Hubris Benchmarking with AmbiGANs: Assessing Model Overconfidence with Synthetic Ambiguous Data

Cátia Teixeira[1][0009−0002−0930−1769], Inês Gomes[1,3][0009−0006−3104−4622], Carlos Soares[1,2][0000−0003−4549−8917], and Jan N. van Rijn[3][0000−0003−2898−2168]

[1] Artificial Intelligence and Computer Science Laboratory, Faculty of Engineering, University of Porto, Portugal
{up200808037}@up.pt, {catia.rds.teixeira}@gmail.com, {ines.gomes,csoares}@fe.up.pt

[2] Fraunhofer Portugal AICOS, Portugal

[3] Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands
j.n.van.rijn@liacs.leidenuniv.nl

Abstract. The growing deployment of artificial intelligence in critical domains exposes a pressing challenge: how reliably models make predictions for ambiguous data without exhibiting overconfidence. We introduce hubris benchmarking, a methodology to evaluate overconfidence in machine learning models. The benchmark is based on a novel architecture, ambiguous generative adversarial networks (AmbiGANs), which are trained to synthesize realistic yet ambiguous datasets. We also propose the hubris metric to quantitatively measure the extent of model overconfidence when faced with these ambiguous images. We illustrate the usage of the methodology by estimating the hubris of state-of-the-art pre-trained models (ConvNext and ViT) on binarized versions of public datasets, including MNIST, Fashion-MNIST, and Pneumonia Chest X-ray. We found that, while ConvNext is on average 3% more accurate than ViT, it often makes excessively confident predictions, on average by 10% points higher than ViT. These results illustrate the usefulness of hubris benchmarking in high-stakes decision processes.

Keywords: Synthetic Data Generation · Overconfidence · Generative Adversarial Networks · Responsible Artificial Intelligence · Computer Vision

1 Introduction

Recently, the European Union High-Level Expert Group on Artificial Intelligence (AI) stated that trustworthy AI systems must be lawful, ethical, and robust [12]. With the growing use of machine learning and AI, responsible AI practices are essential for transparency and accountability [5]. Model evaluation typically focuses on predictive performance metrics such as accuracy, precision, or recall.

While high confidence in predictions is generally desirable, it is important to recognize when models may be unreliable due to inherent data uncertainty. In these cases, overconfident models can mislead decision-makers, posing risks in high-stakes applications. This issue is critical in fields such as medicine, where AI supports the medical expert in distinguishing ambiguous cases, such as a skin lesion or lung scan shadow, that could indicate either benign or malignant outcomes. Typical binary classification models produce output scores that are converted into predictions. In a well-calibrated model, scores should be near 0 or 1 for clear cases, while ambiguous instances should have scores around 0.5, reflecting an equal likelihood of belonging to either class.

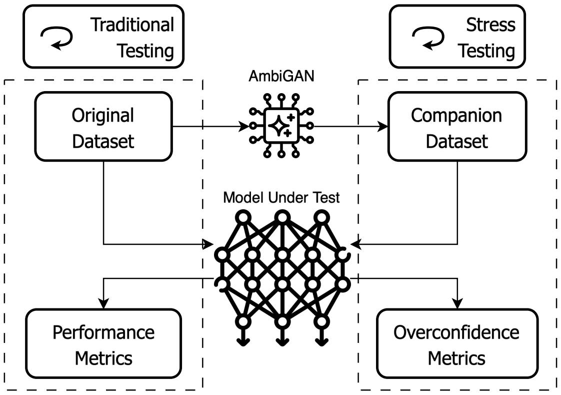

Fig. 1. Schematic view of how the hubris benchmarking methodology complements model testing: on the left side, we have the traditional testing with the original dataset and standard metrics; on the right, we have the stress testing with the companion dataset and overconfidence metrics.

This research proposes *hubris benchmarking*, a method to evaluate the overconfidence of a given classifier when facing ambiguous data. Given that ambiguous data may be scarce in the training set, the first step of our method is to generate a dataset of synthetic, ambiguous and realistic examples. Therefore, we propose a novel generative adversarial network (GAN) framework called ambiguous generative adversarial networks (*AmbiGAN*) to generate such images. To quantify the overconfidence, we propose a new metric called *hubris*, which measures the distance between the expected classifier output (ambiguity) and the true predictions. The metric proposed here is based on Kullback–Leibler (KL) divergence, but other similarity measures could be used as well. Hubris ranges from 0 to 1, where 0 indicates that the model perfectly distinguishes certain cases from uncertain ones. Therefore, the higher the hubris, the higher the model's overconfidence.

