Introducing SDICE: An Index for Assessing Diversity of Synthetic Medical Datasets

Mohammed Talha Alam
mohammed.alam@mbzuai.ac.ae
Raza Imam
raza.imam@mbzuai.ac.ae
Mohammad Areeb Qazi
mohammad.qazi@mbzuai.ac.ae
Asim Ukaye
asim.ukaye@mbzuai.ac.ae
Karthik Nandakumar
karthik.nandakumar@mbzuai.ac.ae

Mohamed bin Zayed University of Artificial Intelligence,
Abu Dhabi,
United Arab Emirates

Abstract

Advancements in generative modeling are pushing the state-of-the-art in synthetic medical image generation. These synthetic images can serve as an effective data augmentation method to aid the development of more accurate machine learning models for medical image analysis. While the fidelity of these synthetic images has progressively increased, the diversity of these images is an understudied phenomenon. In this work, we propose the SDICE index, which is based on the characterization of similarity distributions induced by a contrastive encoder. Given a synthetic dataset and a reference dataset of real images, the SDICE index measures the distance between the similarity score distributions of original and synthetic images, where the similarity scores are estimated using a pre-trained contrastive encoder. This distance is then normalized using an exponential function to provide a consistent metric that can be easily compared across domains. Experiments conducted on the MIMIC-chest X-ray and ImageNet datasets demonstrate the effectiveness of SDICE index in assessing synthetic medical dataset diversity.

1 Introduction

The limited size of medical imaging datasets is often a major roadblock in the development of accurate deep neural network (DNN) models for such domains. While datasets like ImageNet [7], MS-COCO [13], and LAION-400M [19] have been instrumental in the advancement of DNN models, the high costs and expertise required for medical image collection and annotation inhibit the curation of such large-scale medical datasets. Patient privacy concerns and strict regulations such as GDPR [1] and HIPAA [14] further impede the sharing of routine medical datasets within the research community. Latent Diffusion Models (LDMs) such as Stable Diffusion [16] generate high-fidelity synthetic images conditioned on text prompts.

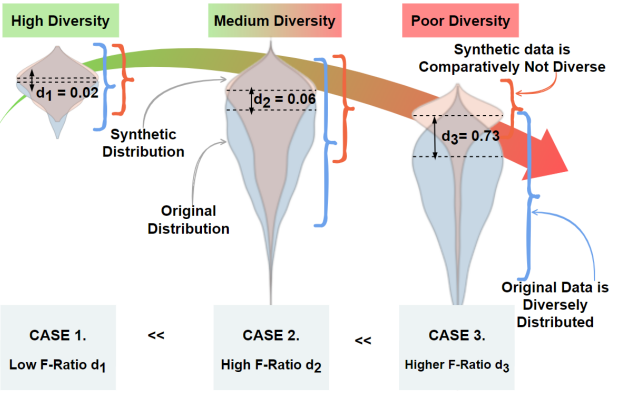

Figure 1: F-ratio between the similarity score distribution of real and synthetic datasets serves as a good **indication** of the diversity within the synthetic dataset.

Several implementations of Stable Diffusion have been proposed in the medical domain including RoentGen [4] for Chest X-ray (CXR) generation, Medical Diffusion [12], and Brain Imaging Generation [15] for MRI and CT image generation. While these works show that high-fidelity synthetic images can be generated, *the ability of these image generation tools to produce synthetic datasets that encompass possible real-world variations is questionable*. Diversity of a synthetic dataset can be broadly defined as the spectrum of features, styles, and semantic variations contained in the generated images. Ensuring diversity in synthetic datasets is essential, as insufficient diversity can impair a model's generalization to real data [10]. Diversity of synthetic datasets is typically evaluated using the Multi-Scale Structural Similarity Index (MS-SSIM) [21] score. A lower average MS-SSIM score is considered as a proxy for good dataset diversity. While the MS-SSIM score allows for an objective diversity assessment of synthetic datasets, it has inherent limitations. Firstly, it is computed at the image level and then extrapolated to the dataset by estimating first and/or second-order statistics. Secondly, it is typically not normalized, which implies that its absolute value is not very meaningful. However, it is still useful for relative comparisons between two competing datasets or methods. [9] shows that the MS-SSIM is a poor indicator of diversity in CXR generation and yields inconclusive correlations. It must be emphasized that the MS-SSIM score was originally formulated as an analytical way to assess the quality of digital pictures by assessing structural similarity.

