architecture extends standard autoencoders

by adopting a probabilistic approach to model complex data distributions. The two networks are trained jointly to minimize a loss function that balances two objectives: the reconstruction error (i.e., difference between original inputs and VAE’s reconstructions), and the Kullback-Leibler divergence [32], which regularizes the learned latent space distribution to align with a prior distribution, e.g., a standard normal distribution. The sampling step during training introduces randomness, enabling VAEs to learn smooth and continuous latent spaces.

In software testing, Kang et al. [24], [29] and Dola et al. [15], [27] used VAEs for generating edge cases to assess the robustness of image classifiers. In particular, their TIGs encode existing images to increase control over generation and perturb the latent vectors returned by the encoder.

B. Generative Adversarial Networks (GANs)

GANs introduced by GoodFellow et al. [28], consist of two neural networks, i.e., a generator and a discriminator, trained simultaneously (Figure 2 (b)). These networks are involved in competitive learning to improve one another performance: the discriminator learns to distinguish real from artificially generated images, while the generator learns to create images from latent vectors sampled from a known distribution that can *fool* the discriminator. While VAEs explicitly model the data distribution and enforce smoothness in the latent space through their probabilistic approach, GANs’ generators implicitly learn to approximate the data distribution by continuously improving their ability to deceive discriminators. This often results in more realistic images than those generated by VAEs, although GANs are more challenging to train due to issues like instability and mode collapse, where the generator learns only a limited subset of features, leading to reduced diversity in the generated images [28]. GANs are useful for testing, as they can generate realistic artificial images that resemble the training distribution and are difficult to distinguish from real data. Moreover, GANs allow test manipulation by perturbing latent space vectors (i.e., input to the GAN generator). In software testing, Dunn et al. [25] adopt conditional Deep Convolutional GANs (cDCGANs), which leverage convolutional layers for improved image generation quality, and incorporate additional information (called *conditions*) such as class labels to better guide the generation of images that meet predefined criteria.

C. Diffusion Models (DMs)

DMs were first introduced by Sohl-Dickstein et al. [33]. The core idea behind these models is to simulate a diffusion process, where information in an image is gradually diffused (i.e., noise is incrementally added) and then progressively recovered through a reverse process, known as denoising (Figure 2 (c)). During training, diffusion models add noise to the data over a sequence of steps, gradually transforming an original vector into a noisy version that resembles a simple Gaussian distribution. The model is then trained to reverse this process and recover the original data, learning how to transform noise back into a complex data distribution, i.e., natural images [34]. This approach enables the generation of highly realistic images, offering significant advantages over other generative models, such as GANs, particularly in terms of training stability and resistance to mode collapse [33].

In text-to-image scenarios, we considered models like Stable Diffusion, which leverages a textual description of the desired input (called prompt) to generate corresponding images. Its architecture consists of three main components: the text encoder, the image information creator, and the image decoder. A noisy latent vector is obtained by combining a noise vector with the latent vector obtained by encoding the prompt. The information image creator iteratively refines this noisy latent vector, guided by the prompt. The final latent vector is then transformed into a high-resolution image by the decoder. This work is the first to integrate Stable Diffusion into a TIG for image classifiers, by manipulating its input latent vector.

III. EVALUATION FRAMEWORK

In this section, we describe the evaluation framework we propose for comparing different GenAI architectures for the generation of misclassification-inducing inputs for DL based image classifiers. Our framework is designed to (1) smoothly integrate different GenAI models, (2) enable the generation and controlled manipulation of novel inputs through the corresponding latent vectors, (3) guide input manipulation towards inducing misclassifications, and (4) return the misclassification-inducing inputs along with statistics on the efficiency of their generation. To explore the latent space produced by the GenAI methods, we adopt search-based optimization using a population-based genetic algorithm, due to its success in testing DL based systems [20], [22], [24], [35]–[37]. Specifically, we define a fitness function that assigns lower values to inputs more likely to induce misclassifications, which the search process seeks to minimize.

Algorithm 1 outlines the steps of our framework. It takes as input the classifier under test C , the latent vector used by the GenAI model to produce an image (i.e., the *seed*), the corresponding expected label (i.e., the class the image should belong to), and the configuration parameters for the test generation process. As output, it returns the first misclassification-inducing input it identifies, along with the number of iterations required to generate it.

The algorithm starts by initializing the perturbation step, which is used to modify the latent vectors of the GenAI

Algorithm 1: Test Generation with GenAI Models

Data: s : seed latent vector, $expLabel$ expected label, G : GenDL model, C : classifier under test, $popSize$: population size, $tshdBest$: threshold for selection operator, N number of iterations, δ_{init} : initial perturbation step, $minBound$, $maxBound$: minimum and maximum values observed during the training of G .

Result: img : misclassification-inducing image, $iter$: # of iterations needed to generate a test input

```

1 perturbation step  $\delta \leftarrow \delta_{init}$ 
2 image  $img \leftarrow G(s)$ 
3 label  $l$ , fitness  $f_{prev} \leftarrow \text{EVALUATE}(img, C, expLabel)$ 
4 if ( $l \neq expLabel$ ) then
5   return  $\emptyset, 0$  /* Return if the original input is misclassified */
6 for  $i = 1 \rightarrow popSize$  do
7   population individual  $P_i \leftarrow \text{MUTATION}(s, \delta)$ 
8 iter  $\leftarrow 0$ 
9 while iter <  $N$  do
10  image population  $I_P \leftarrow G(P)$ 
11  fitness values  $F \leftarrow \text{EVALUATE}(I_P, C, expLabel)$ 
12  best individuals  $P' \leftarrow \text{SELECT}(P, F, tshdBest)$ 
13  best fitness  $f_{min} \leftarrow \min(F)$ 
14  if ( $f_{min} < 0$ ) then
15     $img \leftarrow \text{GETBESTINDIVIDUAL}(P')$ 
16    return  $img, iter$ 
17
18  else if ( $f_{min} == f_{prev}$ ) then
19     $\delta \leftarrow 2 \cdot \delta$ 
20    /* Adaptively increase the mutation extent */
21
22  else
23     $\delta \leftarrow \delta_{init}$ 
24    /* reset to initial mutation extent when fitness improves */
25     $f_{prev} \leftarrow f_{min}$ 
26
27  offspring  $O \leftarrow \emptyset$ 
28  for  $j \leftarrow 1$  to ( $popSize - \text{length}(P')$ ) do
29    parents  $p_{j1}, p_{j2} \leftarrow \text{RANDOMCHOICE}(P', 2)$ 
30    offspring  $o \leftarrow \text{CROSSOVER}(p_{j1}, p_{j2})$ 
31     $o \leftarrow \text{MUTATION}(o, \delta)$ 
32     $o \leftarrow \text{CLAMP}(o, minBound, maxBound)$ 
33     $O \leftarrow O \cup \{o\}$ 
34
35   $P \leftarrow P' \cup O$ 
36  iter  $\leftarrow iter + 1$ 
37
38 return  $\emptyset, iter$ 

```

model G (Line 1). The first image is generated from the seed latent vector, which provides the necessary information for G to produce an image belonging to the target class (Line 2). This first image is evaluated on the classifier under test, determining the predicted label and its fitness score (Line 3). The test generation process terminates if the initial image is misclassified, i.e., it does not belong to the expected class according to the classifier (Lines 4-5). In fact, the goal of a TIG is to find slight perturbations that transform an image predicted as expected into another one predicted differently, as these can highlight weaknesses of the classifier [20]. A population of size $popSize$ is created through the mutation genetic operator, which applies random perturbations to the initial latent vector, thus increasing diversity within the population.

