

Fréchet inception distance under finite sampling between real image sets

Jessica E. Goldring

`jessica.goldring@npl.co.uk`

National Physical Laboratory

Elizabeth A. Cooke

National Physical Laboratory

Carlijn Roozemon

Dutch Expert Centre for Screening

Ruben van Engen

Dutch Expert Centre for Screening

Spencer A. Thomas

National Physical Laboratory

Article

Keywords: Fréchet inception distance, Artificial Intelligence, real images, synthetic images

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10527199/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Fréchet inception distance under finite sampling between real image sets

Jessica E. Goldring^{*1,2}, Elizabeth A. Cooke¹, Carlijn Roozemon³,
Ruben van Engen³, and Spencer A. Thomas¹

¹National Physical Laboratory, Teddington, Middlesex,
TW11 0LW, United Kingdom

²UCL Great Ormond Street Institute of Child Health, University
College London, London, United Kingdom

³Dutch Expert Centre for Screening (LRCB) Nijmegen, The
Netherlands

^{*}jessica.goldring@npl.co.uk

August 13, 2026

Abstract

The Fréchet inception distance (FID) is a widely used metric to quantify differences between image sets, most commonly to compare real images with synthetically generated ones. In this paper, we examine the robustness of the FID metric at quantifying this similarity by applying it to sets of real-images from the same distribution. With the use of 3 benchmark datasets and 3 real world datasets, we evaluate FID scores between distinct sets of real images by partitioning a superset into subsets of balanced class instances, and computing FID across a range of set sizes for both in- and out of distribution data. Here we show that FID values decrease approximately linearly with the inverse of the number of images in each subset. We observe sub-linear gradients for real world images, compared to linear and super-linear for simpler benchmark datasets. In general we observe a clear step change in conditioning when transitioning from an under-determined problem, however this does not appear to correspond to a measurable difference in the FID value, its variance, or the Rank of the covariance matrix used to compute the FID. Extrapolating FID values to extremely large set sizes (10^9 and 10^{10}) of real images from the same superset yields appreciable FID values. This questions the effectiveness of FID at measuring the similarity of real and generated images for practical image set sizes in real world applications.

Keywords: Fréchet inception distance, Artificial Intelligence, real images, synthetic images.

1 Introduction

The Fréchet inception distance (FID) is the current standard metric for quantifying the similarity of two sets of images [1]. Typically, the FID is used as an evaluation metric for the quality of synthetically generated images compared to a reference set of real images. Most applications of the FID have been to benchmark the performance of various generative AI models, such as generative adversarial networks (GANs) [2], diffusion models [3] and transformers [4], at producing high quality synthetic data.

Work by Yu et al. [5] investigated the suitability of the FID measure for the assessment of the generation of synthetic images based on the Modified National Institute of Standards and Technology handwritten digits dataset (MNIST) [6]. It was reported when visually assessing the generated images that the FID score gave a more realistic indicator compared to the regular inception score. However, more generally, Jayasumana et al. [7] highlighted that a limitation of FID as an evaluation metric for modern text-to-image models, was that it did not align with human observation and fails to capture gradual improvements and distortions. A further study [8] looked at inappropriate FID assignment in the context of low perceptual qualities by integrating lower-level features to cover a wider array of visual defects. Moreover, as the FID is based on the latent feature distribution of an Inception-v3 model pretrained on natural images in the ImageNet1K dataset [9], it is known to be sub-optimal for non-natural images

45 due to the data being out of distribution relative to the ImageNet1k training
 46 data [10–13]. A pertinent example of non-natural, out of distribution data are
 47 medical images, where using a pretrained Inception-v3 on natural images may
 48 lead to less informative latent feature representations [14]. Despite these limita-
 49 tions in the latent features representation, there has been a significant focus of
 50 the FID for medical imaging data tasks, in part due to the lack of appropriate
 51 alternative metrics. Medical imaging data typically consist of characteristics
 52 including noisy images, non-natural images, heterogenous samples, and a low
 53 number of instances. The latter is a particular challenge relating to the FID for
 54 medical imaging as natural image benchmarks typically use tens of thousands
 55 of images, significantly more than are typically available for medical imaging
 56 datasets. Previous work looking at sample numbers for the FID between sets of
 57 real and generated images highlight that the FID is a biased estimator sensitive
 58 to the number of instances [15, 16].

59 Many studies have investigated the FID for different applications or com-
 60 pared against other metrics [17, 18]. However, there is a lack of verification
 61 as to whether the FID is a robust measure between two sets of *real* images.
 62 The underlying assumption is that the FID can robustly measure the similar-
 63 ity of sample distributions drawn from the same population, which is critical
 64 for its appropriateness in evaluating the quality of synthetic images. If this
 65 assumption is weak, or indeed false, then the FID is not a suitable metric for
 66 assessing the quality of synthetic images. To the best of our knowledge there
 67 have been no investigations to test this assumption. In this study, we address
 68 this gap by evaluating the FID between samples of real images such as animals
 69 or skin lesions that are each drawn from the same population. This allows us
 70 to evaluate the performance of the FID score at measuring consistent features
 71 in ‘real world’ images. We assess the FID between two sets of real images from
 72 the same distribution over a range of sample sizes to investigate how the score
 73 varies with sample size. Section 2 of this paper outlines the FID experiments
 74 and describes the datasets. Section 3 presents the results and discussion from
 75 our experiments. Finally, Section 4 gives our conclusions.