In this work, we propose a novel approach for diversity quantification of synthetic datasets. Given a sufficiently-diverse reference dataset of real images and a synthetic dataset, it is possible to analyze whether the variations in the synthetic dataset match or exceed those observed in the reference dataset, as shown in Figure 1. Specifically, we characterize the observed variations in a synthetic dataset by analyzing the similarity distributions between images of the same class (intra-class) and images from different classes (inter-class). We assume that the similarity scores are computed based on a contrastive encoder, which is pre-trained to be invariant under different affine/photometric transformations of the same image. We hypothesize that benchmarking of intra- and inter-class synthetic similarity distributions against their counterparts based on a reference dataset is a good proxy for diversity. Based on this hypothesis, we make the following contributions:

- We propose a dataset-level diversity assessment index called SDICE, which characterizes Similarity Distributions Induced by a Contrastive Encoder.

- While the concept of SDICE is generic, we also propose a specific instantiation of the SDICE index using F-ratio as the distance between two distributions and applying an exponential normalization to the resulting distance.

- We demonstrate the utility of the SDICE index by applying it to synthetic datasets generated from two models: (i) RoentGen trained on MIMIC-CXR (Chest X-ray) images and (ii) Stable Diffusion trained on natural images. Our analysis indicates that the generated synthetic CXR dataset has low diversity, especially failing to capture variations within the same class.

Related Work: Saad et al. [18] show high variance in results when assessing the intra-class diversity of generated images in medical and non-biomedical domains using MS-SSIM and cosine distance. Friedman et. al. [8] argue that existing metrics for measuring diversity are often domain-specific and limited in flexibility and propose the Vendi score. They show that even models that capture all the modes of a labeled dataset can be less diverse than the real dataset. Alaa et al. [2] introduce a 3-d metric that characterizes the fidelity, diversity, and generalization performance of any generative model. They quantify diversity in the feature space, while our SDICE index operates in the similarity space. This distinction makes our method more robust, as it captures diversity across all clusters, unlike β-Recall, which is sensitive to the value of β and may struggle with highly multimodal distributions.

2 Proposed SDICE Index

The key intuition underlying the proposed SDICE index is that a synthetic dataset can be considered to have good diversity if the variations in this dataset closely follow or exceed the variations observed in a reference dataset containing sufficiently-diverse real images. Figure 2 provides an overview of the architecture of our proposed SDICE index. However, two main challenges need to be overcome to determine if two datasets (synthetic and real) have similar variations. 1) The variations in a dataset can be caused due to many reasons such as image noise and within and between class differences, and it is essential to capture these variations individually. 2) A good metric is required to capture pair-wise similarities between the images.

Problem Statement: Let, $\mathcal{D}^s = \{\mathbf{x}_i^s, y_i^s\}_{i=1}^{N^s}$ be a synthetic image dataset with N^s samples, where $\mathbf{x} \in \mathcal{R}^{H\times W\times C}$ is an input image (H, W, and C represent the height, width, and number of channels in the input image, respectively) and $y \in \{1,2,\cdots,M\}$ is the class label. Similarly, let $\mathcal{D}^r = \{\mathbf{x}_j^r, y_j^r\}_{j=1}^{N^r}$ be a real image dataset containing N^r samples. Let $\mathcal{F} : \mathcal{R}^{H\times W\times C} \rightarrow \mathcal{R}^{dim}$ be a pre-trained feature extractor that outputs a fixed-length embedding for a given image. Let $\mathcal{S} : \mathcal{R}^{dim} \times \mathcal{R}^{dim} \rightarrow \mathcal{R}$ be a similarity metric that outputs the similarity between two feature embeddings. Given a synthetic dataset $\mathcal{D}^s$, a reference real dataset $\mathcal{D}^r$, a feature extractor $\mathcal{F}$, and a similarity metric $\mathcal{S}$, the goal is to compute a diversity index $\gamma \in [0,1]$ that indicates if the two datasets $\mathcal{D}^s$ and $\mathcal{D}^r$ have similar variations. A higher value of γ indicates better diversity.

