Privacy-preserving datasets by capturing feature distributions with Conditional VAEs

Francesco Di Salvo
francesco.di-salvo@uni-bamberg.de
David Tafler
david-elias.tafler@stud.uni-bamberg.de
Sebastian Doerrich
sebastian.doerrich@uni-bamberg.de
Christian Ledig
christian.ledig@uni-bamberg.de

xAILab Bamberg
University of Bamberg
Bamberg, Germany

Abstract

Large and well-annotated datasets are essential for advancing deep learning applications, however often costly or impossible to obtain by a single entity. In many areas, including the medical domain, approaches relying on data sharing have become critical to address those challenges. While effective in increasing dataset size and diversity, data sharing raises significant privacy concerns. Commonly employed anonymization methods based on the *k*-anonymity paradigm often fail to preserve data diversity, affecting model robustness. This work introduces a novel approach using Conditional Variational Autoencoders (CVAEs) trained on feature vectors extracted from large pre-trained vision foundation models. Foundation models effectively detect and represent complex patterns across diverse domains, allowing the CVAE to faithfully capture the embedding space of a given data distribution to generate (sample) a diverse, privacy-respecting, and potentially unbounded set of synthetic feature vectors. Our method notably outperforms traditional approaches in both medical and natural image domains, exhibiting greater dataset diversity and higher robustness against perturbations while preserving sample privacy. These results underscore the potential of generative models to significantly impact deep learning applications in data-scarce and privacy-sensitive environments. The source code is available at github.com/francescodisalvo05/cvae-anonymization.

1 Introduction

Over the past decade, Deep Neural Networks have made significant progress, impacting a wide range of industries. A key driver of this progress has been the increasing availability of large, well-annotated datasets. However, such datasets are rare in several domains, such as medical image analysis [1]. While data-sharing between institutions offers a potential solution, it also raises privacy concerns over *personal data*. To this extent, data privacy has been a focal point for decades.

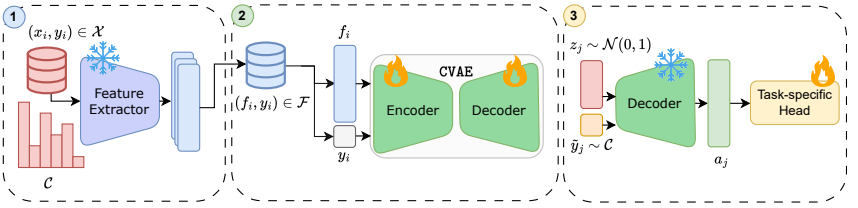

Figure 1: Illustration of our proposed anonymization approach. Given an image dataset $(x_i, y_i) \in \mathcal{X}$ with categorical class distribution $\mathcal{C}$, we first utilize a large pre-trained model to extract and store feature embeddings and corresponding labels $(f_i, y_i) \in \mathcal{F}$. These embeddings capture both local and contextual information while inherently reducing dimensionality. Subsequently, the embeddings are used during training of a Conditional Variational Autoencoder (CVAE) to capture the training distribution conditioned on the respective class labels y_i. Finally, we train a task-specific head while dynamically generating new synthetic feature vectors a_j conditioned on class labels $\tilde{y}_j \sim C$ through CVAE's frozen decoder. This not only ensures data anonymity but also increases data diversity and model robustness.

For instance, the *k-anonymity* paradigm [38] was introduced in 2002 to obscure user data in relational databases, ensuring that no individual could be distinguished from at least $k-1$ others. Building on this idea, the *k*-Same [27] method emerged to address privacy concerns in surveillance technologies. It aggregates disjoint clusters of k data points (*i.e.*, faces) with their centroid, either in the pixel space or in the eigenspace.

With the advent of generative adversarial networks (GANs) [5, 12], new possibilities for generating synthetic images have emerged. However, there is a risk of embedding personal information in the latent space. To address this, methods such as *k*-Same-Net [24] enforced *k*-anonymity within GANs' latent space during training. Furthermore, GANs have struggled with issues such as mode collapse, where models fail to produce diverse outputs, thereby limiting their generalizability [37, 39]. Orthogonally, recent advancements have shifted focus to vision foundation models [2, 29, 35], typically trained in a self-supervised fashion on large unlabeled datasets. These large models using vision transformers [9] excel at capturing high-level complex patterns and contextual information across different domains. Moreover, their robustness against image corruptions and distribution shifts [31] is an established benefit. Researchers are now proposing foundation models for specific domains including healthcare [21, 40, 48], climate [28], geospatial data [22], and more. The rich and low-dimensional feature representation derived from these models makes them a suitable alternative (to the higher dimensional raw images) for pursuing downstream tasks, such as classification.