The main loop is executed up to N times (Lines 9–31), i.e., it stops if a misclassification is found or if the iteration budget is exhausted. At each iteration, G produces images corresponding to the latent vectors in the population P , which

are evaluated by the classifier (Lines 10-11). The selection operator chooses the most promising inputs, i.e., those with lower fitness (Line 12). If an individual exhibits negative fitness (indicating a misclassification), the corresponding image is returned as output and the algorithm terminates (Lines 14–16). Otherwise, the selected individuals are modified using the genetic operators.

The mutation operator is adaptive, i.e., the perturbation step increases if no improvement in fitness is observed during the previous iteration (Lines 18–21). Offsprings are obtained by applying the mutation and crossover operators (Lines 23–29). To prevent the generation of unrealistic images, latent vectors are clamped within the bounds *minBound* and *maxBound* observed during the GenAI model training, as suggested by Dunn et al. [25]. The offsprings and the best individuals from the current iteration form the population for the next iteration (Line 30). In the following, we describe in detail each main component of the search algorithm.

A. Input Representation and Initialization

To represent the input domain, our framework leverages a lower-dimensional latent space synthesized by the GenAI models. The latent space captures underlying patterns of the input data in a more abstract and compressed form compared to the original high-dimensional input space. It consists of a multivariate probability distribution characterized by parameters such as the means and variances of a predefined z -dimensional distribution chosen during training.

Our approach represents test inputs as latent vectors, which are points sampled within the latent space, whose dimensionality is defined by the specific GenAI model adopted by the TIG. Exploring the latent space is more efficient than searching directly in high-dimensional pixel space due to its reduced dimensionality and mapping to a known distribution. Our TIG leverages an initial seed latent vector corresponding to an input image whose label is known and correctly predicted by the classifier under test. From this seed, our TIG produces a population of slight variations, enabling exploration of its neighborhood and introducing diversity into the search process. The starting seed depends on each specific GenAI model.

VAEs use the latent vector produced by the encoder when processing an image from the original test set of the considered dataset, whose ground-truth label is known. This approach, used in the literature by Kang et al. [24] and Dola et al. [27], ensures higher similarity to the original input and, thus, label preservation and better control over the exploration.

As for GANs, the initial seed is a vector sampled from the target distribution used during training. Specifically, we use *conditional* GANs, which accept as input both the latent vector and the target label. This conditioning, also employed by Dunn et al. [25], improves control over image generation.

Also for DMs, the seed is a vector sampled from the training distribution. This GenAI model controls the generation by accepting as input also a textual prompt, which describes both the input domain and the target class. Although no current TIG employs this approach, latent space exploration

for DMs is a well-established practice in the field [38]. Stable Diffusion *guidance scale* parameter significantly influences the image generation process. This parameter determines to what extent the generated image adheres to the prompt. Higher values make the model focus more on the prompt, resulting in images closely aligned with the input text, but may reduce diversity and occasionally lead to lower image quality. Lower values introduce more diversity and creativity, with a looser connection to the prompt. During seed generation, we adopt a relatively higher value (3.5) for the DMs guidance scale parameter compared to latent vector mutation (1.4). This choice, motivated by preliminary experiments, promotes label adherence during mutation and diversity in seed generation, as recommended by the authors’ guidelines [39], [40].

B. Fitness Function

The EVALUATE function calculates the fitness of each individual in the population, based on how the classifier under test predicts the image generated from the corresponding latent vector. In this context, an individual is considered fit if it is likely to cause a misclassification, i.e., the corresponding image should belong to one class but is classified as a different class. To this aim, we leverage the activation levels from the output softmax layer of the image classifier. It provides a surrogate confidence score for each possible class [41] as the softmax output can be interpreted as a probability distribution over the classes, with the highest value indicating the predicted class for a given input. Similar fitness functions have been employed by several existing approaches [16], [17], [20], [25].

Specifically, the fitness value is computed as the difference between the confidence score of the expected class and the highest confidence score for any other class. The formula of the fitness for the input image x is defined as follows:

$$fitness(x) = \sigma_{\text{expected}}(x) - \max_{i \neq \text{expected}} \sigma_i(x) \quad (1)$$

where $\sigma_{\text{expected}}(x)$ represents the softmax output for the expected class, and $\max_{i \neq \text{expected}} \sigma_i(x)$ is the highest softmax output for any other class. The fitness function approaches zero when the confidence scores of the expected class and the second-highest class are similar. A negative value indicates a misclassification, as the highest confidence score is assigned to a class different from the expected one. Thus, this fitness function is designed to be minimized by the test generator to promote inputs that are closer to induce misclassifications.

C. Genetic Operators

Our SELECT operator chooses the most promising individuals for the next generation, i.e., those with the lowest fitness scores. The number of individuals chosen is controlled by the parameter *tshdBEST*, which can be tuned to balance exploration and exploitation. A lower *tshdBEST* value intensifies exploitation by focusing on fewer, high-performing individuals, potentially speeding up failure detection but with the risk of premature convergence to local optima; a higher value,

instead, retains more individuals, promoting greater diversity in the population and allowing for a broader exploration.

Representing inputs as latent vectors enables uniform MUTATION and CROSSOVER operators across all GenAI models. We adopt a single-point CROSSOVER operator that combines two individuals randomly chosen among the ones selected by the selection operator. This operator divides the latent vectors of the parents at a random point, taking the first portion from one parent and the second portion from the other, generating an offspring that inherits characteristics from both. This promotes exploration of new latent space regions by combining information from the most successful individuals.

The MUTATION operator introduces random perturbations to the offspring, encouraging diversity and preventing stagnation in local optima. These perturbations are obtained by multiplying random noise sampled from a normal distribution by the perturbation step, as described in the following formula:

$$\mathbf{z}_{mut} = \mathbf{z}_{orig} + \epsilon \cdot \delta, \quad \epsilon \sim \mathcal{N}(0, 1)^d \quad (2)$$

where \mathbf{z}_{orig} is the initial latent vector of dimensionality d , ϵ is the noise vector of the same dimensionality where each component is sampled from a standard normal distribution, δ is the perturbation step, and \mathbf{z}_{mut} is the perturbed latent vector.

The perturbation step δ is adjusted adaptively based on the fitness of the previous iteration, i.e., if the previous iteration did not improve the best fitness score, then the mutation extent is increased to escape local optima by allowing larger changes to the latent vectors. Conversely, if improvement is observed, the mutation rate is decreased to its default value δ_{init} , in order to exploit promising areas of the solution space.