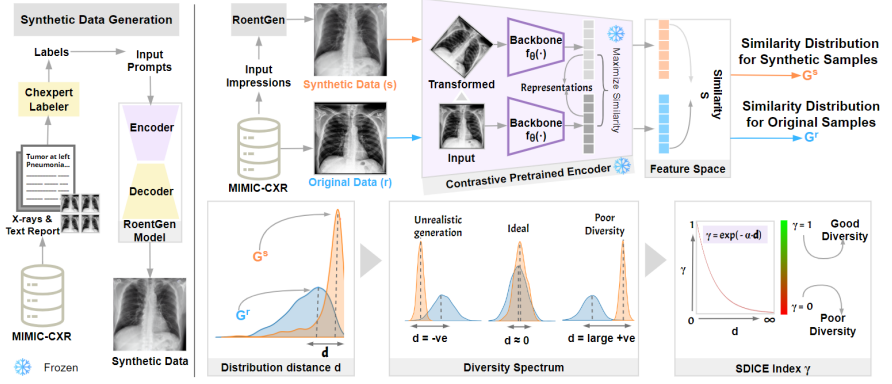

Figure 2: **Overview of the proposed SDICE index.** We input the real and synthetic dataset to the contrastive pretrained encoder to obtain similarity score distributions. The F-ratio between the two distributions after exponential normalization can be used to assess the diversity of the synthetic dataset.

2.1 Generic SDICE Index

The variations in a dataset can be broadly categorized into two types:

1. **Intra-class Variations** (*intra*): These are variations between input images belonging to the same class, e.g., differences between CXR images of the same underlying condition. A well-trained feature extractor will produce embeddings that have high similarity for images of the same class.

2. **Inter-class Variations** (*inter*): These are variations between input images belonging to different classes, e.g., differences between CXR images of patients having different diseases. A good feature extractor will learn to amplify these variations and produce embeddings that have lower similarity scores.

To characterize the above types of variations, we employ the following approach. Let $\mathbf{x}_a^*$ and $\mathbf{x}_b^*$ be a pair of images randomly drawn from dataset $\mathcal{D}^*$, where $* \in \{s, r\}$. Let $S_{ab}^* = \mathcal{S}(\mathcal{F}(\mathbf{x}_a^*), \mathcal{F}(\mathbf{x}_b^*))$ be the similarity between two input images drawn from $\mathcal{D}^*$. Let $\mathcal{G}_{intra}^*$ be the probability distribution of S_{ab}^* when $[y_a^* = y_b^*]$ (i.e., distribution of similarity scores between images of the same class). Similarly, let $\mathcal{G}_{inter}^*$ be the probability distribution of S_{ab}^* when $[y_a^* \neq y_b^*]$ (i.e., distribution of similarity scores between images of different classes). Let $\mathbb{Q} : \mathcal{G} \times \mathcal{G} \rightarrow [0, \infty]$ be a distance measure between two probability distributions. Specifically, let $\mathbb{Q}(\mathcal{G}_0 || \mathcal{G}_1)$ be the distance of a probability distribution $\mathcal{G}_0$ from another distribution $\mathcal{G}_1$. We can compute $d_{intra} = \mathbb{Q}(\mathcal{G}_{intra}^s || \mathcal{G}_{intra}^r)$ and $d_{inter} = \mathbb{Q}(\mathcal{G}_{inter}^s || \mathcal{G}_{inter}^r)$. Larger values of d_{intra} (d_{inter}) indicate that the synthetic similarity distribution is highly dissimilar to the real similarity distribution, indicating low intra-class (inter-class) diversity. On the other hand, smaller values of d_{intra} or d_{inter} indicate good alignment between the score distributions, representing high diversity. Therefore, the diversity index γ should be inversely proportional to the above distance values. One limitation of the above distance values is their unbounded and unnormalized nature, which makes it difficult to interpret these values across domains. To address this issue, we introduce a normalization function $\mathcal{H} : [0, \infty] \rightarrow [0, 1]$