Building on the potential of those highly complex and informative representations, as illustrated in Figure 1, we train a Conditional Variational Autoencoder (CVAE) on the feature space of a pre-trained foundation model to accurately capture the training distribution. The generative process, conditioned on class labels, effectively mimics the original distribution while enhancing privacy and expanding the diversity of the generated feature vectors. After capturing the training data distribution with our CVAE, we demonstrate a paradigm that no longer relies on disclosing the original images (or their embeddings) but simply the frozen CVAE decoder allowing the sampling of any required dataset size. Despite tremendous benefits with respect to data and patient privacy, this also substantially reduces the required data exchange from an order of gigabytes (images) to a few megabytes (decoder weights).

In summary, our contributions are:

- We propose a novel anonymization paradigm using Conditional Variational Autoencoders (CVAEs) trained on the feature space of foundation models, which provides a more informative and compact representation than the traditional image space.
- We demonstrate superior feature diversity and classification performance with our CVAE compared to reference approaches across both medical and natural datasets. This also highlights the effectiveness of foundation model's feature representation, even in areas beyond its initial training focus.
- We show that the diverse feature representation obtained via dynamic sampling and decoding through CVAE's decoder enhances model robustness against perturbations.

2 Related works

As digital data usage expands, the need for robust privacy measures becomes critical. The k-anonymity framework, proposed by Sweeney [38], addresses this by ensuring that an individual cannot be distinguished from at least $k-1$ others within database entries. This method involves generalizing or suppressing specific identifying attributes such as zip code, gender, and other sensitive information. While effective at masking identities, this approach often leads to significant data loss, reducing the utility of the data. With the rise of video surveillance systems in New York, a seminal work [27] introduced k-Same for face de-identification, translating these principles to more complex scenarios, including the *pixel space* and the *eigenspace*. This is achieved by collapsing groups of k elements into a single representative point, namely the centroid. Despite its application in real-world medical systems [11], the inherited data loss and lower data diversity can substantially affect the generalization [15, 45] and robustness [14, 20] of machine learning based solutions, particularly in domains where data is already scarce.

Moreover, building on the capabilities of GANs to generate high-quality synthetic samples, k-Same-Net [24] applies the k-anonymity paradigm within GANs' latent space for face de-identification. Enhancing this further, k-SALSA [18] recently introduced local style alignment to preserve granular details in retinal images, while PLAN [32] introduces an auxiliary identity classifier to prevent sample collisions. However, GANs still face significant challenges due to their high computational demands, training instability, and mode collapse [37, 39], which can affect their efficiency and reliability.

Before the advent of GANs, Variational Autoencoders (VAEs) have already demonstrated their generative capabilities. For face de-identification, Yang et al. [44] generate realistic yet untraceable images by clustering and re-synthesizing facial features. Taehoon et al. [19] enforce privacy by introducing noise into the latent vectors, obscuring identities before reconstruction. Other than the image domain, VAEs have been employed for anonymizing speaker data [33] and tabular data [28], as well. Conditional Variational Autoencoders (CVAEs) enhance the versatility of VAEs by allowing the generative process to be conditioned on additional labels or attributes, effectively tailoring the output to specific data characteristics. Hajihassnai et al. [16] introduce ObscureNet [16], a CVAE-like architecture to anonymize sensor data, that conditions the latent space on private attributes (*e.g.*, age, gender). The architecture consists of one input encoder, *multiple external discriminators*

(one per attribute), and a conditional decoder. Apart from anonymization, the conditional generative process of CVAEs has been used to augment training data [10, 34], addressing issues such as data scarcity and imbalance.