IV. EMPIRICAL SETUP

Our study compares different GenAI models for test input generation. We integrate each of them into our framework (Section III) to create *GenAI TIGs*. We assessed their effectiveness and efficiency in generating valid, label-preserving inputs that cause misclassifications in the classifier under test.

A. Research Questions

RQ₁ (Seed Generation): Which GenAI model generates more correctly classified seed inputs within the same budget?

Effective test generation begins with the identification of reliable seeds, i.e., inputs that produce images that are predicted as expected by the classifier under test [20]. In fact, only these seeds can be used by TIGs to identify input variations that trigger misclassifications.

Metric: For each GenAI model, we compute the ratio of seeds assigned the correct label by the classifier, compared to the total number of generated seeds.

RQ₂ (Effectiveness): Which GenAI TIG generates more misclassification-inducing inputs?

A standard approach to evaluate and compare the effectiveness of TIGs is to count the number of failures triggered within a given budget [8]. A TIG that triggers more misclassifications could potentially expose more weaknesses or bugs in the classifier under test.

TABLE I
DATASETS’ INPUT SIZE AND CLASSIFIERS’ TRAINING ACCURACY.

Dataset	Image Size	Classifier	Accuracy (%)
MNIST	28x28x1	deepconv [24]	99.46
SVHN	32x32x3	VGGNET [24]	95.20
CIFAR-10	32x32x3	VGGNET [24]	86.38
ImageNet	224x224x3	VGG19bn [43]	75.00

Metric: For each TIG, we calculate the number of misclassification-inducing inputs and their ratio over the total number of generated inputs.

RQ₃ (Efficiency): How efficient are GenAI TIGs in triggering misclassifications?

The goal of this research question is to assess the efficiency of each GenAI TIG in triggering misclassifications during the iterative process of our testing framework. Efficiency, in this context, refers to the model’s ability to generate misclassification-inducing inputs with fewer iterations, providing insight into how quickly each TIG can expose weaknesses in the classifier under test.

Metric: We measure the average number of iterations required to trigger a misclassification across all seeds. For seeds where the TIG does not trigger any misclassification, we report the maximum number of iterations, i.e., the search budget.

RQ₄ (Validity): Which GenAI TIG generates more valid, misclassification-inducing inputs, according to humans?

TIGs offer a reliable assessment of classifiers’ quality only when misclassification-inducing inputs are valid, i.e., recognisable by humans as part of the input domain [11], [42]. Invalid inputs may not be worth further analysis by testers, as they refer to images that cannot be observed in the real world and, thus, do not provide meaningful insights into the defects of the DL model under test. In this work, we evaluated each generated input with multiple independent human assessors.

Metric: We calculate the number and ratio of valid inputs (as assessed by human evaluators) over the total number of misclassification-inducing inputs.

RQ₅ (Label Preservation) To what extent the valid misclassification-inducing inputs generated by GenAI TIGs preserve the seed’s label?

A misclassification is detected when the predicted label differs from the expected one, i.e., the seed’s label. For this reason, it is crucial to evaluate whether the images generated by TIGs preserve the expected label.

Metric: For each TIG, we measure the number and ratio of valid, misclassification-inducing, and label-preserving inputs over the total number of generated, valid and misclassification-inducing inputs.

B. Datasets and Image Classifiers

We consider four widely-used image datasets of increasing complexity: MNIST [44], SVHN [45], CIFAR-10 [46], and ImageNet [47]. These datasets cover classification tasks ranging from recognizing 10 classes in small greyscale images

to identifying one of 1,000 classes from large-scale, colored images representing real-world objects. In the following, we describe the considered datasets and classifiers (Table I).

MNIST. A dataset of 70,000 grayscale images of handwritten digits, each with a resolution of 28x28 pixels and pixel values ranging from 0 to 255. The labels correspond to digits from 0 to 9. Due to its simplicity, MNIST is widely used in DL frameworks’ tutorials [48] and research papers [9]. The classifier under test is the convolutional DNN used in the TIG proposed by Kang et al. [24], which includes four convolutional layers, two pooling layers, and two fully connected layers.

SVHN. This dataset contains 600,000 images of house numbers, representing digits from 0 to 9. Unlike MNIST, this dataset poses a more difficult challenge, with larger, colored images (32x32), where the target digit may be surrounded by neighboring digits, adding complexity to the recognition task. The classifier under test is the VGG architecture used by Kang et al. [24], which employs five convolutional blocks.

CIFAR-10. This dataset consists of 60,000 color images, each 32x32 pixels, spanning 10 different classes representing animals and vehicles. This dataset is more complex than SVHN because it involves a wider variety of classes with greater diversity and complexity in terms of backgrounds, textures, lighting, and object orientations. For this dataset, we used the same DL architecture used for SVHN [24].

ImageNet-1k. dataset consists of over 14 million images spanning 1,000 classes. It is widely recognized for its role in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) since 2010, with the 2012 version being a benchmark standard for image classification tasks. Compared to the other three datasets, ImageNet-1k includes high-resolution images and a significantly broader range of categories. We tested the VGG19bn pretrained DL model, available through the PyTorch library [43], consisting of 16 convolutional layers and three fully connected layers.

C. GenAI Models Setup

We compared the GenAI models introduced in Section II. To ensure fairness, each model was trained on the same training set originally used to train the DL classifier under test. The latent space of these GenAI models was leveraged in combination with the TIG described in Section III to generate test images. Table II provides an overview of each GenAI model, including a link to its architecture and details about its training and inference process, offering insights into their respective computational costs and complexities. The models were trained until convergence, on a machine featuring an NVIDIA GeForce RTX 3090 GPU (24 GB VRAM) and CUDA 12.4 for GPU acceleration.

VAE. We trained unconditional VAEs for all the considered datasets. In particular, we trained increasingly more complex architectures, aligning with the difficulty of the classification task. They span from a basic architecture with a single fully connected hidden layer in both the encoder and decoder for MNIST [49], to more sophisticated architectures, incorpo-

TABLE II
CHARACTERISTICS OF THE GENAI MODELS: LATENT VECTOR SIZE, TRAINING TIME UNTIL CONVERGENCE, AVERAGE INFERENCE TIME.

Dataset	Model	LV size	t_{train} (min)	t_{infer} (ms)
MNIST	VAE [49]	400	6	0.27
	GAN [50], [51]	100	9	0.7
	DM [52]	16384	405	960.68
SVHN	VAE [53]	800	93	4.07
	GAN [50], [51]	100	86	1.75
	DM [52]	16384	572	1213.49
CIFAR-10	VAE [53]	1024	423	2.51
	GAN [50], [51]	100	450	1.73
	DM [52]	16384	362	1903.29
ImageNet	VAE [54]	512	2521	11.92
	GAN [55]	128	21600	20.68
	DM [52]	16384	30	1945.77

rating multiple fully connected and convolutional layers to effectively capture the features of ImageNet data [54].