with parameter α to obtain the SDICE index as:

$$\gamma_{intra} = \mathcal{H}_{\alpha}(d_{intra}) \tag{1}$$

$$\gamma_{inter} = \mathcal{H}_{\alpha}(d_{inter}) \tag{2}$$

The tuple $\texttt{SDICE} := (\gamma_{intra}, \gamma_{inter})$ can be used to assess the diversity of a synthetic dataset. If a single diversity index is desired, γ can be defined as:

$$\gamma = \sqrt{\gamma_{intra}^2 + \gamma_{inter}^2} \tag{3}$$

Note that higher values of γ indicate better diversity.

2.2 Practical Implementation of SDICE Index

Four critical design choices must be made to practically implement the proposed SDICE index: (i) feature extractor $\mathcal{F}$, (ii) similarity function $\mathcal{S}$, (iii) probability distance measure $\mathbb{Q}$, and (iv) normalization function $\mathcal{H}$. Before making these design choices, one needs to understand the worst-case scenario for the diversity of a synthetic dataset. A synthetic dataset can be considered to have negligible diversity if all the generated images are either exact copies of each other or minor geometric and/or photometric variations of one another. In this worst-case scenario, the combination of feature extractor and similarity metric must result in a very high similarity value for any pair of images drawn from such a low diversity dataset. This can be achieved by training the feature extractor $\mathcal{F}$ in a self-supervised contrastive manner [5], where different minor transformations (augmentations) of the same image are forced to produce identical feature vectors. This explains our choice of a pre-trained contrastive encoder for $\mathcal{F}$. Since cosine similarity is typically employed in such contrastive encoders, we also choose cosine similarity as the default similarity function.

One common approach to estimate the distance between two probability distributions is to fit parametric density functions based on the available samples and calculate standard probability distance measures. However, for the sake of simplicity, we compute F-ratio between the distributions. Given two distributions $\mathcal{G}_0$ and $\mathcal{G}_1$, whose mean (μ_0 and μ_1, respectively) and standard deviation (σ_0 and σ_1, respectively) values are known, the F-ratio can be computed as:

$$\mathbb{Q}(\mathcal{G}_0||\mathcal{G}_1) = \text{F-ratio}(\mathcal{G}_0, \mathcal{G}_1) = \frac{(\mu_1 - \mu_0)^2}{\sigma_1^2 + \sigma_0^2} \tag{4}$$

Since true mean and standard deviation values are unavailable, we estimate them from the available similarity scores. It is also possible to use other distance measures such as the Earth Mover's Distance (EMD) [17], which is defined as:

$$\mathbb{Q}(\mathcal{G}_0||\mathcal{G}_1) = \text{EMD}(\mathcal{G}_0, \mathcal{G}_1) = \inf_{\nu \in \Gamma(\mathcal{G}_0, \mathcal{G}_1)} \int_{U \times V} |u - v| \, d\nu(u, v) \tag{5}$$

where Γ is the set of all joint distributions ν whose marginals are $\mathcal{G}_0$ and $\mathcal{G}_1$. Finally, the normalization function $\mathcal{H}$ is selected as follows. Following the earlier discussion, in the worst-case scenario, the intra-class similarity distributions of the synthetic dataset would be close to that of the similarity distribution between images that are minor transformations of each other. Let h denote a minor random transformation that can be applied to a real input

image $\mathbf{x}_a^r$ to obtain a transformed image $\mathbf{x}_{a'}^r = h(\mathbf{x}_a^r)$. Let $S_{aa'}^r = \mathcal{S}(\mathcal{F}(\mathbf{x}_a^r), \mathcal{F}(\mathbf{x}_{a'}^r))$ be the similarity between a real input image and its transformed version. Let $\mathcal{G}_{trans}^r$ be the probability distribution of similarity score between a real image and its transformed counterpart. Finally, let $d_{max} = \mathbb{Q}(\mathcal{G}_{intra}^s || \mathcal{G}_{trans}^r)$. When d_{intra} or d_{inter} is closer to d_{max}, it indicates poor diversity. Therefore, we can employ the following exponential normalization function.