In contrast, our approach first enhances the conditional generative process by shifting the generation from the pixel space to the *embedding space*, utilizing large pre-trained foundation models. This transition enables a more efficient feature generation, thanks to the more compact and informative feature representation. In contrast to ObscureNet [16], we employ one conditional encoder rather than multiple discriminators, which allows our architecture to scale more efficiently with the number of classes. Furthermore, instead of simply augmenting our training dataset, we focus on *mimicking* its embedding-based training distribution, aiming to replicate a diverse and privacy-aware distribution. We further integrate this generative process during training of downstream tasks, which allows us to dynamically generate new samples until convergence.

3 Method

The proposed approach to capture and anonymize a given training dataset is illustrated in Figure 1. Initially, we describe the process of feature extraction through vision foundation models. This is followed by a description of CVAE's training procedure. We then introduce two anonymization strategies: the first generates a persistent synthetic replica of the training dataset, while the second continuously generates data on the fly for downstream tasks.

3.1 Feature extraction

We employ a large pre-trained model to extract and subsequently store feature embeddings and their associated labels $(f_i, y_i) \in \mathcal{F}$ for each labeled input image $(x_i, y_i) \in \mathcal{X}$, with an associated categorical class distribution $\mathcal{C}$. The extracted feature representations f_i represent the positional information of x_i within the feature space of the selected feature extractor. Here, close points are more likely to share similar image patterns while distant points indicate semantic differences, potentially across different classes. This descriptive property of the feature space is beneficial for effectively capturing the training distribution. It is much less pronounced in the original pixel space, which suffers from inherent information redundancy and the challenge of calculating meaningful distances in high-dimensional spaces. Additionally, from a storage perspective, while a conventional 2D RGB image of dimensions 224×224 occupies approximately 150,000 pixels, the corresponding image embeddings typically have around 1,000 components, reducing data storage by several orders of magnitude. For our experiments, we rely on DINOv2 ViT-B/14 [29], one of the most recent open-source foundation models that achieved state-of-the-art performance in a variety of tasks, demonstrating the high quality of their latent feature representations.

3.2 CVAE training

Conditional Variational Autoencoders extend traditional Variational Autoencoders by integrating specific attributes, such as class labels, into the data generation process. This extension allows the CVAE to not only produce data that closely resembles the original samples in terms of content but also aligns with the conditioned attributes, thereby improving both the relevance and specificity of the generated data. In our setup, the CVAE is trained to accurately

learn and replicate the distribution of a feature dataset $\mathcal{F}$, conditioned on the corresponding labels, i.e., $(f_i, y_i = c)$. This training allows us to subsequently generate new *anonymous feature vectors* a_j for a chosen class $\tilde{y}_j$ that reflect the original training distribution. We can thus utilize only the pre-trained decoder to sample arbitrarily many samples per class c, adhering to the original categorical class distribution $\mathcal{C}$. This approach ensures that the generated data maintains consistency with the original dataset while enhancing privacy and utility, likewise.

3.3 Offline anonymization

The first way to anonymize our dataset is via storing a persistent anonymized replica of the data. To this end, once the `CVAE` is appropriately trained, we generate synthetic feature vectors that reflect the initial data size and class distribution. This anonymization process is described in Algorithm 1 and requires a pre-trained `CVAE` model, a number of synthetic samples N and a categorical distribution $\mathcal{C}$ representing the class probabilities of the original dataset. This ensures the generation of similar class proportions in the synthetic dataset. For each of the desired N data points (step 2), a class label $\tilde{y}_j$ is sampled from $\mathcal{C}$ (step 3). Next, a latent variable z_j is drawn from a standard normal distribution (step 4) to serve as a stochastic input for data generation. These sampled variables $(z_j, \tilde{y}_j)$ are fed into `CVAE`'s decoder, which generates a new data point a_j conditioned on the class label (step 5). Finally, both the generated data point and its label are then stored (step 6). Upon completion, the algorithm outputs the anonymized datasets, consisting of these synthetically generated feature vectors and their labels, effectively balancing privacy preservation with the utility of the dataset for downstream tasks.