GAN. For MNIST, SVHN and CIFAR-10, we adopted a flexible conditional deep convolutional GAN from the literature [56] and trained it on each dataset separately. As dataset complexity increases, it becomes more challenging for the generator to produce realistic images. For this reason, we adopted for ImageNet the more sophisticated conditional BigGAN architecture [55] used by Dunn et al. [25].

DM. Due to the high complexity and computational demands of the Stable Diffusion model, it was not feasible to train it from scratch on each dataset. For this reason, we fine-tuned the pretrained, robust Stable Diffusion model v1.5 [57] provided by HuggingFace for five epochs on each dataset. We employed the Low-Rank Adaptation (LoRA) technique [58], which reduces memory consumption and accelerates fine-tuning of large models. For each dataset, we defined specific textual prompts to guide the model in learning the concept of each class. Stable Diffusion was fine-tuned on the entire MNIST, SVHN, and CIFAR-10 datasets. Due to the large size and number of classes of ImageNet, we focused on fine-tuning two specific classes separately, i.e., *teddy bear* and *pizza*.

D. Experimental Procedure

We conducted a comprehensive comparison of the considered GenAI models, i.e., VAEs, GANs, and DMs, within our TIG framework. Each model was trained on five tasks across four datasets, resulting in a total of 15 distinct test generators.

For each GenAI model, we generated 100 starting seeds, according to the specific initialization for each model, as described in Section III-A. Each TIG was allocated a budget of 250 iterations with consistent genetic algorithm parameters, i.e., a population size $popSize$ of 25 and a selection threshold $tshdBest$ of 10, maintaining the same ratio as related work [24]. Figure 3 shows examples of misclassification-inducing images generated by our GenAI TIGs. We experimented with two initial perturbation steps δ_{init} that we chose as partitions of latent vector ranges of each GenAI model. After generating 1,000 seeds per model, we computed the range based on the

maximum and minimum values of the latent vectors. In this way, we accounted for each GenAI model’s unique latent space structure. The lower perturbation step ($\delta_{init} = Low$) was obtained by dividing the range by 10^4 , while the higher ($\delta_{init} = High$) by 10^3 . This allowed us to compare the impact of fine-grained vs more substantial perturbations on test generation effectiveness and efficiency.

The validity and label preservation of misclassification-inducing inputs were assessed by independent human evaluators. Automated distribution-based validators [15], [59] were not considered, as they are limited to checking whether inputs are in the same distribution of their training data and overlook the problem of label preservation, as shown by Riccio and Tonella [11]. Instead, humans can provide more accurate judgments on whether the generated images semantically belong to the intended input domain. We used the Amazon Mechanical Turk [60] crowdsourcing platform, which grants access to a diverse and independent group of assessors [61]. This platform is well-suited for qualitative feedback collection, as widely demonstrated in previous studies [11], [20], [62], [63]. Assessors were compensated appropriately [64].

For each image, we asked the assessors to identify which class was represented within the problem domain or if the image did not belong to that domain (i.e., it was invalid). For MNIST, SVHN, and CIFAR-10, the assessors selected from the 10 possible classes or the invalid option (“Not a handwritten digit/house number/real-world object”). Since ImageNet-1K has 1,000 classes, the user could choose between the expected class, the eight most commonly predicted classes by our classifier under test for the corresponding tasks, or the “Another real-world object” and “No real-world objects” options. We implemented two quality control measures to assess the reliability of the responses provided by human assessors. First, we added an Attention Check Question (ACQ) to each survey. Second, we restricted the participation to workers with high reputation, i.e., above 95% approval rate [65]. The ACQ consisted of an image for which the human choice is obvious, and only users passing this check were included in the results.

Surveys consisted of multiple questions for each classification problem, ensuring that each image was shown only once across the surveys. Each survey was answered by 2 assessors, so that each input was assessed by 2 human evaluators, with a total of 182 surveys and 364 human participants. We counted the number of images where both assessors agreed on the validity (i.e., assigned a class within the problem domain) or invalidity. Disagreements were excluded from the analysis. We then compared metrics for each pair of GenAI TIGs within the same classification task and perturbation step.

To determine statistical significance, we performed Fisher’s exact test [66] for binary variables. Since the average number of iterations is a continuous variable, we adopted the Mann-Whitney U-Test [67] when assessing efficiency, measuring the magnitude of the differences with the Cohen’s D [68]. We threshold the p -value to be lower than 0.05, combined with a non-negligible effect size or odds ratio, to assess a statistically significant difference between the compared GenAI TIGs.

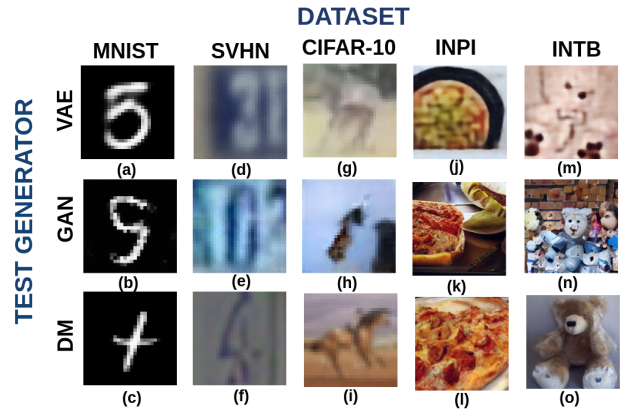


Fig. 3. Misclassification-inducing images generated by GenAI TIGs

V. RESULTS

A. RQ_1 (Seed Generation)

The fourth column of Table III reports the number of seeds generated by each GenAI model that were correctly predicted by the classifiers under test. The values are consistent across both low and high perturbation steps (“Low” and “High”), as we used the same initial seeds for each GenAI TIG.

For MNIST, both VAE and GAN produced a significantly higher percentage of correctly classified seeds compared to DM. This may be due to the fact that the original DM was trained on colored images, and during fine-tuning, it had to adapt to greyscale inputs.

For SVHN and CIFAR-10, however, DMs significantly outperformed VAEs and GANs, demonstrating better adaptability to the complexity of these datasets.

In the ImageNet tasks, GANs performed best for the *pizza* class and similarly to DMs for the *teddy bear* class. This improvement in GAN performance is probably due to the usage of the larger BigGAN architecture [55], which is better suited for high-complexity datasets.

RQ_1 (Seed Generation): VAEs’ performance declined progressively as dataset complexity increased, resulting in only 14 correctly predicted seeds for ImageNet. In contrast, both GANs and DMs maintained consistently high accuracy in seed generation, with more than 66% of seeds correctly predicted for all datasets.

B. RQ_2 (Effectiveness)

The fifth column of Table III presents the percentages and quantities of misclassification-inducing inputs, generated by GenAI TIGs, starting from the correctly predicted seeds (RQ_1).

For low perturbation steps, DMs significantly outperformed or were comparable to other models (e.g., for the *pizza* class), triggering up to $5\times$ more failures for MNIST. Across all models, performance improved as dataset complexity increased. This is likely due to the growing difficulty of the tasks and the corresponding decline in classifier accuracy (Table I).

