# Synthetic Image Learning: Preserving Performance and Preventing Membership Inference Attacks

Eugenio Lomurno[a,*], Matteo Matteucci[a,1]

[a] *Department of Electronics, Information, and Bioengineering,*
*Politecnico di Milano, Milan, 20133, Italy*

**Abstract**

Generative artificial intelligence has transformed the generation of synthetic data, providing innovative solutions to challenges like data scarcity and privacy, which are particularly critical in fields such as medicine. However, the effective use of this synthetic data to train high-performance models remains a significant challenge. This paper addresses this issue by introducing Knowledge Recycling (KR), a pipeline designed to optimise the generation and use of synthetic data for training downstream classifiers. At the heart of this pipeline is Generative Knowledge Distillation, the proposed technique that significantly improves the quality and usefulness of the information provided to classifiers through a synthetic dataset regeneration and soft labelling mechanism. The KR pipeline has been tested on a variety of datasets, with a focus on six highly heterogeneous medical image datasets, ranging from retinal images to organ scans. The results show a significant reduction in the performance gap between models trained on real and synthetic data, with models based on synthetic data outperforming those trained on real data in some cases. Furthermore, the resulting models show almost complete immunity to Membership Inference Attacks, manifesting privacy properties missing in models trained with conventional techniques.

*Keywords:* Generative Deep Learning, Dataset Generation, Classification Accuracy Score, Privacy, Membership Inference Attack, Generative Knowledge Distillation, Knowledge Recycling

## 1. Introduction

The advent of generative deep learning has marked a fundamental technological breakthrough that is rapidly permeating every aspect of society and profoundly affecting the daily lives of every individual. Thanks to this technology, it is now extremely easy to create and interact with high-quality synthetic data, be it images, text, audio or video. This ease of access to artificially generated content makes it increasingly difficult to distinguish between human and algorithmic production. Meanwhile, the applications and innovations of generative models are expanding at a rapid pace, revolutionising many sectors. The implications of this development are profound: on the one hand, new opportunities are opening up, and on the other, ethical and social challenges are emerging in relation to the use and misuse of such technologies.

Today, this technological progress raises problems related to the circulation of images or text documents generated by algorithms and presented as the fruit of human labour. However, it also opens the door to a dual use with immense virtuous potential. It is precisely the difficulty of distinguishing between human and algorithmic production that has led to the use of generative models to enrich real data sets and, more recently, to attempts at total replacement to obtain entire synthetic datasets. However, the creation of entirely synthetic datasets is a complex task that requires models capable of generating large amounts of data in a reasonable amount of time, while carefully balancing the quality and variety of the data generated. Indeed, it is known that training models based solely on synthetic data tends to degrade performance compared to those trained on real data [19]. In addition to these aspects, it is crucial to consider an area of growing importance, that of data privacy, both before and after the data has been learned by the models. This is particularly critical especially in the context of medical data, where the protection of privacy is essential to preserve the relationship of trust between experts and patients. In this scenario, generative technology offers unexploited potential for the secure use of medical data, opening up new opportunities for healthcare research and innovation.

This paper presents the Knowledge Recycling (KR) strategy, a pipeline that aims to improve the generation of synthetic datasets and the training of downstream classifiers using only synthetic images. First, the generator and an auxiliary classifier, named Teacher Classifier, are trained. The optimal checkpoint of the Generator is determined by training a Student Classifier for each checkpoint. These trainings use the proposed technique of Generative Knowledge Distillation, where the Teacher Classifier generates soft labels for the synthetic images, allowing the Student Classifier to learn about uncertainties and class correlations, thus improving its accuracy in predicting both synthetic and real images. After identifying the best checkpoint, the generation parameters are optimised by adjusting the size

[*] Corresponding author; email: eugenio.lomurno@polimi.it

[1] Co-author; email: matteo.matteucci@polimi.it

of the synthetic dataset, the frequency of dataset regeneration during Student Classifier training, and the standard deviation of the Generator. Once the optimal Student Classifier training is completed, its resistance to a Membership Inference Attack is evaluated and compared with the one achieved by the Teacher Classifier [35].