Algorithm 1 `anonymize(CVAE,`$\mathcal{C}$`,`N`)`

1: $\mathcal{A} \leftarrow [\,]$ {initialize anonymized dataset}
2: **for** $j = 1, 2, \ldots, N$ **do**
3: $\quad \tilde{y}_j \sim \mathcal{C}$ {sample $\tilde{y}_j$ from class categorical distribution $\mathcal{C}$}
4: $\quad z_j \sim \mathcal{N}(0,1)$ {standard normal sampling}
5: $\quad a_j \leftarrow$ `CVAE.decoder`$(z_j, \tilde{y}_j)$ {conditional decoding}
6: $\quad \mathcal{A}$`.append`$((a_j, \tilde{y}_j))$ {append the generated tuple}
7: **return** $\mathcal{A}$

3.4 Online anonymization

Another way to ensure data anonymization is to eliminate the use of persistent datasets entirely. Our proposed `CVAE` model exemplifies this approach by replacing the traditional dataset with the `CVAE`'s pre-trained decoder. Specifically, we iteratively generate new data on the fly during the training of our task-specific head, meaning no persistent data is required anymore. This method not only overcomes the need to store or send large volumes of sensitive data but also enables the sharing of a simple model (`CVAE`'s decoder) capable of replicating the training data distribution at an arbitrary incidence rate without actual data exchange. Moreover, in the context of model sharing, such as in federated learning [36, 46], our approach offers additional security benefits. While federated learning allows for task-specific models trained on private data to be shared without exchanging the data itself, it is not without risks; model weights can potentially reveal training data characteristics [13, 26].

Our approach iteratively generates anonymized and stochastic data. This can introduce an additional layer of security, further mitigating the risks associated with traditional data- and model-sharing approaches.

4 Experimental results

Initially, we outline the process of feature extraction and the training of our Conditional Variational Autoencoder (CVAE). Subsequently, we evaluate the classification performance and feature diversity using our anonymized method compared to k-Same [27], the foundational anonymization method that influenced machine-learning-related applications, such as GANs [24]. Additionally, we evaluate the classification performance under inference perturbations, showcasing the robustness of the generated feature representation. Lastly, we perform a qualitative analysis of the initial feature space and the anonymized one, in order to visually comprehend the representativeness of the proposed method.

4.1 Feature extraction and CVAE training

For feature extraction, we used the DINOv2 ViT-B/14 foundation model [29], which has been trained on 142 million natural images and outputs image embeddings of size 768. To optimize the training process of our CVAE, these embeddings were pre-generated and stored on disk [7, 25]. For datasets lacking official validation splits, we extracted a stratified sample (10%) from the training data to ensure representativeness. Note that both validation and test sets were kept non-anonymized, preserving their semantic integrity.

The CVAE architecture comprises a conditional encoder and a conditional decoder. The encoder first concatenates input features with one-hot class labels and then processes it through two linear layers having 256 and 100 dimensions, respectively. The latent variable z is again concatenated with the one-hot class labels and passed to a symmetric decoder that mirrors the encoder's architecture. We train the CVAE with the Adam optimizer, learning rate of 0.001, and early stopping. In addition to the MSE loss, a KL-divergence loss is employed to enforce a standard normal prior in the latent space weighted with $\beta = 0.1$, prioritizing the reconstruction. These hyperparameters are standardized across all experiments to ensure consistent comparisons.

4.2 Feature diversity and downstream performance

Evaluation metrics The k-Same method simplifies anonymization by collapsing groups of k data points into their centroid. While effective for anonymization, this approach often results in significant information loss and increased data sparsity. Conversely, CVAE effectively maintains the initial distribution's representation, enabling the generation of more samples than originally used, without the compromise on diversity. Yu et al. [45] estimate data diversity through the *convex hull* and the *maximum dispersion*, defined as the sum of all pairwise distances within the (anonymized) dataset. Although calculating the convex hull is impractical for datasets with fewer samples than dimensions, it is trivial to demonstrate that CVAE achieves higher maximum dispersion compared to k-Same, as the latter's regression to the centroid inherently reduces pairwise distances.

