POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Adversarial and Generative Deep Learning for Data Privacy in Human-Centered Artificial Intelligence

Doctoral Programme in
Information Technology

Eugenio Lomurno

Advisor: Prof. Matteo Matteucci
Tutor: Prof. Francesco Amigoni
A.Y. 2023-2024

Abstract

Artificial Intelligence is growing rapidly in a highly interconnected world, providing solutions to problems that were unimaginable just a few years ago, while at the same time opening the door to existential risks and dangers for humanity. Keeping its development under control and respecting the individual is one of the main goals of Human-Centred Artificial Intelligence, a branch of computer science that has emerged in the last decade and aims to make the research, production and use of Artificial Intelligence algorithms transparent, credible, safe and ethical. With the advent of cyber-attacks against such algorithms, regulation and protection have become imperative. Through the use of certain Artificial Intelligence models, it is indeed possible to extract the information learned by third party algorithms, showing how the training data is present in these architectures, albeit in the form of a latent representation. Data, whatever its nature or form, is thus one of the most important and debated resources, being on the one hand the essential ingredient for learning algorithms, and on the other hand an asset to be protected and kept private.

This thesis begins by examining the current landscape of privacy preservation techniques in deep learning, revealing significant challenges in balancing model performance with data protection. Existing methods, including Differential Privacy, often result in substantial compromises with respect to privacy guarantees and model performance, limiting their practical ap-

plication in real-world scenarios. In response to these challenges, this research introduces a series of novel contributions aimed at enhancing both privacy and performance in deep learning systems. Initially, it explores regularisation techniques as a means to improve privacy protection whilst maintaining model performance. This approach proves to be a promising alternative to more computationally intensive methods, offering a better balance between privacy and utility. Building upon this foundation, the work presents Discriminative Adversarial Privacy (DAP), a new strategy that leverages adversarial training to simultaneously optimise for task performance and privacy protection. This approach demonstrates significant improvements over traditional methods, offering a more favourable balance between model accuracy and privacy guarantees. The thesis then investigates the potential of federated learning as a privacy-preserving technique for collaborative model development. Recognising the vulnerabilities inherent in traditional approaches, it proposes Synthetic Generative Data Exchange (SGDE). This innovative method leverages generative models to produce synthetic data for exchange within a federated learning context, significantly enhancing privacy protections whilst maintaining or even improving model performance.

Expanding on the concept of synthetic data, a comprehensive pipeline called Gap Filler (GaFi) is developed to optimise the quality and utility of synthetic datasets for downstream tasks. This approach significantly narrows the performance gap between models trained on synthetic versus real-world data across various domains. Additionally, the research explores the adaptation of Stable Diffusion 2.0 for synthetic dataset generation, incorporating techniques such as transfer learning and fine-tuning. Building upon these advancements, the Knowledge Recycling (KR) pipeline is introduced, which integrates and refines the insights from GaFi and the Stable Diffusion experiments. KR employs advanced generative techniques

to further enhance the effectiveness of synthetic data in model training, demonstrating its potential to surpass real data in certain scenarios. In the context of collaborative learning, this research proposes Federated Knowledge Recycling (FedKR). This novel approach enables secure and effective collaboration across institutions without compromising data privacy. By leveraging locally generated synthetic data and sophisticated aggregation mechanisms, it offers enhanced security and improved model performance compared to traditional federated learning techniques. In conclusion, this thesis presents a series of methodologies and techniques that contribute to the ongoing development of privacy-preserving deep learning. The proposed approaches offer potential solutions to some of the current challenges in balancing data utility and privacy in machine learning applications.

Keywords: Human-Centred Artificial Intelligence, Privacy Preserving Deep Learning, Synthetic Data Generation, Federated Learning, Adversarial Training

Contributions

[1] On the utility and protection of optimization with differential privacy and classic regularization techniques
E. Lomurno, M. Matteucci – Conference
International Conference on Machine Learning, Optimization, and Data Science (LOD), 2022
Online resources: *Paper*

[2] SGDE: Secure Generative Data Exchange for Cross-Silo Federated Learning
E. Lomurno, A. Archetti, L. Cazzella, S. Samele, L. Di Perna, M. Matteucci – Conference
International Conference on Artificial Intelligence and Pattern Recognition (AIPR), 2022
Online resources: *Paper*

[3] Bridging the Gap: Enhancing the Utility of Synthetic Data via Post-Processing Techniques
A. Lampis, **E. Lomurno**, M. Matteucci – Conference
British Machine Vision Conference (BMVC), 2023
Online resources: *Paper*, *Poster*, *Video*, *Supplementary Material*, *Code*

[4] Discriminative Adversarial Privacy: Balancing Accuracy and Membership Privacy in Neural Networks
E. Lomurno, F. Ausonio, A. Archetti, M. Matteucci – Conference
British Machine Vision Conference (BMVC), 2023
Online resources: *Paper*, *Poster*, *Video*, *Supplementary Material*, *Code*

[5] Stable Diffusion Dataset Generation for Downstream Classification Tasks
E. Lomurno, M. D'Oria, M. Matteucci – Conference
European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2024
Online resources: *Paper*

[6] Synthetic Image Learning: Preserving Performance and Preventing Membership Inference Attacks
E. Lomurno, M. Matteucci – Journal
Pattern Recognition Letters - Special Issue on Synthetic Images to Support Computer-Aided Diagnosis Systems, 2025
Online resources: *Paper*, *Supplementary Material*

[7] Federated Knowledge Recycling: Privacy-Preserving Synthetic Data Sharing
E. Lomurno, M. Matteucci – Journal
Pattern Recognition Letters - Special Issue on Synthetic Images to Support Computer-Aided Diagnosis Systems, 2025
Online resources: *Paper*, *Supplementary Material*

Funding

This work has been partially funded by AI-SPRINT: AI in Secure Privacy-pReserving computINg conTinuum (European Union H2020 grant agreement No. 101016577), ESSENCE: Empathic platform to personally monitor, Stimulate, enrich, and aSsist Elders aNd Children in their Environment (European Union H2020 grant agreement No. 101016112) and FAIR: Future Artificial Intelligence Research (NextGenerationEU, PNRR-PE-AI scheme, M4C2, investment 1.3, line on Artificial Intelligence).

I thank every living being with whom I have had the opportunity to interact, regardless of the level and positivity or negativity of that interaction. I thank every contingency, every event, whether directly related to me or not. I thank every variation in my perception of the present, every interpretation of reality, and every phase of change as well as stagnation. I thank curiosity, passion, every act of will as well as their total denial.

In this sense, I thank life.

"Thoughts without content are empty,
intuitions without concepts are blind."
Immanuel Kant, Critique of Pure Reason

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AOP	Accuracy Over Privacy
AUC	Area Under the Curve
CAS	Classification Accuracy Score
CLIP	Contrastive Language-Image Pretraining
DAP	Discriminative Adversarial Privacy
DDPM	Denoising Diffusion Probabilistic Models
DDA	Dynamic Datasets Aggregation
DP	Differential Privacy
DP-SGD	Differentially Private Stochastic Gradient Descent
DP	Denoising Step
FedAVG	Federated Averaging
FedKR	Federated Knowledge Recycling
FID	Fréchet Inception Distance
FL	Federated Learning
GaFi	Gap Filler
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GKD	Generative Knowledge Distillation
HCAI	Human-Centered Artificial Intelligence

IS	Inception Score
KR	Knowledge Recycling
MIA	Membership Inference Attack
RBF	Radial Basis Function
SD	Stable Diffusion
SGD	Stochastic Gradient Descent
SGDE	Secure Generative Data Exchange
UGS	Unconditioned Guidance Scal
VAE	Variational Auto-Encoder
XAI	Explainable AI

Introduction

Deep learning is now the vanguard of artificial intelligence. Thanks to continuous technological advances, it has now penetrated deeply into our society through a process of capillarisation that shows no signs of slowing down. And this process is leading to exciting developments: generative art makes it possible to create masterpieces with simple commands, enabling everyone to express their creativity in ways that were unimaginable just a few years ago; personalised medicine is making great strides thanks to specialised neural systems, improving early diagnosis and treatment of various diseases; precision agriculture optimises crop management through computer vision, increasing productivity and reducing environmental impact. On the other hand, significant challenges are emerging: the proliferation of increasingly sophisticated deepfakes threatens the integrity of information, making it difficult to distinguish the true from the false; profiling algorithms raise serious questions about digital privacy, with the massive collection of personal data raising ethical concerns; the use of deep learning to amplify online polarisation, as seen in social media echo chambers, risks undermining democratic processes and social cohesion, and influencing public debate in potentially damaging and complex ways.

This rapid evolution is not the result of coincidence, but of an unprecedented acceleration over the last decade, fuelled by a synergy of factors: intense interest from the scientific community, huge public and private in-

vestment, and technological advances that have made previously unimaginable computing power available. Although the roots of deep learning go further back in history, a pivotal moment in its development occurred at the beginning of the last decade. This period saw the convergence of three technological milestones: the creation of ImageNet, a dataset of annotated images of unprecedented size; the development and refinement of Convolutional Neural Networks (CNN); and the innovative use of graphics accelerators to parallelise the learning of these networks [8, 9]. This combination gave rise to the modern concept of deep learning, understood as a process of hierarchical, layered and automatic extraction of information using parametric layers. This approach has significantly outperformed previous techniques and revolutionised the field of computer vision in particular. The resulting advances have opened new frontiers in areas such as object recognition, semantic segmentation, image compression and search, laying the foundation for the advanced applications that permeate our daily lives today.

In 2014, another important milestone was reached with the invention of Generative Adversarial Networks (GAN). GANs introduced an innovative paradigm for the generation of high-quality multimedia content. Through a process of competition between a generating and a discriminating network, GANs are able to produce synthetic data that matches the distribution of real data, while being completely novel [10]. This approach has made it possible not only to replicate or rework existing data, but also to "invent" new plausible examples, thus expanding the range of creative possibilities in areas such as image, video and audio generation. The ability of GANs to operate in the latent space of data has opened up new frontiers in the interpolation and semantic manipulation of generated content.

In 2017, the emergence of Transformers represented a decisive revolutionary step forward. Originally designed for natural language processing

tasks, Transformers have demonstrated unprecedented versatility, extending their applicability far beyond the textual domain [11]. The attention mechanism, the cornerstone of this architecture, has enabled the capture of long-range dependencies in data sequences of different types, innovating not only the understanding and generation of text, but also the analysis of images, video and structured data. The impact of Transformers was such that it laid the foundations for the development of large-scale language models capable of generating coherent text, solving complex problems and, to some extent, emulating human reasoning processes.

More recently, around 2020, Denoising Diffusion Probabilistic Models have entered the deep learning landscape, taking image generation to previously unimaginable levels of quality and control [12]. These models have not only perfected the creation of realistic visual content, but have also opened up new frontiers in the manipulation and editing of existing images, with applications ranging from the restoration of artworks to the creation of customised content on a large scale. Subsequently, the Contrastive Language-Image Pre-training (CLIP) model, introduced in 2021, marked another significant advance, bridging the domains of text and images [13]. CLIP paved the way for systems capable of understanding and generating multimodal content, facilitating communication between seemingly distant forms of data. This intermodal 'translation' capability has greatly expanded the application possibilities of deep learning, from visual search based on textual descriptions to image generation driven by linguistic prompts. CLIP has demonstrated an amazing capacity for generalisation, allowing classification tasks to be performed on datasets never seen during training, simply through textual descriptions.

The convergence of these technologies, fuelled by an increasingly active scientific community, evolving hardware and massive investment, has led to the creation of multimodal models of extraordinary versatility. These

systems are capable of processing and generating multiple types of content in an integrated and synergistic manner. The potential and large-scale social impact of these developments has triggered a real race for innovation, with both large corporations and the open source community playing leading roles, each with their own approaches and philosophies.

The focus of research and development remains on perfecting, adapting and "democratising" these technologies for a humanity in a paradoxical position: projected into an unprecedented technological future, yet still anchored in mental patterns and social structures that struggle to adapt to the rapidity of change. This dichotomy raises crucial questions about our collective ability to assimilate and make ethical and constructive use of the innovations produced to date, and highlights the growing gap between the rhythm of technological progress and the speed of societal adaptation.

Nevertheless, efforts are focused on the integration of technologies that allow these multimodal models to adapt to different needs and applications in an increasingly capillary and contextualised way. The desired direction seems to be the development of systems capable not only of processing information, but also of generating new knowledge through processes of abstraction and generalisation that come as close as possible to what we think we have understood about the functioning of the human intellect. These approaches aim to overcome the current limitations of models based on statistical patterns in order to achieve deeper understanding and more articulate reasoning. The very latest frontiers of research are beginning to explore concepts towards the earliest forms of general artificial intelligence and brain-machine interfaces, opening up scenarios that until recently were the domain of science fiction, and raising profound questions about the future of human-machine interaction and the nature of intelligence itself.

Human-Centered Artificial Intelligence

Human-Centred Artificial Intelligence (HCAI) is an emerging paradigm in the field of artificial intelligence (AI) that has developed in parallel with the rise of deep learning. This innovative approach aims to design and implement AI systems that put people at the centre, both as the primary beneficiaries and as active collaborators in the process of developing and using intelligent technologies. HCAI aims to balance technological progress with ethical values and human needs, maximising the benefits of AI while minimising potential risks and negative impacts on society. Despite the various definitions and different points of view, it is possible to identify some key common factors and to state that the HCAI is based on four key principles to guide and control the development and implementation of AI systems [14]:

- **Explainability**. The decision-making processes and internal operations of AI systems should be comprehensible and verifiable. This clarity is valuable for both developers and end-users. Current research in Explainable AI (XAI) explores methods to improve the interpretability of deep learning models, potentially enhancing our understanding of AI technologies [15].
- **Fairness**. The minimisation of bias and discrimination in automated systems is a key concern. This objective requires careful consideration of diversity and inclusion throughout the AI development process, from data collection to algorithm implementation. Fairness in AI seeks to reduce disparate impacts on different demographic groups and promote equitable outcomes across diverse populations [16].
- **Accountability**. The assignment of ethical and legal responsibilities in AI utilisation is an important consideration. This principle involves developing governance mechanisms that enable traceabil-

ity of AI system decisions and establish clear lines of responsibility. Accountability measures aim to ensure responsible use of AI technologies within regulatory frameworks [17].

- **Privacy**. The protection of personal data and adherence to users' privacy rights are significant concerns in AI development. Implementing appropriate data security measures is necessary to safeguard sensitive information. Techniques such as federated learning and differential privacy offer potential approaches for training AI models on distributed data whilst maintaining data confidentiality, addressing the balance between data utility and individual privacy [18].

The practical implementation of these principles has stimulated the development of specific policies and laws by governments and international organisations, which, in response to these challenges, have begun to regulate the development and use of AI.

The European Union's "General Data Protection Regulation" (GDPR), which came into force in 2018, laid the foundations for the protection of personal data in the digital age and has had a significant impact on the development of AI systems. The GDPR introduced concepts such as the "right to explanation" for automated decisions, pushing for greater algorithmic transparency [19]. More recently, the European Union approved the "AI Act", an ambitious attempt to create a comprehensive regulatory framework for AI. The AI Act proposes a risk-based approach, classifying AI applications into categories and imposing stricter requirements for systems considered high-risk [20].

In the United States, while specific federal legislation on AI is still under discussion, some states have taken the initiative. California, for instance, has introduced legislation to regulate specific aspects, such as the use of chatbots and the manipulation of digital content. The California Bot Dis-

closure Law (SB-1001) exemplifies these state-level efforts to address AI-related challenges. China has also been proactive in developing guidelines on ethical AI. The "Ethical Norms for New Generation Artificial Intelligence", published by the Ministry of Science and Technology, outlines China's approach to ensuring the responsible development of AI technologies [21]. Similarly, India launched its "National Strategy for Artificial Intelligence" in 2018, aiming to promote AI adoption across various sectors whilst considering ethical implications. Russia adopted its "National Strategy for the Development of Artificial Intelligence" in 2019, setting objectives and priorities for AI development in the country until 2030 [22]. Brazil, too, has joined the global effort with its "Brazilian Artificial Intelligence Strategy", published in 2021, which aims to guide federal government actions in promoting research, development, and ethical use of AI. At the global level, organisations such as UNESCO have developed ethical guidelines for AI to promote a human-centred approach internationally. UNESCO's "Recommendation on the Ethics of Artificial Intelligence" provides a comprehensive framework for ensuring that AI technologies benefit humanity as a whole.

However, the rapid development of AI technology poses a constant challenge to regulation. Regulators often find themselves chasing innovation, trying to balance the need to protect citizens' rights with the desire not to stifle technological development, which is often faster than their own understanding. In addition, the practical application of concepts such as algorithmic "fairness" and "bias removal" collides with the complexity of social systems and different cultural interpretations of fairness and justice, principles that are often presented as universal but which in practice are highly specific to place and historical moment. HCAI thus stands at the intersection of technological innovation, ethics and legislation. While it promises to create AI systems that are more in tune with human val-

ues and needs, its practical implementation remains a decidedly complex challenge.

Thesis Outline

Following this introductory chapter, the thesis is structured as follows:

1. Chapter 1 critically examines existing privacy preservation techniques in deep learning, with particular emphasis on differential privacy and its inherent limitations. It offers novel insights into regularisation methods and introduces two significant contributions: Discriminative Adversarial Privacy (DAP), an innovative technique designed to enhance model performance whilst maintaining robust privacy guarantees, and Accuracy Over Privacy (AOP), a novel metric that simultaneously captures performance and resilience against membership inference attacks.