This work aims to demonstrate the possibility of obtaining classifiers trained on synthetic data with comparable performance to those trained on real data, while providing superior resistance to Membership Inference Attacks. The main contributions of this research are:

- The introduction of Knowledge Recycling, a novel pipeline for optimised generation and use of synthetic data in the context of classifier training.
- The development of Generative Knowledge Distillation, a technique that improves the quality of information transferred from synthetic data to classifiers, thereby reducing the performance gap with models trained on real data.
- Demonstrate the effectiveness of the proposed pipeline in producing models that are nearly immune to Membership Inference Attacks, resulting in a positive trade-off between performance and resistance to this category of privacy attacks.

## 2. Related Works

The state of the art in image generation is currently contested between the Generative Adversarial Networks (GAN) family and the Denoising Diffusion Probabilistic Models (DDPM) [12, 15]. Although using different mechanisms, both families of models are in fact capable of producing and manipulating images with very high resolution by being conditioned in a variety of ways [17, 28]. In parallel to this line of research, which aims to produce high quality single images, a second line of research has been developed with the aim of exploiting this generative power to create fully synthetic datasets or to enrich existing ones.

The first attempts in this direction concerned contexts where it is very complex and time-consuming to collect and label new data, such as in the medical field. Frid-Adar et al. used GANs to generate synthetic images of liver lesions. This work showed that adding synthetic images to the source dataset improved the performance of classification models for diagnosing liver lesions [11]. Subsequently, Sedigh et al. and Islam et al. also used GAN models to generate synthetic images of skin cancer and brain PET, respectively. In both cases, enriching the dataset with real images led to improved classification performance [34, 16].

In addition, studies have been undertaken not to enrich existing image datasets, but to generate entiraly new ones and evaluate their properties via downstream machine learning problems [22, 8, 32]. From early work, it has become clear that the semantic information contained in synthetic data is not in itself sufficient for a model trained on such data to perform well when making inferences on real data [30]. Techniques have been developed to make the most of the information that can be extracted from generative models, as well as the potentially unlimited number of images that can be generated. It has been shown that both recycling the synthetic dataset in the training phase and creating synthetic datasets with higher cardinality than the dataset used to train the generator are very beneficial to performance [4, 23]. Filtering techniques were also proposed to discard synthetic images that were classified incorrectly or with low confidence by an auxiliary classifier [9]. This also allowed sampling from sparser distributions, further enriching the information of the synthetic datasets [19].

### *2.1. Privacy Threats and Countermeasures*

As research on new models and techniques continues and their applications increase, the attack surface on such models and the importance of privacy protection continue to grow. Among the most common and popular attacks are the families of Membership Inference Attacks (MIAs), Model Inversion, Model Extraction, and Data Poisoning, which can be applied depending on the context and the ability to access and interact with the attacked model [35, 10, 37, 5]. MIAs are the easiest family of attacks to use, as they can be executed in black-box contexts and from logits alone. Their purpose is to guess whether the sample given as input to the model to be attacked was present in its training set or not. From this attack, and thus once the presence of a particular sample learned by the attacked model is known, it is possible, for example, to refine Model Inversion or Model Extraction attacks, or to proceed with inference attacks aimed at extracting more refined information. Recently, the Likelihood Ratio Attack (LiRA) [7] has emerged as a state-of-the-art approach for membership inference, achieving significantly better results at low false-positive rates compared to previous methods. LiRA works by training shadow models to estimate per-example distributions of model confidences and performing a careful likelihood ratio test, making it particularly effective at confidently identifying training set members while maintaining low false positive rates.

Many defensive mechanisms have been developed to deal with this threat, many of which rely on relaxed forms of Differential Privacy [1]. Although such mechanisms are very effective in preventing MIAs, they often require very long training times and lead to performance degradation of the protected model [24]. Recently, however, in addition to empirical metrics to measure the trade-off between performance and resistance to MIAs, alternative techniques have been proposed to protect models by adversarial training or individual private training steps instead of the entire training [21, 36].

## 3. Method

This section presents the Knowledge Recycling (KR) pipeline for the creation of synthetic datasets and their subsequent use for training downstream classifiers. The pipeline starts with a preliminary step where an auxiliary classifier, called **Teacher Classifier**, and a data generator, called **Generator**, are trained on the same real dataset. The first proper step, called **Checkpoint Optimisation**, aims at identifying the best checkpoint of the Generator. During this step, a classifier is trained for each checkpoint of the Generator. These classifiers, called **Student Classifiers**, have the same architecture as the Teacher Classifier and are trained using the same training technique. For each checkpoint, the generation of the synthetic datasets is performed using the proposed technique called Generative Knowledge Distillation (GKD), which is explained in detail in the Subsection 3.4. Once the optimal checkpoint is identified, the **Tuning** step follows, in which the generation parameters are optimised and the final Student Classifier is trained. Finally, the last step, called **Membership Inference Attack**, tests the robustness of the Student Classifier against the homonymous privacy attack.