To provide a more nuanced comparison, we analyze the *average nearest neighbor distance* $\mathcal{D}$ between the original feature set $\mathcal{F}$ and the anonymized feature set $\mathcal{A}$ (both of size N). This is formally defined in Equation 1:

$$\mathcal{D}(\mathcal{A},\mathcal{F}) = \frac{1}{N}\sum_{j=1}^{N} d_{\min}(a_j,\mathcal{F}) \tag{1}$$

where $d_{\min}(a_i,\mathcal{F})$ is the minimum Euclidean distance between $a_j \in \mathcal{A}$ and all $f_i \in \mathcal{F}$. Intuitively, this is closely related to the *mean of the minimum LPIPS* [47] used in PLAN [32]. In fact, the latter measures the average minimum distance between generated samples and their closest real ones, using the activations of a pre-trained network. Moreover, an equally important goal of any anonymization technique is to preserve downstream performance while ensuring data privacy. In our evaluation, considering image classification as the downstream task, we use the *area under the receiver operating curve* (AUC) as the evaluation metric. This allows us to take into account different class imbalance ratios across the evaluated datasets. Following prior works in feature space evaluation [2, 29], we train a linear layer on top of the anonymized embeddings using the Adam optimizer with a learning rate of 0.001 and early stopping. To fairly compare our method against k-Same, we replicate the *same exact class proportions* of the original dataset.

Reference method and datasets Consistent with prior works [18, 32], we evaluate the performance of our `CVAE` against k-Same using $k \in \{2,5,10,15\}$. We exclude the previously mentioned GAN-based methods such as k-SALSA [18] and PLAN [32], since they applied k-Same within GAN's latent space to generate anonymous synthetic *images*. Our analysis begins with a wide range of real-world, small-sized, medical datasets to underscore the clinical relevance of our method. From the MedMNIST+ collection [8, 43], we use BreastMNIST with resolution 224×224 (breast ultrasound, binary, 780 samples). Additionally, we include datasets from the MedIMeta collection [42]: Skin Lesion (dermatoscopy, multi-class, 1011 samples) and Axial Organ slices (CT, multi-class, 1645 samples). Another challenging dataset is OCTDL [6] (OCT imaging, multi-class, 2064 samples), which exhibits a severe class imbalance. To extend our evaluation to the natural image domain, we selected multi-class datasets from the `torchvision.datasets` module [1], focusing on those with moderate sample sizes. This includes STL-10 [4], DTD [3], Oxford-Pets [30], and FGVC-Aircraft [23], allowing us to cover a wide spectrum of general imaging tasks.

Results Figure 2 presents a scatterplot with AUC on the x-axis and average nearest neighbor distance $\mathcal{D}$ on the y-axis, positioning the most effective methods towards the top-right corner. To ensure the reliability of our findings, all results are averaged across three different seed runs. `CVAE` consistently exhibits both high diversity and high downstream performance ($\mathcal{D}\uparrow$, AUC $\uparrow$). In contrast, as expected, lower k-Same configurations ($k=2,5$) achieve comparable downstream performance but lack diversity ($\mathcal{D}\downarrow$, AUC $\uparrow$), whereas higher k-Same settings ($k=10,15$) increase diversity at the expense of performance ($\mathcal{D}\uparrow$, AUC $\downarrow$). Interestingly, consistent results are observed with the OCTDL dataset, which is particularly challenging. Specifically, it presents a severe class imbalance, with a majority class of 752 samples and a minority of just 13. Furthermore, the medical results hold true on natural datasets as well, although the differences in classification performance are less pronounced, partly because these datasets were included in the DINOv2 training set [29].

[1] https://pytorch.org/vision/stable/datasets.html

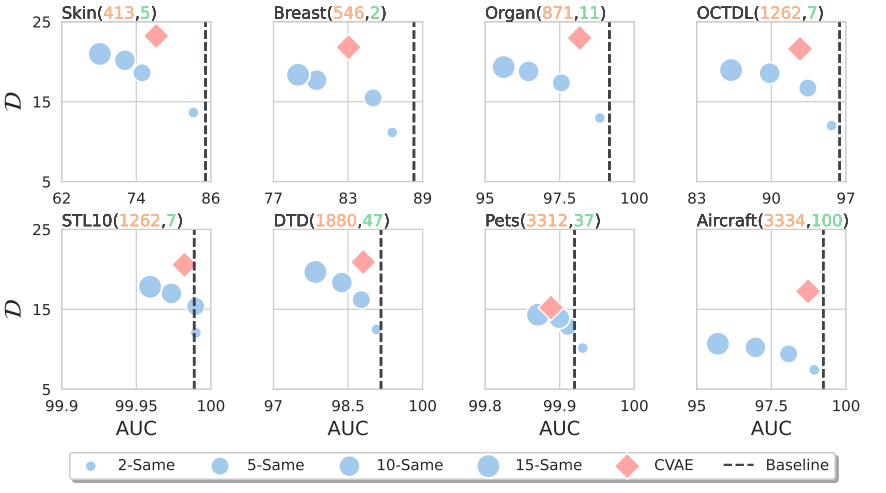

Figure 2: Classification performance (AUC) and average nearest neighbor distance ($\mathcal{D}$) on medical (top row) and non-medical (bottom row) datasets. We report in brackets the number of (training samples,classes). Our objective is to maximize both metrics (top-right corner). The vertical line represents the *baseline* performance achieved without anonymization.