76 2 Methods

77 2.1 FID Calculation

78 The FID is a measure of distance between two sets of images as an extension of
 79 the Fréchet distance between two distributions modelled as multivariate Gaus-
 80 sians with mean μ, μ' and covariance Σ, Σ' . For the FID, a convolutional neural
 81 network, Inception-v3 [19], pre-trained on ImageNet [4] is used to extract la-
 82 tent features from the respective sets of images, which are then are modelled as
 83 multivariate Gaussian distributions. The corresponding means and covariances
 84 are then used to compute the FID as

$$\text{FID}(X, X') = \|\mu - \mu'\|_2^2 + \text{Tr} \left(\Sigma + \Sigma' - 2(\Sigma\Sigma')^{1/2} \right), \quad (1)$$

where Tr is the trace operation. By comparing distributions of image attributes, FID scores are generally used to measure similarities between real and synthetic images to determine the effectiveness of a trained network at generating the synthetic images. Similar images are expected to have a low score so, in the extreme case, an image set paired with itself gives a FID of zero. The FID increases as the image features in the sets diverge, such that images with different sets of features have a high score. There lacks, however, an assessment or validation of this assumption, that the FID is effective and robust for measuring similarities between samples from the same distribution, i.e. real images of the same entity. Throughout this work we consider *real* images as an image acquired from a device or sensor that derives from a physical or digital measurement such as a camera, x-ray, magnetic resonance imaging scanner, microscope, etc. We then consider *generated* images as the output of any model or process that simulates or synthesises image acquisition such as an AI system, mathematical model, digital twin or simulation.

Several studies have noted the dependence of FID on sample size [15, 16] with varying sample sizes used in the literature [11, 18, 20]. Mathematically, there is a lower limit for robustly computing the FID based on the size of the covariance matrix in equation (1), which is set by the output of the Inception-v3 model as 2048. Fewer instances than this limit lead to an under-determined solution which may be numerically unstable. This limit is pertinent for medical imaging datasets which are typically in the hundreds or thousands rather than the 10's of thousands in some natural imaging datasets.

2.2 Experiments

To evaluate the FID between samples of real images, we take a collection of real images and split them into equal sets consisting of comparable class instances, and compute the FID for varying set sizes. This provides a best case, or lower bound, for the FID between these real images from the same distribution. If the FID between these sets of real images is non-zero, it is impractical to expect the FID between a real and generated set to be close to zero, except for the case where the generated images are trivially direct copies of the real images.

To conduct these experiments, we use several well-known open-source datasets (Section 2.3) to benchmark the FID calculations. We repeat the experiments for a range of set sizes to identify dependence of the FID score on the number of images, which we can compare to the study of Chong et al [16] as we extrapolate our results as the investigators recommend in order to avoid bias that they report. For a set of real images, X , of C classes, we split each class into two equal size subsets, X_1^c and X_2^c , where $c \in \{1 \dots C\}$, and compute the FID between these image sets. In order to assess the uncertainty of these calculations we randomly split the complete set of images, X , five times and compute the mean and variance of the score. For assessing the dependence on set size, we randomly select instances from X_1^c and X_2^c independently to reach the required set size, again computing mean and variance for the FID. As we compute the FID for each class, we refer to these experiments as *Class-Wise* FID.

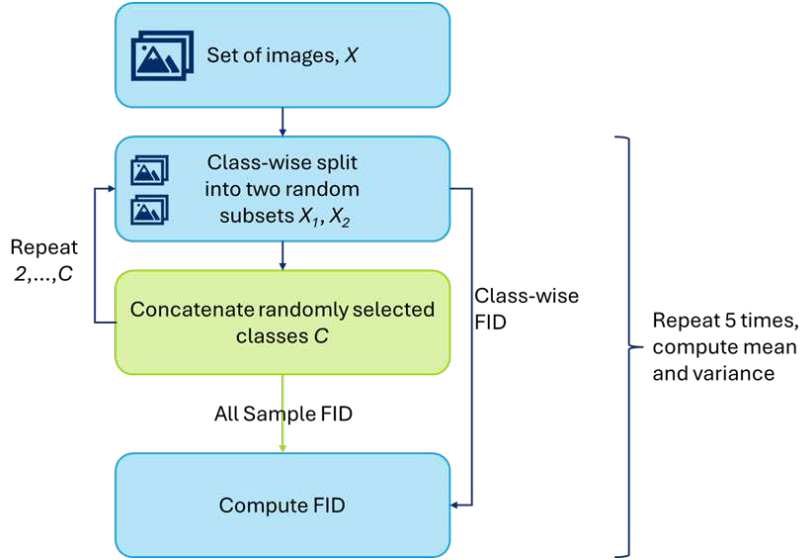


Figure 1: **Experiment Setup** Overview of experimental setup for the Class-wise and All Sample FID calculations

129 In order to extend the analysis of set size in this work we perform addi-
130 tional experiments that combine multiple classes together. In these additional
131 experiments, we subset the data as above and combine all classes together (it-
132 eratively increasing the number of instances from each class). For example,
133 for the MNIST dataset we split the ‘0’ class into two sets, x_1^0 and x_2^0 , and
134 split the ‘1’ class into x_1^1 and x_2^1 , etc. then concatenate these sets, such that
135 $X_1 = [x_1^0, x_1^1, \dots, x_1^9]$ and $X_2 = [x_2^0, x_2^1, \dots, x_2^9]$. The size of x_1^c and x_2^c are
136 balanced and iteratively increased as the *Class-wise* FID experiments prior to
137 concatenation. Although there are differences between the two classes, both are
138 equally represented in the subsets X_1 and X_2 and therefore the distribution of
139 image features in both subsets is consistent and should not affect the FID score.
140 As we combine classes in this experiment we refer to these experiments as *All*
141 *Sample FID*. An illustration of both experiments is given in Fig. 1.