2. Chapter 2 explores federated learning approaches as a paradigm for collaborative model development. It presents the Secure Generative Data Exchange (SGDE) method, which leverages generative models to produce synthetic data for exchange within a federated learning context. This approach significantly enhances privacy protections whilst maintaining, and in some cases improving, model performance.

3. Chapter 3 investigates advanced techniques for generating high-utility synthetic datasets. It introduces the Gap Filler (GaFi) pipeline, an innovative framework that integrates multiple post-processing techniques to substantially enhance the performance of generative models. Furthermore, it explores the adaptation of state-of-the-art models, such as Stable Diffusion 2.0, for downstream classification tasks, pushing the boundaries of synthetic data utility.

4. Chapter 4 presents the Knowledge Recycling (KR) pipeline, incorporating the novel Generative Knowledge Distillation (GKD) technique. This approach significantly narrows the performance gap between models trained on synthetic versus real data. The chapter demonstrates the pipeline's effectiveness across a diverse range of datasets, with particular emphasis on its applicability to medical imaging data.

5. Chapter 5 introduces Federated Knowledge Recycling (FedKR), a novel approach that combines improved synthetic dataset generation with federated learning principles. This method is specifically designed to enhance both privacy and performance in cross-silo environments, addressing key challenges in collaborative learning for sensitive domains, with a particular focus on healthcare applications.

The thesis concludes with a comprehensive discussion of the findings and contributions to the field of privacy preserving deep learning. This final chapter examines the broader implications of this research for future developments in artificial intelligence and data privacy, and delineates promising directions for further investigation.

1 | Privacy Preserving Deep Learning

In recent years, the widespread adoption of deep learning models across various industries has underscored the paramount importance of safeguarding data privacy, particularly in light of the sensitive information frequently utilised during model training. Differential Privacy (DP) has emerged as a prominent and widely acknowledged method for ensuring data protection within artificial intelligence frameworks, predominantly through techniques such as Differentially Private Stochastic Gradient Descent (DP-SGD). Despite the robust defence DP offers against various privacy threats, particularly membership inference attacks, its implementation often leads to significant reductions in model performance and heightened computational demands, thus raising questions about its viability in practical, real-world applications.

This chapter addresses these challenges through a dual-faceted approach. Initially, an in-depth evaluation is conducted to compare the efficacy of DP-SGD with standard optimisation techniques that incorporate regularisation, focusing on the resulting model performance and susceptibility to privacy breaches. The analysis exposes the limitations inherent in DP and empirically demonstrates the potential of regularisation as a viable alternative for privacy preservation.

Subsequently, in response to the challenges posed by DP, a novel strategy termed Discriminative Adversarial Privacy (DAP) is introduced. DAP seeks to harmonise privacy protection with model accuracy and training efficiency by leveraging adversarial training alongside an innovative loss function designed to concurrently minimise prediction errors and elevate the error rates of membership inference attacks. To measure the balance between performance and privacy, a new metric—Accuracy Over Privacy (AOP)—is proposed. Extensive comparisons between DAP and various DP methodologies reveal that DAP outperforms standard DP approaches, achieving a 15% enhancement in model accuracy and a 30% reduction in training time, all while maintaining comparable levels of privacy protection.

1.1. Introduction

The rapid expansion and diversification of deep learning applications within modern society have been primarily propelled by advancements in computational resources, the availability of extensive datasets, and the continuous evolution of neural network architectures alongside optimisation algorithms. The efficacy of these models is tightly interwoven with both the volume and quality of the data employed during training, prompting an ongoing pursuit to amass increasingly larger datasets in order to secure reliable and precise outcomes. Nevertheless, this data-centric approach is accompanied by significant challenges, particularly concerning data security and privacy.

In an era defined by rigorous data protection regulations, such as the European General Data Protection Regulation (GDPR), the safeguarding of sensitive data utilised in the training of deep learning models has emerged as a critical issue [23]. Ensuring the security and integrity of models trained

on such data is essential, especially in the context of data sharing. Traditional security measures, although robust, have proven vulnerable to the rapidly evolving landscape of cyber threats. The rise of deep learning has not only amplified the potency of pre-existing attack vectors but has also introduced novel vulnerabilities that specifically exploit these models.

Among the most notable and perilous techniques are membership inference attacks, first introduced by Shokri *et al.* [24]. These attacks aim to ascertain whether a specific data instance was included in the training set of a compromised model. Equally concerning is the model inversion attack, which seeks to reconstruct input data from accessible or leaked model outputs. These attacks pose substantial risks and form the groundwork for more sophisticated and potentially devastating security breaches.

To counteract the effectiveness of such attacks on artificial neural networks, a prevalent strategy involves integrating noise into the training process to enforce differential privacy. Differential privacy is widely regarded as the standard mechanism for preserving privacy, attributed to its theoretical soundness and robustness, as evidenced by the work of Abadi *et al.* [25]. The primary advantage of this approach lies in its ability to provide a quantifiable privacy budget, thereby ensuring that privacy leakage remains both constrained and measurable post-training.

Nonetheless, the implementation of this technique is not devoid of limitations. The high level of privacy afforded by differential privacy often incurs a substantial cost. The noise injection necessary to attain the desired privacy budget can significantly hinder the model's training efficiency, impacting both time and utility. It has been demonstrated that enforcing a stringent level of differential privacy typically results in considerable performance degradation and elongated training periods, rendering it impractical for many real-world scenarios, and sometimes even for basic simulations [26].

The delicate balance between model performance and privacy represents a formidable challenge for the deep learning community. As the threats and potential for data breaches evolve at a pace that frequently surpasses regulatory frameworks, the development of robust, efficient, and practical privacy-preserving techniques becomes increasingly crucial. These techniques must adeptly balance the necessity for high-quality training data with the imperative to protect sensitive information, thereby ensuring that the sustained advancement and application of deep learning technologies do not come at the cost of individual privacy and data security.

1.1.1. Main Contributions

This chapter delves into two complementary approaches for privacy preservation within deep learning frameworks, both aimed at mitigating the risks posed by membership inference attacks while ensuring the retention of model utility and training efficiency.

The first investigation focuses on the role of differential privacy in deep learning models, with particular attention given to the application of differentially private stochastic gradient descent (DP-SGD) as a defensive measure. This study not only evaluates the robustness of models against privacy breaches but also considers the effects on model accuracy and training duration. Furthermore, the analysis integrates two widely recognised regularisation techniques, Dropout and L2 regularisation, which are known to enhance a model's generalisation capabilities. These techniques are examined in relation to the established correlation between a model's vulnerability to privacy attacks and its degree of overfitting. The empirical findings suggest that L2 regularisation and Dropout can offer privacy-preserving benefits that are comparable to or even exceed those provided by DP-SGD, all while maintaining both model utility and training efficiency.

The second line of research introduces an innovative privacy-preserving technique termed Discriminative Adversarial Privacy (DAP). This approach leverages the framework of membership inference attacks to implement multi-objective adversarial learning. In DAP, a discriminator is trained using shadow models within the context of membership inference attacks and is employed to introduce a regularisation component into a novel loss function. This loss function is designed to simultaneously minimise prediction error, akin to standard loss optimisation, and maximise the error of an attacker. The resultant models achieve privacy levels on par with differential privacy, yet they do so with significantly less performance degradation and reduced training times.

The principal contributions and findings of these studies are summarised as follows:

- A comparative evaluation of DP-SGD, Dropout, and L2 regularisation as privacy-preserving techniques, revealing that traditional regularisation methods can attain privacy outcomes that are comparable to or superior to those achieved with DP-SGD, while preserving model utility.
- The introduction of Discriminative Adversarial Privacy (DAP), a novel learning technique that merges adversarial learning principles with membership inference attacks to effectively balance model performance, training efficiency, and privacy preservation.
- The development of a new loss function tailored for DAP, specifically designed to optimise the trade-off between prediction accuracy and privacy protection.
- The proposal of a novel metric, termed Accuracy Over Privacy (AOP), to better capture and manage the trade-off between model performance and privacy preservation.

- A thorough empirical validation demonstrating that DAP provides a comparative advantage over differentially private optimisation in terms of model performance, training time, and the preservation of privacy.

1.2. Related Works

The issue of privacy in machine learning has garnered significant attention in recent years. Although deep learning models are adept at solving complex problems by learning from large datasets, they remain susceptible to various forms of adversarial attacks [27–29]. These vulnerabilities expose users to varying degrees of risk, with membership inference and model inversion attacks standing out as some of the most critical and concerning threats [30].

1.2.1. Membership Inference Attacks

Among the numerous threats to deep learning models, membership inference attacks are particularly notable, having spurred considerable research into both offensive strategies and defensive techniques [31]. These attacks are designed to determine whether a specific data point was included in the training set of a target model, thereby compromising the confidentiality of the training data.

The foundational work by Shokri *et al.* [24] remains one of the most impactful contributions to this field. Their method capitalises on the tendency of overparameterised models to retain information about individual training instances, which extends beyond mere generalisation of the task at hand. The core technique involves the construction of multiple shadow models, which are trained on surrogate datasets that mimic the target model's training data in both structure and distribution. The outputs of

these shadow models, combined with their corresponding labels, are used to create an attack dataset. This dataset then trains a metamodel that infers the membership status of data points based on the behaviour of the shadow models. However, the effectiveness of this approach is constrained by the shadow models' fidelity in replicating the target model and the assumptions made about the adversary's knowledge concerning the target model's architecture and training data distribution.

In response to these limitations, Salem *et al.* [32] introduced several attack methods that operate under less stringent assumptions regarding the adversary's knowledge. The most prominent of these is the threshold-based attack, where a simple binary classifier evaluates the maximum posterior probability from the target model's prediction vector against a predefined threshold. If this maximum exceeds the threshold, the data point is classified as part of the training set. This method offers significant advantages, including independence from the target model's specifics, elimination of the need for shadow models, and avoidance of attack model training.

The scope of membership inference attacks has continued to expand with recent advancements. For instance, Chen *et al.* employed data poisoning techniques to enhance the accuracy of membership inference attacks while mitigating the impact on test-time performance [33]. He *et al.* demonstrated the feasibility of these attacks against self-supervised learning models and investigated early stopping as a possible defence mechanism [34]. Researchers have also assessed membership inference attacks across various model types, such as GANs [29, 35], diffusion models [36, 37], recommender systems [38, 39], semantic segmentation models [40, 41], and text-to-image models [42]. Additionally, Hui *et al.* proposed the innovative BlindMI technique, which enhances the attacker's capabilities by leveraging differential comparison and data generation to extract membership semantics from the target model [43].

1.2.2. Model Inversion Attacks

Model inversion attacks are a sophisticated class of techniques aimed at reconstructing training data from trained deep learning models. These methods have evolved significantly, from initial white-box approaches to more advanced black-box strategies.

In the early stages of the field, researchers developed reconstruction techniques that relied on extensive model knowledge. These early approaches, known as maximum a-posteriori (MAP) attacks, exploited full access to model parameters, output labels, and prior feature distributions to estimate sensitive feature values [44]. However, it was found that the effectiveness of MAP attacks decreases significantly when applied to high-dimensional feature spaces.

To address these limitations, subsequent research has reframed the attack paradigm as an optimisation problem [45]. This novel approach uses gradient descent to reconstruct the original samples, with the objective function depending on the output of the target model. The versatility of this method allows it to be applied in both white-box and black-box scenarios, depending on the level of access to the target model's architecture.

Further advances in the field have led to the development of attack methods that operate without specific knowledge of the model or its training data [46]. Instead, these approaches rely on a general understanding of the distribution of the training data and the output format of the model. Building on this foundation, many researches have adopted generative adversarial networks (GANs) to learn representations of the training data, thereby facilitating more sophisticated inversion attacks [10, 47].

The exploitation of artificial intelligence explainability tools has emerged as another avenue for achieving high fidelity input data reconstructions [48]. This line of research focuses on identifying the explanation methods that

are most advantageous to potential attackers by assessing the degree of information leakage associated with the target data.

In the context of federated learning, novel techniques have been developed to reconstruct user input data from leaked gradients [49]. These methods involve the generation of dummy gradients from randomly initialised inputs, followed by an iterative refinement process to minimise the discrepancy between these dummy gradients and the real ones, building on previous work in this area [50].

A significant advance in this area is the GradInversion technique, which demonstrates the ability to accurately recover individual images from gradients of neural networks, even when trained with substantial batch sizes [51]. This approach reformulates the input reconstruction task as an optimisation problem where synthetic images generated from random noise are iteratively adapted. The innovation of GradInversion lies in its ability to recover labels on a batch-wise basis and apply auxiliary losses, ensuring both the fidelity of the reconstructed images and consistency across output groups.

1.2.3. Differential Privacy

The concept of differential privacy was originally developed for database query contexts, wherein adjacent databases are characterised by a single entry difference. A formalisation of this concept involves a randomised mechanism $M\text{:}\ D \rightarrow R$, where D represents the domain and R the range. This mechanism achieves ε-differential privacy if, for any pair of adjacent inputs $d,\ d' \in D$ and any output subset $S \subseteq R$, the following inequality is satisfied:

$$Pr[M(d) \in S\ ] \leq e^{\varepsilon} Pr[M(d') \in S\ ]. \tag{1.1}$$

In this formulation, ε denotes the privacy budget, which serves as a metric for the acceptable degree of information disclosure. A lower ε value correlates with enhanced privacy protection.

The significance of differential privacy in the realm of privacy preserving techniques stems from its provision of three essential attributes:

- **Composability:** This attribute facilitates the development of intricate mechanisms whilst preserving differential privacy. The principle operates under the condition that each component adheres to the privacy requirements. Composability's scope encompasses both sequential and parallel structural arrangements, enabling the construction of complex privacy preserving systems.
- **Group privacy:** In scenarios involving datasets with interrelated information, such as multiple entries from a single source, this feature ensures a measured and controlled diminution of privacy safeguards. Group privacy prevents an abrupt collapse of protective measures, instead allowing for a gradual reduction in privacy assurances as correlations increase.
- **Robustness:** This characteristic guarantees the persistence of privacy levels despite potential external knowledge. The robustness property ensures that privacy guarantees remain steadfast, regardless of any auxiliary information an adversary might possess, thereby maintaining the integrity of the privacy preserving mechanism in diverse informational contexts.

Whilst differential privacy offers robust safeguards, its stringent nature presents substantial theoretical challenges. To enhance its applicability in practical scenarios, researchers have proposed various relaxations of the privacy budget ε. Two particularly notable relaxations are (ε, δ)-differential privacy and Rényi differential privacy [25, 52].

The framework of **(ε, δ)-differential privacy** is defined by a randomised mechanism *M:* $D \rightarrow R$, which satisfies the following condition for any pair of adjacent inputs *d, d'* $\in D$ and any output subset $S \subseteq R$:

$$Pr[\ M(d) \in S\ ] \leq e^{\varepsilon} Pr[\ M(d') \in S\ ] + \delta, \tag{1.2}$$

In this context, δ represents the probability of breaching pure ε-differential privacy. The property of composability is preserved even when multiple mechanisms are employed, albeit with a more intricate formulation that accounts for the cumulative privacy loss incurred by each component. Building upon this composability, Abadi *et al.* introduced the moments accountant, a function designed to monitor the privacy cost associated with each data access, thereby enabling the computation of the overall privacy loss for the mechanism [25]. In the same research, they proposed the differentially private stochastic gradient descent (DP-SGD) algorithm, which has subsequently emerged as one of the most widely adopted optimisers for implementing differential privacy in deep learning systems.

Rényi Differential Privacy. The concept of Rényi differential privacy represents a nuanced relaxation of traditional privacy measures, grounded in the mathematical framework of Rényi divergence. This approach offers a more flexible and potentially powerful tool for quantifying and ensuring privacy in complex data environments.

At the core of this methodology lies the Rényi divergence, a measure of dissimilarity between probability distributions. For any two probability distributions P and Q defined over a set R, the Rényi divergence of order α (where $\alpha > 1$) is formally expressed as:

$$D\alpha(P \parallel Q) \triangleq \frac{1}{\alpha - 1} \log Ex \sim Q \left(\frac{P(x)}{Q(x)} \right)^{\alpha}. \tag{1.3}$$

The interconnection between differential privacy and Rényi divergence can be elucidated through a mathematical relationship. A randomised mechanism *M: D* $\rightarrow$ *R* achieves ε-differential privacy if and only if its distribution, when applied to any pair of adjacent inputs *d, d'* $\in$ *D*, satisfies the following inequality:

$$D_{\infty}(M(d) \parallel M(d')) \leq \varepsilon. \tag{1.4}$$

Building upon this foundation, the notion of (α, ε)-Rényi differential privacy is introduced. This refined concept stipulates that a randomised mechanism *M: D* $\rightarrow$ *R* adheres to ε-Rényi differential privacy of order α if, for all adjacent inputs *d, d'* $\in$ *D*, the following criterion is met:

$$D_{\alpha}(M(d) \parallel M(d')) \leq \varepsilon. \tag{1.5}$$

Notably, research has demonstrated that (α, ε)-Rényi differential privacy preserves several crucial properties associated with traditional differential privacy. These include composability, robustness to auxiliary information, and group privacy, as evidenced in the work of Mironov *et al.* [52]. The preservation of these essential characteristics underscores the potential of Rényi differential privacy as a valuable tool in the ongoing development of privacy preserving data analysis techniques.