4.3 Robustness of adaptive data sampling

This experiment evaluates the robustness of our proposed CVAE method, which dynamically generates new samples a_j of class $\tilde{y}_j \sim \mathcal{C}$ at every batch. Here, the inherent data diversity provided by our method is expected to contribute positively to model robustness. To evaluate this claim, a common approach involves testing against image corruptions that simulate real-world distortions [17]. However, these corruptions in the pixel space might already be mitigated by the inherent robustness of foundation models [31]. Therefore, we propose a perturbation test by injecting Gaussian noise into the test feature embeddings. We systematically apply this noise with zero mean and a gradually increasing standard deviation $\sigma = \{1,2,3\}$. Here, we evaluate the performance of CVAE against k-Same. Therefore, we first train our CVAE and then use its decoder to continuously generate new samples during training. Instead of only sampling from the latent space with a standard Gaussian distribution $(\mu = 0, \sigma^2 = 1)$ as reported in step 4 of Algorithm 1, we sample with variance $\sigma^2 = \{0.5, 1.0, 1.5\}$. This allows us to vary the prototypicality of the generated vectors. Specifically, since we enforce a standard normal distribution $(\sigma^2 = 1)$ during training, a lower sampling variance during data generation yields samples closer to the class prototypes, and vice versa. The classification pipeline follows the previous settings, reporting the average AUC over three seed runs.

Results Table 1 shows the AUC observed on the clean test feature vectors $(\sigma = 0)$ and on those subjected to varying levels of Gaussian noise. CVAE consistently ranks among the top performers. Although CVAE may initially show slightly lower performance in a noise-free environment, it substantially outperforms k-Same as the noise level increases. This is particularly pronounced in the Skin Lesion and OCTDL, where CVAE $(\sigma^2 = 0.5)$ outperforms 2-Same at $\sigma = 3$ by 4.2% and 3.2%, respectively. Notably, this lower sampling

variance remarkably improves performance and robustness across all datasets, suggesting that a higher prototypicality might be beneficial in settings with limited and imbalanced data. These results underscore the ability of our approach to enhance model robustness across diverse domains, each presenting unique challenges related to data size and data imbalance.

	AUC ↑															
	Skin				Breast				Organ				OCTDL			
σ	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
2-Same	**83.2**	**78.9**	**72.3**	**67.2**	**86.6**	**83.5**	77.2	72.4	**98.8**	**98.2**	**96.0**	**92.2**	**95.6**	**92.3**	**84.7**	77.9
5-Same	74.9	71.7	66.9	63.1	**85.0**	**82.3**	76.6	72.2	97.6	96.9	94.8	**91.2**	**93.4**	90.0	82.2	75.4
10-Same	72.2	69.8	65.6	62.2	80.5	78.4	72.8	68.7	96.5	95.6	93.3	89.6	89.8	87.8	82.6	76.8
15-Same	68.1	66.5	63.5	60.7	79.0	76.7	71.8	67.9	95.6	94.7	92.1	88.3	86.2	84.5	79.4	74.5
CVAE$\|\sigma^2 = 0.5$	**82.4**	**80.5**	**75.9**	**71.4**	82.9	81.2	**77.7**	**74.2**	98.2	97.6	**95.3**	91.1	93.3	**91.2**	**86.5**	**81.1**
CVAE$\|\sigma^2 = 1.0$	78.9	75.4	69.9	65.4	82.6	81.2	**77.8**	**73.8**	98.5	97.8	95.2	90.5	92.6	90.2	84.5	**78.4**
CVAE$\|\sigma^2 = 1.5$	73.2	69.1	63.9	60.3	83.5	81.9	76.9	72.0	**98.6**	**98.0**	**95.3**	90.4	91.0	88.2	82.1	75.7

Table 1: Area Under the Receiver Operating Curve (AUC ↑) calculated on the clean test embeddings ($\sigma = 0$) and across three replicas subject to Gaussian noise with zero mean and standard deviation $\sigma = \{1, 2, 3\}$. The top two results are displayed in **bold**.