142 We compute the rank and conditioning of the covariance matrix in the FID
143 calculation for all experiments, which along with the deviation across multiple
144 runs, provides an indication of confidence on the FID values. The Rank shows
145 the number of independent rows in the covariance matrix, which provides in-
146 sight about the amount of ‘useful’ information used to compute the FID. The
147 condition of the matrix is a measure of stability of the solution, and how much
148 it can change under small perturbations in the data. Our analysis showed that
149 the Rank and condition number are approximately equal for Σ and Σ' hence we
150 only report the Rank and Condition of the covariance matrix for brevity.

2.3 Datasets

We use several datasets in our experiments: three benchmark datasets and three real-world medical imaging datasets to demonstrate the generality of our findings. For benchmark data we use the widely used MNIST [6] dataset, containing a total of 70,000 images of handwritten digits sorted into 10 classes, the CIFAR-10 [21] collection, also with 10 classes and containing 60,000 images of objects and animals, and the ImageNet1K dataset with over 1.2 million images for 1000 classes of natural images [4]. For real-world datasets, we use examples ranging across several imaging modalities and number of classes. These include: the HAM-10,000 dataset of 10,015 dermoscopic images of eight imbalanced diagnosis classes [22], the PathMNIST [23] dataset which includes 107,180 colon pathology images divided into nine classes, and a mammography dataset of 4,486 mammograms collated by the Dutch Expert Centre for Screening (LRCB) [24].

For the mammography data all experimental protocols were approved by the LRCB and were carried out in accordance with applicable EU and Dutch guidelines and regulations governing mammography screening. The dataset consisted solely of fully anonymised mammography images collected for research purposes with patient consent. All personally identifiable information had been removed prior to access by the investigators, and no patient-identifiable data were used in this study. The data were originally collected before the implementation and subsequent tightening of Dutch healthcare data governance requirements under the Wabvz framework and were handled in accordance with the regulations applicable at the time of collection, as well as current GDPR requirements for anonymised research data.

We note that as the Inception-v3 model used to obtain the feature representations was pre-trained on the ImageNet1K dataset, some of these datasets are out of distribution relative to the training data. The ImageNet1K data is, by definition, in distribution (note that as we are only extracting feature representations and not performing downstream tasks, ‘overfitting’ is not an issue), and we consider CIFAR-10 also as in distribution. The remaining datasets are all out of distribution relative to ImageNet1K and therefore the feature embeddings of the Inception-v3 model are sub-optimal for the calculation of the FID (see equation 1). Our experiments therefore also explore the impact of out of distribution data on the computation of the FID.

All images are resized to $299 \times 299 \times 3$ pixels in order to be compatible with the Inception-v3 model, and are also normalised to $[0, 1]$ to achieve the same overall distribution. For the mammogram dataset, the image pre-processing is performed using proprietary software on the instrument as used in a clinical setting. All experiments were performed in Matlab 2024a, with the mammography latent features extracted in Python prior to analysis in Matlab.

We use sample sizes starting from either 50 or 100, with increments varying depending on the number of instances in the full dataset and coverage of range or sample sizes. For larger datasets we compute up to 30 k images in each of X_1 and X_2 . As a single class size enforces an upper limit on the number of instances we use, we also use the All-Sample FID calculations to provide larger,

196 balanced sets of images.

197 HAM-10,000 contains the ‘Melanocytic Nevi’ class that corresponds to 67 %
198 of the overall dataset so this is the only individual class compared to the All
199 Samples experiment. The PathMNIST dataset contains sufficient samples in
200 each class and we observed insignificant difference between classes so it was
201 decided that an All Samples group was not necessary, and for the mammography
202 dataset there were no class labels to compute *class-wise* FIDs.

203 3 Results and discussion

204 In all our results the FID is plotted in relation to set size as well as the rank
205 and the condition of the covariance matrix in equation (1). Broadly, the FID
206 values follow the same dependence with sample size across all benchmark and
207 real world datasets, exhibiting an initially rapid decrease in FID with increasing
208 set size N until converging asymptotically close to zero to a FID value in the
209 limit $N \rightarrow \infty$. Both the class-wise and All Sample FID scores demonstrate
210 this behaviour, with overlap between the class-wise and All Sample FIDs for
211 MNIST, CIFAR-10 and HAM-10,000, Figure 2. The dashed line depicts where
212 the set size equals the dimensionality of the feature vector produced by the
213 Inception-v3 model (2048), notably affecting the computation of the covariance
214 in equation (1). This threshold for set size, below which the problem is under-
215 determined. Interestingly, there is no notable difference in variance of the FIDs
216 across the five repeats of the experiment (other than an expected reduction with
217 increasing samples) in the different regimes (observed across all datasets). The
218 underdetermined regime (left of the dashed lines in Figure 2 shows rapid reduc-
219 tion of FID with increasing set size compared to the alternative (determined)
220 regime, however, this could primarily be down to sample-wise variation within
221 the subsets which would be significant at small set sizes. Due to the small
222 number of images, the mammography dataset has only a few points outside the
223 underdetermined regime.