1.2.4. Alternative Privacy Preservation Approaches

Differential privacy has become a prominent approach for ensuring privacy in machine learning, owing to its rigorous guarantees and the explicit quantification of the privacy budget [53–55]. However, research by Bagdasaryan *et al.* has highlighted a significant drawback: the application of

differential privacy via DP-SGD can adversely affect model performance, particularly at lower values of ε [26]. This performance degradation arises because stronger privacy necessitates the introduction of more noise during the training process, coupled with a reduction in the number of training iterations, which can impair the model's ability to learn effectively. Furthermore, this trade-off aligns with the findings of Salem *et al.*, who demonstrated a direct correlation between the degree of overfitting in deep learning models and their susceptibility to membership inference attacks[32]. This relationship is intuitive: the noise introduced by DP-SGD functions as a form of regularisation, and differential privacy serves as a key defence against membership inference attacks by reducing overfitting.

In response to these challenges, alternative techniques have been proposed to enhance privacy protection in deep learning without compromising model performance as severely. For instance, Jain *et al.* investigated the inherent differential privacy properties of Dropout and assessed its effectiveness in defending against membership inference attacks [56, 57]. Additionally, Ermis *et al.* introduced a variant of differential privacy inspired by the Bayesian interpretation of Gaussian Dropout, offering another potential solution to the privacy-utility trade-off [58].

Another line of research includes adversarial regularisation, proposed by Nasr *et al.*, as a defence mechanism to improve the robustness of models against privacy attacks [59]. Yang *et al.* further extended this idea by developing a technique called prediction purification, which uses adversarial learning to protect models from both membership inference and model inversion attacks [60].

Chen *et al.* introduced a novel approach called RelaxLoss, which offers an alternative to differential privacy by modifying the entropy loss function in training [61]. This method aims to reduce the generalisation gap while simultaneously mitigating privacy leakage, presenting a promising direction

for balancing privacy and performance.

In another related study, Kaya and Dumitras conducted an extensive evaluation of various data augmentation techniques to determine their effectiveness in reducing the risk of membership inference attacks, particularly in image classification tasks [62]. Their analysis covered seven different privacy preserving mechanisms, including differential privacy. While they found that data augmentation could enhance model utility, it did not necessarily lower the risk of membership inference attacks. Moreover, their study revealed a paradoxical finding: label smoothing, a commonly used regularisation technique, may actually increase the vulnerability of models to such attacks, contrary to its intended purpose of improving generalisation.

1.3. Regularisation Privacy Properties

This research presents a comprehensive examination of deep learning models, focusing on the integration and evaluation of privacy preserving techniques. Particular attention is given to differentially private stochastic gradient descent (DP-SGD) and regularisation methods, assessing their efficacy in safeguarding against privacy threats whilst maintaining model accuracy and training efficiency.

The implementation of DP-SGD utilised the TensorFlow Privacy library, adhering to the (ε, δ)-differential privacy framework proposed by Abadi *et al.*[25]. To ensure equitable comparison, a consistent model architecture was maintained throughout all experiments, as depicted in Figure1.1. The chosen architecture comprised a straightforward convolutional neural network (CNN), employing ReLU activation functions in hidden layers and a softmax function in the output layer.

The evaluation of the model's resilience against membership inference at-

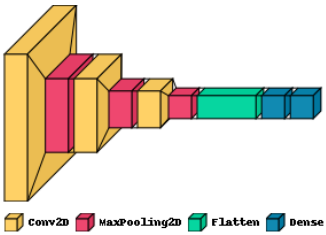

Figure 1.1: Architecture of the target model's neural network [1].

tacks employed two distinct strategies, implemented using a tool from TensorFlow Privacy based on the work of Salem *et al.* [32]. These strategies encompassed a threshold-based approach and a shadow model technique, both operating without prior knowledge of data distribution. The tool autonomously selected the most effective attack strategy, utilising the Area Under Curve (AUC) metric to quantify attack efficacy.

In addition to membership inference, a model inversion attack was developed. This attack methodology leveraged activation maps to reconstruct the target model's training data, as illustrated in Figure 1.2. Post-training of the target model, a specific layer was identified to facilitate input reconstruction. The architecture of the attacking network is delineated in Figure 1.3, where the sequential layer corresponds to the immutable segment of the target model. Following extensive optimisation, the mean squared error (MSE) was selected as the loss function, with SiLU serving as the activation function for all hidden layers. The Adam optimiser was employed with a learning rate of 10^{-3}. The output layer consisted of a convolutional layer with a sigmoid activation function, corresponding to the number of channels in the target model's input data. The efficacy of the inversion attack was evaluated using the MSE metric, quantifying the fidelity of the reconstructed data.

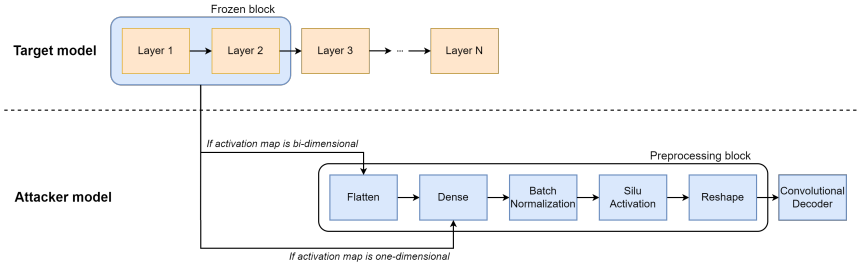

Figure 1.2: Conceptual scheme of the proposed model inversion technique. The process reconstructs input data from intermediate or final layer outputs, depending on the degree of accessibility of the system [1].

To incorporate differential privacy into the target model's training process, the standard optimiser was supplanted by DP-SGD, whilst maintaining the original architecture and parameter configuration. To comprehensively assess the impact of privacy-preserving techniques, experiments were conducted across three distinct privacy budget levels: $\varepsilon = 2$, $\varepsilon = 4$, and $\varepsilon = 8$, with a constant privacy leakage probability of $\delta = 10^{-5}$. Discrete models were trained for each privacy budget level.

Concurrently, additional target models were trained utilising various regularisation techniques to explore their impact on model performance. The study focused on prevalent anti-overfitting strategies, specifically dropout and weight decay (L2 regularisation). Models were trained with each technique independently and in combination. Dropout was applied following each convolutional layer and before and after the initial dense layer, whilst L2 regularisation was confined to the final dense layer with softmax activation.

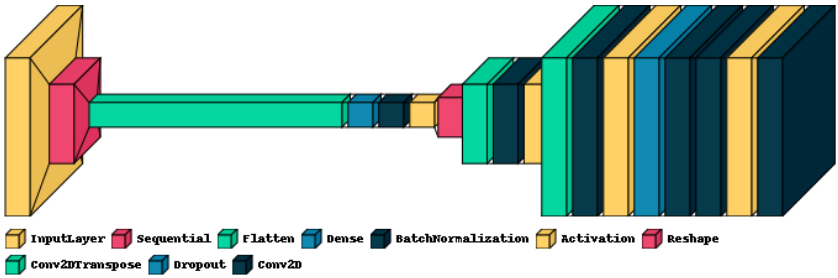

Figure 1.3: Architecture of the adversary model used in the model inversion attack [1].

1.3.1. Experiments and Results

The experimental phase was conducted on a system equipped with an Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz and an Nvidia GeForce GTX TITAN X GPU. The analysis encompassed three image datasets – CIFAR10 [63], MNIST [64], and FashionMNIST [65] – each subjected to a normalisation pre-processing step. For detailed dataset information, refer to Appendix A.

To ensure experimental consistency, specific hyperparameters were standardised across different target model configurations. The privacy budget parameter, ε, was typically calculated a-posteriori to assess the privacy level achieved by models utilising a differentially private optimiser. As elucidated by Abadi *et al.* [25], ε is a function of the number of data samples, micro-batches, noise multiplier, and training epochs. In the experimental setup, differentially private target models were trained on identical data as non-private models, with a fixed batch size of 200. To optimise utility, the number of micro-batches was equated to the number of mini-batches. Furthermore, the number of training epochs was fixed at the optimal value determined for the target model without privacy-preserving

Table 1.1: Test accuracy of the target models across datasets (higher values indicate better performance).

Dataset	Baseline	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	$DP_{\varepsilon=8}$	L2	Dropout	L2+Dropout
CIFAR10	66.41	51.80	53.87	53.65	64.97	**69.22**	64.81
MNIST	99.25	94.27	96.38	96.85	99.18	**99.64**	99.42
FashionMNIST	**90.07**	82.05	82.37	82.92	89.66	89.04	87.93

techniques, enabling control of the ε privacy budget through variation of the noise multiplier parameter. Similar considerations were applied to regularisation techniques, maintaining constant batch size, data sample number, and optimiser. Following extensive tuning, the Dropout rate was established at 20

The performance results of each target model on the test datasets, summarised in Table 1.1, corroborate the findings of Bagdasaryan *et al.* [26]. A performance decrease inversely proportional to the privacy budget ε is evident. This behaviour aligns with the expectation that heightened levels of differential privacy, entailing greater noise injection during training, would result in diminished model accuracy. Notably, in the most privacy-preserving scenario ($\varepsilon = 2$), the target model exhibited an accuracy reduction ranging from 5.0% to 14.6%. Other differentially private configurations displayed slightly lower, yet still significant, performance losses. Regarding regularisation techniques, L2 regularisation yielded results nearly identical to the baseline, whilst Dropout consistently enhanced performance, particularly on CIFAR10. However, the combination of L2 and Dropout resulted in a marginal performance degradation, potentially due to excessive regularisation.

Table 1.2 delineates the time required by each target model configuration to complete an equivalent number of epochs across the datasets. Models trained with DP-SGD exhibited significantly extended training dura-

Table 1.2: Average training time (in seconds) for each model and configuration.

Dataset	Baseline	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	$DP_{\varepsilon=8}$	L2	Dropout	L2+Dropout
CIFAR10	73.30	3029.81	2923.17	2897.12	69.11	**67.4**	70.47
MNIST	39.57	1886.24	1870.25	1859.8	31.20	**29.56**	29.67
FashionMNIST	**77.46**	3876.72	3952.14	3967.11	78.20	83.0	78.55

Table 1.3: AUC scores of membership inference attacks on the proposed models across datasets (lower values indicate better privacy protection).

Dataset	Baseline	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	$DP_{\varepsilon=8}$	L2	Dropout	L2+Dropout
CIFAR10	68.92	**52.71**	54.52	53.64	61.95	57.16	55.14
MNIST	73.81	58.53	59.84	58.34	56.75	61.47	**55.51**
FashionMNIST	64.24	**54.12**	56.46	54.97	56.48	55.24	57.54

tions compared to other configurations, with increases ranging from 41 to 50 times for $\varepsilon = 2$. This substantial temporal augmentation can be attributed to the specific implementation requirements of DP-SGD. The methodology necessitates an expansion of the tensor input to the network, incorporating an additional micro-batch channel. The quantity of micro-batches, which is constrained to fall between unity and the total number of mini-batches, represents a critical compromise between computational efficiency and model accuracy. An increase in micro-batch quantity results in prolonged computation times but enhanced performance, whilst the converse holds true for a reduction in micro-batches. In this configuration, prioritising performance inevitably led to a noticeable drop in accuracy. Conversely, models employing regularisers demonstrated training durations analogous to the baseline.

An examination of the results of membership inference attacks against the target models, quantified by the AUC metric and presented in Table 1.3,

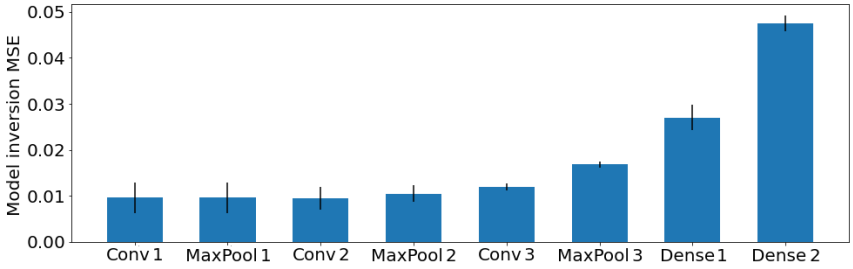

Figure 1.4: Mean reconstruction MSE for model inversion attacks against the baseline target model, averaged across all datasets and stratified by network layer, illustrating attack effectiveness at different depths [1].

provides crucial insights into the efficacy of various privacy-preserving techniques. Models trained using DP-SGD exhibited remarkable resilience, with performance metrics approximating those of a random guessing attack. This approximation to random guessing is indicative of robust privacy preservation. Notably, when $\varepsilon = 2$, the attacker experienced a significant average reduction in AUC of 13.86%, underscoring the effectiveness of DP-SGD in mitigating membership inference vulnerabilities. Target models incorporating regularisation also demonstrated enhanced protection against membership inference attacks compared to the baseline. L2 regularisation outperformed DP-SGD in scenarios where the latter was less effective, achieving superior results on the MNIST dataset and an average AUC reduction of 10.63%. Dropout exhibited similarly competitive privacy guarantees with an average AUC reduction of 11.07%. The combination of L2 and Dropout further improved the average score, resulting in an AUC reduction of 12.93%. These results suggest that regularisation techniques can provide privacy protection comparable to that achieved by differential privacy optimisers.

Figure 1.4 presents the outcomes of model inversion attacks against the

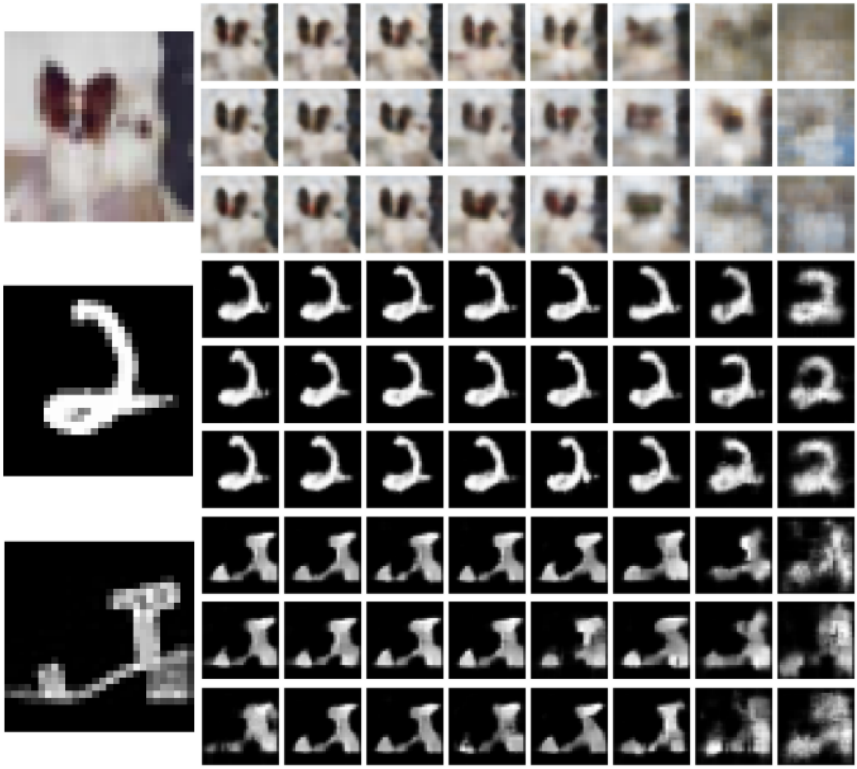

Figure 1.5: Visual comparison of model inversion attack reconstructions for CIFAR10, MNIST, and FashionMNIST datasets across baseline, $DP_{\varepsilon=2}$, and L2+Dropout model configurations, showing reconstructions from each network layer [1].

Table 1.4: Percentage variation in model inversion reconstruction MSE relative to the baseline model, stratified by network layer and averaged across all datasets, for different privacy preserving techniques.

Dataset	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	$DP_{\varepsilon=8}$	L2	Dropout	L2+Dropout
Conv1	2.13	2.51	-1.44	-0.74	**22.01**	10.80
MaxPool1	3.11	0.69	1.73	-0.52	**19.04**	-0.21
Conv2	4.60	0.72	3.21	-2.99	**18.26**	1.61
MaxPool2	3.42	-0.50	-1.31	-3.28	**28.21**	4.23
Conv3	4.20	-9.35	-10.53	-7.13	**11.42**	2.74
MaxPool3	-14.63	-15.64	-17.27	-11.51	**8.60**	-3.64
Dense	-21.71	-25.83	-25.20	17.12	17.3	**21.33**
Prediction	-17.72	-18.25	-15.83	**73.89**	-1.84	64.21

baseline target model. Each bar represents a layer, displaying the reconstruction Mean Squared Error (MSE) averaged across all evaluated datasets. The findings indicate that model inversion attacks demonstrate greater efficacy when executed closer to the network input. This observation aligns with logical expectations, given the increasing complexity of transformations that the attacker model would need to reverse when targeting deeper layers. Notably, the attack exhibited approximately four times greater effectiveness when initiated from the final dense layer.

Table 1.4 delineates the percentage change in reconstruction error for each layer and approach, relative to the baseline. The results reveal that models trained with DP-SGD exhibited a reduction in MSE across all final layers. This decrease in reconstruction error suggests an advantage for the attacker, indicating that, at least in this particular scenario, differential privacy might inadvertently enhance the effectiveness of model inversion attacks. In contrast, models employing regularisation techniques demonstrated markedly different behaviour. L2 regularisation substantially mitigated the effects of model inversion in the layer where it was applied, yielding an average improvement in reconstruction error of 73.89%. Dropout

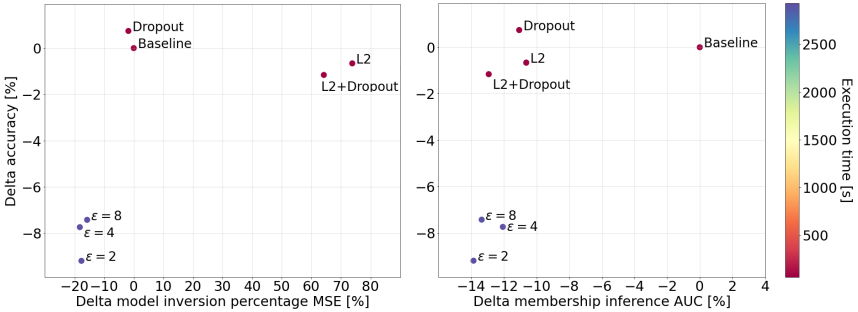

Figure 1.6: Comparative analysis of model performance and privacy protection efficacy in black-box scenarios, averaged across all datasets. Left: Model inversion attack resistance (MSE variation) vs classification accuracy. Right: Membership inference attack resistance (AUC variation) vs classification accuracy. Colour indicates relative training time [1].

exhibited a similar effect, enhancing the protection of layers prior to its application. However, the inability to apply Dropout after the prediction layer limits its protective capacity. The combination of L2 and Dropout resulted in an amalgamation of the two techniques' effects, marginally improving the protection of hidden layers while augmenting the protection of the prediction layer.