The proposed AmbiGAN is based on generative adversarial stress test networks (GASTeN) [9, 27] — a GAN-based framework that generates synthetic realistic images close to the decision boundary of a given classifier. With AmbiGANs, we provide a generative framework that generates a universally ambiguous dataset, referred to as *companion dataset.* We consider a dataset universally ambiguous when different classifiers agree that a given data point has low confidence, as opposed to situations where we stress test models on samples generated specifically for that model [14]. This new companion dataset can test the hubris of machine learning models. We demonstrate the utility of our approach by creating companion datasets for subsets of three image classification benchmarks: MNIST [20], Fashion-MNIST [31], and Pneumonia Chest X-ray datasets [19]. We then evaluate two public, pre-trained models, ConvNext [22] and a Visual Transformer (ViT) [10]. Our results show that these models are overconfident when tested on our AmbiGAN-generated dataset, with hubris values close to 1.00. Most importantly, the results show how hubris benchmarking can provide an alternative perspective on evaluating models: while ConvNext is generally more accurate than ViT, it is also more overconfident, which could make it less interesting in some scenarios (e.g., medical diagnosis). These findings demonstrate how hubris benchmarking can be used to uncover model vulnerabilities by exposing its overconfidence in ambiguous scenarios. By providing a structured method to measure such vulnerabilities, this approach aids in building AI systems that are more transparent, reliable, and safer for deployment in high-stakes fields. The contributions in this article are the following:

1. the hubris benchmarking methodology, which can be applied to any binary image classification problem. It includes a novel hubris metric to evaluate the overconfidence of a model in ambiguous images;
2. the AmbiGAN architecture, which generates companion ambiguous datasets for binary classification purposes without the need for a human-in-the-loop;
3. an illustrative application of our methodology on three datasets, used to evaluate two pre-trained models: ConvNext [22] and ViT [10];
4. publicly available companion datasets for subsets of Pneumonia Chest X-ray[4], MNIST[5], and Fashion-MNIST[6];
5. AmbiGAN source code available for reproducibility and future usage[7].

2 Related Work

We review generative adversarial networks, focusing on those designed for stress testing, and discuss metrics for evaluating synthetic and ambiguous images.

2.1 Generative Adversarial Networks

Generative adversarial networks (GANs), introduced by Goodfellow *et al.* [15], model data distributions through a generator (G) and a discriminator (D) in

[4] https://huggingface.co/datasets/crdsteixeira/AmbiGAN-XRay

[5] https://huggingface.co/datasets/crdsteixeira/AmbiGAN-MNIST

[6] https://huggingface.co/datasets/crdsteixeira/AmbiGAN-Fashion

[7] https://github.com/crdsteixeira/Hubris-AmbiGANs

an adversarial setup. G produces synthetic samples to mimic real data, while D distinguishes real from generated samples [28]. Both are typically multilayer networks with convolutional and fully connected layers [8]. GANs aim for a Nash equilibrium where G generates data indistinguishable from the true distribution. This is formalized as a minimax optimization in Equation 1, with D maximizing the value function $V(D, G)$ by classifying real samples $x \sim p_{\text{data}}$ and fake ones $G(z)$ with $z \sim p_z$, while G minimizes it by generating complex instances for D.

$$\begin{aligned} \mathscr{L}_{D,G}^{GAN} &= \min_{G} \max_{D} V(D, G) \\ &= \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \end{aligned} \tag{1}$$

Unlike models such as Boltzmann machines [1], GANs avoid Markov chains, improving efficiency. Still, issues remain with diversity, image quality, and training stability [29]. Variants such as WGAN [3], DCGAN [25], and InfoGAN [7] address these through improved architectures, adversarial [24], task-specific [11], and multi-objective loss functions [2].

2.2 Generative Adversarial Stress Test Networks

GASTeN [9] is a method that uses GANs to stress test a given model by generating ambiguous yet realistic inputs. It uses a DCGAN [25] to synthesize data near the classifier's decision boundary in binary classification, aiming to evaluate the performance and reliability of a model under ambiguous conditions. GASTeN follows a two-step training: first, pre-training a DCGAN; then, refining the generator using classifier (C) feedback through a modified loss, shown in Equation 2. The equation includes a confusion distance term (cd), scaled by α, to measure the distance from the prediction to the decision threshold. Average confusion distance (ACD) quantifies the confusion distance (cd) of all the generated images.

$$\mathscr{L}_{G}^{GASTeN} = \mathscr{L}_{G}^{GAN} + \alpha \cdot cd(C(G(z))) \tag{2}$$

However, GASTeN often faced trade-offs between quality (measured by the Fréchet inception distance (FID)) and low confidence (measured by ACD) with instability and mode collapse during training. GASTeNv2 [27] addressed these by restoring the original DCGAN architecture, replacing earlier linear-layer modifications, and introducing a new loss. The updated loss combines GAN loss with a Gaussian negative log likelihood (GNLL) term, shown in Equation 3. This term guides the generator to produce samples near the decision boundary (0.5) under the assumption of classifier calibration. The variance (σ^2) is treated as a hyperparameter, adjusting tolerance around the boundary. The updated loss improved both training stability and the ability to generate ambiguous examples.