### *3.1. Teacher Classifier*

The Teacher Classifier plays a key role in the KR pipeline, as it is not only the core of the GKD technique, but also the benchmark against which the Student Classifiers can be compared in terms of accuracy performance and resistance to privacy attacks. In order to have a fair and robust comparison, the architecture and training technique of the Teacher Classifier is also replicated for the Student Classifiers. Having to balance performance and training speed, since each checkpoint requires a whole one, the chosen architecture is a ResNet14 model [13]. Training is done in Mixed Precision for 500 epochs with SGD optimiser, initial learning rate of 0.5, cosine annealing scheduler and TrivialAugment and MixUp as main augmentation [25, 27, 40]. For more details see Section B of Supplementary Materials.

### *3.2. Generator*

In the field of image generation, Generative Adversarial Networks (GAN) and Denoising Diffusion Probabilistic Models (DDPM) currently represent the state of the art [12, 15]. Although they differ significantly in their operation, both approaches offer high and comparable performance in generating different types of media content that can be conditioned in different ways. GANs are known for their fast inference, but suffer from instability during training. DDPMs, on the other hand, offer more stable training but require longer generation times. Despite recent developments, DDPM models can drastically reduce the number of denoising steps, but their generation times are still too long to compete with GANs in generating large amounts of data [33].

For this work, a GAN-based approach was chosen, which favours the speed of inference. In particular, a modified version of BigGAN-Deep was chosen, a model that represents a milestone in the development of GAN [6] models. Indeed, BigGAN introduced several important innovations, including the use of conditional batch normalisation, the use of a truncation trick to control the trade-off between quality and generation diversity, combined with advanced optimisation techniques to handle large networks, such as spectral normalisation [26]. The proposed implementation modifies the original BigGAN-Deep model in several aspects. The hinge loss is replaced by a logistic loss, and the tanh activation is replaced by a sigmoid. In addition, regularisation techniques such as label smoothing have been introduced to improve the quality of the discriminator, and the AdamW optimiser with a weight decay of 0.0005 has been adopted. These changes aim at improving the stability of the training and the quality of the generated images. The model was trained for 500 epochs, with a 4:1 ratio between discriminator and generator updates. To ensure the robustness of the model, we implemented a system of saving checkpoints at regular intervals of 5 epochs. A detailed description of the implementation and a comparison between the vanilla model and our modified version can be found in Section A of Supplementary Materials.

### *3.3. Evaluation Metrics*

In the evaluation of image generators, the most widely used metrics in the literature are the Inception Score (IS) and the Fréchet Inception Distance (FID) [31, 14]. The IS aims to quantify the quality of the distribution generated by assessing both the clarity and diversity of the images produced. The FID, on the other hand, provides a more comprehensive measure by comparing the generated distribution with the actual distribution used to train the Generator, thus capturing both the quality and fidelity of the synthetic images.

However, recent studies have highlighted the limitations of using these metrics to assess the usefulness of generated images in downstream learning contexts. A lack of correlation was observed between IS and FID and the effectiveness of the generated data for subsequent classification tasks [19]. Furthermore, a trade-off between the quality of individual images and the diversity of the generated distribution was identified [2]. The maximisation of IS and FID tends to favour the quality of the generated images at the expense of the variety, which is crucial for the creation of synthetic datasets that favour the generalisation of the models trained on them.

In this study, the Classification Accuracy Score (CAS) is adopted as the main metric. The CAS measures the validation accuracy on real data of a classifier trained on synthetic datasets [29]. This metric helps to identify the training epoch that produces the most effective synthetic dataset and, like IS and FID, helps to prevent the mode collapse of the Generator.