224 The effect of small sample size is clearly seen in the condition of the co-
225 variance matrix (see Figure 3), which is unsurprisingly large and variable in
226 the underdetermined region. The right side of the dashed line exhibits a step
227 change when the problem is no longer underdetermined where the condition
228 reduces by several orders of magnitude and the variance is indistinguishable
229 in most plots. Interestingly, the change in the rank of the covariance matrix
230 with sample size differs across the datasets. MNIST approaches full rank in the
231 underdetermined region, possibly due to the simplicity of the dataset, whereas
232 CIFAR-10, ImageNet1K and HAM-10,000 show the rank of the covariance ma-
233 trix (Figure 4) increasing considerably even for a large number of images. For
234 ImageNet1K and HAM-10,000, the rank continues to increase beyond the 10,000
235 and 5,000 (respectively) subset sizes used in our experiments. However, for the
236 PathMNIST data, the covariance matrix is at full rank at the dashed line and
237 therefore does not change with increasing subset size. For the mammography
238 dataset, we observe the step change in covariance matrix conditioning, whereas

239 the rank appears to be increasing beyond the available subset size. The trend
240 of Rank with subset size appears to vary across the datasets with no observable
241 similarity for in or out of distribution datasets respectively.

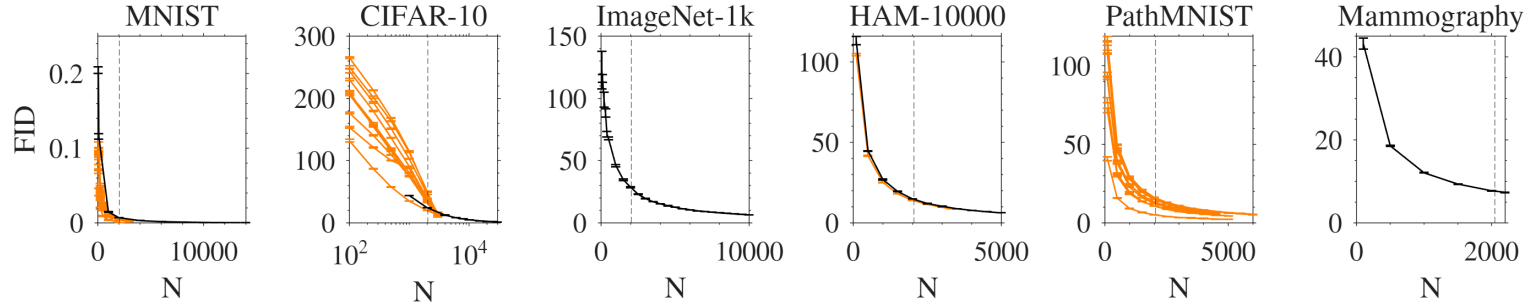


Figure 2: **FID Dependence on Sample Size** FID values dependence on the number of images in each subset for the different datasets (columns). Orange lines denote FID for the class-wise experiments and black lines are for the All Samples experiment (classes are merged and equally represented in each subset). The vertical dashed line indicates where the number of images equals the size of the covariance matrix in equation (1). Error-bars represent the standard deviation over the independent runs. N.B. some plots are on a log scale for clear visualization

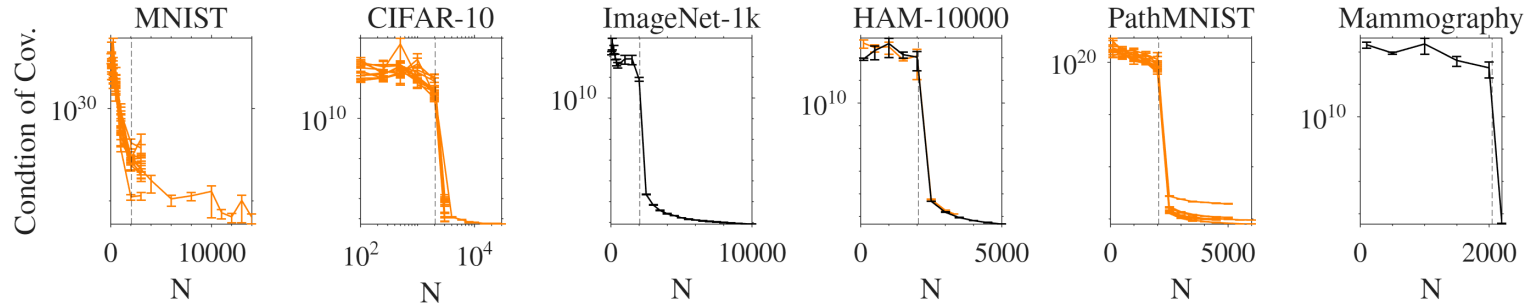


Figure 3: **Condition of covariance matrix** The condition of the covariance matrix in equation (1) dependence on the number of images in each subset for the different datasets (columns). Orange lines denote FID for the class-wise experiments and black lines are for the All Samples experiment (classes are merged and equally represented in each subset). The vertical dashed line indicates where the number of images equals the size of the covariance matrix in equation (1) Error-bars represent the standard deviation over the independent runs. N.B. some plots are on a log scale for clear visualization