TABLE III
COMPARISON BETWEEN GENAI TIGS ACROSS DIFFERENT DATASETS AND MUTATION EXTENTS IN TERMS OF VIABLE SEEDS, MISCLASSIFICATION-INDUCING INPUTS, NUMBER OF ITERATIONS TO GENERATE FAILURE, INPUT VALIDITY, AND LABEL PRESERVATION. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, WHILE THE UNDERLINED VALUES ARE NOT STATISTICALLY DIFFERENT FROM THE BEST.

Dataset	Pert. Step (δ_{init})	Model	% Seeds	% Misclass. (#)	# Iterations	% Validity (#)	% Preserved (#)
MNIST	Low	VAE	99	4.04 (4)	245.41	50.00 (2)	100.00 (2)
		GAN	99	8.08 (8)	242.05	75.00 (6)	83.33 (5)
		DM	87	50.57 (44)	164.61	45.45 (20)	30.00 (6)
	High	VAE	99	100.00 (99)	62.92	73.74 (73)	<u>49.32</u> (36)
		GAN	99	96.97 (96)	107.46	<u>69.79</u> (67)	62.69 (42)
		DM	87	100.00 (87)	26.77	40.23 (35)	34.29 (12)
SVHN	Low	VAE	66	50.00 (33)	178.20	51.52 (17)	41.18 (7)
		GAN	84	42.86 (36)	182.22	30.56 (11)	<u>45.45</u> (5)
		DM	95	69.47 (66)	131.04	39.39 (26)	57.69 (15)
	High	VAE	66	100.00 (66)	27.00	39.39 (26)	30.77 (8)
		GAN	84	<u>98.81</u> (83)	39.00	36.14 (30)	50.00 (15)
		DM	95	100.00 (95)	13.23	23.16 (22)	18.18 (4)
CIFAR-10	Low	VAE	39	<u>82.05</u> (32)	118.90	<u>53.13</u> (17)	29.41 (5)
		GAN	69	<u>66.67</u> (46)	140.32	45.65 (21)	19.05 (4)
		DM	87	89.66 (78)	85.63	60.26 (47)	61.70 (29)
	High	VAE	39	100.00 (39)	<u>19.51</u>	30.77 (12)	33.33 (4)
		GAN	69	100.00 (69)	<u>25.78</u>	31.88 (22)	22.73 (5)
		DM	87	100.00 (87)	14.18	62.07 (54)	68.52 (37)
ImageNet (Teddy Bear)	Low	VAE	14	100.00 (14)	13.57	78.57 (11)	81.82 (9)
		GAN	<u>85</u>	100.00 (85)	98.27	74.12 (63)	36.51 (23)
		DM	87	<u>98.85</u> (86)	48.45	91.86 (79)	49.37 (39)
	High	VAE	14	100.00 (14)	1.36	100.00 (14)	64.29 (9)
		GAN	85	100.00 (85)	26.38	83.53 (71)	32.39 (23)
		DM	87	100.00 (87)	6.63	<u>94.25</u> (82)	<u>56.10</u> (46)
ImageNet (Pizza)	Low	VAE	25	100.00 (25)	12.96	92.00 (23)	<u>91.30</u> (21)
		GAN	99	88.00 (87)	172.88	88.51 (77)	46.75 (36)
		DM	73	<u>97.26</u> (71)	83.60	98.59 (70)	92.86 (65)
	High	VAE	25	100.00 (25)	2.60	80.00 (20)	<u>75.00</u> (15)
		GAN	99	100.00 (99)	47.93	86.87 (86)	51.16 (44)
		DM	73	100.00 (73)	12.53	100.00 (73)	86.30 (63)

For high perturbation steps, all TIGs achieved high misclassification rates, with the VAE performing worst at 96.97%. This indicates that larger perturbation steps are effective in triggering misclassifications, even within a constrained budget.

RQ₂ (Effectiveness): DMs emerged as the most effective in generating misclassification-inducing inputs, especially at low perturbation steps, outperforming VAEs and GANs TIGs in most cases. Under high perturbation steps, all models showed considerable success in triggering misclassifications, indicating that increased perturbations uniformly promote misbehaviors across different architectures and datasets.

C. RQ₃ (Efficiency)

The sixth column of Table III reports the average number of iterations required by each TIG to trigger misclassifications.

For both low and high perturbation steps, DMs required significantly fewer iterations than other GenAI TIGs for MNIST, SVHN, and CIFAR-10. However, for ImageNet tasks, VAEs caused misclassifications in the fewest iterations. This

result reveals an interesting trade-off with VAEs: although they perform poorly in terms of the number of usable seeds and misclassification-inducing inputs (RQ₁ and RQ₂), they compensate by requiring fewer iterations to trigger misclassifications. Specifically, VAEs generate images with lower quality than more sophisticated models, particularly for complex datasets such as ImageNet, resulting in fewer acceptable test inputs. On the other hand, VAEs' training process tend to produce a smoother and more continuous latent space, which allows for more efficient exploration, i.e., changes in the latent vector tend to have a direct impact on the generated image.

RQ₃ (Efficiency): DMs were the most efficient for MNIST, SVHN, and CIFAR-10. In contrast, VAEs demonstrated the highest efficiency for ImageNet tasks. Across all GenAI TIGs, increasing the perturbation step consistently reduced the number of iterations needed to cause misclassifications, highlighting that larger perturbations accelerate the process of uncovering classifier weaknesses.

D. RQ₄ (Validity)

The seventh column of Table III shows the percentage and number of misclassification-inducing inputs deemed valid by human validators, excluding those where assessors disagreed.

For simpler datasets (MNIST and SVHN), VAEs either outperformed or performed comparably to other models at high perturbation steps. In these cases, DMs produced more valid inputs at low perturbation steps, despite showing lower percentages overall, which is partially due to their higher number of viable seeds (as discussed in RQ₁). For more complex datasets (CIFAR-10 and ImageNet), DMs generate significantly more valid inputs. For instance, DMs produced nearly 6× more valid misclassification-inducing inputs than VAEs for the *teddy bear* class in ImageNet.

Interestingly, we observed that higher perturbation steps within the same GenAI TIG do not always compromise input validity. For instance, for the ImageNet *teddy bear* class, both the number and percentage of valid inputs increased across all GenAI TIGs. This suggests that larger perturbations can still produce semantically valid inputs while being more effective at inducing misclassifications (as seen in RQ₃).

RQ₄ (Validity): According to humans, DMs excel at generating valid misclassification-inducing inputs for complex datasets like CIFAR-10 and ImageNet. For simpler datasets, GenAI TIGs demonstrate different trade-offs between the number and ratio of valid inputs, depending on the chosen perturbation step.

E. RQ₅ (Label Preservation)

The eighth column of Table III reports the percentage and number of preserved labels, which represent the valid misclassification-inducing inputs that maintained the ground-truth labels of their corresponding seeds. These inputs are particularly valuable for software testers.