$$\begin{aligned} \mathcal{L}_{G}^{GASTeNv2} &= \mathcal{L}_{G}^{GAN} + \alpha \cdot \mathcal{L}_{GNLL} \\ &= \mathcal{L}_{G}^{GAN} + \alpha \cdot \left(\frac{1}{2} \log(\sigma^2) + \frac{(C(G(z)) - 0.5)^2}{2\sigma^2} \right) \end{aligned} \tag{3}$$

2.3 Metrics for Synthetic Images

Synthetic image evaluation commonly uses FID [17] to measure image quality and realism. FID quantifies the similarity between real and generated data by comparing Gaussian feature statistics, capturing both realism and diversity [6, 17]. An alternative approach is GIQA [16], which evaluates realism using two metrics: a quality score (QS), which measures perceptual fidelity, and a diversity score (DS), which measures sample variety. A higher quality score implies better visual quality; a higher diversity score indicates more diverse outputs. GIQA has been applied in generative frameworks for tasks such as image editing, segmentation, and synthesis [13,26,32], supporting more nuanced evaluation of image generation performance.

2.4 Ambiguous Images in Classification

Machine learning models rely on high-quality data, yet real-world scenarios often present ambiguous inputs, which are underrepresented in existing datasets [4,30]. Weiss *et al.* define true ambiguity as occurring when a single input yields nonzero probabilities for multiple classes: "x is truly ambiguous if and only if $P(c|x) > 0$ for more than one class c" [30]. Unlike model-specific ambiguity, true ambiguity is inherent to the data and independent of a classifier's decision boundary [21, 30]. This aligns with aleatoric uncertainty, which arises from intrinsic data variability and cannot be reduced even with improved models [18,30]. In contrast, epistemic uncertainty stems from limited model knowledge and can be reduced through better training or data [18].

Various approaches explored the generation of ambiguous samples. GASTeN (Section 2.2) synthesizes ambiguity based on a model's decision boundary, linking it to model-specific uncertainty. In contrast, AMBIGUESS [30] constructs a model-agnostic dataset, ensuring that ambiguity arises from the input itself rather than a classifier. Similarly, AmbiguousMNIST [23] extends the standard MNIST dataset to include samples with multiple plausible labels, reflecting real-world ambiguity.

3 Hubris Benchmarking

Ensuring that AI systems meet the standards of transparency requires tackling a critical blind spot in traditional evaluation: how models handle ambiguity. Figure 2 shows a case of distinguishing between the handwritten digits "8" and "9". These images are inherently ambiguous, even for human observers, yet conventional empirical evaluation fails to assess how models respond to these uncertainties.

A typical binary classification model generates output scores that are subsequently converted into predictions. Ideally, these scores should be near the extremes (e.g., close to 0 or 1) when the label is clear, while for ambiguous cases, the scores should remain closer to 0.5, assuming proper model calibration.

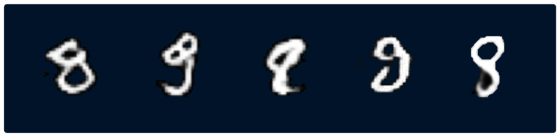

Fig. 2. Ambiguous digits (with features from 9 and 8) generated with AmbiGAN.

At the same time, some models show high certainty for these ambiguous images. This is what is defined as model overconfidence. Without tools to measure and mitigate overconfidence in these scenarios, AI systems risk making unwarranted, high-certainty predictions.

3.1 Companion Datasets with Ambiguous Data

As illustrated in Figure 1, model training and evaluation typically rely on standard datasets and metrics such as accuracy, precision, or recall. However, these metrics fail to capture model behavior in ambiguous scenarios. To address this, we generate synthetic, ambiguous images using AmbiGANs (see Section 4), forming a *companion dataset* that systematically challenges the model under test. This dataset enables the analysis of model predictions in ambiguous cases, revealing overconfidence not detected by traditional metrics. To quantify this, we introduce the *hubris* metric (Section 3.2), which measures the extent of overconfidence. These tools offer a complementary evaluation framework focused on model reliability in uncertain conditions.