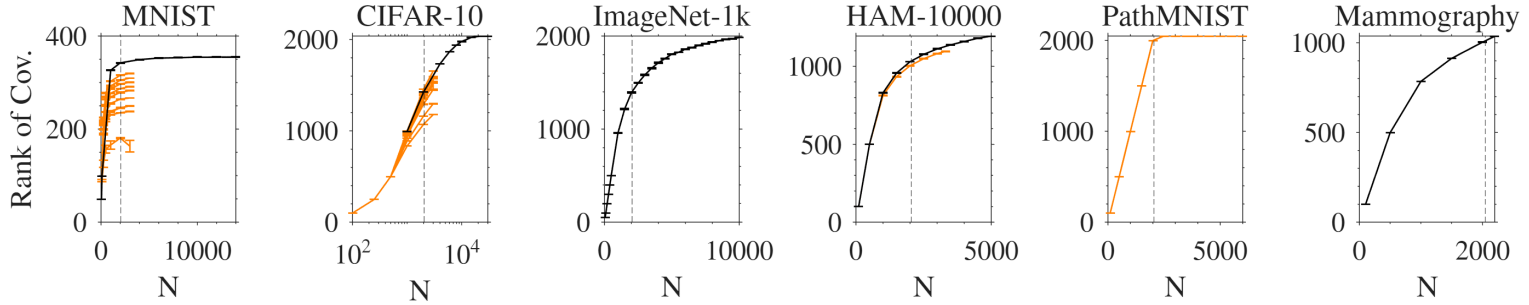


Figure 4: **Rank of the covariance matrix** The rank of the covariance matrix in equation (1) dependence on the number of images in each subset for the different datasets (columns). Orange lines denote FID for the class-wise experiments and black lines are for the All Samples experiment (classes are merged and equally represented in each subset). The vertical dashed line indicates where the number of images equals the size of the covariance matrix in equation

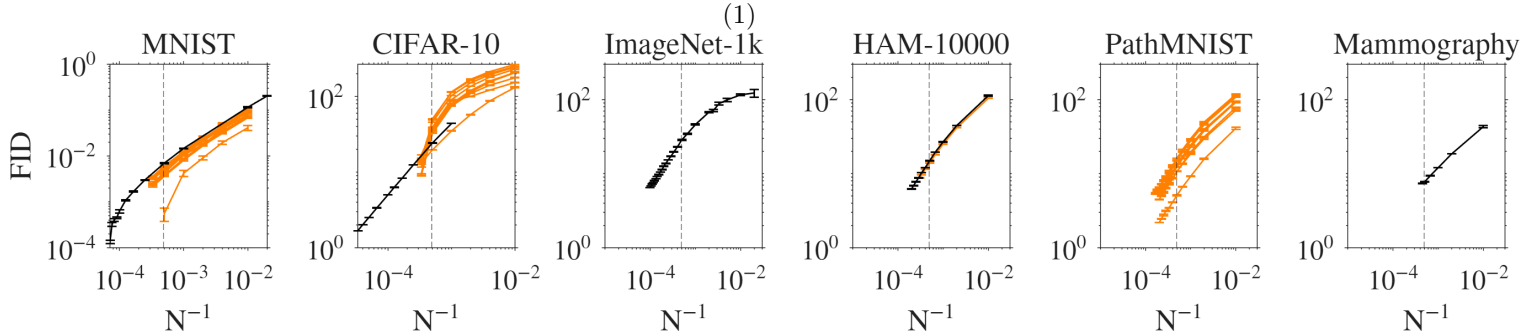


Figure 5: **FID dependence on inverse sample size** FID dependence on inverse sample size for the different datasets (columns) in this study. Orange lines denote FID for the class-wise experiments (each line represents a different class) and black lines are for the All Samples experiment (classes are merged and equally represented in each subset). The vertical dashed line indicates where the number of images equals the size of the covariance matrix in equation (1). Error-bars represent standard deviation over the independent runs. N.B. Plots are on a log scale for clear visualization

Chong et al. [16] demonstrated that the FID is a biased estimator and recommended extrapolating FID versus $1/N$ as $N \rightarrow \infty$. Figure 5 presents our results in this format. Except for small sample sizes, there is no obvious change in FID dependence with N on the left or right sides of the dashed line. In their studies, Chong et al. [16] observed a linear decrease in FID with $1/N$ when comparing real and generated images. However we only observe near linear trends for large samples sizes (i.e. low $1/N$) when comparing sets of real images. Due to the large dynamic range of our experiments we present these plots on a log-log scale for visualisation. We note that the right side of Figure 5 is the underdetermined region due to the $1/N$ x -axis. We observe differing FID values across the different data modalities indicating that the FID is not comparable across applications. Interestingly, we also observe that the FID is lower for out of distribution data (HAM-10,000, PathMNIST) than in-distribution data (Imagenet1K) at the same sample size. This indicates that the subsets from out of distribution data are more similar to each other than the subsets from in distribution data, though this could be due to the lack of model optimisation for distinguishing features in these (out of distribution) domains, effectively homogenising the features in out of distribution data. For the sample-wise FID assessments (MNIST, CIFAR-10 and PathMNIST), the FID values at the same subset size can differ by a factor of two or three between classes regardless of whether they are in distribution (CIFAR-10) or out of distribution (PathMNIST). This provides further evidence that FID values are not comparable between datasets, even subsets drawn from the same distribution.