### *3.4. Checkpoint Optimisation*

Once both the Teacher Classifier and the Generator have been defined, trained on real training data and frozen, it is possible to proceed to the Checkpoint Optimisation step. The goal of this first step is to identify the optimal checkpoint to maximise the performance of the downstream Student Classifier models. For each Generator checkpoint, a Student Classifier is trained using a strategy similar to that of the Teacher Classifier, but with a reduced number of epochs – 100 instead of 500 – for efficiency reasons. At the beginning of each training session, a synthetic dataset of the same cardinality as the real one is generated using the current checkpoint. The input noise is sampled from a multivariate Gaussian distribution with a standard deviation of 1.0. To maintain data diversity, the synthetic data set is fully regenerated every 10 epochs during training.

Previous studies have demonstrated the effectiveness of filtering the generated data to improve CAS. Dat et al. used a model similar to the Teacher Classifier to exclude images with inconsistent predictions [9], while Lampis et al. introduced an additional filtering step based on the confidence of the predictions [19].

In the KR pipeline, the technique of **Generative Knowledge Distillation** (GKD) is proposed and adopted. In contrast to filtering methods, the Teacher Classifier is used to evaluate the generated images and produce soft labels for the Student Classifier. These probability labels are more informative than binary ones, as they capture uncertainties and correlations between classes, leading to a significant improvement in CAS, as detailed in Section D of Supplementary Materials. This approach optimises both the quality of the information passed to the Student Classifier and the efficiency of the synthetic dataset generation process, allowing the desired dataset cardinality to be achieved faster than with filter-based techniques.

### *3.5. Tuning*

After the identification of the optimal checkpoint with respect to the CAS metric, it is possible to proceed with the Tuning step to optimise also the generation parameters. These parameters, held constant during the Checkpoint Optimisation step, are now re-computed to further improve Student Classifiers performance. The parameters being optimised are:

- The regeneration rate of the synthetic dataset: previously fixed at 10 epochs, is now varied between 1 and 10 epochs.
- The scale of the cardinality of the generated dataset: previously set at 1, is now made to vary between 1 and 10.
- The standard deviation used during sampling from the multivariate Gaussian distribution: previously equal to 1.0, is now made to vary between 1.0 and 2.5.

It has been shown in previous works that regenerating the dataset more frequently and creating more numerous datasets contributes to the improvement of CAS [19, 23]. With regard to standard deviation, this approach, in the opposite direction to the Truncation Trick implemented in the BigGAN-Deep vanilla model, aims to favour a more varied generation, even at the expense of the perceptual quality of the generated data [6, 19]. The Tuning step is carried on via a Tree-structured Parzen Estimator associated with a Hyperband pruning mechanism [3, 20]. The optimisation proceeds for 50 iterations. In each iteration, a Student Classifier is trained with the same procedure used in Checkpoint Optimisation but using the data generated with the current parameter configuration, with the aim of maximising CAS. At the end of the search, the optimal parameter configuration is used to train the final Student Classifier for 500 epochs.

### *3.6. Membership Inference Attack*

The final step in the KR pipeline involves a resistance test for the Student Classifier against a Membership Inference Attack (MIA). This type of attack aims to compromise privacy by identifying the training data stored within the attacked model. In the context of this study, sensitive training data is never directly exposed to the Student Classifier, but is only used for training the Generator and Teacher Classifier. The objective of this step is therefore to assess the effectiveness of the MIA in identifying the data used to train the Generator via the Student Classifier and to compare this effectiveness with that of the same attack carried out against the Teacher Classifier. To perform the MIA, the Likelihood Ratio Attack (LiRA) proposed by Carlini et al. is adopted [7]. This state-of-the-art approach frames membership inference as a hypothesis testing problem, where the goal is to distinguish between two distributions: one where the model was trained on the target example, and one where it was not. The implementation involves:

1. Creation of 256 shadow models, identical in architecture and training technique to the Teacher Classifier.
2. Utilisation of the validation dataset with 50/10/40 splits to obtain training, validation and test sets for the shadow models, with different splits for each model.
3. For each example $(x, y)$, collection of model confidences in logit scale:

$$\phi(p) = \log\left(\frac{p}{1-p}\right), \text{ where } p = f(x)_y$$

4. Estimation of Gaussian distributions $\tilde{Q}_{in}$ and $\tilde{Q}_{out}$ for each example, representing the distribution of logit-scaled confidences when the example is included or excluded from training, respectively.
5. Implementation of both online and offline variants:
   - Online: estimating means ($\mu_{in}$, $\mu_{out}$) and variances ($\sigma^2_{in}$, $\sigma^2_{out}$) for both distributions
   - Offline: estimating only $\mu_{out}$ and $\sigma^2_{out}$ to perform one-sided hypothesis testing
6. Application of both global and per-example variance estimation, selecting the most effective approach for each attack scenario.