3.2 Hubris Metric

We propose *hubris* as a metric to quantify model overconfidence in ambiguous samples. Hubris measures the distance between a model's prediction distribution, $\hat{y}$, and an ideal, unbiased reference prediction of 0.50. In binary classification, a balanced prediction of 0.50 ideally indicates no overconfidence since, in the case of an unbiased and calibrated model, it captures characteristics equally from both classes. In our approach, hubris leverages KL-Divergence to measure these distribution distances. *KL-Hubris* is defined in Equation 4, with $\hat{y}$ being the prediction of the model under test. In KL-Hubris, *Ref* represents a Dirac distribution centered around 0.50, meaning any sampled value from *Ref* will always be exactly 0.50. U is a uniform distribution representing random predictions from 0 to 1. The result is scaled from 0.00 (no overconfidence) to 1.00 (total overconfidence), where lower KL-Hubris values indicate that the model treats ambiguous samples with predictions close to the ideal decision boundary.

$$\textit{KL-Hubris} = 1 - e^{-\left(\frac{KL(\hat{y}|Ref)}{KL(U|Ref)}\right)} \tag{4}$$

We define absolute hubris (H_A) as the measure of overconfidence relative to the ideal ambiguous output distribution centered at 0.50. Hubris can also measure overconfidence relative to any ambiguity estimation methodology, enabling comparisons of probability distributions across models. When used in this context, we refer to it as relative hubris (H_R), which will be used in Section 6.3.

4 Ambiguous Generative Adversarial Networks

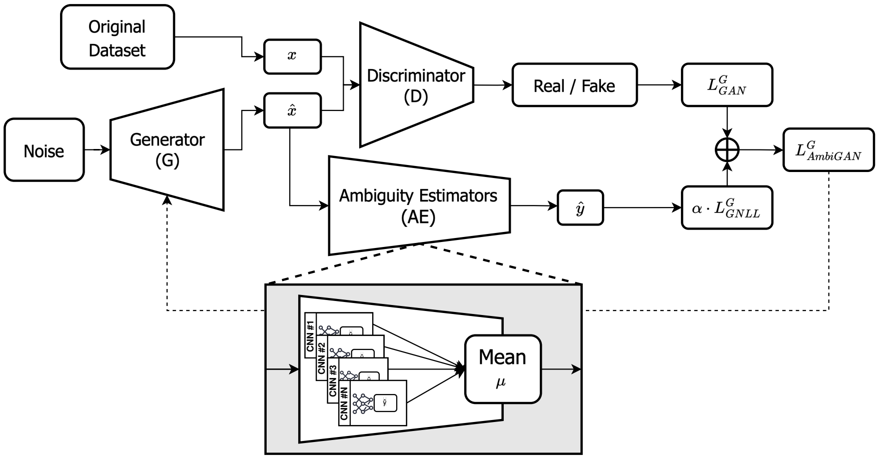

Fig. 3. AmbiGANs Architecture: a GAN-based architecture incorporating the ambiguity estimator (AE); AE averages the individual classifier's predictions; Generator loss combines both original GAN loss and Gaussian loss.

We propose hubris benchmarking, which assesses the overconfidence of models in datasets that contain only ambiguous data. However, these are limited in the standard datasets used to evaluate models in computer vision and other learning tasks. We address this by introducing ambiguous generative adversarial networks (*AmbiGANs*) to generate universally ambiguous synthetic samples.

We define *Universal Ambiguity* as equally challenging samples for machine learning models, i.e., where models predict equal scores for all labels (e.g., 0.5, assuming binary classification tasks). We approximate universal ambiguity by leveraging diversity among classifiers. We generate a set of diverse classifiers (by varying the architecture and training data). Samples that are ambiguous for all those classifiers are considered universally ambiguous. Hence, we implement this concept in AmbiGANs by integrating multiple diverse classifiers acting as the *ambiguity estimator* (AE), where the mean of the individual classifier's predictions is the final output prediction of the ambiguity estimator.

Figure 3 presents a summary of the proposed architecture. We maintain the approach of GASTeN (Section 2.2) regarding the multi-objective loss function, which minimizes both the standard GAN loss and Gaussian loss, to guide the generation to ambiguity. AmbiGANs balance realism and ambiguity, training the generator to produce realistic images that are universally ambiguous, according to the definition above. The ambiguity estimator output prediction is necessary to calculate the Gaussian loss of the generated images. The final loss propagated to the generator is the sum of the original GAN loss and Gaussian loss with a weighting factor (α), shown in Equation 6.

$$\begin{aligned} \mathcal{L}_G^{AmbiGAN} &= \mathcal{L}_G^{GAN} + \alpha \cdot \mathcal{L}_{GNLL} && (5) \\ &= \mathcal{L}_G^{GAN} + \alpha \cdot \left(\frac{1}{2} \log(\sigma^2) + \frac{(AE(G(z)) - 0.5)^2}{2\sigma^2} \right) && (6) \end{aligned}$$

5 Experimental Setup

To illustrate hubris benchmarking, we describe the AmbiGAN training process and companion dataset generation, followed by the evaluation of two pre-trained models. All experiments were conducted using a machine equipped with one Tesla T4 GPU.