GANs achieved remarkable results for simpler datasets (MNIST and SVHN), particularly at high perturbation steps, with up to 7× more label-preserving inputs than other GenAI models. For more complex datasets like CIFAR-10 and ImageNet, GANs exhibited a performance decline, yielding the worst results, while DMs consistently outperformed the others, preserving up to 32× more labels than GANs for CIFAR-10 under high perturbation steps. DMs achieved a high percentage of preserved labels (i.e., > 86%) for the ImageNet *pizza* class, while nearly half of the labels were not preserved for the *teddy bear* class, highlighting variability in label preservation across different tasks even within the same dataset.

RQ₅ (Label Preservation): DMs achieve superior label preservation for complex datasets, achieving up to 92.86% preserved labels. For simpler datasets, GANs frequently provided better or comparable performance than/to other GenAI TIGs.

F. Threats to Validity

Internal Validity. To mitigate threats due to uncontrolled variables, we integrated all GenAI models into a unified TIG framework, ensuring consistent parameters across all experiments, e.g., number of iterations and configuration of the selection operators. Moreover, we studied the potential impact of the perturbation step by performing experiments with low and high perturbation steps, obtained following a standardized procedure. Another possible threat may arise from survey participants providing unreliable answers. We addressed this threat by incorporating an ACQ in each survey, and limiting participation to workers with high reputations.

External Validity. A possible threat is the choice of models and datasets. To mitigate it, we chose four popular datasets of increasing complexity, including ImageNet-1K, which contains high-resolution images from 1,000 different classes. We also considered widely recognized GenAI models from the literature, although our results may not generalize to all architectures. We aim to expand our comparison in the future. **Reproducibility.** We have made our experimental data and models publicly available [30].

VI. LESSONS LEARNED AND KEY INSIGHTS

A. Advanced GenAI Models Excel in Complex Tasks, but Their Superior Performance Comes at a Higher Cost

VAEs and GANs performed well on less complex tasks, despite their simpler architectures and training processes compared to DMs. For instance, GANs produced the highest number of valid label-preserving misclassification-inducing inputs for MNIST and SVHN at high perturbation steps. Instead, DMs struggled with simpler datasets, as their lower resolution and limited variation do not offer enough complexity to fully exploit the diffusion process. However, when considering more realistic datasets, such as CIFAR-10 and ImageNet, DMs clearly outperform other models, producing more viable seeds and up to 32× more label preserving inputs.

This superior performance comes with a trade-off in model cost. As emerges from Table II, VAEs and GANs have much faster inference times than DMs, which is crucial for automated testing, especially for search-based TIGs that generate multiple inputs for several iterations. DMs consistently take longer to generate an image compared to VAEs and GANs, mainly because the former architecture involves multiple steps of adding and removing noise. This process is time-consuming and demands more memory and processing resources, making it less efficient for tasks that require quick results.

For these reasons, testers should carefully assess task complexity and their available budget before beginning a testing campaign: in some cases, a simpler GenAI model may be more suitable than the latest, most advanced, architectures.

B. Higher Perturbation Steps Speed Up Test Generation Without Compromising Input Validity or Label Preservation

Latent space exploration, guided by our fitness function, effectively directed test generation to trigger misclassifications. As is common in search-based testing, increasing the strength

of the mutation, i.e., the perturbation step, tends to improve efficiency. However, a frequent concern is that disruptive changes in inputs may compromise their validity and reduce their usefulness for testing [11].

In our experiments, higher perturbation steps consistently reduced the number of iterations needed to trigger failures. However, we found no clear evidence that increased perturbation steps negatively impacted input validity or label preservation across all GenAI models.

One explanation for this phenomenon is the presence of mechanisms ensuring the adherence of the generative process to the target distribution, as well as the encoding of the desired input features (e.g., label conditions in conditional GANs or textual prompts in DMs). This capability highlights the potential of GenAI models for producing novel and meaningful test inputs. Moreover, our TIG framework enforces the generation of latent vectors that remain within the observed distribution ranges reconstructed by GenAI models through the clamping operation. Clamping restricts the values of mutated latent vectors to a predefined range observed in the training data, mitigating the risk of exceeding boundaries that could lead to invalid inputs. Preliminary experiments confirmed that this mechanism is essential for maintaining input validity, as also highlighted by recent studies [11], [25].

C. Latent Vectors Should Be Carefully Constrained and Carefully Manipulated

Although constraining latent vectors was essential for effective test generation, our study identified limitations of GenAI TIGs. For example, even advanced TIGs generated at most 15 valid, label-preserving inputs out of 100 SVHN seeds. For the ImageNet *pizza* class, nearly half of the valid inputs did not preserve the expected label.

Latent vectors constraints are less intuitive and interpretable than preconditions in traditional software, which are typically derived from domain-specific requirements. Moreover, latent space exploration remains a challenging and open research area. Advancements like formulating constraints over the geometry of the latent space [27], latent space regularization [69], or anomaly detection [59] should be generalized to more complex architectures and integrated into TIGs.

VII. RELATED WORK

In the literature, TIGs for DL based image classification have been largely compared on test effectiveness. These works define a certain TIG more effective than others if it can expose a higher number of misclassifications, regardless of the validity of such inputs. Researchers also proposed and adopted adequacy criteria specific to DL systems, such as neuron-based coverage criteria [16], [18], [70], which assess the extent to which test inputs exercise specific sets of neurons or DL model layers. Despite neuron-based coverage criteria have been extensively used to evaluate TIGs [15], [17], [71]–[73], empirical results showed that higher neuron coverage may lead to the generation of invalid inputs [74]. Other studies [75]

adopted mutant adequacy, e.g., the statistical notion of mutation adequacy introduced by Jahangirova and Tonella [76]. These studies assess whether TIGs can expose DL model mutations, i.e., artificially injected faults that simulate real faults [77]. The aforementioned works do not consider GenAI TIGs and mostly overlook the notion of test input validity and label preservation, which may influence their results.

On the other hand, more and more recent studies are considering the validity assessment of generated synthetic inputs [11], [15], [42], [78]. Automation of validity assessment has been achieved by measuring the reconstruction error of VAEs [15], [59], [78]. However, such automated validation does not consider label preservation, primarily focusing on outlier detection as a proxy for validity. Other research has involved human assessors in evaluating TIGs. Tian et al. [79] demonstrated, through human assessment, that DL image classifier predictions are often unreliable, as they are influenced more by the surrounding context than by the predicted object. Attaoui et al. [80] involved industry practitioners to evaluate their feature extraction and clustering techniques for DL systems. Riccio et al. and Zhang et al. conducted studies on input validity, involving both automated validators and human assessors [11], [42], [81]. They adopted a human evaluation similar to ours to perform a comprehensive comparison of different TIGs from the literature. Unlike their work, we consider a broader range of classification tasks and focus specifically on TIGs based on GenAI models, including the latest advancements, i.e., diffusion models. While their work only focus on validity and label preservation, we also consider effectiveness and efficiency, by providing a framework to fairly compare the impact of these models on test generation.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper, we present a comprehensive comparison of various GenAI models based on their ability to generate misclassification-inducing inputs for testing image classifiers. To achieve this, we introduced a search-based test generation framework that integrates different GenAI models and manipulates inputs by perturbing their latent space.