The attack computes the likelihood ratio between these distributions to determine membership, with the final score given by:

$$\Lambda = \frac{p(\phi(f(x)_y)|\mathcal{N}(\mu_{in}, \sigma^2_{in}))}{p(\phi(f(x)_y)|\mathcal{N}(\mu_{out}, \sigma^2_{out}))}$$

This approach is applied to both the Teacher Classifier and the Student Classifier, leveraging the enhanced ability of LiRA to achieve high true-positive rates at very low false-positive rates, making it particularly effective for privacy auditing.

The evaluation of resistance to MIAs is based on two metrics: the AUC, typically used to evaluate this type of attack, and the Accuracy Over Privacy (AOP), which provides an estimate of the trade-off between performance – measured as test accuracy – and resistance to MIAs [21]. The objective of this step is to examine whether the proposed training from synthetic data may constitute an additional layer of privacy, making the attack less effective.

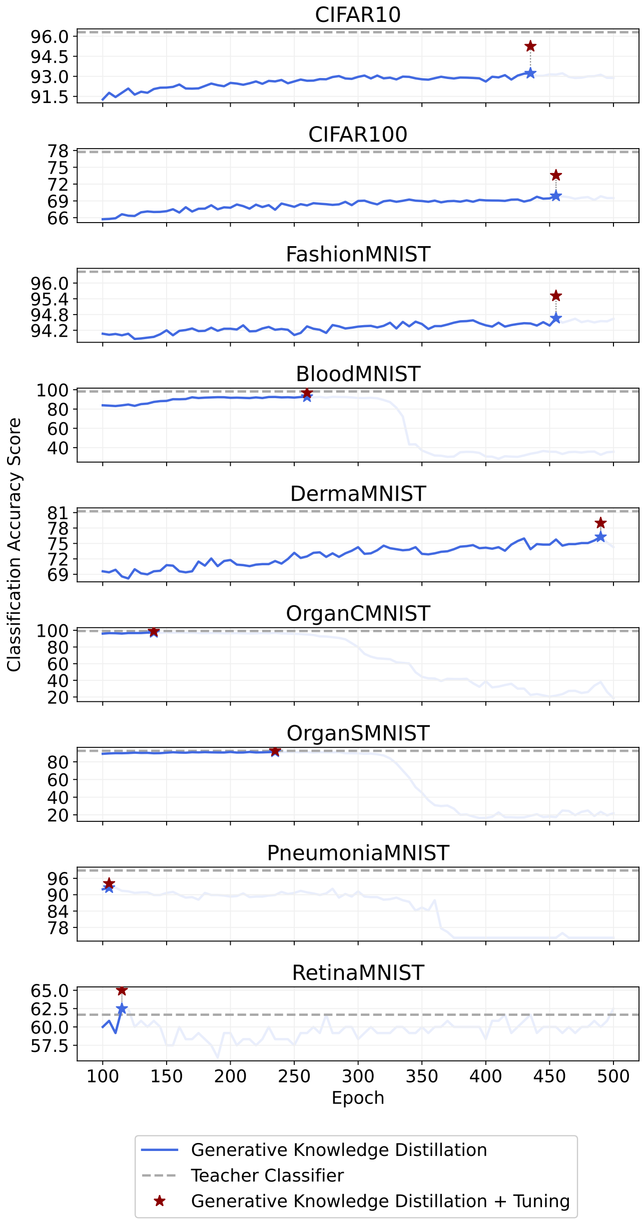

Figure 1: The Classification Accuracy Score (CAS) of the validation calculated for each checkpoint of the Generator for the considered datasets. The continuous blue line represents the CAS obtained during the Checkpoint Optimisation step using the Generative Knowledge Distillation technique. The best checkpoint is marked with a blue star. The dashed grey line represents the best validation Accuracy obtained with the Teacher Classifier. The red star indicates the optimal checkpoint CAS of the validation after training with Generative Knowledge Distillation with parameters found during the Tuning step.