Figure 1.5 illustrates three examples of comprehensive model inversion attacks for the primary target model configurations. The images demonstrate the progressive deterioration in the quality of the attacker's reconstructions as the reconstruction process initiates from deeper layers. Consistent with the numerical results, reconstructions from DP-SGD target models tend to exhibit fewer artefacts compared to those derived from the baseline or the combined L2 and Dropout regularisation approach.

Figure 1.6 presents a comprehensive summary of all obtained results. The

left-hand figure depicts the percentage variation of the MSE for black-box model inversion (i.e., an attack on the final dense layer) as a function of accuracy variation. In this representation, the least favourable results are situated in the lower left region, while the most favourable outcomes are in the upper right region. The right-hand plot illustrates the percentage variation in membership inference AUC as a function of accuracy variation, with the optimal region in the upper left and the least desirable in the lower right. The scatter plot's colour scheme clearly distinguishes between training times with and without the DP-SGD optimiser.

In conclusion, despite the undeniable advantages in terms of protection against membership inference attacks and the capacity to quantify privacy, differential privacy techniques may not invariably represent the most suitable option. In the context of this study, this mechanism demonstrated a suboptimal trade-off between protection guarantees, utility, and training time. Conversely, the investigated regularisation techniques exhibited promising performance, positioning them as viable alternatives in scenarios where training time and performance requirements are subject to stringent constraints.

1.4. Discriminative Adversarial Privacy

Discriminative Adversarial Privacy. This work presents a novel learning framework for privacy-preserving deep learning, called Discriminative Adversarial Privacy (DAP) and depicted in Figure 1.7. The DAP framework facilitates the efficient training of high-performance deep learning models that exhibit significant resilience against membership inference attacks (MIA). At the core of DAP is a deep neural network classifier, denoted $\mathcal{C}_{base}$, which is trained using the hold-out method on the dataset *Data*. In parallel, K shadow models, denoted as $\mathcal{C}_{S,k}$, are also trained us-

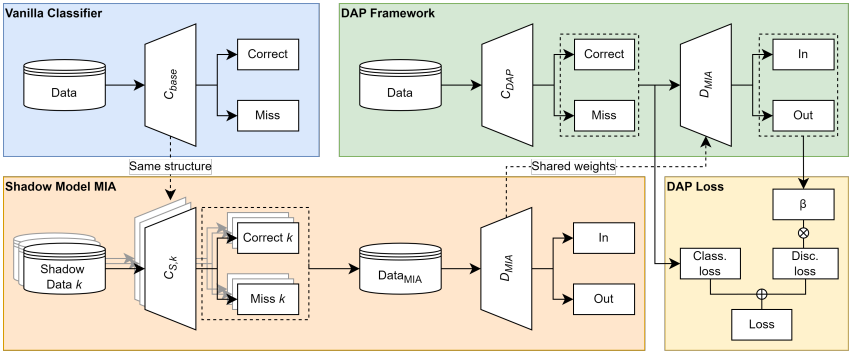

Figure 1.7: Schematic representation of the Discriminative Adversarial Privacy framework [4].

ing the hold-out technique, following the methodology outlined by Shokri *et al.* [24]. For each shadow model, the ground truth, predictions and associated loss values are recorded for both the training and test samples. Each sample is labelled with a binary value indicating its membership in the training or test set. These labelled data points are then used to construct the adversarial binary classification dataset $Data_{MIA}$, which is then used to train the binary discriminator $\mathcal{D}_{MIA}$. Once the discriminator is fully trained, its weights are kept in a fixed state, ensuring consistent application in subsequent phases of the framework.

At this stage, adversarial training is performed using the discriminator $\mathcal{D}_{MIA}$ alongside a new classifier, $\mathcal{C}_{DAP}$, which has the same architecture as the baseline model $\mathcal{C}_{base}$. In the DAP framework, the classifier $\mathcal{C}_{DAP}$ is initially trained to minimise the categorical cross-entropy loss, thereby maximising the probability of correctly classifying the training examples. This error is used to update the weights across all layers of the classifier. Subsequently, for each batch of data, the predictions from $\mathcal{C}_{DAP}$ are

collected along with their corresponding ground truth labels and associated loss values. This batch is then fed through the discriminator $\mathcal{D}_{MIA}$, and the prediction error of the discriminator is calculated with the aim of maximising this error in accordance with the min-max adversarial training strategy outlined by Goodfellow *et al.* [10]. The secondary error generated by this process is then used to update the final dense layer of $\mathcal{C}_{DAP}$. The goal of this update is to reduce the extent to which the classifier's outputs can be accurately distinguished by a potential adversary, thereby increasing the model's privacy preserving capabilities. The optimisation procedure of DAP framework can be described as follows:

$$\min_{\mathcal{C}} \max_{\mathcal{D}} \mathcal{L}(\mathcal{C}, \mathcal{D}, t) = \mathbb{E}_{x \sim p(x)}[\log(\mathcal{C}(x, t))] + \beta \mathbb{E}_{x,y \sim p(x,y)}[\log(1 - \mathcal{D}(\mathcal{C}(x, t), y))]. \tag{1.6}$$

In Equation (1.6), the variables x and y denote the training inputs and the ground truth labels respectively. The symbol t represents the current epoch, while β serves as a dynamic parameter that balances the loss terms. The role of β is fundamental in maintaining stability throughout the learning process. This need arises from the fact that the losses, due to their different characteristics, can vary significantly in magnitude depending on both the stage of training and the particular distribution of the data. Without proper compensation, these differences could lead to instability in the learning process. The parameter β is dynamically adjusted throughout the training process and is determined by the following equation:

$$\beta(\mathcal{C}, \mathcal{D}, t, r) = \begin{cases} \frac{\mathbb{E}[\log(\mathcal{C}(x, t-1))]_v}{\mathbb{E}[\log(1-\mathcal{D}(\mathcal{C}(x, t-1), y))]_v} \cdot r & \text{if } t > 0 \\ 1 & \text{otherwise} \end{cases}. \tag{1.7}$$

As expressed in the Equation (1.6), the value of β at time step t is proportional to the ratio between the classification loss and the discrimination loss observed

on the validation set at the previous time step, $t-1$. This ratio is further modulated by a hyperparameter r, which controls the influence of the discriminator. Importantly, for the initial time step $t = 0$, β is uniformly set to 1. The complete DAP framework is illustrated in Figure 1.7.

Accuracy Over Privacy. In the context of evaluating machine learning models within a privacy-preserving framework, it is imperative to consider both the effectiveness of the model and the privacy of the training data. Balancing these two aspects poses significant challenges, particularly due to the inherent difficulty in quantifying privacy and the complexity of comparing metrics across different domains. To address these challenges, a novel metric called Accuracy Over Privacy (AOP) is introduced. This metric provides a concise assessment of both the accuracy and privacy of a given model, encapsulating the trade-off between these critical factors. In the domain of MIAs, the effectiveness of the attack model is commonly evaluated using the area under the curve (AUC) of the receiver operating characteristic curve. Conversely, when evaluating the performance of a classifier, the top-1 accuracy (ACC) metric is typically used. Consequently, the AOP is calculated by taking into account both the ACC of the classifier and the AUC of the MIA, denoted AUC_{MIA}:

$$\text{AOP}(\lambda) = \frac{\text{ACC}}{(2\max(\text{AUC}_{\text{MIA}}, 0.5))^{\lambda}}. \tag{1.8}$$

The Equation 1.8 provides a comprehensive metric that combines both the ACC and the Area Under the AUC_{MIA} into a single metric. The parameter $\lambda \geq 1$ is crucial in this metric as it modulates the emphasis placed on the privacy aspect when calculating the AOP. The AOP metric is defined within the interval $[0, 1]$, reflecting its bounded nature. If a model exhibits either poor accuracy or high susceptibility to membership inference attacks, the AOP value will tend towards 0. Conversely, a model that excels in both accuracy and robustness to such attacks will yield an AOP value close to 1. The parameter λ has a significant impact on the metric, adjusting the balance between the accuracy and privacy components. The denominator in the AOP formula contains a max operator to

ensure that the AUC never falls below that of a random guessing model, which is 0.5. This feature of the denominator is crucial as it guarantees that for models with optimal privacy, the AOP directly reflects the classification accuracy.

1.4.1. Experiments and Results

To ensure the transparency and reproducibility of this study, a detailed description of the experimental procedures and configurations is provided. The proposed DAP algorithm operates in two different settings. The first setup, termed $\mathbf{DAP}_t$ (test DAP), involves training shadow models using the test dataset. This setup simulates a scenario where both the attacker and the victim have access to an external dataset that could potentially be publicly available. By using DAP_t, deep learning engineers are able to preemptively mitigate possible attacks by using the same dataset that could be exploited by the attacker. The second setting, referred to as $\mathbf{DAP}_v$ (validation DAP), involves training shadow models on the validation dataset. This scenario represents a more typical situation where the attacker's data distribution does not match that of the victim. In both configurations, a total of 10 shadow models were used and the parameter r was optimised over a uniform range from 0 to 1, with increments of 0.025.

In order to provide a fair and comprehensive evaluation, the performance of the DAP algorithm was compared with several alternative approaches. Firstly, a **Baseline** model was created, consisting only of the base classifier without any defensive mechanisms. Subsequently, a model called **Reg** was included, following the methodology outlined in Section 1.3. More precisely, this model applies Dropout regularisation to each intermediate classifier weight and L2 regularisation to the model output. The Dropout probability is fine-tuned over values of 0.2, 0.33 and 0.5, while the L2 regularisation weight was optimised over values of 0.1, 0.01 and 0.001. In addition, the study is extended to models incorporating differential privacy called **DP**, specifically (ε,δ)-differential privacy. Here, the δ parameter was fixed at 10^{-5}, and the ε privacy budget is varied by adjusting the number of training epochs. Four different models were tested, corresponding to ε values of 0.5, 1, 2 and 4.

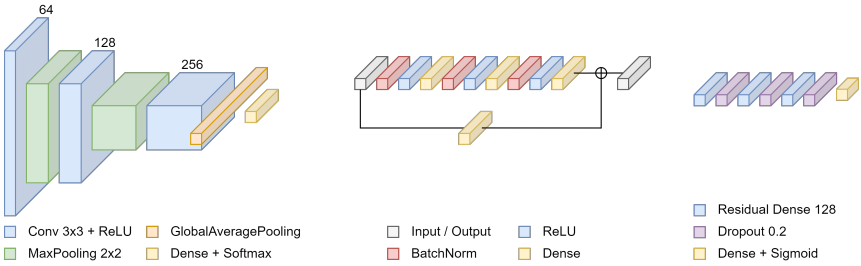

Figure 1.8: The experiments used neural network architectures: (left) CNN for classifiers and shadow models, (middle) custom residual block for DAP discriminator, (right) overall DAP discriminator architecture [4].

To limit the number of free parameters within the experiments, a unique architecture was used for each classifier across all configurations, as shown on the left side of Figure 1.8. This choice was driven by the intrinsic spatial complexity associated with differential privacy training. The residual architecture used for the discriminator in both the DAP_t and DAP_v models is also shown in Figure 1.8, with the proposed residual block in the middle and the overall structure shown on the right. All models were optimised using the Adam algorithm with a learning rate of 10^{-5}, 10^{-4} or 10^{-3} and a batch size of 32. Training was performed until convergence, using early termination with a patience parameter of 25 epochs, based on validation accuracy. However, for the differentially private optimisation models, the number of training epochs was predetermined and scaled according to ε. The performance of the proposed models is evaluated using several metrics: top-1 accuracy for classification, area under the curve (AUC) for membership inference attacks (MIA) – measured using the TensorFlow Privacy toolkit – and time per training epoch. In addition, the just-defined metric, Accuracy Over Privacy (AOP), is measured incorporating different values of λ to rigorously evaluate the trade-off between performance and privacy, especially as the emphasis on privacy increases. The results presented are averages derived from five independent replications of each experiment.

The analysis includes eight different datasets, namely CIFAR10 [63], CIFAR100 [63],

Table 1.5: Test set accuracy for various privacy-preserving models across multiple datasets. Best results are in **bold**, second-best underlined.

Dataset	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
CIFAR10	78.44	**81.23**	31.31	37.42	41.94	41.50	62.26	61.45
CIFAR100	47.80	**53.34**	4.03	8.15	9.30	6.97	31.40	27.41
FashionMNIST	92.89	**92.88**	60.09	70.18	73.98	77.28	86.95	86.66
EuroSAT	95.38	**95.06**	31.07	59.05	67.95	64.68	90.00	89.78
TinyImageNet	36.78	**37.74**	3.12	2.86	2.82	2.87	26.23	21.84
OxfordFlowers	56.84	**65.60**	3.46	5.02	9.08	14.18	28.79	25.86
STL10	65.74	**64.72**	8.29	14.63	24.66	29.07	48.20	37.95
CINIC10	67.31	**70.47**	28.31	34.01	38.67	40.58	58.11	58.98
Average	67.62	**70.10**	21.18	28.93	33.51	34.61	53.98	51.25

FashionMNIST [65], EuroSAT [66], TinyImageNet [67], OxfordFlowers [68], STL10 [69] and CINIC10 [70]. Detailed descriptions of these datasets can be found in Appendix A. The experiments were performed on a system equipped with an Intel(R) Xeon(R) Gold 6238R CPU running at 2.20GHz and an Nvidia Quadro RTX 6000 GPU.

Table 1.5 summarises the accuracy scores obtained for each model across the different datasets. As expected, the (DP) models show the lowest accuracy scores, even with a relatively high privacy budget ($\varepsilon = 4$). In contrast, the Reg model consistently achieves high accuracy, sometimes even surpassing the performance of the baseline model. In particular, the proposed DAP method consistently outperforms the DP models in terms of accuracy. This is particularly evident in the EuroSAT dataset, where the DAP_t and DAP_v configurations show significant accuracy improvements of 22% and 21%, respectively, over the best performing DP model. In summary, the Reg model provides the highest average performance in terms of classification accuracy, closely followed by DAP_t and DAP_v.

Table 1.6 summarises the results of MIAs on the target models. The DAP_t and DAP_v methods yield average AUCs of 51.37% and 50.80% respectively, suggesting that both approaches provide robust protection against MIAs. This per-

Table 1.6: Area Under the Curve metrics for membership inference attacks on privacy-preserving models. Best results are in **bold**, second-best underlined.

Dataset	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
CIFAR10	64.87	63.12	50.57	52.63	51.94	**50.34**	50.75	50.58
CIFAR100	60.34	62.16	**50.10**	51.57	50.73	50.67	51.61	50.69
FashionMNIST	55.21	56.25	**50.24**	**50.24**	50.48	50.59	50.74	50.63
EuroSAT	54.48	52.83	50.56	50.24	**50.07**	50.21	50.14	50.14
TinyImageNet	60.37	59.22	51.43	**50.18**	52.18	50.42	51.65	50.96
OxfordFlowers	76.14	76.59	54.31	53.74	52.67	53.23	53.85	**52.19**
STL10	60.49	56.37	50.25	52.47	50.55	**50.19**	50.82	50.63
CINIC10	57.29	61.48	**50.13**	51.47	51.15	50.76	51.39	50.76
Average	61.15	61.00	50.94	51.57	51.22	**50.80**	51.37	50.82

formance is almost equivalent to random guessing, demonstrating competitive results comparable to DP models. Conversely, the Reg model shows a similar level of privacy to the baseline. The observed discrepancy between these results and those reported in Section1.3 is due to different experimental setups; in particular, the current experiments were conducted until convergence, balancing the privacy budget with more noise injection, rather than limiting the number of epochs. In conclusion, the DAP framework proves to be a robust training method capable of producing models that are significantly resilient to MIAs.

Table 1.7 presents the results for the proposed AOP metric, illustrating the effectiveness of DAP in achieving private models that maintain competitive performance. DAP_t and DAP_v significantly outperform DP and Reg techniques in terms of the accuracy-privacy trade-off. Although the Reg model is more vulnerable to MIAs, it remains a viable intermediate option due to its superior accuracy. On the other hand, DP models, while highly effective in preventing MIAs, suffer from reduced accuracy, resulting in suboptimal performance. The AOP metric also aligns with the privacy budget ε, with the most private model (DP with $\varepsilon = 0.5$) performing worst due to the significant noise introduced during gradient updates.

Table 1.7: Mean Accuracy Over Privacy metric for varying λ values across privacy-preserving models. Best results are in **bold**, second-best underlined.