We extrapolate the lines in Figure 5 by fitting with a power law, $y = Ax^\alpha$, via $\log y = \log A + \alpha \log x$, extrapolating towards $1/N \rightarrow 0$. Table 1 contains the FIDs from the largest subset size in each dataset used, $\text{FID}(N_{max})$, and an extrapolation of FID at very large samples as an estimation of the unbiased value FID^* . Extrapolated values are reported as median estimates with 95 % confidence intervals estimated via non-parametric bootstrap resampling (with replacement) of the data using 5,000 replicates. With the exception of MNIST and CIFAR-10, even at set sizes many orders of magnitude larger than typical experiments, and practically available, the estimate of FID^* are appreciably larger than zero. Interestingly the out of distribution data (HAM-10,000 and PathMNIST) are an order of magnitude lower than for ImageNet1k, though as previously stated this may be due to the homogenisation of latent features from a sub-optimal embedding and therefore misleading.

The gradient α of the fitted lines characterises the growth rate of the FID with increasing number of images, and further highlight the variation with datasets. Some in distribution data (ImageNet-1k) have the same sub-linear growth rate as out of distribution data (Mammography), where as others (CIFAR-10) are approximately linear. The growth rate for out of distribution data can vary drastically from sub-linear (Mammography, HAM-10000, PathMNIST) to super-linear (MNIST).

Table 1: **Extrapolation of Results** Extrapolating FID versus $1/N$, $N \rightarrow \infty$ (Fig. 5) to $\text{FID}(N^*)$, where N^* is 10^9 or 10^{10} samples depending on dataset, as an estimation for the unbiased FID value $\text{FID}^* \approx \text{FID}(N^*)$. $N_{\max}$ is the maximum set size used for a dataset and α is the fitted gradient on the log log scale plots. Calculations are quoted with 95 % bootstrap confidence intervals (CI). [†]PathMNIST values are averaged over the class-wise FID lines

Dataset	$\text{FID}(N_{\max})$	FID^* [95 % CI]	α [95 % CI]
MNIST	1.30×10^{-4}	$2.72 [0.00, 20.30] \times 10^{-11}$	$1.19 [1.08, 1.73]$
CIFAR-10	1.69	$2.65 [1.83, 3.41] \times 10^{-6}$	$0.97 [0.95, 0.99]$
ImageNet-1k	6.48	$1.29 [0.20, 3.75] \times 10^{-3}$	$0.63 [0.56, 0.76]$
HAM-10000	6.22	$2.02 [0.14, 4.16] \times 10^{-4}$	$0.71 [0.67, 0.90]$
PathMNIST [†]	5.67	$2.45 [1.45, 3.44] \times 10^{-4}$	$0.75 [0.73, 0.76]$
Mammography	7.38	$2.67 [1.17, 3.46] \times 10^{-3}$	$0.58 [0.56, 0.63]$

4 Conclusions

In this study we assess whether the FID is a useful metric for evaluating the quality of synthetically generated images by computing the FID for real (i.e., non-generated) images across several benchmark and real-world datasets. We examine the impact of set size on the FID value, the effect of out of distribution data and class variation, as well as the robustness of the FID computation.

Our results show that FID score decreases approximately linearly on a log-log scale with the inverse of the number of images in each subset. We observe comparable gradients across classes within real world datasets, with different intercepts indicating class-wise heterogeneity for FID values. We observe this for natural benchmark images that are in distribution, relative to the pre-trained Inception-v3 model used to compute the FID values, and for complex medical images that are out of distribution for the model. Interestingly we observe the similar sub-linear gradient for in distribution data (ImageNet1k) and out of distribution data (HAM-10000, PathMNSIT, Mammography) providing set sizes were sufficient. However other datasets MNIST (out of distribution) and CIFAR-10 (in distribution) exhibit a super-linear and linear gradients respectively, indicating behaviour is dataset dependent.

Even when comparing extremely large subsets of non-generated images derived from the same distribution, the FID produces non-zero values, limiting its usefulness when evaluating the quality of generated images compared to a ground truth set. Moreover, we find the FID is generally lower for out of distribution data compared to in distribution images at the same sample size, which we hypothesise is due to a sub-optimal feature embedding for the out of distribution data, effectively homogenising the classes. This potentially makes low FID values misleading when evaluating the quality of generated data that is out of distribution.

Despite a clear step change in conditioning of the problem when moving

from an under-determined problem, we observe no other impacts on the computation of the FID value. The FID value, its variance and the Rank of the Covariance matrix in equation (1) exhibit no discernible impact or change when moving from an under-determined problem. The dependence of Rank on set size appears to be model dependent which may be due to the complexity of the problem and the effectiveness of the model at embedding the image features. In any case further investigations are required here.