5.1 AmbiGAN Training

Following the standard GASTeN training strategy (Section 2.2), AmbiGAN training is split into two stages. In the first stage, the generator and discriminator are pre-trained over 100 epochs for stability. In the second stage, over 50 epochs, the ambiguity estimator is introduced into the loss (Equation 6, Section 4), influencing generator updates based on its predictions. The ambiguity estimator, frozen during the second stage, is built from 50 classifiers trained on the original dataset. To generalize ambiguity, classifiers vary in architecture (layer counts, feature sizes, and initializations) and are trained on non-overlapping data subsets. Final outputs use sigmoid activation to approximate universal ambiguity.

AmbiGAN architectures differ by dataset: for MNIST and Fashion-MNIST, the generator and discriminator both use 3 layers, with 512 features in their final and initial layers, respectively, and a latent space of 256. Due to larger images, a deeper 5-layer architecture is used for the Chest X-ray dataset, with 128 features and a latent space dimension of 512. Learning rates follow the ones proposed for DCGAN [25], set to 0.0002 for MNIST and Fashion-MNIST, and 0.001 for Chest X-ray. All models use Adam optimizer ($\beta_1 = 0.000$, $\beta_2 = 0.999$), and class balancing is enforced via random over-sampling. We tested α values of 0.5, 1.0, and 1.5 to control the trade-off between realism and ambiguity, and σ^2 values of 0.01 and 0.1 to set tolerance around the decision boundary. A lower variance leads to closer predictions of 0.5, while a higher one allows more variance. Three runs were conducted for each dataset, combining all α and σ^2 values, yielding six trained models per run. Hyperparameter selection relied on GASTeN metrics: classification accuracy, FID (image realism), and ACD (ambiguity). While lower values for both FID and ACD are ideal, FID is prioritized to ensure realism, provided ACD remains acceptable. Ambiguity estimator accuracy and ACD are calculated by averaging predictions across classifiers.

5.2 Companion Datasets

We generated companion datasets for MNIST [20], Fashion-MNIST [31], and Chest X-ray [19]. As in GASTeN, AmbiGANs are limited to binary classification [9], so we curated binary subsets: *9 vs. 8* in MNIST, and *Dress vs. T-Shirt* in Fashion-MNIST, selecting visually ambiguous class pairs for easier inspection. Each dataset subset includes grayscale 28x28 images, allowing for compact AmbiGAN architectures. The Chest X-ray dataset originally consisted of 640x640 RGB images labeled as Pneumonia or Normal. We down-scaled them to 128x128 pixels to match our model constraints. Unlike the other datasets, Chest X-ray is already binary but class-imbalanced.

Each companion dataset contains 5 000 generated samples, produced in three runs with different random seeds per dataset, totaling nine companion datasets. We made all companion datasets publicly available on HuggingFace.

To evaluate generation quality, we used FID, quality score, and diversity score, alongside visual inspection for realism, ambiguity, and GAN collapse. We compared each companion dataset against a baseline generated during the initial GAN training stage (before introducing the ambiguity estimator), which lacks deliberate ambiguity and serves as a reference. We also compared image quality with AmbiguousMNIST [23] (see Section 2.4) to benchmark against state-of-the-art ambiguity generation.

5.3 Evaluated Models

We used the generated companion datasets for hubris benchmarking of two pretrained models, ConvNext [22] and ViT [10], originally designed for multi-class tasks. We fine-tuned the models for binary classification by modifying the final layer to output a single value, using a batch size of 64, a learning rate of 0.001, and the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). We first evaluated model accuracy on the original datasets. Then, using the hubris and ACD metrics, we assessed overconfidence on each companion dataset and on AmbiguousMNIST [23]. ACD was computed based on the model's output predictions.

We calculated absolute hubris (H_A), which compares model outputs to an ideal ambiguous distribution, and relative hubris (H_R), which compares outputs to an ambiguity estimator not used during AmbiGAN training. H_R was not computed for AmbiguousMNIST, as its ambiguity differs in nature. While our subsets focus on the ambiguity between specific pairs, AmbigousMNIST combines features from multiple classes across the dataset. For a fair comparison, we filtered AmbiguousMNIST to include only samples labeled with our subset.

6 Results and Discussion

Our experimental workflow consists of three phases: tuning AmbiGAN hyperparameters for each dataset, generating companion datasets, and evaluating pretrained models.

6.1 AmbiGAN Training

In Table 1, we present the results with lower FID obtained during all runs for all datasets. The weight and variance for Gaussian loss used in the best run are also presented for each one.

Table 1. AmbiGAN results (for 2048 generated samples, averaged across 3 runs).