λ	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
1	56.43	**58.69**	21.19	28.59	33.50	33.90	53.14	50.66
2	47.61	49.67	20.77	27.58	32.22	33.69	**52.31**	50.10
5	30.24	31.10	20.88	26.35	31.11	33.08	**48.96**	48.79
10	15.27	16.79	19.65	25.31	29.56	32.09	44.70	**45.71**
20	4.79	6.73	18.06	22.49	26.71	29.24	38.78	**41.25**
50	-0.18	1.20	15.39	17.80	21.68	23.51	26.06	**30.52**
Average	25.71	27.41	19.31	24.67	29.16	30.97	44.03	**44.50**

The parameter λ is crucial in determining the suitability of the models. When $\lambda = 1$, both DAP and DP are less advantageous compared to the Reg model, which not only emerges as the superior option, but also outperforms the baseline model in AOP due to its balanced resilience and performance. This result suggests that in scenarios where privacy and accuracy are of equal importance, the adoption of highly private models that compromise accuracy is suboptimal. As λ increases to 2, both DAP models quickly become the optimal choice and continue to dominate as λ increases. The proposed metric also shows that the DP models become preferable to the Baseline and Reg models as the emphasis on privacy increases, especially when λ reaches 10. This trend is consistent with expectations, given the loss of accuracy associated with DP-SGD. Nevertheless, empirical evidence suggests that DAP models consistently outperform in privacy-sensitive contexts, although it should be noted that unlike DP models, DAP lacks a formal privacy guarantee.

Table 1.8 summarises the training time per epoch for each model and shows that the Baseline and Reg models are the most time efficient, while the DP models take about eight times longer. The DAP models strike a balance, achieving robust results in both privacy and accuracy, while requiring only twice the time of the Baseline model.

Table 1.8: Per-epoch training time (in seconds) for each model across multiple datasets. Best results are in **bold**, second-best underlined.

Dataset	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
CIFAR10	5.6	**5.9**	46.5	46.5	46.5	46.5	17.8	17.7
CIFAR100	8.9	**9.5**	48.2	48.2	48.2	48.2	17.9	17.9
FashionMNIST	12.1	**11.7**	53.2	53.2	53.2	53.2	23.4	23.5
EuroSAT	4.7	**5.5**	72.7	72.7	72.7	72.7	10.2	9.5
TinyImageNet	31.7	**32.6**	338.5	338.5	338.5	338.5	62.4	61.4
OxfordFlowers	1.2	**1.8**	20.8	20.8	20.8	20.8	2.3	2.4
STL10	1.7	**2.5**	34.9	34.9	34.9	34.9	**2.5**	2.8
CINIC10	30.0	**22.1**	106.5	106.5	106.5	106.5	41.2	42.8
Average	12.0	**11.5**	90.2	90.2	90.2	90.2	22.2	22.2

In conclusion, considering accuracy, privacy, AOP and training time, DAP presents a superior trade-off compared to DP and regularisation. The DAP_t configuration yields high performance models with moderate privacy, while DAP_v emphasises strong privacy with a slight trade-off in accuracy. Both configurations outperform DP in dealing with the performance-privacy trade-off, and do so in significantly less time.

1.5. Conclusions

This chapter thoroughly examined the advantages, disadvantages, and limitations of differentially private optimisation techniques. Through empirical analysis, it was demonstrated that a differentially private stochastic gradient descent optimiser, despite its privacy budget guarantees, cannot be universally regarded as a comprehensive protection mechanism for deep learning models. Although effective in mitigating membership inference attacks, this approach was found to significantly impair model utility, extend training duration, and enhance the success rate of model inversion attacks.

Furthermore, the investigation revealed that the combined application of l2 and dropout regularisation techniques presents a compelling alternative. This com-

bination not only preserves model utility but also maintains manageable training times while concurrently offering resistance to both membership inference and model inversion attacks.

In addition, the Discriminative Adversarial Privacy (DAP) framework and the Accuracy Over Privacy (AOP) metric were introduced. DAP was designed to develop deep learning models with robust defences against Membership Inference Attacks (MIAs), while the AOP metric succinctly encapsulates the privacy-accuracy trade-off into a single evaluative measure. Experimental results highlighted the superior performance of DAP in achieving an optimal balance between model accuracy and privacy, surpassing models reliant on Differential Privacy (DP). The AOP metric effectively captured these outcomes, providing a precise and comprehensive assessment criterion. Moreover, DAP was shown to require significantly less computational resources, thereby reducing training times compared to DP-based methods.

2 | Synthetic Data Sharing in Federated Learning

Data protection laws, such as the GDPR, establish transparency and security as fundamental principles for data processing algorithms. Within this regulatory framework, federated learning has emerged as a highly influential approach to privacy preserving distributed machine learning, demonstrating remarkable success in a variety of natural language processing and computer vision tasks. To mitigate the risk of private data leakage to unauthorised entities and malicious actors, many federated learning systems incorporate differential privacy mechanisms. However, the standard federated learning paradigm has been shown to be vulnerable to both poisoning and inference attacks, raising significant privacy concerns. In response to these challenges, this chapter introduces SGDE, a generative data exchange protocol designed to enhance user security and machine learning performance within a cross-silo federation. The central innovation of SGDE is the use of data generators trained on private data with strong differential privacy guarantees to produce synthetic data instead of directly transmitting gradient information. These generators are capable of producing a virtually unlimited amount of data that, while preserving the essential characteristics of the original private samples, is sufficiently different to mitigate privacy risks. The effectiveness of SGDE is demonstrated through experiments conducted within a cross-silo federated network using beta-variational autoencoders as the underlying data generation models on both image and tabular datasets. The results indicate that the integration of SGDE not only improves task accuracy and fair-

ness, but also significantly strengthens the resilience of the system against the most potent attacks on federated learning.

2.1. Introduction

The formalisation of stringent privacy regulations, such as the European GDPR and the Chinese Cyber Security Law, has necessitated the embedding of privacy as a core principle in the design of data processing algorithms [23, 71]. This imperative is particularly relevant in the field of artificial intelligence, where the development of secure and robust algorithms is essential to protect the rights of data owners while extracting valuable insights.

Federated Learning (FL) has emerged as a prominent framework in distributed machine learning, designed with an inherent focus on privacy [72]. In FL, multiple clients, each with their own private data, collaborate to train a shared machine learning model facilitated by a central server. The foundational algorithm, FedAVG, introduced by McMahan *et al.*, involves the central server initialising a model and distributing it to the clients [73]. Each client then runs a limited number of stochastic gradient descent steps on its local data and returns the updated model weights to the server. The server aggregates these updates, typically weighted by the number of data samples involved in the local training, and the process is iterated until model convergence.

Throughout this process, private data remains on the client devices, theoretically preserving user privacy. However, the iterative exchange of model updates creates potential vulnerabilities, as these communications can be exploited by adversarial entities within the federation. Various attacks, including those using generative deep learning techniques, have been developed to extract confidential information from the gradients shared during training [74]. These attacks can lead to the reconstruction of private data or the manipulation of the central model through strategically crafted updates.

The increasing frequency of data breaches in FL systems has raised doubts about their practical security, leading to a surge in research aimed at developing de-

fensive strategies. These include methods for robust model aggregation, model pruning, and gradient encryption [75–77]. Despite these efforts, the complexity and scale of FL systems introduce unique vulnerabilities, and existing countermeasures are often insufficient to ensure robust security [78].

2.1.1. Main Contributions

In order to enhance data security and ensure user privacy in a federated learning environment, this chapter introduces the Secure Generative Data Exchange (SGDE) framework, which is designed for the secure exchange of data through differentially private data generators (Figure 2.1). SGDE is divided into three distinct phases: *Subscribe*, *Push* and *Pull*. During the *Subscribe* phase, clients express their intention to participate in the data exchange by establishing communication with the server. Subsequently, in the *Push* phase, clients locally train data generators that adhere to strict differential privacy standards according to the server's requirements, and these generators are then submitted to the server. In the final *Pull* phase, clients gain access to the collection of generators maintained by the server, allowing them to locally generate synthetic data and train machine learning models on this data. The SGDE framework is particularly suited to cross-silo federated learning, where clients are typically large institutions such as hospitals, universities or corporations. This context is characterised by a limited number of clients – often only in the hundreds – each with sufficient computational resources to train generative models independently.

SGDE enables customers to produce large volumes of synthetic data that faithfully replicate the characteristics of real-world data, and make this synthetic data available for offline machine learning tasks. This approach provides the transparency and flexibility of a centralised dataset while significantly reducing communication overhead. In addition, in supervised classification scenarios, SGDE facilitates the generation of balanced samples across all classes, thereby improving fairness for underrepresented classes in the original dataset.

From a security perspective, SGDE provides robust protection against common federated learning attacks, such as poisoning and inference attacks. The offline

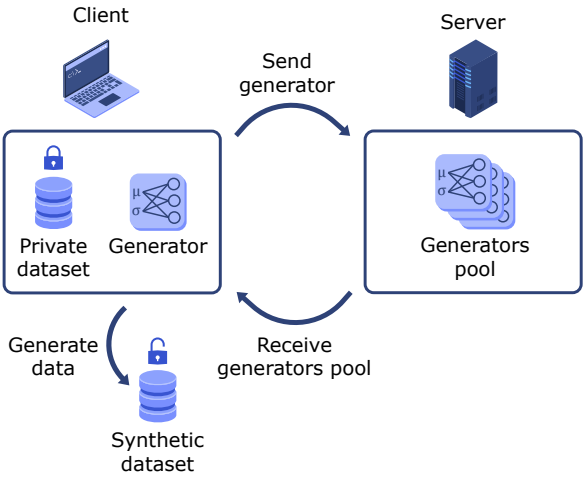

Figure 2.1: Overview of the SGDE framework: clients exchange differentially private data generators with a server and gain access to a pool of generators, enabling the creation of a large synthetic dataset for offline use [2].

nature of the synthetic dataset inherently reduces the attack surface, making malicious interference more detectable and manageable [79]. The absence of a centralised model and global task further reduces the opportunities for attackers, thereby enhancing the security posture of the framework.

The contributions of this work are as follows:

- The introduction of the SGDE framework, along with a thorough examination of its benefits in a federated learning context;
- A detailed discussion of how the use of differential privacy compliant data generators within SGDE contributes to significant security improvements;
- Empirical validation of SGDE's effectiveness through experimental evaluations using both tabular and image datasets, demonstrating performance

benefits from the perspective of federation members;

- Demonstration of SGDE's ability to address and mitigate issues related to distribution bias and discrimination within datasets.

2.2. Related Works

Federated Learning (FL) [72, 80] was introduced as a privacy preserving framework for distributed machine learning that could scale to millions of devices. In a standard FL setting, a central server coordinates K clients, each with a local dataset, to jointly minimise a global loss function F with respect to a model's parameters w:

$$\min_{w} F(w) = \min_{w} \sum_{k=1}^{K} p_k F_k(w). \tag{2.1}$$

Here F_k is the loss function for client k, calculated over n_k local samples, with p_k acting as a weighting factor to ensure $\sum_{k=1}^{K} p_k = 1$. Common choices for p_k include $p_k = \frac{n_k}{n}$, where $n = \sum_{k=1}^{K} n_k$, or $p_k = \frac{1}{K}$.

One of the fundamental challenges in FL has been managing the inherent heterogeneity across a large client network. This includes variations in computing power and connectivity, as the network may include devices with different hardware architectures and memory constraints. Communication channels may be unreliable, resulting in delayed or lost updates, which, if ignored, could bias the global model towards clients with more stable connectivity. Numerous strategies were proposed to mitigate communication costs and address issues related to lost updates [81, 82]. According to Lim *et al.*, the primary methods for reducing bandwidth requirements in mobile networks included improved edge computation, model compression, and importance-based updating [83].

In addition, energy efficiency was identified as critical, particularly in networks with IoT and battery-powered devices, given the high energy demands of wireless communications [84, 85]. Statistical heterogeneity was another major concern,

as the assumption of independent and identically distributed (IID) data often failed in real-world FL scenarios. Techniques such as Agnostic FL were introduced to address this by optimising the central model for any combination of client updates, thereby improving fairness [86]. FedProx extended the FedAvg algorithm by incorporating a regularisation term that ensured convergence in non-IID environments [72, 73].

While the standard FL framework typically provided a single model for each user, this approach could limit performance for local inference tasks. Personalisation of local models was proposed as a solution to address non-IID data distributions [87]. For example, Wu *et al.* discussed personalisation techniques for IoT devices, including transfer learning, meta-learning, federated multitask learning, and federated distillation [88]. In the context of meta-learning, methods such as MAML were widely adopted to personalise models for individual users [89–91]. In addition, Huang *et al.* introduced attentive message passing to improve personalisation by aggregating customers with similar characteristics [92].

2.2.1. Differential Privacy and Generative Models

As detailed in Chapter 1, Differential Privacy (DP)[93] provided a rigorous mathematical framework for quantifying a system's resistance to the disclosure of sensitive information associated with individual data points. A mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ was defined to be (ε)-differentially private if, for any pair of adjacent inputs $x, y \in \mathcal{X}$ and any subset of outputs $S \subseteq \mathcal{Y}$, the following condition held:

$$\Pr[\mathcal{M}(x) \in S] \leq \exp^{\varepsilon} \Pr[\mathcal{M}(y) \in S]. \tag{2.2}$$

In Equation (2.2), ε represents the privacy budget, quantifying the potential information leakage. Lower values of ε correspond to stronger privacy guarantees. DP was particularly advantageous in iterative optimisation methods, such as stochastic gradient descent, due to its composability, group privacy, and robustness to auxiliary information.

However, the strict constraints of ε-DP could make it impractical for many real-

world applications. To address this, several relaxed versions were proposed, including f-DP, concentrated DP, and R'enyi DP [52, 94, 95]. The most common relaxation was (ε, δ)-DP, where a mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ was (ε, δ)-differentially private if for any adjacent inputs $x, y \in \mathcal{X}$ and any subset of outputs $S \subseteq \mathcal{Y}$, the following held [25]:

$$\Pr[\mathcal{M}(x) \in S] \leq \exp^{\varepsilon} \Pr[\mathcal{M}(y) \in S] + \delta. \tag{2.3}$$

In Equation (2.3), δ accounts for the probability of violating ε-DP by a small margin. In FL systems, (ε, δ) DP techniques were often used to improve the privacy of data stored on client devices. A common approach involved the use of differentially-private optimisers, such as differentially private SGD, where gradient clipping and the addition of Gaussian noise at each iteration ensured compliance with DP standards [25]. PATE, another DP-aware framework, used a student-teacher model where a central "student" model learned from a set of private "teacher" models via noisy voting [96].

DP measures were also crucial in synthetic data generation by deep generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) [10, 97]. For example, differentially-private VAEs [98] and GANs [99] integrated DP by introducing noise into the gradients during training. Advanced GAN architectures were modified to comply with DP standards, including InfoGAN [100] and DP-Conditional GAN [101], which used the R'enyi DP model [52, 102].

One notable framework was DP-auto-GAN [103], which was designed to generate synthetic data from unlabelled, mixed-type datasets by combining GANs with autoencoders for dimensionality reduction. Similarly, PATE-GAN [104] extended the PATE framework by using a set of private teacher discriminators to train a student discriminator against a shared generator, ensuring differential privacy of the synthetic data. In [105], the authors proposed a two-step approach: first, K-means clustering with a differentially private kernel was applied to sensitive data, followed by training K generative models, one for each cluster,

achieving greater data utility than a single-model approach.

2.2.2. Threats to Federated Learning

Although FL pipelines often incorporated differential privacy techniques to protect customer data, numerous studies showed that targeted attacks could still compromise sensitive information. A comprehensive review of the major FL attacks was provided in [74], where they were categorised into poisoning attacks and inference attacks. Poisoning attacks involved the deliberate manipulation of training data or model parameters by a malicious actor to distort the learning process away from its intended goal. Inference attacks, on the other hand, aimed to infer whether a particular data sample was involved in the training process (membership inference [24, 106]) or to reconstruct the input data from a given model and its output (input inference or model inversion [45]). Another critical attack vector in FL was the iterative gradient exchange that underpinned the training process. Gradient leakage attacks exploited this vulnerability, allowing adversaries to extract information about raw private data samples via GAN-based gradient reconstruction [107]. It was shown that even partial access to intermediate updates could lead to significant data leakage [108].

2.3. Method

This section introduces SGDE, a novel data exchange framework utilising generative models. SGDE is designed to provide robust privacy whilst addressing significant security challenges inherent in federated learning (FL). Unlike traditional learning protocols, SGDE does not involve optimising a target model or solving a predefined task. Instead of exchanging model gradients, SGDE allows each client within a cross-silo federation to access differentially private data generators capable of synthesising an arbitrary number of data samples. These generators encapsulate the characteristics of their respective private datasets without revealing explicit information, thus safeguarding against both curious and malicious entities. Once distributed, clients can autonomously generate synthetic data for any local machine learning task. This data generation can be

performed entirely on the client side, offering the flexibility to either train models privately on synthetic data or engage in federated iterative processes with a shared model, all whilst ensuring enhanced privacy, as synthetic data inherently lacks sensitive information.

Although this study focuses on supervised classification, the SGDE framework can be extended to various other machine learning tasks. The advantages of SGDE in a cross-silo federation are manifold:

- **Flexibility**: SGDE grants clients complete control over synthetic data, both in terms of generation and utilisation. Clients have the freedom to select the most appropriate generators to construct their datasets and can determine the size of these datasets according to their needs. Additionally, the synthetic data generated by SGDE is task-agnostic, allowing its use in different machine learning applications without the need for central coordination between federation participants.
- **Security**: In SGDE, private data remains within the client environment, and only data generators are exchanged. Moreover, private training parameters remain local, with the exception of the privacy level of the generator. This significantly reduces the attack surface for both data poisoning and inference attacks.
- **Fairness**: Clients can generate an unlimited number of samples for each predefined label, ensuring balanced class representation within the synthetic dataset. This capability helps mitigate distributional biases, as demonstrated in Section2.4.
- **Communication Efficiency and Robustness**: SGDE eliminates the need for iterative data exchange, as training is performed entirely on the client device, without requiring Internet communication. Generators are exchanged only once between the client and the central server, resulting in significant bandwidth savings.