Our findings demonstrate that GenAI models can successfully generate valid, label-preserving and failure-inducing inputs across all considered classification tasks. Notably, simpler GenAI models perform well on less complex tasks such as MNIST and SVHN, while more advanced models are needed for more challenging ones, i.e., CIFAR-10 and ImageNet. Additionally, we found that increasing the latent vector perturbation step accelerates test generation without compromising input validity or label preservation.

This study and the proposed framework open up several avenues for future research. We plan to conduct a broader evaluation with more datasets and GenAI architectures and extend our framework to incorporate more sophisticated search algorithms and additional testing objectives, such as input diversity and mutation killing.

REFERENCES

- [1] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," *Pattern Recognition Letters*, vol. 141, pp. 61–67, 2021.
- [2] D. Sarwinda, R. H. Paradisa, A. Bustamam, and P. Anggia, "Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer," *Procedia Computer Science*, vol. 179, pp. 423–431, 2021.
- [3] G. Lou, Y. Deng, X. Zheng, M. Zhang, and T. Zhang, "Testing of autonomous driving systems: where are we and where should we go?" in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2022, pp. 31–43.
- [4] S. Tang, Z. Zhang, Y. Zhang, J. Zhou, Y. Guo, S. Liu, S. Guo, Y.-F. Li, L. Ma, Y. Xue *et al.*, "A survey on automated driving system testing: Landscapes and trends," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 5, pp. 1–62, 2023.
- [5] N. Neelofar and A. Aletti, "Identifying and explaining safety-critical scenarios for autonomous vehicles via key features," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 4, 2024.
- [6] A. Guerriero, R. Pietrantuono, and S. Russo, "Operation is the hardest teacher: estimating DNN accuracy looking for mispredictions," in *Proceedings of the 43rd International Conference on Software Engineering*. IEEE Press, 2021, p. 348–358.
- [7] T. Zohdinasab, V. Riccio, and P. Tonella, "DeepAtash: Focused Test Generation for Deep Learning Systems," in *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*. New York, NY, USA: Association for Computing Machinery, 2023, p. 954–966.
- [8] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Transactions on Software Engineering*, vol. 48, no. 1, pp. 1–36, 2022.
- [9] V. Riccio, G. Jahangirova, A. Stocco, N. Humbatova, M. Weiss, and P. Tonella, "Testing machine learning based systems: a systematic mapping," *Empirical Software Engineering*, vol. 25, 2020.
- [10] H. B. Braiek and F. Khomh, "On testing machine learning programs," *Journal of Systems and Software*, vol. 164, 2020.
- [11] V. Riccio and P. Tonella, "When and why test generators for deep learning produce invalid inputs: an empirical study," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2023.
- [12] L. Baresi and M. Young, "Test oracles," *Technical Report CIS-TR-01-02*, 2001.
- [13] M. Pezze and C. Zhang, "Automated test oracles: A survey," in *Advances in computers*. Elsevier, 2014, vol. 95, pp. 1–48.
- [14] G. Jahangirova, D. Clark, M. Harman, and P. Tonella, "Test oracle assessment and improvement," in *Proceedings of the 25th International Symposium on Software Testing and Analysis*. New York, NY, USA: Association for Computing Machinery, 2016.
- [15] S. Dola, M. B. Dwyer, and M. L. Soffa, "Distribution-aware testing of neural networks using generative models," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 2021.
- [16] K. Pei, Y. Cao, J. Yang, and S. Jana, "DeepXplore: Automated whitebox testing of deep learning systems," in *proceedings of the 26th Symposium on Operating Systems Principles*, 2017, pp. 1–18.
- [17] J. Guo, Y. Jiang, Y. Zhao, Q. Chen, and J. Sun, "Dlfuzz: Differential fuzzing testing of deep learning systems," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2018.
- [18] L. Ma, F. Juefei-Xu, F. Zhang, J. Sun, M. Xue, B. Li, C. Chen, T. Su, L. Li, Y. Liu *et al.*, "Deepgauge: Multi-granularity testing criteria for deep learning systems," in *Proceedings of the 33rd ACM/IEEE international conference on automated software engineering*, 2018.
- [19] H. B. Braiek and F. Khomh, "Deepevolution: A search-based testing approach for deep neural networks," in *2019 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 2019.
- [20] V. Riccio and P. Tonella, "Model-based exploration of the frontier of behaviours for deep learning system testing," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020.
- [21] T. Zohdinasab, V. Riccio, A. Gambi, and P. Tonella, "Deephyperion: exploring the feature space of deep learning-based systems through illumination search," in *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2021, pp. 79–90.
- [22] R. Ben Abdesslem, S. Nejati, L. C. Briand, and T. Stifter, "Testing vision-based control systems using learnable evolutionary algorithms," in *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*, May 2018, pp. 1016–1026.
- [23] H. Fahmy, F. Pastore, M. Bagherzadeh, and L. Briand, "Supporting deep neural network safety analysis and retraining through heatmap-based unsupervised learning," *IEEE Transactions on Reliability*, vol. 70, no. 4, pp. 1641–1657, 2021.
- [24] S. Kang, R. Feldt, and S. Yoo, "Sinvad: Search-based image space navigation for dnn image classifier test input generation," in *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*, 2020, pp. 521–528.
- [25] I. Dunn, H. Pouget, D. Kroening, and T. Melham, "Exposing previously undetectable faults in deep neural networks," in *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2021, pp. 56–66.
- [26] A. Aletti, "Software testing of generative ai systems: Challenges and opportunities," in *2023 IEEE/ACM International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE)*, 2023.
- [27] S. Dola, R. McDaniel, M. B. Dwyer, and M. L. Soffa, "Cit4dnn: Generating diverse and rare inputs for neural networks using latent space combinatorial testing," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.
- [28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, 2020.
- [29] S. Kang, R. Feldt, and S. Yoo, "Deceiving humans and machines alike: Search-based test input generation for dnns using variational autoencoders," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 4, pp. 1–24, 2024.
- [30] Maryam, M. Biagiola, A. Stocco, and V. Riccio, "Replication package," https://github.com/Maryammaryam877/genai_tigs, 2024.
- [31] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [32] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, 1951.
- [33] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015.
- [34] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, 2020.
- [35] R. B. Abdesslem, A. Panichella, S. Nejati, L. C. Briand, and T. Stifter, "Testing autonomous cars for feature interaction failures using many-objective search," in *Proceedings of ASE '18*, ser. ASE 2018. New York, NY, USA: ACM, 2018.
- [36] R. Ben Abdesslem, S. Nejati, L. C. Briand, and T. Stifter, "Testing advanced driver assistance systems using multi-objective search and neural networks," in *31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2016.
- [37] A. Gambi, M. Mueller, and G. Fraser, "Automatically testing self-driving cars with search-based procedural content generation," in *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2019, pp. 318–328.
- [38] I. Stenbit, F. Chollet, and W. Luke, "A walk through latent space with stable diffusion," https://keras.io/examples/generative/random_walks_with_stable_diffusion/, 2023, accessed: 10-09-2024.
- [39] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," 2022.
- [40] Getimg.ai, "Interactive guide to stable diffusion guidance scale parameter," accessed: 1 January 2024. [Online]. Available: <https://getimg.ai/guides/interactive-guide-to-stable-diffusion-guidance-scale-parameter>
- [41] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70. PMLR, 2017.
- [42] J. Zhang, J. Keung, X. Ma, X. Li, Y. Xiao, Y. Li, and W. K. Chan, "Enhancing valid test input generation with distribution awareness for deep neural networks," in *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2024, pp. 1095–1100.
- [43] PyTorch Contributors, "Source code for torchvision.models.vgg," https://pytorch.org/vision/0.12/_modules/torchvision/models/vgg.html#vgg19, 2024, accessed: 2024-01-01.