$$\mathcal{H}_\alpha(d) = \exp\left(\ln(\alpha)\frac{d}{d_{max}}\right) \tag{6}$$

The above normalization function ensures that when $d \approx d_{max}$, $\mathcal{H}_\alpha(d) \approx \alpha$ and $\mathcal{H}_\alpha(d) \to 0$ when $d \to 0$. Here, α is usually set to a small value, say 10^{-4}.

We use the samples from the given datasets $\mathcal{D}^s$ and $\mathcal{D}^r$ to empirically estimate the required distributions. For example, in the intra-class scenario, we select n samples from a class and compute the similarities between all possible $n(n-1)/2$ pairs to obtain $\mathcal{G}_{intra}^*$. Since there are M classes in the dataset, the number of possible similarity scores for the inter-class distribution will be $n^2 * (M^2 - M)$. Finally, to estimate $\mathcal{G}_{trans}^r$, we obtain a total of nk similarity values for each class, where k is the number of random transformations applied per image.

3 Experimental Results

Datasets and Generators: We use the MIMIC-CXR dataset, comprising 377,110 CXRs and associated reports, selecting representative samples from subsets p11, p12, and p13. The 'impression' sections of the reports were analyzed with the CheXpert labeler to generate 14 diagnostic labels. We generated synthetic CXRs using the RoentGen [4] with prompts crafted from CheXpert labels. Additionally, we matched 14 ImageNet classes with MIMIC-CXR classes for a broader evaluation, generated using UniDiffuser [3] (Figure 8). Our experiments included three distinct prompt types for each dataset to investigate their impact on synthetic image quality and relevance. For MIMIC-CXR, P_1 = '`CLS`', P_2 = *'An image of a chest x-ray showing* `CLS`', and P_3 = *'A realistic image of a chest x-ray showing* `CLS`', where `CLS` is the class name. For ImageNet, P_1 = '`CLS`', P_2 = *'An image of a* `CLS`', and P_3 = *'A realistic image of a* `CLS`'.

Feature Extractor and Similarity Function: We employed a ResNet50 backbone pre-trained on CXR using self-supervised contrastive learning for computing pairwise embeddings [6]. For ImageNet, a pre-trained ResNet50 was utilized to leverage its strong representation capabilities. Since our aim was to obtain representative embeddings, not to train or test the model, the potential bias concern is mitigated. The embeddings were used solely for similarity evaluation. Classifiers trained on synthetic ImageNet samples showed a significant drop in accuracy when tested on real data. Notably, there was a fourfold drop in accuracy when classifiers trained on synthetic CXRs were tested on real CXRs. This highlights the importance of evaluating the diversity of synthetic datasets. Our `SDICE` index correlates with these performance declines, offering valuable insights into the diversity of synthetic samples for downstream tasks.