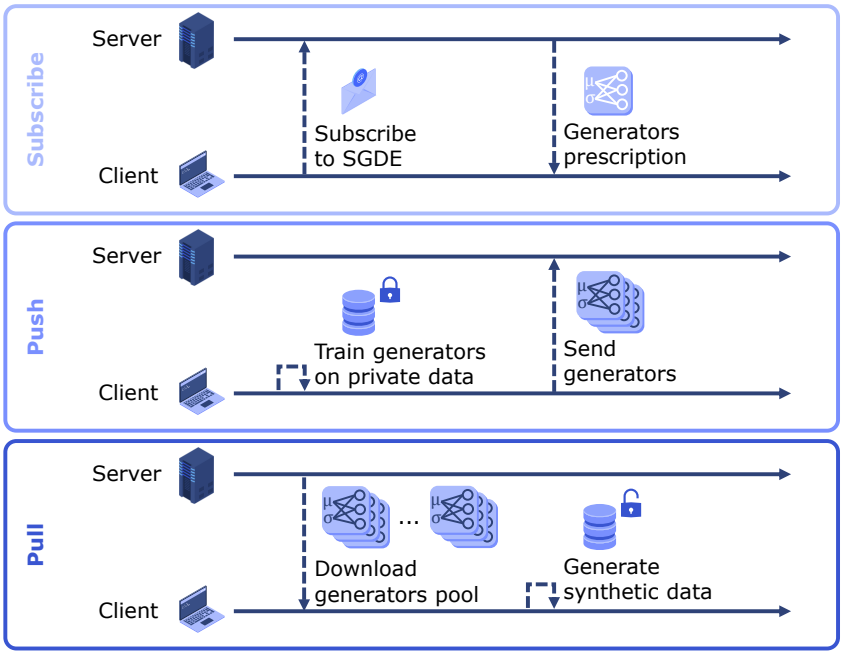

Figure 2.2: Schematic representation of the tripartite structure of the SGDE protocol. The process begins with an initial exchange phase to establish communication between client and server entities. The protocol then proceeds to a model development phase, where the client constructs a generative model using its own data sets, adhering to parameters specified by the server, before submitting the resulting model to the server infrastructure. The protocol culminates in a resource retrieval phase that allows the client to access the server's repository of generative models [2].

2.3.1. The Steps of SGDE

Consider a federated network consisting of K clients, where each client k owns a private dataset $\mathcal{D}_k = \{x_1, x_2, \ldots, x_{n_k} | x_i \in \mathcal{X}\}$, where $\mathcal{X}$ represents the data domain. Although the discussion is framed in the context of supervised classification, SGDE can be generalised to other machine learning paradigms. In this scenario, each data point x_i is associated with a class label y_i from a finite set $\mathcal{Y}$. Furthermore, it is assumed that each client k desires access to a larger dataset than $\mathcal{D}_k$ and is therefore motivated to join a federation with other like-minded clients.

In this federated environment, SGDE provides a secure solution by requiring each client to generate a set of generators $\mathcal{G}_k$, where each generator g_k^y corresponds to a class $y \in \mathcal{Y}$. These generators are then collected by a central server and made available to other clients within the federation.

As depicted in Figure 2.2, SGDE encompasses three distinct phases: *Subscribe*, *Push*, and *Pull*:

1. **Subscribe**: Participant k signals its intention to join the protocol by communicating with the central server. The server responds by providing a comprehensive set of requirements for the data generators. These specifications may delineate the nature of the synthetic data, the internal architecture of the generators, or the minimum differential privacy parameters (ε, δ) that must be satisfied.
2. **Push**: For each class $y \in \mathcal{Y}$, participant k undertakes the training of a generator $g_k^y : \mathcal{Z} \to \mathcal{X}$. This generator is capable of synthesising samples $\hat{x}$ that correspond to label y, utilising a noise vector $z \in \mathcal{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ as its initial input. The resulting set of generators $\mathcal{G}_k$ is subsequently transmitted to the central server, accompanied by the (ε, δ) differential privacy levels measured during the training process. Notably, the server does not require any additional information about the training process, thereby preserving the participant's privacy.
3. **Pull**: Participant k is granted access to the generator repository hosted

on the central server, which contains generators from various participants. This access allows participant k to examine the parameters, structure, and privacy levels of the available generators, facilitating an informed selection for download. These acquired generators can be employed to sample an unlimited number of synthetic data instances, enabling the participant to construct datasets of any desired size.

2.3.2. Threat Analysis

This section evaluates the resilience of SGDE against common attack vectors in federated learning, with a particular focus on internal threats from malicious entities within the federation, as outlined in [74].

The first category of attack considered is data poisoning. In SGDE, the primary vulnerability lies in the design of the generators. However, unlike centralised FL settings where a single point of failure could compromise the entire system, the decentralised nature of SGDE prevents an attacker from undermining the entire federation. A malicious actor could potentially create generators that produce poisoned data, leading to inaccurate model predictions. However, from the perspective of an honest client, such an attack could be detected by performance degradation and mitigated by excluding the compromised generators during the data sampling process. Furthermore, as discussed in [79], poisoning is less effective against a synthetic dataset such as that generated by SGDE, as direct heuristic analysis facilitates the detection of malicious activity, a capability that does not exist in standard FL scenarios where clients do not have visibility into others' data.

The second category of attacks comprises inference threats. SGDE significantly reduces the attack surface compared to traditional FL setups by eliminating the iterative exchange of gradients, thereby neutralising gradient leakage attacks aimed at reconstructing private data.

It is argued that inference attacks targeting data generators, such as membership inference and model inversion, would also be ineffective against SGDE. Differ-

ential privacy built into the SGDE framework provides protection against these types of attacks, as shown in [109]. Strict differential privacy settings in model training enhance resistance to membership inference attacks. Similarly, model inversion attacks, which attempt to reverse-engineer data from a differentially private generator, would at best recover the latent noise vector rather than any meaningful data.

Regarding external threats, SGDE also provides enhanced protection against network-level attacks, which typically target system availability rather than the learning protocol itself. By reducing communication to the exchange of data generators, SGDE minimises the information transmitted over the Internet, thereby reducing susceptibility to server availability attacks that exploit the continuous gradient exchange in traditional FL.

In conclusion, traditional FL and SGDE can be combined to further enhance security guarantees. A distributed learning system operating on a synthetic dataset generated via SGDE need not concern itself with the confidentiality of the synthetic data. As private data never leaves the client's premises, novel machine learning algorithms can be applied to public synthetic datasets that retain the essential characteristics of their private counterparts, thus maintaining a high level of privacy and security.

2.4. Experiments and Results

The Secure Generative Data Exchange (SGDE) protocol facilitates the exchange of class-specific data generators amongst clients in a federated learning environment. This approach enables the training of machine learning models on synthetic data, effectively expanding the range of available training data beyond the limited private datasets on individual devices.

In the SGDE experiments, clients train local machine learning models using an optimal number of synthetic samples generated by the exchanged SGDE generators. The synthetic dataset comprises samples uniformly selected from each available generator. Tables 2.1 and 2.2 present the average performance met-

rics of these local models, evaluated against the clients' private datasets and a separate test set, respectively. The datasets used in this set of experiments – Titanic [110], Breast Cancer [111], Mushrooms [110], Adult [110], Wine Quality [112], MNIST [64] and FashionMNIST [65] – are described in Appendix A.

To ensure a rigorous and unbiased comparison, the experiments employ well-established models from the machine learning literature as classifiers. For tabular datasets, logistic regression serves as the primary classifier. Image datasets utilise the first eight pre-trained layers of VGG16, adapted through transfer learning [113]. In the latter case, the final dense layer of the VGG16 architecture, comprising 256 neurons with LeakyReLU activation function followed by a softmax classifier, remains the sole trainable component. The VGG16 model undergoes training using the Adam optimiser with a learning rate of 0.001. All experiments are conducted on a system equipped with an Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz and an Nvidia Quadro RTX 6000 GPU.

Data generators play a pivotal role in the SGDE protocol, as they must meet stringent security standards whilst producing synthetic data with high utility for machine learning tasks. Although the generated samples need not appear perceptually realistic, they must still provide significant utility. Figure 2.3 illustrates synthetic images from a single client's class-specific generators, showcasing visibly noisy images that do not closely resemble real samples. Nevertheless, the experimental results demonstrate a high level of utility for these synthetic images, with performance sometimes surpassing that of the real data.

From a security perspective, data generators must not reveal any information about the private data used in their training, thus safeguarding client privacy. SGDE permits the exchange of any data generator that satisfies these security requirements.

The empirical investigations employed a bespoke implementation of the β-VAE architecture, trained using a differentially private variant of the Adam optimiser [25, 114]. To achieve optimal resistance against inference attacks, rendering such attacks no more effective than random guessing, each participant is required to train its generators to attain a final (ε, δ)-DP level characterised by

Figure 2.3: Examples of synthetic data generated from SGDE generators. The figure presents synthetic images for MNIST (first four rows) and Fashion MNIST (remaining four rows). Each column contains images from a single generator trained for a specific class. The images exhibit noticeable background noise and content distortion, resulting from differential privacy techniques. These visually identifiable synthetic images are not linked to any privacy-protected real samples from a client's dataset, thus preserving data privacy [2].

Table 2.1: Performance evaluation on local data splits. The table compares the average performance of local models trained on local data (*Baseline*), a single global model trained with FedAvg (*FedAvg*), and local models trained on synthetic data generated via the SGDE protocol (*SGDE*). The *Baseline* columns present results from 10-fold cross-validation, while *FedAvg* columns show performance on private validation splits. *SGDE* columns report performance on entire local datasets. The table highlights the average improvements achieved through federated learning and SGDE compared to local training.

Dataset	Accuracy ↑			F1 Score ↑			AUC ↑		
	Baseline	FedAvg	SGDE	Baseline	FedAvg	SGDE	Baseline	FedAvg	SGDE
Titanic	75.67	76.67	**80.87**	19.43	**69.89**	63.37	75.70	69.38	**78.35**
Breast Cancer	89.67	89.50	**97.09**	93.37	84.16	**97.81**	99.17	98.75	**99.27**
Mushrooms	92.93	91.51	**93.49**	92.43	91.32	**93.14**	96.23	95.84	**96.61**
Adult	**80.64**	76.01	79.65	49.69	47.95	**61.64**	83.30	79.81	**83.73**
Wine Quality	93.46	90.44	**98.54**	82.98	85.81	**97.10**	99.44	99.10	**99.49**
MNIST	98.20	**99.40**	98.72	98.16	**99.39**	98.71	99.02	**99.66**	99.31
FashionMNIST	88.47	**91.75**	89.30	88.32	**91.65**	89.22	93.87	**95.76**	94.76
Avg. Improvement		-0.54	**+2.66**		+6.54	**+10.94**		-1.20	**+0.68**

$\varepsilon \leq 1.5$, $\delta \ll \frac{1}{|\mathcal{D}_c|}$, and $RDP \geq 9$. Here, $|\mathcal{D}_c|$ denotes the number of samples in class c of dataset $\mathcal{D}$, and RDP represents the R'enyi-DP value [106, 109]. Each participant retains solely the decoder, responsible for generating synthetic samples, whilst excluding the encoder to preclude potential privacy breaches.

An ablation study was conducted to optimise the hyperparameters of the β-VAE architecture, striking a balance between generation performance and the noise introduced by differential privacy. This study encompassed a grid search across two architectural configurations, tailored for tabular and image data, as elucidated in Table 2.3. For tabular data, each dense layer is succeeded by a LeakyReLU activation function. Conversely, for image data, each convolutional layer is followed by a Swish activation function [115]. The latent space dimension and β value were meticulously fine-tuned for each dataset.

Table 2.2: Performance evaluation on test sets. The table compares the average performance of local models trained on local data (*Baseline*), a single global model trained with FedAvg (*FedAvg*), and local models trained on synthetic data generated via the SGDE protocol (*SGDE*). All models are evaluated on a hold-out set. The table highlights the average improvements achieved through federated learning and SGDE compared to local training.

Dataset	Accuracy ↑			F1 Score ↑			AUC ↑		
	Baseline	FedAvg	SGDE	Baseline	FedAvg	SGDE	Baseline	FedAvg	SGDE
Titanic	71.83	**74.81**	74.01	29.70	**71.61**	56.00	77.14	70.85	**77.43**
Breast Cancer	89.42	91.86	**93.02**	92.25	90.82	**94.78**	99.60	99.36	**99.76**
Mushrooms	92.56	91.27	**93.49**	91.92	91.20	**93.14**	96.30	96.07	**96.61**
Adult	**80.87**	76.47	79.00	50.14	47.70	**60.21**	84.02	81.24	**84.08**
Wine Quality	92.57	90.23	**97.79**	82.42	85.10	**95.70**	98.63	98.50	**98.65**
MNIST	97.76	**99.08**	98.49	97.71	**99.08**	98.49	99.02	**99.50**	99.19
FashionMNIST	85.97	87.94	**88.13**	85.81	87.99	**88.04**	92.65	93.68	**94.13**
Avg. Improvement		+0.10	**+1.85**		+6.22	**+8.06**		-1.17	**+0.36**

The primary concern for individual clients is whether participation in the SGDE protocol offers tangible benefits in terms of utility. As shown in Table 2.1, participation in SGDE enabled each member of the federated network to improve average classification accuracy and AUC by 2.66% and 0.68% respectively across all datasets. This suggests that synthetic data from SGDE-shared generators is more effective for learning classification tasks than using the client's local data alone. Consequently, synthetic data can serve as a viable substitute for privacy-sensitive local data in machine learning processes.

Moreover, the F1 score exhibited an average improvement of 10.94% across all experiments. This enhancement suggests that the uniform generation of data for each class mitigates distributional bias, particularly in cases of underrepresented classes in unbalanced datasets, as participants gain access to more comprehensive information about minority classes. Thus, engaging in the SGDE protocol and training a model on synthetic data proves beneficial not only for accuracy but also for achieving more equitable classification performance.

Table 2.3: Hyperparameters of the β-VAE architecture for tabular and image data. The table outlines the structure of the encoder and decoder networks, specifying the number of neurons for dense layers and the number of filters, kernel size, and stride for convolutional layers.

Architecture	Layer	Tabular Data	Image Data
Encoder	1^{st} Layer	`Dense(64)`	`Conv2D(128,3,2)`
	2^{nd} Layer	`Dense(32)`	`Conv2D(256,3,2)`
	3^{rd} Layer	-	`Conv2D(512,3,2)`
Decoder	1^{st} Layer	`Dense(64)`	`Conv2DT(128,3,2)`
	2^{nd} Layer	`Dense(128)`	`Conv2DT(256,3,2)`
	3^{rd} Layer	-	`Conv2DT(512,3,2)`

The robustness of these findings is further corroborated by the results presented in Table 2.2, where the models are evaluated on the test set. Across the Baseline and SGDE experiments, there is an average improvement in accuracy and AUC of 1.85% and 0.36% respectively. The F1 score increases by an average of 8.06%, reaffirming the significant fairness benefit of participating in the SGDE generator exchange.

The main focus of the analysis so far has been on the comparative performance of training on synthetic samples versus training on real local data. The next logical investigation concerns the comparison between training on synthetic samples and federated learning, the preeminent privacy-preserving approach for training machine learning models on distributed data. In particular, SGDE demonstrates advantages over federated learning, not only in terms of reduced communication costs due to its non-iterative nature, but also in terms of superior classification performance in numerous scenarios. The experiments show that SGDE performs comparably to, or even better than, the FedAvg algorithm [73], especially in scenarios with a small number of participants and unbalanced data distributions. This is most evident in the first five rows of tables 2.1 and 2.2, where the experiments involve small tabular data sets and logistic regression as the global machine learning model.

In conclusion, the experiments demonstrate that SGDE provides a secure mechanism for sharing the knowledge embedded in private data by pooling data generators and making them publicly accessible. These findings establish SGDE as a potent tool for enhancing the performance of machine learning tasks for participants in the generator exchange. Furthermore, by enabling the generation of a transparent local dataset, SGDE eliminates the need for the iterative model exchange process typical of federated learning, thereby reducing communication overhead. SGDE also outperforms standard federated learning in terms of accuracy, fairness, and transparency, whilst ensuring robust protection of participants' private data. In a broader context, this research advocates the development of secure technologies that leverage publicly available synthetic data. Such collaboration in a secure environment is posited as key to increasing value and knowledge in a manner that satisfies privacy requirements.

2.5. Conclusions

This study presents SGDE, a data exchange framework that uses data generators to ensure robust privacy protection. The benefits of a generative approach to data sharing are fully analysed, particularly in contexts where strict privacy requirements are paramount. Generators are able to preserve the essential characteristics of private data while allowing the production of an unlimited set of synthetic samples that are public, reproducible and fair. Furthermore, centralised datasets in this framework show strong resilience to poisoning and inference attacks, which are significant threats to conventional federated learning systems.

The effectiveness of SGDE is demonstrated through various experimental scenarios requiring high confidentiality, using differentially private β-VAEs as data generators. The results show that training on synthetic data not only preserves individual privacy, but also achieves superior performance compared to training on real private data. SGDE consistently outperforms federated averaging in the scenarios considered, which is widely regarded as the leading technique for privacy-preserving machine learning on distributed data. In particular, the

generative approach used by SGDE offers improved communication efficiency, transparency and overall effectiveness, especially in cases where data distribution is uneven across clients. The research community's growing recognition of the benefits of generative methods in privacy-aware federations points to a promising direction for future research. This approach has significant potential to reduce barriers to the development of more user-centric, fair and secure large-scale machine learning systems.