- [44] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [45] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng *et al.*, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- [46] A. Krizhevsky, "Learning multiple layers of features from tiny images," in *Master's thesis, University of Toronto*, 2009. [Online]. Available: <https://api.semanticscholar.org/CorpusID:18268744>
- [47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.
- [48] TensorFlow, "Keras example," https://www.tensorflow.org/datasets/keras_example, accessed: 2022-08-29.
- [49] S. Kang, R. Feldt, and S. Yoo, "Vae training code for mnist," <https://github.com/coinse/SINVAD/blob/master/vae/train.py>, 2020, accessed: August 1, 2023.
- [50] PyTorch, "Dcgan example in pytorch," <https://github.com/pytorch/examples/blob/main/dcgan/main.py>, accessed: September 2, 2023.
- [51] Zokovie, "Cdcgan-mnist," <https://github.com/zokovi/cDCGAN-MNIST>, accessed: September 3, 2023.
- [52] B. Maltais, "Kohya ss - lora gui code," https://github.com/bmaltais/kohya_ss/blob/master/kohya_gui/lora_gui.py, accessed: November 30, 2023.
- [53] S. Kang, R. Feldt, and S. Yoo, "Vae convolutional training code," https://github.com/coinse/SINVAD/blob/master/vae/train_conv.py, accessed: September 29, 2023.
- [54] M. Bhandari, "Training vae on imagenet with pytorch," <https://www.kaggle.com/code/maunish/training-vae-on-imagenet-pytorch>, accessed: April 2024.
- [55] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2019.
- [56] J. Luo, J. Huang, and H. Li, "A case study of conditional deep convolutional generative adversarial networks in machine fault diagnosis," *Journal of Intelligent Manufacturing*, vol. 32, no. 2, pp. 407–425, 2021.
- [57] "Stable diffusion v1.5 repository," <https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5>, accessed: 2023-12-15.
- [58] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.
- [59] A. Stocco, M. Weiss, M. Calzana, and P. Tonella, "Misbehaviour prediction for autonomous driving systems," in *Proceedings of the ACM/IEEE 42nd international conference on software engineering*, 2020.
- [60] Amazon, "Mechanical turk," <https://www.mturk.com>.
- [61] T. S. Behrend, D. J. Sharek, A. W. Meade, and E. N. Wiebe, "The viability of crowdsourcing for survey research," *Behavior research methods*, vol. 43, pp. 800–813, 2011.
- [62] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2008.
- [63] J. Heer and M. Bostock, "Crowdsourcing graphical perception: Using mechanical turk to assess visualization design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: ACM, 2010.
- [64] F. Pastore, L. Mariani, and G. Fraser, "Crowdoracles: Can the crowd solve the oracle problem?" in *2013 IEEE Sixth International Conference on Software Testing, Verification and Validation*, March 2013, pp. 342–351.
- [65] E. Peer, J. Vosgerau, and A. Acquisti, "Reputation as a sufficient condition for data quality on amazon mechanical turk," *Behavior research methods*, vol. 46, pp. 1023–1031, 2014.
- [66] A. W. Edwards, "Ra fischer, statistical methods for research workers, (1925)," in *Landmark writings in western mathematics 1640-1940*. Elsevier, 2005, pp. 856–870.
- [67] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, 1945.
- [68] J. Cohen, *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J: L. Erlbaum Associates, 1988.
- [69] M. Weiss, A. G. Gómez, and P. Tonella, "Generating and detecting true ambiguity: a forgotten danger in dnn supervision testing," *Empirical Software Engineering*, vol. 28, no. 6, p. 146, 2023.
- [70] L. Ma, F. Juefei-Xu, M. Xue, B. Li, L. Li, Y. Liu, and J. Zhao, "Deepct: Tomographic combinatorial testing for deep learning systems," in *2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2019, pp. 614–618.
- [71] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deepest: Automated testing of deep-neural-network-driven autonomous cars," in *Proceedings of the 40th international conference on software engineering*, 2018.
- [72] S. Demir, H. F. Eniser, and A. Sen, "Deepsmartfuzzer: Reward guided test generation for deep learning," *arXiv preprint arXiv:1911.10621*, 2019.
- [73] X. Xie, L. Ma, F. Juefei-Xu, M. Xue, H. Chen, Y. Liu, J. Zhao, B. Li, J. Yin, and S. See, "Deephunter: a coverage-guided fuzz testing framework for deep neural networks," in *Proceedings of the 28th ACM SIGSOFT international symposium on software testing and analysis*, 2019, pp. 146–157.
- [74] F. Harel-Canada, L. Wang, M. A. Gulzar, Q. Gu, and M. Kim, "Is neuron coverage a meaningful measure for testing deep neural networks?" in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, pp. 851–862.
- [75] V. Riccio, N. Humatova, G. Jahangirova, and P. Tonella, "Deepmetis: Augmenting a deep learning test set to increase its mutation score," in *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2021, pp. 355–367.
- [76] G. Jahangirova and P. Tonella, "An empirical evaluation of mutation operators for deep learning systems," in *2020 IEEE 13th International Conference on Software Testing, Validation and Verification (ICST)*. IEEE, 2020, pp. 74–84.
- [77] N. Humatova, G. Jahangirova, G. Bavota, V. Riccio, A. Stocco, and P. Tonella, "Taxonomy of real faults in deep learning systems," in *Proceedings of the ACM/IEEE 42nd international conference on software engineering*, 2020, pp. 1110–1121.
- [78] Z. Jiang, H. Li, R. Wang, X. Tian, C. Liang, F. Yan, J. Zhang, and Z. Liu, "Validity matters: Uncertainty-guided testing of deep neural networks," *Software Testing, Verification and Reliability*, p. e1894, 2024.
- [79] Y. Tian, S. Ma, M. Wen, Y. Liu, S.-C. Cheung, and X. Zhang, "To what extent do dnn-based image classification models make unreliable inferences?" *Empirical Software Engineering*, vol. 26, no. 5, p. 84, 2021.
- [80] M. Attaoui, H. Fahmy, F. Pastore, and L. Briand, "Black-box safety analysis and retraining of dnns based on feature extraction and clustering," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 3, 2023.
- [81] T. Zohdinasab, V. Riccio, and P. Tonella, "An empirical study on low- and high-level explanations of deep learning misbehaviours," in *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 2023, pp. 1–11.