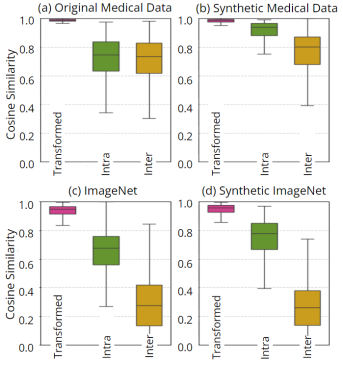

Figure 3: Qualitative analysis of distribution change across cases

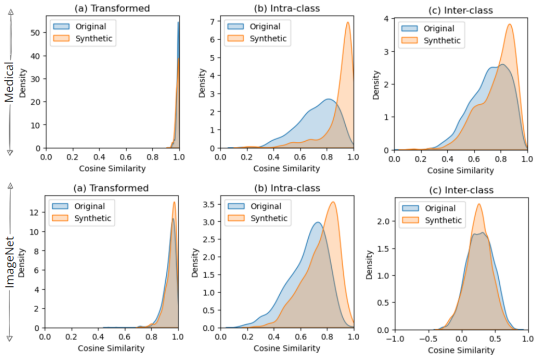

Figure 4: Distribution variation and overlap between real and synthetic samples

Table 1: SDICE index using F-ratio and EMD for intra (γ_{intra}) and inter (γ_{inter}) cases, along with the influence of sample size (a) and prompt type (b) on the SDICE index. Additional results on FairFace dataset [11] are provided in the supplementary material.

Dataset	γ	SDICE index (γ)		Sample size			Prompt type		
		F-ratio	EMD	n	$2n$	$4n$	P_1	P_2	P_3
MIMIC [9]	γ_{intra}	0.11	0.01	0.26	0.36	0.47	0.63	0.36	0.37
	γ_{inter}	0.84	0.44	0.74	0.84	0.99	0.91	0.84	0.89
ImageNet [7]	γ_{intra}	0.47	0.26	0.37	0.47	0.56	0.47	0.83	0.89
	γ_{inter}	0.98	0.74	0.98	0.99	0.99	0.80	0.95	0.98

3.1 Diversity evaluation

Firstly, we observe that $\mathcal{G}^r_{trans}$ consistently hovers close to 1.0 (see Figure 3), which is expected because the feature extractor and similarity computation are designed to ignore differences between an image and its transformed version. However, intra-class variations depict a greater range in $\mathcal{D}^r$ than in $\mathcal{D}^s$ for CXRs, whereas inter-class variations present comparable extents across both distributions, as shown in Figure 4. To quantitatively assess these observations, we computed the SDICE index using both F-ratio and EMD for each case under study. As detailed in Table 1, the γ_{intra} for CXRs is significantly lower, indicating a lack of intra-class diversity compared to those of ImageNet. An analysis using different sample sizes and prompts is also presented in Table 1 and discussed in detail in section 3.3.

Further investigation of γ_{intra} was done by measuring the diversity within individual classes of both datasets as shown in Figure 5 and detailed in Table 2. We observe that several classes in the MIMIC-CXR synthetic dataset do not have the same range of diversity as its real counterpart. We observe poor diversity in classes with niche domain-specific names (such as 'Atelectasis' and 'Enlarged Cardiomegaly') as opposed to more general ones ('Pneumonia' and 'Fracture'). We hypothesize that the generative model [9] possibly fails to capture the true variations within the esoteric classes due to limited training. Inter-class variation in the generated data is similar to that of the reference data for both datasets. To confirm the observed trend in γ_{inter} across datasets, we divided the dataset into 'q' subsets where 'q' is the number of classes in the dataset. Each subset gradually included more

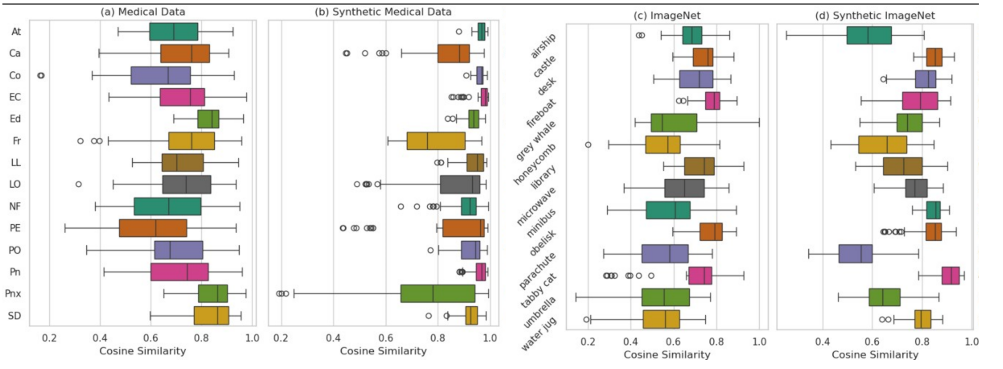

Figure 5: **Qualitative analysis of distribution differences between the real and synthetic samples in terms of individual classes.** (a) and (b) depict the class-wise distributions for the MIMIC-CXR dataset, while (c) and (d) illustrate the same for the 14 classes of ImageNet.

classes, starting with only the first class in the initial subset and progressing to the last set with all available classes. Our findings (Figure 6) reveal a gradual increase in γ_{inter}, indicating growing diversity. However, this upward trend reaches a saturation point suggesting that inter-class diversity does not increase beyond a certain threshold of class inclusion.