3 | Downstream Task Oriented Dataset Generation

The acquisition and annotation of high-quality datasets for training deep learning models presents significant challenges, often necessitating laborious and time-intensive efforts. These difficulties can substantially impede research advancements. Recently, generative models have emerged as a viable alternative for producing synthetic datasets that either supplement or entirely replace real-world data. Despite their promise, synthetic datasets frequently fall short of adequately capturing the intricate variability and richness inherent in real-world data, which limits their efficacy.

This chapter offers two distinct contributions aimed at addressing this issue.

The first contribution examines the application of Generative Adversarial Networks (GANs) for the creation of synthetic datasets, which are subsequently employed to train classifiers evaluated on real-world imagery. To enhance both the quality and the diversity of the synthetic datasets, three novel post-processing methodologies are proposed: Dynamic Sample Filtering, Dynamic Dataset Recycle, and the Expansion Trick. Furthermore, a pipeline referred to as Gap Filler (GaFi) is introduced, which integrates these techniques in a systematic and optimised manner, with the goal of maximising classification accuracy when applied to real-world data. Experimental results demonstrate that GaFi success-

fully narrows the gap in Classification Accuracy Score to error margins of 2.03%, 1.78%, 3.99%, 3.33%, and 2.04% on the FashionMNIST, CIFAR10, CIFAR100, CINIC10, and DermaMNIST datasets, respectively. These findings establish a new benchmark in Classification Accuracy Score and underscore the pivotal role of post-processing techniques in improving the utility of synthetic datasets.

The second contribution of this work focuses on the adaptation of the Stable Diffusion 2.0 model for synthetic dataset generation, employing methods such as Transfer Learning, Fine-Tuning, and the optimisation of generation parameters to enhance the dataset's applicability for downstream classification tasks. A class-conditional variant of the model is presented, which incorporates a Class-Encoder alongside the optimisation of key generation parameters. The proposed approach yielded synthetic datasets that, in one-third of cases, enabled the development of models that surpassed the performance of those trained on actual datasets.

3.1. Introduction

In recent years, deep generative models have evolved to the point where they are capable of producing data that is nearly indistinguishable from real-world samples, whether in the form of images, videos, or other multimedia content. This notable progress has led to significant interest in exploring whether these synthetic datasets could be utilised as effective substitutes for real data in various machine learning applications. In particular, the potential to reduce the costs and logistical challenges of data collection has gained attention, especially in contexts where data cannot be freely shared due to privacy concerns or the sensitive nature of the information. Additionally, generative models may offer solutions when original datasets are too large or cumbersome, as they provide a compressed representation of the real data, thus facilitating more efficient data handling and analysis.

However, while synthetic data generation offers considerable advantages, several challenges remain. A central concern is the reduced informational richness of syn-

thetic datasets compared to real-world data, which can hinder the performance of models trained exclusively on such data. Furthermore, training generative models, especially those relying on diffusion techniques, often demands substantial computational resources and time, particularly when generating large volumes of synthetic samples. These limitations have raised questions about the utility of synthetic data in practical machine learning applications.

In response to these concerns, efforts have been made to formalise methods for evaluating the effectiveness of synthetic data in training machine learning models. Ravuri *et al.* introduced the Classification Accuracy Score (CAS), a metric designed to assess the performance of a classifier trained solely on synthetic data by measuring its accuracy on a real-world test set [116]. Surprisingly, despite the advances in the perceptual quality of synthetic data and the ability of generative models to produce vast quantities of samples, models trained on synthetic data often underperform when compared to their counterparts trained on real data. This discrepancy highlights the need for further research into both improving the quality of synthetic datasets and understanding the underlying factors that contribute to the gap in performance between synthetic and real data.

3.1.1. Main Contributions

This thesis explores two complementary approaches to enhancing the utility of synthetic data generated by generative models, focusing on improving their applicability to downstream tasks. The first part of the study addresses the gap between synthetic and real-world data performance, as measured by the CAS metric, by proposing novel post-processing techniques and introducing an innovative pipeline for optimising synthetic data generation. The second part investigates the adaptation of a pre-trained diffusion model for class-conditioned image generation, aiming at efficiently producing high-quality synthetic datasets for classification tasks.

In the first half of this chapter, existing post-processing methods from the literature are critically analysed, and new strategies are introduced to elevate the quality of synthetic data. A key contribution is the development of the Gap

Filler (GaFi) pipeline, which integrates multiple post-processing techniques to significantly enhance the performance of any generative model, without requiring changes to the model's architecture or training methodology. The primary contributions of this section are as follows:

- Two improved post-processing techniques are proposed: Dynamic Sample Filtering and Dynamic Dataset Recycle, alongside a novel approach termed the Expansion Trick.
- The GaFi pipeline is introduced as a flexible, model-agnostic framework aimed at maximising the CAS metric for synthetic data generation.
- Empirical results demonstrate that the GaFi pipeline yields synthetic data with CAS values approaching the upper bound of real data accuracy, establishing a new state of the art in the field of synthetic data generation for classification.

The second part of the study focuses on adapting the pre-trained Stable Diffusion 2.0 model to generate synthetic datasets conditioned on class vectors, specifically targeting the generation of data with high informational content for downstream classification [117]. This class-conditioned adaptation is achieved through a multi-step pipeline involving transfer learning, fine-tuning, and Bayesian optimisation of key hyperparameters. The contributions of this approach include:

- A class-conditioned version of the Stable Diffusion 2.0 model is presented, capable of generating synthetic datasets that provide substantial utility for training classifiers.
- An adaptation pipeline is proposed, incorporating transfer learning and fine-tuning, alongside Bayesian optimisation, to refine dataset generation for diffusion models.
- The effectiveness of the approach is demonstrated by showing incremental improvements in dataset quality and a reduction in per-sample generation times. In some instances, classifiers trained on the synthetic data outperform those trained on real-world data.

3.2. Related Works

The past decade has witnessed remarkable advancements in deep learning, particularly in the realm of generative models. These models have demonstrated an increasing capacity to produce synthetic data that closely mimics real one. At the forefront of this progress are three key architectural frameworks: Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Denoising Diffusion Probabilistic Models (DDPMs) [10, 12, 97]. The field of computer vision has been significantly impacted by these models, which have found extensive application in image generation tasks. To assess the perceptual quality of synthetically generated images, researchers have developed various metrics. Among these, the Inception Score (IS) and the Fréchet Inception Distance (FID) have emerged as the most widely accepted and empirically validated measures [118, 119]. Both metrics utilise the Inception network architecture and employ statistical methods to quantify the similarity between generated images and their real counterparts [120].

The generation of synthetic datasets, either as a substitute for or complement to real data, has garnered increasing attention within the machine learning community. Synthetic data offers numerous advantages, including the ability to generate large-scale datasets with well-defined characteristics, thereby mitigating the costs associated with data collection and annotation. Furthermore, it addresses concerns related to privacy and restricted access to real-world data. The application of generative models for creating synthetic datasets has found utility across diverse domains, including semantic segmentation [121–124], optical flow estimation [125–127], human motion analysis [128–131], image classification [116, 132, 133], and recently even neural architecture generation [134, 135].

A significant contribution to this alternative application of generative models was made by Ravuri *et al.*, who introduced the Classification Accuracy Score (CAS) metric to evaluate GAN performance[116]. This metric assesses the efficacy of synthetic data in supporting downstream tasks by training a classifier on synthetic images and evaluating its performance on real images. Ideally, if a generative model accurately captures the underlying data distribution, classi-

fiers trained on synthetic data should exhibit performance comparable to those trained on real data. However, achieving such performance parity remains a substantial challenge, despite ongoing efforts to bridge this gap.

One promising approach to address this issue is the **Sample Filtering** technique proposed by Dat *et al.* [136]. This method aims to enhance the quality of generated data by utilising an auxiliary classifier, trained on real data, to predict labels for synthetic samples. Samples that are incorrectly classified or exhibit low prediction confidence are discarded. The same researchers suggest that employing multiple generative models can further improve the fidelity of synthetic data by more accurately capturing the real data distribution. An alternative strategy for narrowing the performance gap is the **Dataset Smoothing** technique introduced by Besnier *et al.* [133]. This approach involves maintaining a dynamic dataset, where only a subset of the synthetic training data is replaced with new samples at each training epoch, thereby ensuring both diversity and gradual adaptation. Building upon these foundational methods, researchers have proposed further refinements and innovative techniques to more effectively minimise the CAS gap.

Concurrently, other generative models have been employed to create datasets through the use of textual prompts for conditioning, or by distilling the essence of real datasets into a limited number of representative samples [137, 138]. These approaches have expanded the utility of synthetic data across various machine learning contexts, further demonstrating the potential of generative models in advancing the field.

3.3. Gap Filler

This research presents a novel post-processing framework engineered to improve the Classification Accuracy Score (CAS). The proposed pipeline exhibits remarkable adaptability, accommodating a wide range of generative methods. Its development is based on a careful analysis of the existing literature, identifying the most effective approaches and refining them to increase their versatility. In

addition, the framework incorporates novel, bespoke methods designed to further optimise performance.

3.3.1. Post-Processing Techniques

Dynamic Sample Filtering. Extending the seminal work of Dat *et al.*, this research has refined their sampling filtering methodology by implementing an adaptive approach tailored to the specific dataset and generative model under investigation. Their ablation study demonstrated that the use of a dynamic filtering threshold yields superior synthetic datasets compared to static filtering approaches [136].

The newly developed technique, called Dynamic Sample Filtering, works in two distinct phases. First, a pre-trained classifier is used to predict the labels of the generated samples, discarding those that are incorrectly classified. Next, a series of filtering thresholds are set, ranging from 0 to 0.9. For each threshold, a separate data set is generated, consisting only of the correctly classified samples with confidence levels above the set threshold. Data generation continues until each synthetic dataset reaches the desired size. A novel classifier is then trained from scratch on each dataset and the threshold corresponding to the dataset with the highest CAS is retained. This approach effectively eliminates poor quality samples that might otherwise affect the performance of the downstream classifier.

Dynamic Dataset Recycle. Inspired by Besnier *et al.*'s Dataset Smoothing technique, a novel approach called Dynamic Dataset Recycle has been formulated. In contrast to the original technique, which only updates parts of the dataset during each iteration, this method replaces the entire synthetic dataset at each iteration.

Empirical analysis using ablation studies has shown that recycling the entire dataset significantly improves performance with respect to CAS. To mitigate the computational complexity, which scales with the size of the generated dataset, the proposed method generalises the recycling process by performing it at every

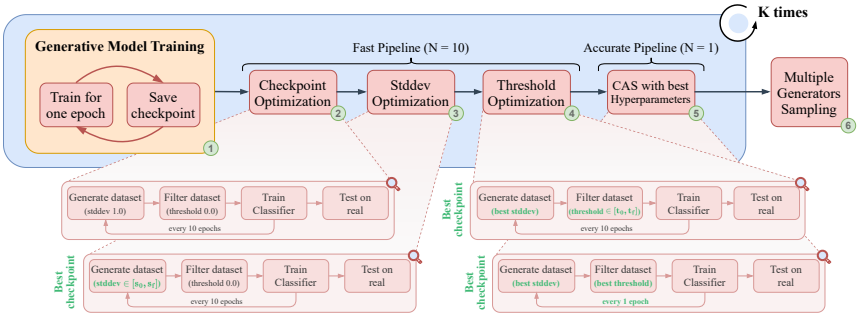

Figure 3.1: Schematic representation of the Gap Filler (GaFi) pipeline, illustrating the sequential application of post-processing techniques to optimise the Classification Accuracy Score (CAS).

N epochs during classifier training, thus optimising both time and performance.

Expansion Trick. A novel approach, termed the Expansion Trick, has been developed as a counterpoint to the Truncation Trick proposed by Brock *et al.* [139]. Instead of restricting the noise space, the Expansion Trick expands it by sampling from a normal distribution with a standard deviation larger than that used in model training. By expanding the noise space, the model is encouraged to explore previously under-sampled regions, producing more diverse and novel images – particularly beneficial in scenarios where diversity is prioritised over image fidelity.

While expanding the noise distribution increases variability, it also reduces the quality of individual samples. To overcome this limitation, the Expansion Trick is most effective when combined with filtering techniques, which serve to eliminate low-quality samples and ensure that only the most relevant examples are used for classifier training.

3.3.2. Pipeline Implementation

The Gap Filler (GaFi) pipeline implements a sequential application of the above post-processing techniques. The effectiveness of this pipeline depends on the appropriate sequencing of these techniques and the careful selection of hyperparameters. Figure 3.1 outlines the overall structure of the pipeline, which includes the following stages:

1. **Generative Model Training.** The initial phase comprises the training of the selected generative model, with model checkpoints stored at the end of each epoch for subsequent evaluation. It is noteworthy that the specific architecture of the generative model remains flexible and unconstrained by the pipeline.
2. **Checkpoint Optimisation.** To improve the performance of downstream classifiers, it is essential to identify the most effective checkpoint from those retained during training. Each checkpoint is evaluated by calculating its CAS, and the checkpoint with the highest score is retained. This phase uses a fixed set of hyperparameters: a standard deviation of 1.0 and a filtering threshold of 0.0, whereby only samples predicted to be incorrect are discarded, regardless of prediction confidence. The Dynamic Dataset Recycle parameter N is set to 10, meaning that the dataset is updated every 10 training epochs, forming the so-called "Fast Pipeline".
3. **Stddev Optimisation.** Having identified the optimal checkpoint, the following step involves fine-tuning the standard deviation parameter for the Expansion Trick. This is achieved by using the Fast Pipeline to compute the CAS while modulating the standard deviation between two predefined limits, s_0 and s_f. In the present study, s_0 is set at 1.0 and s_f is set at 2.0, with an increment of 0.05.
4. **Threshold Optimisation.** Following the determination of the optimal standard deviation, the next step is to optimise the filter threshold for the Dynamic Sample Filtering technique. Analogous to the previous optimisation step, the Fast Pipeline is used to compute the CAS, with the filtering

threshold varying between two values, t_0 and t_f. In the experiments carried out, t_0 is set at 0.0 and t_f at 0.9, with increments of 0.1.

5. **Final CAS Optimisation.** Once the optimal hyperparameters have been determined, the classifier is trained using the "Accurate Pipeline". In this phase, the Dynamic Dataset Recycle parameter N is set to 1, facilitating the use of a highly diverse dataset to achieve optimal classifier performance.

6. **Multiple Generators Sampling.** The final stage of the GaFi pipeline incorporates multiple generative models to improve the sampling process. This step involves iterating the previous stages of the pipeline K times, with each iteration training multiple identical generative models with different random seeds on the same dataset, as proposed by Dat *et al.* [136]. This approach allows different aspects of the data distribution to be captured, culminating in the creation of a composite synthetic dataset by uniformly sampling the outputs of these models at each epoch of classifier training.

3.3.3. Experiments and Results

This section elucidates the experimental protocol and configuration to ensure replicability and transparency. The generative model employed in this research was the BigGAN-Deep architecture, implemented via StudioGAN with minor modifications to the residual block arrangements [139, 140]. Both generative and discriminative networks were initialised using Orthogonal Initialisation and optimised via Adam, with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a fixed learning rate of 2×10^{-4}[141, 142]. An exponential moving average was applied to the generator weights update with a 0.9999 decay rateBrock et al. [139]. Data augmentation was constrained to random horizontal flipping of the training corpus. All models utilised a batch size of 192, with a 3:1 ratio of discriminator to generator updates.

The downstream classifier used the ResNet20 architecture, known for its effectiveness [143]. The width was set to 64, and conventional training techniques were used, including cross-entropy loss, a batch size of 128, 100 epochs of train-

Table 3.1: Impact of Dynamic Sample Filtering on Classification Accuracy Score (CAS) for various filtering thresholds across the five datasets. Best results are in **bold**.

Dataset	No Filter	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
FashionMNIST	88.70	89.88	90.01	89.59	89.98	89.89	90.05	**90.21**	89.86	90.12	89.90
CIFAR10	87.11	88.45	88.95	88.75	88.45	88.67	**89.06**	88.72	88.96	88.09	88.41
CIFAR100	57.74	59.13	58.82	**59.39**	59.35	59.20	59.06	58.79	58.76	57.28	55.52
CINIC10	75.58	76.70	77.10	77.85	76.83	78.08	77.62	77.80	77.94	**78.50**	77.11
DermaMNIST	67.48	67.58	67.38	67.08	66.88	67.18	**67.93**	67.53	66.48	67.08	66.38

ing, and SGD optimisation with an initial learning rate of 0.1, momentum of 0.9, and weight decay of 1×10^{-4}. The learning rate was reduced by an order of magnitude at epochs 60 and 80. To augment the synthetic training set, which matched the real dataset in cardinality and class balance, a simple augmentation technique was implemented: zero-padding the input image to 40×40, followed by randomly extracting a 32×32 section for the final input and random horizontal flip.

The experimental setting involved five datasets: FashionMNIST [65], CIFAR10 [63], CIFAR100 [63], CINIC10 [70], and DermaMNIST [144]. Comprehensive dataset descriptions are available in the Appendix A. To standardise image dimensions, FashionMNIST images were zero-padded to 32×32, while DermaMNIST images were resized to 32×32 using the Lanczos algorithm due to their RGB nature. The experiments were performed on a machine equipped with an Intel(R) Xeon(R) Gold 6238R CPU @ 2.20GHz and an Nvidia Quadro RTX 6000 GPU. The training time for a single ResNet20 model ranged from 1 to 2.5 hours, depending on the specific post-processing techniques used, while the BigGAN-Deep training took approximately 48 hours.