Acknowledgements

The authors would like to thank Ciaran Bench (NPL) for additional coding support, Chris Barnes and Philip Aston (NPL) for feedback on the manuscript.

Author Contribution

Conceptualization, S.A.T. and J.E.G.; Methodology, S.A.T. and J.E.G.; Software, S.A.T., E.A.C. and J.E.G.; Validation, all authors; Formal Analysis, S.A.T., E.A.C. and J.E.G.; Investigation, J.E.G., E.A.C., and S.A.T.; Resources, C.R. and R.v.E.; Writing – Original Draft Preparation, J.E.G. and S.A.T.; Writing – Review & Editing, E.A.C. S.A.T., R.v.E. and C.R.; Visualization, E.A.C., J.E.G. and S.A.T.; Supervision, S.A.T.; Project Administration, S.A.T.; Funding Acquisition, all authors. The authors thank Ciaran Bench (NPL) for providing the mammography feature embeddings.

Conflicts of Interest

The authors declare no conflicts of interest for this work.

Data Availability

Five of the datasets used are open-source and available as follows:

MNIST; <https://www.kaggle.com/datasets/hojjatk/mnist-dataset>, cifar-10; <https://www.kaggle.com/c/cifar-10>, HAM-10,000; <https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000> and PathMNIST; <https://www.kaggle.com/datasets/dingchengtian/pathmnist>, <https://www.kaggle.com/datasets/sautkin/imagenet1k1>.

In the case of the LRCB dataset, Access may be granted by the data custodian (LRCB) upon reasonable request and subject to the necessary ethical and governance approvals.

Funding Declaration

The project 22HLT05 MAIBAI has received funding from the European Partnership on Metrology, co-financed from the European Union’s Horizon Europe Research and Innovation Programme and by the Participating States. Funding for the UK partners was provided by Innovate UK under the Hori-

359 zon Europe Guarantee Extension.

360 **Ethics Statement**

361

362 All experimental protocols were approved by the Dutch Expert Centre for
363 Screening (LRCB) and were carried out in accordance with applicable EU
364 and Dutch guidelines and regulations governing mammography screening. The
365 dataset consisted solely of fully anonymised mammography images collected for
366 research purposes with patient consent. All personally identifiable information
367 had been removed prior to access by the investigators, and no patient-identifiable
368 data were used in this study. The data were originally collected before the im-
369 plementation and subsequent tightening of Dutch healthcare data governance
370 requirements under the Wabvpz framework and were handled in accordance with
371 the regulations applicable at the time of collection, as well as current GDPR
372 requirements for anonymised research data.

373 **References**

- 374 1. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S.
375 *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash*
376 *Equilibrium in Advances in Neural Information Processing Systems* (eds
377 Guyon, I. *et al.*) **30** (Curran Associates, Inc., 2017).
- 378 2. Esser, P. *et al.* *Scaling rectified flow transformers for high-resolution image*
379 *synthesis in Proceedings of the 41st International Conference on Machine*
380 *Learning* (JMLR.org, Vienna, Austria, 2024).
- 381 3. Lawton, G. *What is Fréchet inception distance (FID)?* [https://www.](https://www.techtarget.com/searchenterpriseai/definition/Frechet-inception-distance-FID)
382 [techtarget.com/searchenterpriseai/definition/Frechet-inception-](https://www.techtarget.com/searchenterpriseai/definition/Frechet-inception-distance-FID)
383 [distance-FID](https://www.techtarget.com/searchenterpriseai/definition/Frechet-inception-distance-FID) (2025).
- 384 4. Deng, J. *et al.* *ImageNet: A large-scale hierarchical image database in 2009*
385 *IEEE Conference on Computer Vision and Pattern Recognition* (2009),
386 248–255.
- 387 5. Yu, Y., Zhang, W. & Deng, Y. *Frechet Inception Distance (FID) for Eval-*
388 *uating GANs* Sept. 2021.
- 389 6. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning
390 applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324
391 (1998).
- 392 7. Jayasumana, S. *et al.* *Rethinking FID: Towards a Better Evaluation Metric*
393 *for Image Generation in 2024 IEEE/CVF Conference on Computer Vision*
394 *and Pattern Recognition (CVPR)* (IEEE, June 2024), 889.
- 395 8. Nunn, E. J., Khadivi, P. & Samavi, S. Compound Frechet Inception Dis-
396 tance for Quality Assessment of GAN Created Images. *CoRR* **abs/2106.08575**.
397 arXiv: 2106.08575 (2021).
- 398 9. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge.
399 *International Journal of Computer Vision (IJCV)* **115**, 211–252 (2015).