4.4 Qualitative results

In addition to quantitative metrics, we conduct a qualitative analysis of the feature space, illustrated in Figure 3, using t-distributed stochastic neighbor embedding (t-SNE) [41]. Notably, our feature extractor was not originally trained on medical data, therefore it may encounter difficulties with tight class separation due to the subtle nature of diagnostic features. For instance, in the BreastMNIST dataset, which consists of breast ultrasound images, the fine-grained details necessary for accurate diagnosis are not always clearly represented. Nevertheless, subtle differences within the class still seem to be detectable, as indicated by the shape of small clusters scattered throughout the 2-dimensional feature space. With respect to the evaluated anonymization techniques, the limitations of the *k*-Same method, such as information loss and data sparsity, become particularly evident. Although DINOv2 ViT-B/14 may not well separate classes in such datasets, our CVAE approach manages to maintain the original distribution. Clearly, unlike *k*-Same, our method avoids information loss, ensuring our privacy-aware feature distribution closely mirrors the initial one.

5 Discussion and conclusion

Limitations While our method does not directly expose training data, we did not rigorously investigate formal privacy guarantees. Furthermore, the effectiveness of CVAEs heavily depends on the capabilities of the chosen feature extractor. In addition, this approach could further benefit from domain-specific foundation models, as they may be able to capture more nuanced patterns. Lastly, our data-sharing framework assumes that all parties involved use the same feature extractor. However, with the increasing interest over *general foundation models* [40], this assumption becomes more reasonable and can be justified by the considerable benefits and enhanced downstream performance.

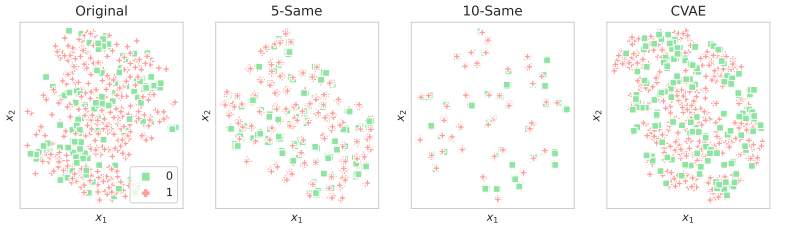

Figure 3: Class distribution of the *BreastMNIST* dataset and its anonymous counterparts through k-Same $(5, 10)$ and CVAE. Clearly, while CVAE faithfully preserves data diversity, k-Same tends to agglomerate information, increasing data sparsity and losing precious information, especially on limited-size datasets.

Conclusion We demonstrate that conditional generative models, such as CVAEs, effectively address challenges related to privacy and data utility across a diverse range of datasets. This approach further shows high robustness even with small sample sizes and severe class imbalances. Future research can build upon our insights, investigating the sensitive role of the sampling variance, exploring further conditional generative models, or extending our approach to be trainable in an end-to-end manner for downstream tasks.

Acknowledgments

This study was funded through the Hightech Agenda Bayern (HTA) of the Free State of Bavaria, Germany.

References

[1] Fouzia Altaf, Syed MS Islam, Naveed Akhtar, and Naeem Khalid Janjua. Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access*, 7:99540–99572, 2019.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[4] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[5] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[6] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022.

[7] Sebastian Doerrich, Tobias Archut, Francesco Di Salvo, and Christian Ledig. Integrating knn with foundation models for adaptable and privacy-aware image classification. In *2024 IEEE 21th International Symposium on Biomedical Imaging (ISBI)*, 2024.

[8] Sebastian Doerrich, Francesco Di Salvo, Julius Brockmann, and Christian Ledig. Rethinking model prototyping through the medmnist+ dataset collection. *arXiv preprint arXiv:2404.15786*, 2024.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[10] Val Andrei Fajardo, David Findlay, Charu Jaiswal, Xinshang Yin, Roshanak Houmanfar, Honglei Xie, Jiaxi Liang, Xichen She, and David B Emerson. On oversampling imbalanced data with deep conditional generative models. *Expert Systems with Applications*, 169:114463, 2021.