## 4. Experiments and Results

The experiments were performed on nine image datasets, all rescaled to 32x32.

CIFAR10, CIFAR100 and FashionMNIST were used both for the final comparisons and to calibrate and test the Knowledge Recycling (KR) pipeline, as described in detail in Section A and Section D of Supplementary Materials [18, 38]. The six medical datasets - BloodMNIST, DermaMNIST, OrganCMNIST, OrganSMNIST, PneumoniaMNIST and RetinaMNIST - contain real images from the MedMNIST v2 benchmark [39]. These medical datasets represent the primary field of application for the proposed technique. The KR pipeline, having been calibrated on the three aforementioned datasets, is subsequently applied to these medical datasets without further specific adaptations. This approach allows for the evaluation of the technique's effectiveness and robustness in a more specialised and complex context, distinct from that on which it was initially calibrated.

Further details of all the datasets used can be found in Section C of Supplementary Materials. Experiments were run on four NVIDIA Quadro RTX 6000 GPUs.

The Figure 1 illustrates the results of the Checkpoint Optimisation and Tuning phases, expressed as Classification Accuracy Score (CAS) on the respective validation sets, compared with the optimal Accuracy performance of the Teacher Classifier on the same set. The importance of the selection of the optimal checkpoint and its evaluation via CAS is evident both for Generators with more stable checkpoints (e.g. DermaMNIST, RetinaMNIST) and for those subject to mode collapse and consequent drop in performance (e.g. BloodMNIST, OrganSMNIST). The application of the Generative Knowledge Distillation (GKD) technique alone demonstrates to be sufficient to obtain results close to those of the Teacher Classifier. In the case of the RetinaMNIST dataset, a more accurate model is even obtained from the synthetic data alone. The Tuning step turns out to be beneficial overall, increasing the validation CAS from a minimum of 0.85% for FashionMNIST to a maximum of 4.03% for BloodMNIST, as reported in Table 1. This improvement is due to two factors. The first is the increased availability of information due to the higher cardinality of the generated datasets and their higher recycling frequency. The second is the increased diversity of data due to sampling with a larger standard deviation, which, in combination with the GKD technique, also makes it possible to exploit images that would be uninformative if associated with the hard label used to generate them. These images would likely be filtered out and discarded if coupled with another training technique from synthetic data.

The Table 2 presents the results of the final comparison between Teacher Classifier and Student Classifier. The testing CAS of the Student Classifiers approaches the testing Accuracy of the Teacher Classifiers on average, exceeding it in the cases of PneumoniaMNIST and RetinaMNIST. With regard to resilience to Membership Inference Attacks (MIA), the Student Classifiers demonstrate *significantly greater resilience*, with attack performance close to random guessing, despite the use of a more powerful state-of-the-art attack method [7]. In contrast, the attacks on the Teacher Classifiers are much more effective, resulting in a larger gap in AUC$MIA$ between the two approaches. The Min, Mean, and Max improvements in AUC$MIA$ between the Teacher and Student Classifiers have significantly increased, with the Mean Improvement reaching -11.17% and the Max Improvement reaching -28.18%, as shown in Table 2. This highlights the substantial increase in privacy protection provided by the Student Classifiers under the more powerful attack.

The Accuracy Over Privacy (AOP) metric, which measures the trade-off between performance and resilience to MIAs, shows that Student Classifiers consistently outperform Teacher Classifiers, with an average improvement of 28.60%. This implies that the slight margin of loss in CAS on thest set is positively compensated by the increased and almost total resilience to MIAs.

## 5. Discussion and Limitations

The Knowledge Recycling (KR) technique proposed in this study has been shown to be effective in creating Student Classifiers with comparable performance to the corresponding Teacher Classifiers, while maintaining considerable resistance to Membership Inference Attacks (MIA). This approach, initially calibrated on standard datasets such as CIFAR10, CIFAR100 and FashionMNIST, and subsequently applied to six medical image datasets from the MedMNIST v2 benchmark, establishes a new state-of-the-art in this field.