Dataset	**Loss Params**		**ACD**	**FID**
	Weight	**Var**		
MNIST (9 vs. 8)	0.50	0.10	0.08 ± 0.01	18.83 ± 4.23
Fashion-MNIST (Dress vs. T-Shirt)	1.50	0.10	0.06 ± 0.00	37.37 ± 0.77
Chest X-ray (Pneumonia vs. Normal)	1.50	0.10	0.08 ± 0.01	20.21 ± 2.41

The results highlight low ACD values (≤ 0.08), indicating effective confusion generation, and FID scores ranging from 18.83 to 37.37, reflecting a balance between realism and ambiguity in the generated samples. These results demonstrate the adaptability of AmbiGAN architecture to diverse datasets while maintaining quality and ambiguity. ACD and FID had no significant deviation in all runs, and we saw no GAN collapse in any of the experiments performed.

6.2 Companion Datasets

After selecting the AmbiGAN variant with optimal hyperparameters for each dataset specified in Table 1, we generate the complete companion datasets for each of the three original datasets. Table 2 presents the image quality evaluation metrics for the companion datasets. The baseline metrics obtained during the AmbiGAN pre-training phase (before incorporating the ambiguity estimator) are also included for comparison. With this reference, we can analyze the compromise between realism and ambiguity. We aim to keep realism but also expect to improve ambiguity significantly.

While FID scores are higher in the companion datasets than the baseline, quality (QS) and diversity (DS) metrics remain consistent or improve slightly. This balance demonstrates AmbiGAN's ability to generate visually realistic and diverse samples while introducing ambiguity across datasets such as MNIST, Fashion-MNIST, and Chest X-ray. The increase in FID reflects the added variability introduced by the ambiguity estimator loss factor. This fact is caused by the introduced variability in the new samples that reside within the decision boundary of the estimator. We also note that for AmbiGAN-Fashion and AmbiGAN-XRay, the DS decreases for the companion dataset when compared with the baseline. On the other hand, the QS increases for these companion datasets. This indicates that our companion ambiguous dataset has less diversity than the original dataset. For AmbiGAN-MNIST, both QS and DS increase, providing confidence that the generated samples are realistic, which we confirmed by manual inspection. However, in all cases, the differences are small, which

Table 2. Baseline and companion datasets metrics (for 5,000 generated samples) including FID, QS, and DS (Section 2.3), in comparison to the state-of-the-art dataset.

Companion Dataset	Baseline			Companion		
	FID	QS	DS	FID	QS	DS
Ambiguous-MNIST [23] (9 vs. 8)	–	–	–	74.40	0.83	0.80
AmbiGAN-MNIST (9 vs. 8)	3.19 ±0.06	0.82 ±0.03	0.78 ±0.03	17.67 ±0.18	0.83 ±0.01	0.79 ±0.00
AmbiGAN-Fashion (Dress vs. T-Shirt)	11.77 ±0.16	0.74 ±0.01	0.83 ±0.02	33.96 ±0.40	0.80 ±0.02	0.79 ±0.02
AmbiGAN-XRay (Pneumonia vs. Normal)	17.32 ±0.26	0.79 ±0.02	0.87 ±0.00	18.64 ±0.25	0.83 ±0.01	0.83 ±0.01

indicates that the promotion of ambiguity does not reduce the quality of the generated data. Examples of generated images are presented in Figure 4. Especially in those of AmbiGAN-MNIST and AmbiGAN-Fashion, which are images of common concepts, it is clear that they are realistic and ambiguous.

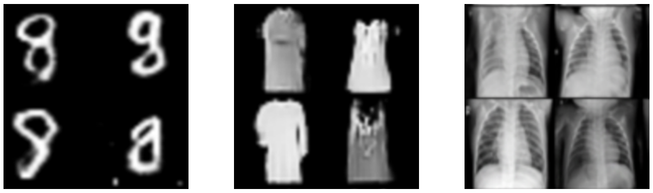

Fig. 4. Examples of generated companion datasets: in the left, AmbiGAN-MNIST (9 vs. 8); in the middle, AmbiGAN-Fashion (Dress vs. T-Shirt); in the right, AmbiGAN-XRay (Pneumonia vs. Normal).

When compared to AmbiguousMNIST [23] dataset, the AmbiGAN-MNIST companion dataset demonstrates significantly lower FID scores (17.67 vs. 74.40), indicating improved realism in the generated samples. Additionally, the quality (QS) and and diversity (DS) are comparable (0.83 vs 0.83 and 0.79 vs. 0.80), suggesting that AmbiGAN maintains competitive levels of quality and diversity. A QS value around 0.8 indicates that the generated samples retain high visual realism, while a DS value near 0.8 reflects sufficient diversity to ensure a broad range of variability.

6.3 Hubris of ConvNext and ViT

We applied hubris benchmarking to pre-trained ConvNext and ViT models using the AmbiGAN-generated companion datasets and AmbiguousMNIST [23]. Results are summarized in Table 3.