[11] Aris Gkoulalas-Divanis, Grigorios Loukides, and Jimeng Sun. Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of biomedical informatics*, 50:4–19, 2014.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[13] Rémi Gosselin, Loïc Vieu, Faiza Loukil, and Alexandre Benoit. Privacy and security in federated learning: A survey. *Applied Sciences*, 12(19), 2022. ISSN 2076-3417.

[14] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

[15] Ahsan Habib, Chandan Karmakar, and John Yearwood. Impact of ecg dataset diversity on generalization of cnn model for detecting qrs complex. *IEEE Access*, 7:93275–93285, 2019.

[16] Omid Hajihassnai, Omid Ardakanian, and Hamzeh Khazaei. Obscurenet: Learning attribute-invariant latent representation for anonymizing sensor data. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, IoTDI '21, pages 40–52, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383547.

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[18] Minkyu Jeon, Hyeonjin Park, Hyunwoo J Kim, Michael Morley, and Hyunghoon Cho. k-salsa: k-anonymous synthetic averaging of retinal images via local style alignment. In *European Conference on Computer Vision*, pages 661–678. Springer, 2022.

[19] Taehoon Kim and Jihoon Yang. Latent-space-level image anonymization with adversarial protector networks. *IEEE Access*, 7:84992–84999, 2019.

[20] Stefan Larson, Anthony Zheng, Anish Mahendran, Rishi Tekriwal, Adrian Cheung, Eric Guldan, Kevin Leach, and Jonathan K. Kummerfeld. Iterative feature mining for constraint-based data collection to increase data diversity and model robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8097–8106, Online, November 2020. Association for Computational Linguistics.

[21] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[22] Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–4, 2022.

[23] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[24] Blaž Meden, Žiga Emeršič, Vitomir Štruc, and Peter Peer. k-same-net: k-anonymity with generative deep neural networks for face deidentification. *Entropy*, 20(1):60, 2018.

[25] Kengo Nakata, Youyang Ng, Daisuke Miyashita, Asuka Maki, Yu-Chieh Lin, and Jun Deguchi. Revisiting a knn-based image classification system with high-capacity storage. In *European Conference on Computer Vision*, pages 457–474. Springer, 2022.

[26] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

[27] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17 (2):232–243, 2005.

[28] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[30] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[31] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022.

[32] Matteo Pennisi, Federica Proietto Salanitri, Giovanni Bellitto, Simone Palazzo, Ulas Bagci, and Concetto Spampinato. A privacy-preserving walk in the latent space of generative models for medical applications. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 422–431. Springer, 2023.

[33] Juan M Perero-Codosero, Fernando M Espinoza-Cuadros, and Luis A Hernández-Gómez. X-vector anonymization using autoencoders and adversarial training for preserving speech privacy. *Computer Speech & Language*, 74:101351, 2022.

[34] Mehran Pesteie, Purang Abolmaesumi, and Robert N Rohling. Adaptive augmentation of medical data using independently conditional variational auto-encoders. *IEEE transactions on medical imaging*, 38(12):2807–2820, 2019.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[36] Nuria Rodríguez-Barroso, Daniel Jiménez-López, M Victoria Luzón, Francisco Herrera, and Eugenio Martínez-Cámara. Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion*, 90:148–173, 2023.

[37] Divya Saxena and Jiannong Cao. Generative adversarial networks (gans) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, 54(3):1–42, 2021.

[38] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

[39] Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020.

[40] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[42] Stefano Woerner, Arthur Jaques, and Christian F Baumgartner. A comprehensive and easy-to-use multi-domain multi-task medical imaging meta-dataset (medimeta). *arXiv preprint arXiv:2404.16000*, 2024.

[43] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[44] Yuanzhe Yang, Zhiyi Niu, Yuying Qiu, Biao Song, Xinchang Zhang, Yuan Tian, and Ran Guo. A cluster-based facial image anonymization method using variational autoencoder. In *International Conference on Big Data and Security*, pages 621–633. Springer, 2022.

[45] Yu Yu, Shahram Khadivi, and Jia Xu. Can data diversity enhance learning generalization? In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[46] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[48] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.