- 400 10. Konz, N. *et al.* Fréchet Radiomic Distance (FRD): A Versatile Metric for
401 Comparing Medical Imaging Datasets. *arXiv preprint arXiv:2412.01496*
402 (2024).
- 403 11. Wu, Y. *et al.* A Pragmatic Note on Evaluating Generative Models with
404 Fréchet Inception Distance for Retinal Image Synthesis. *arXiv preprint*
405 *arXiv:2502.17160* (2025).
- 406 12. Truong, T., Mohammadi, S. & Lenga, M. *How transferable are self-supervised*
407 *features in medical image classification tasks?* in *Machine Learning for*
408 *Health* (2021), 54–74.
- 409 13. Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T. & Lehtinen, J. The
410 role of imagenet classes in Fréchet inception distance. *arXiv:2203.06026*
411 (2022).
- 412 14. Bench, C., Desai, V., Roozemon, C., Engen, R. v. & Thomas, S. A. *In-*
413 *vestigation into using stochastic embedding representations for evaluating*
414 *the trustworthiness of the Fréchet Inception Distance* arXiv:2601.21979
415 [cs.LG]. Jan. 2026. <http://arxiv.org/abs/2601.21979> (2026).
- 416 15. Bińkowski, M., Sutherland, D. J., Arbel, M. & Gretton, A. *Demystifying*
417 *MMD GANs* en. Jan. 2018. <https://arxiv.org/abs/1801.01401v5>
418 (2026).
- 419 16. Chong, M. J. & Forsyth, D. *Effectively Unbiased FID and Inception Score*
420 *and where to find them* arXiv:1911.07023 [cs.CV]. June 2020. [http://](http://arxiv.org/abs/1911.07023)
421 arxiv.org/abs/1911.07023 (2026).
- 422 17. Chan, D. A. & Sithungu, S. P. *Evaluating the Suitability of Inception Score*
423 *and Fréchet Inception Distance as Metrics for Quality and Diversity in Im-*
424 *age Generation in Proceedings of the 2024 7th International Conference on*
425 *Computational Intelligence and Intelligent Systems* (Association for Com-
426 puting Machinery, 2025), 79–85. ISBN: 9798400717437. [https://doi.org/](https://doi.org/10.1145/3708778.3708790)
427 [10.1145/3708778.3708790](https://doi.org/10.1145/3708778.3708790).
- 428 18. Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T. & Lehtinen, J. *The*
429 *Role of ImageNet Classes in Fréchet Inception Distance* arXiv:2203.06026
430 [cs]. Feb. 2023. <http://arxiv.org/abs/2203.06026> (2025).
- 431 19. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking
432 the Inception Architecture for Computer Vision, 2818–2826 (2016).
- 433 20. Veeramacheneni, L., Wolter, M., Kuehne, H. & Gall, J. *Fréchet Wavelet*
434 *Distance: A Domain-Agnostic Metric for Image Generation in Interna-*
435 *tional Conference on Learning Representations* (eds Yue, Y., Garg, A.,
436 Peng, N., Sha, F. & Yu, R.) **2025** (2025), 41235–41265.
- 437 21. Krizhevsky, A. & Hinton, G. *Learning multiple layers of features from tiny*
438 *images* tech. rep. 0 (University of Toronto, Toronto, Ontario, 2009). [https:](https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf)
439 [//www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf](https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf).

- 440 22. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 Dataset: A
441 Large Collection of Multi-Source Dermatoscopic Images of Common Pig-
442 mented Skin Lesions. *CoRR* **abs/1803.10417**. arXiv: 1803.10417. <http://arxiv.org/abs/1803.10417> (2018).
443
- 444 23. Yang, J. *et al.* MedMNIST v2 - A large-scale lightweight benchmark for
445 2D and 3D biomedical image classification. *Scientific data* **10** (2023).
- 446 24. LRCB. *Dutch Expert Centre for Screening* <https://www.lrcb.nl/en/>
447 (2026).

448 **Figure 1: Experimental Setup** Overview of experimental setup for the
449 Class-wise and All Sample FID calculations

450 **Figure 2: FID Dependence on Sample Size Effect** FID values (top row),
451 and the condition (middle row) and rank (bottom row) of the covariance ma-
452 trix in equation (1) dependence on the number of images in each subset for the
453 different datasets (columns) in this study. Purple lines denote FID for the class-
454 wise experiments (each line represents a different class) and black lines are for
455 the All Samples experiment (classes are merged and equally represented in each
456 subset). The vertical dashed line indicates where the number of images equals
457 the size of the covariance matrix in equation (1). Errorbars represent standard
458 deviation over the independent runs. Note the log x-axis for the second column
459 (CIFAR-10 dataset) due to the large range of sample sizes.

460 **Figure 3: FID dependence on inverse sample size** FID dependence on
461 inverse sample size on a log-log plot for the different datasets (columns) in this
462 study. Purple lines denote FID for the class-wise experiments (each line repre-
463 sents a different class) and black lines are for the All Samples experiment (classes
464 are merged and equally represented in each subset). The vertical dashed line
465 indicates where the number of images equals the size of the covariance matrix
466 in equation (1). Errorbars represent standard deviation over the independent
467 runs.

468 **Table 1: Extrapolation of Results** Extrapolating FID versus $1/N$, $N \rightarrow \infty$
469 (Fig. 5) to $\text{FID}(N^*)$, where N^* is 10^9 or 10^{10} samples depending on dataset,
470 as an estimation for the unbiased FID value $\text{FID}^* \approx \text{FID}(N^*)$. $N_{\max}$ is the
471 maximum set size used for a dataset and α is the fitted gradient on the log log
472 scale plots. Calculations are quoted with 95 % bootstrap confidence intervals
473 (CI). [†]PathMNIST values are averaged over the class-wise FID lines