Table 3. Pre-trained models evaluation with companion datasets. Accuracy of model with the original dataset is also presented.

Model	**Companion Dataset**	**Accuracy**	**ACD**	H_A[a]	H_R[b]
ConvNext [22]	Ambiguous-MNIST [23] (9 vs. 8)	99.95%	0.49 ±0.00	0.97 ±0.00	–
	AmbiGAN-MNIST (9 vs. 8)		0.48 ±0.00	0.97 ±0.00	0.91 ±0.00
	AmbiGAN-Fashion (Dress vs. T-Shirt)	98.30%	0.48 ±0.00	0.96 ±0.00	0.95 ±0.00
	AmbiGAN-XRay (Pneumonia vs. Normal)	95.38%	0.49 ±0.00	0.97 ±0.00	0.94 ±0.00
ViT [10]	Ambiguous-MNIST [23] (9 vs. 8)	99.07%	0.41 ±0.03	0.92 ±0.02	–
	AmbiGAN-MNIST (9 vs. 8)		0.39 ±0.02	0.89 ±0.03	0.79 ±0.03
	AmbiGAN-Fashion (Dress vs. T-Shirt)	93.97%	0.33 ±0.03	0.80 ±0.05	0.77 ±0.06
	AmbiGAN-XRay (Pneumonia vs. Normal)	90.63%	0.41 ±0.00	0.92 ±0.00	0.86 ±0.00

[a] Absolute hubris: KL-Hubris (ref = 0.50)

[b] Relative hubris: KL-Hubris (ref = Ambiguity estimator prediction)

Both models achieved high accuracy ($> 90\%$), with ConvNext outperforming ViT on MNIST (99.95%) and Chest X-ray (95.38%), confirming its superior predictive performance. However, hubris benchmarking reveals that accuracy alone can be misleading in ambiguous scenarios. ConvNext consistently showed high absolute hubris ($H_A \approx 1.00$) across all AmbiGAN datasets, indicating strong overconfidence. ViT had lower H_A on simpler datasets (0.89 on AmbiGAN-MNIST and 0.80 on AmbiGAN-Fashion) but approached ConvNext's hubris (0.92) on the more complex AmbiGAN-XRay. ACD values support this trend: ConvNext had values close to 0.50, while ViT achieved lower values (close to 0.36) on simpler datasets, reflecting less confident predictions. Relative hubris (H_R) further highlights overconfidence, particularly in ConvNext. Both models deviated significantly from the ambiguity estimator's more balanced outputs. This suggests that deeper architectures such as ConvNext may amplify confidence by reinforcing features across layers, leading to overconfident predictions in ambiguous cases. AmbiguousMNIST results mirror these findings. ConvNext reached $H_A = 0.97$, while ViT was slightly lower at 0.92, showing reduced overconfidence. These results highlight the limitations of relying solely on accuracy. While ConvNext appears superior by standard metrics, ViT's lower hubris scores suggest it may be more reliable in ambiguity-sensitive contexts. Companion datasets generated by AmbiGANs provide a valuable tool for revealing these differences.

6.4 Limitations

We identify some limitations that may hinder the conclusions drawn from this study. The image metrics used focus on assessing realism and quality but do not explicitly account for ambiguity, as they do not consider cases where a single image exhibits features from multiple realistic samples. In companion datasets, such as AmbiGAN-XRay, human labeling could further validate the realism and ambiguity of the generated images. Our approach to universal ambiguity assumes the ambiguity estimator represents all binary classifiers, which may be an oversimplification. Since models generalize differently, our estimator may not fully capture all their decision boundaries.

7 Conclusions

We introduce hubris benchmarking, a methodology to evaluate model overconfidence in machine learning models when faced with ambiguous data. The framework provides a new metric, hubris, to quantify how confidently models make predictions on ambiguous samples. This is relevant for applications such as medical diagnosis and autonomous driving, where overconfidence can be harmful.

Since ambiguous data is rarely available, we introduce AmbiGANs to generate realistic, ambiguous companion datasets. To enable systematic overconfidence evaluation, we generated companion datasets to MNIST, Fashion-MNIST, and Chest X-ray. Then, we benchmarked ConvNext and ViT models, finding that while ConvNext achieves 3% higher accuracy on average, it is also 10 percentage points more overconfident than ViT.

Our results confirm that AmbiGANs effectively support hubris benchmarking. The provided codebase generalizes to other datasets for binary image classification, positioning hubris benchmarking as a practical standard for evaluating overconfidence in ambiguous settings.