3.2 Comparison with SSIM and FID

A further analysis was conducted to highlight the variation obtained by the SDICE index as compared to the SSIM and FID scores (Supp: Table 3). Our analysis shows mean FID scores hover around 0.0082 and 0.0099 for intra and inter-class distributions, respectively. The FID score shows poor resolution as compared to the SDICE index, as the latter benefits from a domain-specific contrastive encoder. Similarly, SSIM values also fail to provide a clear separation between intra and inter-class diversity with mean SSIM scores of 0.68 and 0.60, respectively. The SDICE index effectively highlights the contrast in diversity across intra and inter-class categories, providing clear insights that FID and SSIM metrics may overlook. This highlights the benefits of SDICE index in domain-specific dataset analysis.

3.3 Ablation Studies

3.3.1 Impact of number of samples on γ_{intra}

Table 1(a) outlines how γ_{intra} values evolve as we increase sample sizes from n to $2n$, and further to $4n$. This progression reveals that γ_{intra}, or the measure of diversity within classes, tends to rise with larger sample sets. Observed in both the MIMIC and ImageNet datasets, this trend suggests that expanding the dataset by introducing a wider variety of examples within each class enhances the overall diversity. The initial sample size was 350 for both MIMIC-CXR and ImageNet datasets, meaning 25 images per class. We found that a balanced sample size yielded better results in terms of diversity assessment compared to imbalanced samples.

MIMIC-CXR		ImageNet	
Class	γ_{intra}	Class	γ_{intra}
At	**1.56e-08**	airship	3.53e-01
Ca	4.29e-01	castle	1.19e-03
Co	1.68e-05	desk	8.07e-02
EC	1.43e-04	**fireboat**	**9.97e-01**
Ed	1.74e-03	grey whale	9.01e-02
Fr	**7.93e-01**	honeycomb	3.05e-01
LL	3.25e-05	library	9.94e-01
LO	3.91e-01	microwave	1.06e-01
NF	3.38e-03	**minibus**	**5.52e-05**
PE	5.24e-02	obelisk	5.32e-01
PO	9.95e-04	parachute	9.54e-01
Pn	3.14e-04	tabby cat	4.53e-03
Pnx	5.56e-01	umbrella	3.21e-01
SD	1.10e-01	water jug	1.36e-04

Table 2: γ_{intra} values where 'Atelectasis' in **MIMIC-CXR** exhibits the least diversity, while 'Fracture' demonstrates the highest diversity. In **ImageNet**, 'minibus' class has the least diversity, and 'fireboat' stands out as the most diverse class.

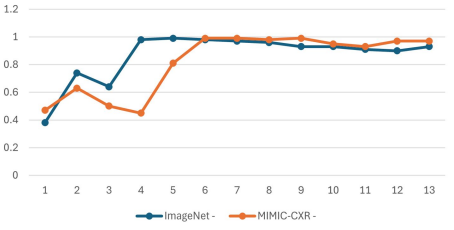

Figure 6: **Progression of γ_{inter} with increasing class inclusion.** This illustrates the increase in diversity within MIMIC-CXR and ImageNet as more classes are added, leveling off to indicate a maximum inter-class diversity threshold.