The average performance gap between Teacher Classifiers and Student Classifiers was reduced to -1.24% in terms of the Classification Accuracy Score (CAS) over the test sets, a significant improvement on previous results. This progress is particularly remarkable considering the use of a single Generator, in contrast to previous works. Dat et al. had achieved an average gap of -10.08% with a single Generator and -5.81% with six, while Lampis et al. had achieved -3.87% with a single Generator and -2.63% with six [19]. The approach proposed in this study exceeds these results, suggesting potential for improvement through the use of multiple Generators in parallel.

The inclusion of a metric to empirically measure one of the privacy-related aspects, such as resistance to MIAs, proved to be crucial for a richer and more multifaceted evaluation of the proposed method, especially if the data under analysis are medical images with potential sensitivities to violations of their privacy. Under the stronger, state-of-the-art attack method proposed by Carlini et al. [7], the Teacher Classifiers, despite being trained with regularization and augmentation techniques, showed increased vulnerability to MIAs. In contrast, the Student Classifiers maintained almost complete resistance to these attacks, with attack performances close to random guessing. This significant increase in the gap of AUC$_{MIA}$ between Teacher and Student Classifiers reinforces the privacy advantage of the proposed approach.

The main limitations of this study concern the small size of the images used (32x32 pixels) and the choice of models that are efficient but not comparable in performance with the current state of the art in their respective tasks. These decisions were dictated by computational efficiency considerations, given the onerous nature of the KR pipeline. The use of higher resolution images and more complex models, both for the Classifier (ResNet14) and the Generator (BigGAN-Deep), could lead to further performance improvements. In particular, upgrading the Generator model could further reduce the performance gap between Teacher and Student Classifiers, potentially outperforming the Teacher. The scalability of the proposed approach, both in terms of the number of Generators and the cardinality and frequency of generation, offers exciting prospects for future developments. With continued hardware advancement, it is plausible that in the near future it will be possible to apply this technique with more complex models and on larger datasets, opening up new possibilities in the field of private learning and the generation of high-quality synthetic data.

## 6. Conclusions

In this paper, the Knowledge Recycling pipeline was presented, demonstrating how synthetic data can be generated and used to train downstream classifiers. It has been shown how the Generative Knowledge Distillation technique, used within the pipeline, improves the quality of information transferable to such downstream classifiers compared to techniques previously proposed in the literature. It was possible to simultaneously reduce the gap between the performance obtainable from real data alone and that obtainable from generated data, setting a new state of the art, as well as to obtain models from synthetic data that manifest privacy properties such that Membership Inference Attacks are ineffective. This was tested on real medical image datasets, demonstrating how it is possible to simultaneously preserve performance and reduce privacy attack surfaces.

## 7. Acknowledgements

This paper is supported by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence).

## References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[2] P. Astolfi, M. Careil, M. Hall, O. Mañas, M. Muckley, J. Verbeek, A. R. Soriano, and M. Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models. *arXiv preprint*, 2024.

[3] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, 2011.

[4] V. Besnier, H. Jain, A. Bursuc, M. Cord, and P. Pérez. This dataset does not exist: training models from generated images. In *International Conference on Acoustics, Speech and Signal Processing*, 2020.

[5] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *Proceedings of the International Conference on Machine Learning*, 2012.

[6] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018.

[7] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022.

[8] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[9] P. T. Dat, A. Dutt, D. Pellerin, and G. Quénot. Classifier training from a generative model. In *International Conference on Content-Based Multimedia Indexing*, 2019.

Table 1: The optimum generation parameters identified after the Tuning step for each dataset under consideration. The Δ CAS column represents the improvement of the validation Classification Accuracy Score compared to the performance obtained from the default generation parameters.

| Hyperparameter | Standard Deviation | Regeneration Rate | Cardinality Scale | Δ CAS |
|---|---|---|---|---|
| CIFAR10 | 1.40 | 9 | 8 | +2.02 |
| CIFAR100 | 1.44 | 7 | 9 | +3.67 |
| FashionMNIST | 1.58 | 1 | 6 | +0.85 |
| BloodMNIST | 2.23 | 1 | 10 | +4.03 |
| DermaMNIST | 1.23 | 10 | 8 | +2.69 |
| OrganCMNIST | 2.33 | 2 | 10 | +1.42 |
| OrganSMNIST | 2.42 | 7 | 10 | +1.22 |
| PneumoniaMNIST | 2.15 | 3 | 5 | +1.53 |
| RetinaMNIST | 1.61 | 2 | 7 | +2.50 |