Acknowledgments. This work was partially funded by projects AISym4Med and CRAI and LIACC. European Union under the Horizon Europe Framework Programme Grant Agreement N$^{\underline{o}}$: 101095387; Agenda "Center for Responsible AI", nr. C645008882-00000055, investment project nr. 62, financed by the Recovery and Resilience Plan (PRR) and by European Union — NextGeneration EU. Funded by the European Union — NextGenerationEU; UID/00027 of the LIACC — Artificial Intelligence and Computer Science Laboratory — funded by Fundação para a Ciência e a Tecnologia, I.P./ MCTES through the national funds. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. The computational resources of the Google Cloud Platform were provided by the project CPCA-IAC/AF/594904/2023.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for boltzmann machines. Cognitive Science **9**(1), 147–169 (1985)
2. Albuquerque, I., Monteiro, J., Doan, T., Considine, B., Falk, T.H., Mitliagkas, I.: Multi-objective training of generative adversarial networks with multiple discriminators. In: Proceedings of the 36th International Conference on Machine Learning. pp. 202–211 (2019)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning. pp. 214–223 (2017)
4. Aroyo, L., Paritosh, P.: Uncovering unknown unknowns in machine learning (2021), https://ai.googleblog.com/2021/02/uncovering-unknown-unknowns-in-machine.html, online; Google AI Blog
5. Arrieta, A.B., Rodríguez, N.D., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)
6. Baraheem, S.S., Le, T., Nguyen, T.V.: Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook. Artificial Intelligence Review **56**, 10813–10865 (2023)
7. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems 29. pp. 2172–2180 (2016)
8. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE Signal Processing Magazine **35**, 53–65 (2018)
9. Cunha, L., Soares, C., Restivo, A., Teixeira, L.F.: Gasten: Generative adversarial stress test networks. In: Advances in Intelligent Data Analysis XXI — 21st International Symposium on Intelligent Data Analysis. pp. 91–102 (2023)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR (2021)
11. Dunn, I., Pouget, H., Melham, T.F., Kroening, D.: Adaptive generation of unrestricted adversarial inputs. CoRR (2019)
12. European Commission: Ethics guidelines for trustworthy AI (2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
13. Geng, Z., Yang, B., Hang, T., Li, C., Gu, S., Zhang, T., Bao, J., Zhang, Z., Li, H., Hu, H., et al.: Instructdiffusion: A generalist modeling interface for vision tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12709–12720 (2024)
14. Gomes, I., Teixeira, L.F., van Rijn, J.N., Soares, C., Restivo, A., Cunha, L., Santos, M.: Finding Patterns in Ambiguity: Interpretable Stress Testing in the Decision Boundary. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2024)
15. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680 (2014)

16. Gu, S., Bao, J., Chen, D., Wen, F.: GIQA: generated image quality assessment. In: Computer Vision - ECCV 2020 - 16th European Conference. pp. 369–385 (2020)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems 30. pp. 6626–6637 (2017)
18. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine learning **110**(3), 457–506 (2021)
19. Kermany, D.: Labeled optical coherence tomography (oct) and chest x-ray images for classification. Mendeley data (2018)
20. LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: Advances in Neural Information Processing Systems 2. pp. 396–404 (1989)
21. Liu, Y., Feng, L., Wang, X., Zhang, S.: Deepboundary: A coverage testing method of deep learning software based on decision boundary representation. In: 22nd IEEE International Conference on Software Quality, Reliability, and Security. pp. 166–172 (2022)
22. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
23. Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr;, P.H., Gal, Y.: Deterministic neural networks with inductive biases capture epistemic and aleatoric uncertainty. CoRR (2021)
24. Pan, Z., Yu, W., Wang, B., Xie, H., Sheng, V.S., Lei, J., Kwong, S.: Loss functions of generative adversarial networks (GANs): Opportunities and challenges. IEEE Transactions on Emerging Topics in Computational Intelligence **4**, 500–522 (2020)
25. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: 4th International Conference on Learning Representations (2016)
26. Tang, Z., Gu, S., Bao, J., Chen, D., Wen, F.: Improved vector quantized diffusion models. CoRR (2022)
27. Teixeira, C., Gomes, I., Soares, C., van Rijn, J.N.: GASTeNv2: Generative adversarial stress testing networks with gaussian loss. In: Proceedings of the 23rd International Conference on Artificial Intelligence — EPIA 2024 (2024)
28. Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., Wang, F.: Generative adversarial networks: introduction and outlook. IEEE/CAA Journal of Automatica Sinica **4**, 588–598 (2017)
29. Wang, Z., She, Q., E.Ward, T.: Generative adversarial networks in computer vision: A survey and taxonomy. ACM Computing Surveys **54**, 37:1–37:38 (2022)
30. Weiss, M., Gómez, A.G., Tonella, P.: Generating and detecting true ambiguity: a forgotten danger in DNN supervision testing. Empirical Software Engineering **28**, 146 (2023)
31. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. CoRR (2017)
32. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023)