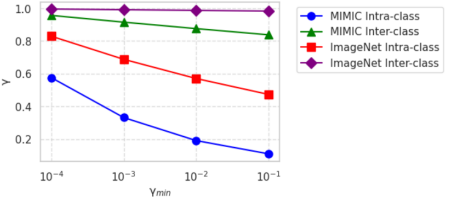

Figure 7: **SDICE Index Variation with γ_{min} in MIMIC-CXR and ImageNet.** This displays a marked decrease in MIMIC-CXR's intra-class diversity with increasing γ_{min}, in contrast to ImageNet's consistent inter-class diversity.

3.3.2 Impact of different prompts on γ_{intra}

Table 1(b) shows that the complexity of prompts affects the diversity of the generated images. In the case of CXR images, less detailed prompts, such as P_1, appear to encourage a wider diversity, perhaps due to the generative model having broader interpretative freedom. For ImageNet, descriptive prompts such as P_3 lead to more diverse outputs, which implies that the detailed nature of these prompts provides useful guidance to the model, enabling it to capture the extensive variability inherent across ImageNet's classes. This suggests that the level of detail in prompts should be carefully considered to match the desired diversity of the dataset being synthesized.

3.4 Parameter Sensitivity Analysis

We investigate how variations in the parameter γ_{min} affect the SDICE index (γ) in both the MIMIC-CXR and ImageNet datasets. Figure 7 illustrates the sensitivity of the SDICE index to changes in γ_{min}. The figure reveals a clear downward trend in intra-class scenarios for both ImageNet and MIMIC-CXR, indicating lower intra-class diversity compared to inter-

class diversity. Notably, MIMIC-CXR consistently exhibits significantly lower intra-class diversity values compared to ImageNet across different γ_{min} values. In the inter-class scenario, MIMIC-CXR shows a more pronounced downward trend compared to the stable and consistently diverse inter-class diversity observed in ImageNet (around $\gamma = 1.0$) across various γ_{min} values. This emphasizes how the SDICE index is sensitive to parameter changes, revealing distinct diversity characteristics within datasets.

4 Conclusion

This work introduced the SDICE index for evaluating the diversity of synthetic medical image datasets. Leveraging the power of contrastive encoders, the SDICE index characterizes the similarity distributions observed in the reference and synthetic datasets and provides a normalized measure to assess and compare dataset variability. Our experiments on MIMIC-CXR and ImageNet confirm its efficacy, revealing particularly low diversity in synthetic CXRs, highlighting areas where generative models may need refinement. Moving forward, we will focus on reducing the computational complexity of the similarity score computation by exploring more efficient methods, such as approximate nearest neighbors. These improvements aim to enhance the scalability and practicality of our approach, further solidifying the SDICE index in the evaluation of synthetic medical image datasets.

References

[1] General data protection regulation (gdpr). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679. Accessed on November 17, 2023.

[2] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.

[3] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. *arXiv preprint arXiv:2303.06555*, 2023.

[4] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] Kyungjin Cho, Ki Duk Kim, Yujin Nam, Jiheon Jeong, Jeeyoung Kim, Changyong Choi, Soyoung Lee, Jun Soo Lee, Seoyeon Woo, Gil-Sun Hong, et al. Chess: Chest x-ray pre-trained model via self-supervised contrastive learning. *Journal of Digital Imaging*, pages 1–9, 2023.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

[9] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[10] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

[11] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[12] Firas Khader, Gustav Mueller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baessler, Sebastian Foersch, et al. Medical diffusion–denoising diffusion probabilistic models for 3d medical image generation. *arXiv preprint arXiv:2211.03364*, 2022.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[14] U.S. Department of Health & Human Services. Health insurance portability and accountability act (hipaa) of 1996. *U.S. Department of Health & Human Services*, 1996. Available at https://www.hhs.gov/hipaa/index.html.

[15] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 117–126. Springer, 2022.

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[17] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121, 2000.

[18] Muhammad Muneeb Saad, Mubashir Husain Rehmani, and Ruairi O'Reilly. Assessing intra-class diversity and quality of synthetically generated images in a biomedical and non-biomedical setting. *arXiv preprint arXiv:2308.02505*, 2023.

[19] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021.

[20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. Scipy 1.7.3. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html, 2023. [Online; accessed November 20, 2023].

[21] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.