Table 2: The comparison of Accuracy, $\text{AUC}_{MIA}$ and AOP performance between Teacher Classifier and Student Classifier calculated on the test set – for the Student Classifier the Accuracy is intended as Classification Accuracy Score. The best score is highlighted with **bold**.

| | Accuracy ↑ | | $\text{AUC}_{MIA}$ ↓ | | AOP ↑ | |
|---|---|---|---|---|---|---|
| Model | Teacher Classifier | Student Classifier | Teacher Classifier | Student Classifier | Teacher Classifier | Student Classifier |
| CIFAR10 | **96.24** | 95.83 | 63.10 | **51.98** | 60.42 | **88.80** |
| CIFAR100 | **77.65** | 74.92 | 80.76 | **55.46** | 29.76 | **60.90** |
| FashionMNIST | **95.94** | 95.21 | 57.80 | **51.20** | 71.72 | **90.81** |
| BloodMNIST | **97.49** | 96.26 | 56.31 | **50.99** | 76.81 | **92.59** |
| DermaMNIST | **79.50** | 76.46 | 79.75 | **51.57** | 31.25 | **71.91** |
| OrganCMNIST | **93.16** | 90.23 | 69.99 | **59.29** | 47.54 | **64.13** |
| OrganSMNIST | **79.78** | 78.76 | 66.19 | **58.18** | 45.52 | **58.10** |
| PneumoniaMNIST | 86.54 | **86.70** | 58.88 | **57.36** | 62.49 | **65.90** |
| RetinaMNIST | 54.25 | **55.00** | 56.13 | **52.33** | 43.02 | **50.22** |
| Min Imp | - | -3.04 | - | -1.52 | - | 12.58 |
| Mean Imp | - | -1.24 | - | -11.17 | - | 28.60 |
| Max Imp | - | 0.75 | - | -28.18 | - | 40.66 |

[10] M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2015.

[11] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 2018.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 2017.

[15] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[16] J. Islam and Y. Zhang. Gan-based synthetic brain pet image generation. *Brain informatics*, 2020.

[17] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[18] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Technical Report of Toronto University*, 2009.

[19] A. Lampis, E. Lomurno, and M. Matteucci. Bridging the gap: Enhancing the utility of synthetic data via post-processing techniques. *British Machine Vision Conference*, 2023.

[20] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 2018.

[21] E. Lomurno, A. Archetti, F. Ausonio, and M. Matteucci. Discriminative adversarial privacy: Balancing accuracy and membership privacy in neural networks. *British Machine Vision Conference*, 2023.

[22] E. Lomurno, A. Archetti, L. Cazzella, S. Samele, L. Di Perna, and M. Matteucci. Sgde: Secure generative data exchange for cross-silo federated learning. In *International Conference on Artificial Intelligence and Pattern Recognition*, 2022.

[23] E. Lomurno, M. D'Oria, and M. Matteucci. Stable diffusion dataset generation for downstream classification tasks. *arXiv preprint arXiv:2405.02698*, 2024.

[24] E. Lomurno and M. Matteucci. On the utility and protection of optimization with differential privacy and classic regularization techniques. In *International Conference on Machine Learning, Optimization, and Data Science*, 2022.

[25] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2022.

[26] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[27] S. G. Müller and F. Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[28] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

[29] S. Ravuri and O. Vinyals. Classification accuracy score for conditional generative models. *Advances in Neural Information Processing Systems*, 2019.

[30] S. Ravuri and O. Vinyals. Seeing is not necessarily believing: Limitations of biggans for data augmentation. *International Conference on Learning Representations*, 2019.

[31] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 2016.

[32] M. B. Sarıyıldız, K. Alahari, D. Larlus, and Y. Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[33] A. Sauer, F. Boesel, T. Dockhorn, A. Blattmann, P. Esser, and R. Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*, 2024.

[34] P. Sedigh, R. Sadeghian, and M. T. Masouleh. Generating synthetic medical images by using gan to improve cnn performance in skin cancer classification. In *International Conference on Robotics and Mechatronics*, 2019.

[35] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *Symposium on Security and Privacy*, 2017.

[36] T. Steinke, M. Nasr, and M. Jagielski. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems*, 2024.

[37] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing machine learning models via prediction {APIs}. In *USENIX Security Symposium*, 2016.

[38] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[39] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 2023.

[40] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.