Dynamic Sample Filtering Table 3.1 demonstrates the efficacy of the Sample Filtering technique across all examined datasets. However, the optimal threshold exhibits significant dataset dependency. For instance, the CAS for FashionMNIST and CIFAR10 remains relatively stable across thresholds, whereas CIFAR100 exhibits substantial classifier performance degradation at higher thresh-

Table 3.2: Effect of Dynamic Dataset Recycle frequency on Classification Accuracy Score (CAS) for the five datasets. Best results are in **bold**.

Dataset	No Recycle	N=10	N=5	N=1
FashionMNIST	88.70	89.29	89.88	**90.16**
CIFAR10	87.11	89.72	90.25	**90.42**
CIFAR100	57.74	59.68	60.57	**61.38**
CINIC10	75.58	79.37	80.55	**82.57**
DermaMNIST	67.48	68.03	68.33	**68.43**

olds.

This behaviour is hypothesised to stem from dataset complexity variations. In simpler datasets, such as FashionMNIST and CIFAR10, generators produce images closely resembling the original dataset. Consequently, classifiers pre-trained on real images exhibit high prediction confidence, with most low-quality images already eliminated due to incorrect labelling. Conversely, CIFAR100 presents a more complex challenge, with a tenfold increase in class count. This heightened complexity results in generated images more susceptible to rejection by the pre-trained classifier when a high filtering threshold is applied. On average, the Dynamic Sample Filtering technique enhances the CAS by 1.7% relative to the baseline CAS (represented by the "No Filter" column in the table).

Dynamic Dataset Recycle Table 3.2 illustrates that the proposed dataset recycling technique substantially improves CAS across all five datasets. The results indicate that even with a relatively infrequent recycling period, such as $N = 10$, there is an accuracy increase ranging from 0.55% to 3.79%, depending on the dataset. Notably, reducing the recycling period, i.e., generating new synthetic data more frequently during training, leads to further performance enhancement. As anticipated, the accuracy improvement is more pronounced with increasing dataset complexity. This is attributed to the generative model potentially requiring multiple attempts to produce meaningful data, particularly for classes learned with lower effectiveness.

Expansion Trick Figure 3.2 illustrates the impact of the Expansion Trick on

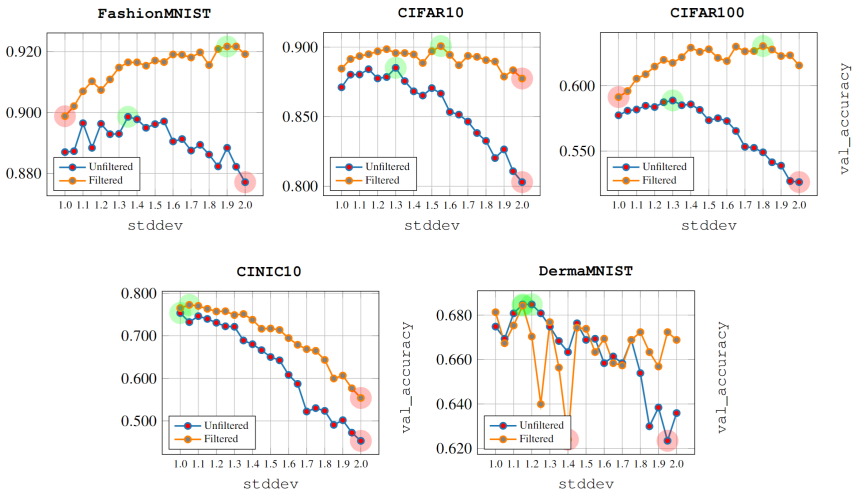

Figure 3.2: Impact of the Expansion Trick on Classification Accuracy Score (CAS) for filtered and unfiltered datasets with various standard deviations.

the CAS. The results demonstrate that for unfiltered datasets, a modest performance increase is observed when using a standard deviation slightly exceeding 1, with the exception of CINIC10. However, as the standard deviation increases, a performance decline is noted. This outcome is not unexpected, as a higher standard deviation leads to more diverse images at the expense of image quality. Consequently, at a certain point, the images become excessively degraded for effective downstream classifier training. Further information is available in Appendix C.

Conversely, when the Expansion Trick is employed in conjunction with a sample filtering technique, significantly higher standard deviation values can be achieved without compromising performance. This is due to the retention of only "good" images meeting the filtering criteria - in this instance, correctly classified images. This ensures a more diverse dataset while maintaining higher image quality than the unfiltered dataset, resulting in improved classification accuracy.

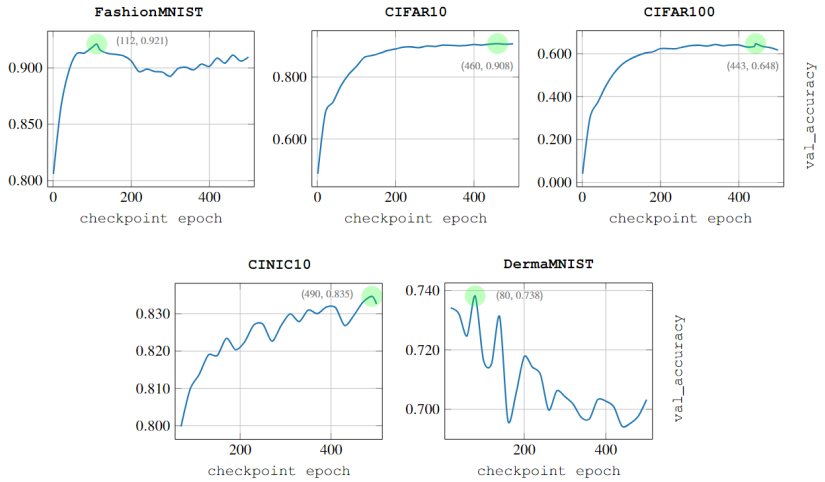

Figure 3.3: Evolution of Classification Accuracy Score (CAS) across generative model checkpoints for different datasets.

The final application of the Expansion Trick technique enhances CAS by 3.47%, 2.96%, 5.29%, 1.69%, and 0.95% on the FashionMNIST, CIFAR10, CIFAR100, CINIC10, and DermaMNIST datasets, respectively. These improvements underscore the effectiveness of the Expansion Trick in augmenting the generative model's capacity to produce informative samples, thereby enhancing downstream classifier performance.

Checkpoint Optimization Figure 3.3 depicts the evolution of the CAS metric as a function of the generative model checkpoint. It is important to note that each point corresponds to the complete training of a ResNet20 classifier, underscoring the computational intensity of this step. The objective is to identify the optimal checkpoint with respect to the CAS metric.

Analysis of the graphs reveals that the optimal epochs for FashionMNIST and DermaMNIST are 112 and 80 respectively, while for CIFAR10, CIFAR100, and CINIC10, the optima are reached at epochs 460, 443, and 490 respectively. A

Table 3.3: Optimal hyperparameter configuration and resultant Classification Accuracy Score (CAS) using the Accurate Pipeline for each dataset.

Dataset	Checkpoint	Standard Deviation	Filtering Threshold	CAS
FashionMNIST	112	2.00	0.0	94.03
CIFAR10	460	1.60	0.3	92.60
CIFAR100	443	1.70	0.1	68.92
CINIC10	490	1.25	0.0	84.37
DermaMNIST	80	1.30	0.4	73.66

key observation is that the CAS increases with the number of epochs up to a certain point. This behaviour aligns with that of GANs from the perspective of perceptual quality of generated images, which tend to deteriorate after a certain number of training iterations.

The optimal point clearly varies with the dataset and its complexity. Given the uniform image size across the considered datasets, this complexity is understood in terms of the number of classes, image channels, and the cardinality of the datasets themselves. This assertion is corroborated by the CAS convergence points occurring in the first quarter of the available epochs for the simplest datasets and in the last quarter for the more complex ones. Overall, the Checkpoint Optimisation step is critical, enabling subsequent steps in the GaFi pipeline to commence from the optimal generative model.

Accurate Pipeline After determining the optimal configuration, summarised in Table 3.3, the classifier can be retrained using the "Accurate Pipeline". In this phase, the recycling period N is set to 1, regenerating the dataset at each training epoch in order to obtain the optimal CAS through the GaFi pipeline. It is evident that the Expansion Trick, combined with the Dynamic Sample Filtering technique, played a crucial role in achieving the optimal CAS. This is supported by the shifted values towards the standard deviation of 2 in each configuration compared to the use of the Expansion Trick alone.

This approach achieves superior results with improvements over the baseline of 5.33% for FashionMNIST, 6.09% for CIFAR10, 14.21% for CIFAR100, 10.14%

Table 3.4: Comparative analysis of Classification Accuracy Score (CAS) for classifiers trained on generated data, comparing the GaFi pipeline with previous methods and real data. Best results from generative datasets are in **bold**, second-best underlined.

#	Method	FashionMNIST	CIFAR10	CIFAR100	CINIC10	DermaMNIST
-	Real Data	96.01	94.98	75.64	89.05	77.25
	Baseline	88.70	87.11	57.74	75.58	67.48
1	Dat *et al.*	-	88.25	62.22	-	-
	GaFi	**94.03**	92.60	68.92	84.37	73.66
2	Dat *et al.*	-	89.68	64.33	-	-
	GaFi	93.98	92.74	70.22	85.42	75.06
4	Dat *et al.*	-	90.68	67.22	-	-
	GaFi	93.99	93.02	71.75	85.62	74.71
6	Dat *et al.*	-	91.14	67.56	-	-
	GaFi	93.98	**93.20**	**71.95**	**85.72**	**75.21**

for CINIC10 and 7.73% for DermaMNIST. Furthermore, this pipeline achieves higher accuracy even when using only one generator compared to the best configuration of Dat *et al.* with six generators. These results show that the proposed post-processing techniques, and the way they are applied in the GaFi pipeline, lead to superior classifiers trained on more general and useful data.

It is noteworthy that the gap between synthetic and real data has been significantly reduced. Specifically, for the FashionMNIST, CIFAR10, CIFAR100, CINIC10 and DermaMNIST datasets, the gap with respect to the baseline has been reduced from 7.31%, 7.87%, 17.90%, 13.47% and 9.77% to 2.03%, 1.78%, 3.99%, 3.33% and 2.04% respectively. This remarkable result demonstrates the undeniable effectiveness of the proposed approach.

3.4. Stable Diffusion 2.0 Adaptation

This section explains the methodology used to adapt the Stable Diffusion (SD) 2.0 model, initially pre-trained on ImageNet-1K, for the generation of synthetic datasets tailored to classification tasks. Since SD 2.0 was originally designed for text-to-image generation, the first step in facilitating class-conditional generation is to replace the text-conditional embedding. Consequently, the text encoder is replaced by a fully connected class encoder, which is tasked with linearly mapping one-hot encoded class vectors into the dimensional space previously used for text conditioning in the SD 2.0 architecture. The proposed adaptation pipeline consists of four distinct phases:

1. **Transfer Learning for Class-Encoder.** The class encoder is specifically trained on the target dataset, while the rest of the architecture retains its original pre-trained weights. The training scheme spans 50 epochs, maintaining a consistent batch size of 64. The optimisation process uses the Adam algorithm with a weight decay coefficient of 4×10^{-3}, global gradient norm clipping at 10, and a cosine annealing schedule for learning rate decay initialised at 1×10^{-4}. At the end of each epoch, a checkpoint of the class encoder is preserved. The primary goal of this phase is to achieve class-conditional generation with minimal changes to the existing model framework.
2. **Initial Hyper-Parameters Optimisation.** The SD generation process is highly dependent on two critical hyperparameters: the number of denoising steps (DS) and the unconditioned guidance scale (UGS). Their precise calibration is crucial to achieve an optimal balance between image quality and intra-class diversity. As the default values of the original model (DS=50 and UGS=7.5) were optimised for text-conditioned single image quality, their direct application in this context would lead to sub-optimal results. Therefore, an exhaustive search for the optimal combination of these parameters is performed in parallel with the identification of the most effective Transfer Learning epoch. The optimisation strategy uses the Tree-Structured Parzen Estimator as a Bayesian optimisation tech-

nique, enhanced with Hyperband Pruning [145]. Each optimisation iteration selects a combination of the three hyperparameters – (DS∈[5,50], UGS∈[0,7.5], Epoch∈[1,50]) – and generates a compact dataset of 4000 images with a uniform class distribution. This dataset is then used to train a ResNet20 architecture and compute the Classification Accuracy Score (CAS), which quantifies the accuracy achieved by a model trained exclusively on generated data when evaluated against a real test set [116]. This process is iterated 50 times to maximise the CAS metric.

3. **Fine-Tuning of the Diffusion Model.** Using the optimal parameters identified in the previous phase, the SD model is fine-tuned, leaving all other components, including the class encoder, unchanged. This phase uses a training strategy analogous to the first step, with the following modifications: the initial learning rate is reduced to 1×10^{-5}, the number of epochs is limited to 10, and the batch size is adjusted to 16. At the end of each epoch, a checkpoint of the SD model is preserved for subsequent analysis.

4. **Final Hyper-Parameters Refinement.** Similar to the second phase, this step involves a simultaneous search for the optimal adaptation epoch and refinement of the DS and UGS hyperparameters. The main difference lies in the reduced number of fine-tuning checkpoints (10), the restricted range of DS values (between 5 and the optimal value from phase 2) and the extended range for UGS (between 0 and twice the previously identified optimal value).

Upon identification of the optimal configuration, synthetic datasets are generated to replicate the size of the target dataset, scaled by factors ranging from 1 to 10 in integer increments. In this final stage, the class distribution is carefully maintained to reflect that of the real data. For each synthetic dataset, the CAS is calculated and then compared to the test accuracy of a ResNet20 model trained exclusively on the authentic training set.

Table 3.5: Quantitative assessment of the adaptation pipeline efficacy. Top-1 Accuracy and Generation Time computed subsequent to each pipeline phase utilising a synthetic dataset comprising 4000 images. The optimal score for each dataset and metric is denoted in **bold**.

Dataset	Classification Accuracy Score ↑ (%)				Generation Time ↓ (s)			
	After 1.	After 2.	After 3.	After 4.	After 1.	After 2.	After 3.	After 4.
CIFAR10	39.52	47.05	51.94	**59.30**	30000	**18600**	**18600**	**18600**
CIFAR100	15.50	17.48	17.35	**27.36**	30000	21000	21000	**16800**
PathMNIST	39.95	63.06	59.18	**78.95**	67500	**64800**	**64800**	**64800**
DermaMNIST	19.10	62.64	61.84	**65.43**	5250	5250	5250	**2835**
BloodMNIST	73.89	74.97	78.98	**86.20**	9000	8460	8460	**7920**
RetinaMNIST	37.00	40.75	41.74	**47.45**	810	600	600	**486**

3.4.1. Experiments and Results

In order to evaluate the effectiveness of the proposed methodology, a series of experiments were performed using six 32x32 RGB datasets: CIFAR10 [63], CIFAR100 [63], PathMNIST [144], DermaMNIST [144], BloodMNIST [144] and RetinaMNIST [144]. Detailed descriptions of these datasets can be found in Appendix A.

The only pre-processing step was to adapt these datasets to the input requirements of the SD model. Specifically, each image underwent bilinear interpolation to achieve a resolution of 128x128 pixels, followed by normalisation to the [-1, 1] range. The ResNet20 models were consistently trained over 100 epochs using an Adam optimiser with an initial learning rate of 1×10^{-3}, label smoothing set to 0.1, and a batch size of 256. A learning rate scheduler was implemented to reduce the rate upon reaching a plateau, with a patience of 10 epochs and a reduction factor of 0.1. Early stopping was enforced with a patience of 25 epochs. In particular, no data augmentation techniques were used to ensure a fair comparison between the information content of real and generated images. All experiments were performed on a single Nvidia Quadro RTX 6000 GPU.

Table 3.5 presents the performance metrics achieved during each phase of the

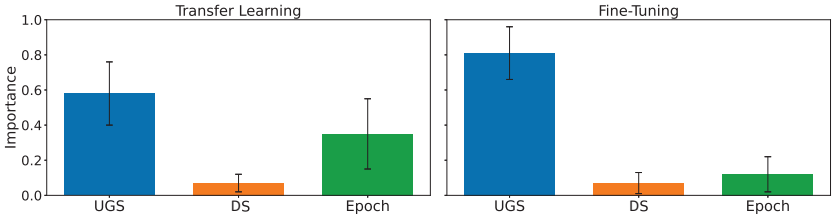

Figure 3.4: Comparative analysis of hyperparameter significance. Average importance of hyperparameters across the dual optimisation phases, as determined through functional ANOVA methodology.

adaptation pipeline by a ResNet20 model trained on 4000 synthetically generated images, evenly stratified by class. The first evaluation, after the Transfer Learning phase (After 1.), serves as a baseline for both the CAS of the classifier and the generation time of the SD model. Substantial improvements are observed after the initial hyperparameter optimisation (After 2.), with CAS increases ranging from 1.08% to 43.54% and generation time reductions of up to 38%, facilitated by the identification of the optimal IS value. The fine tuning phase (After 3.) gives mixed results in terms of CAS, suggesting that the parameters considered optimal in the previous phase may not be universally applicable. However, on completion of the final pipeline phase (After 4.), further improvements in CAS are recorded, with increases from 10.45% to 46.33%, accompanied by further reductions in generation times.

Figure 3.4 illustrates the average importance of the hyperparameters across the two optimisation phases, using functional ANOVA to assess the importance of individual hyperparameters and their interactions [146]. While the analysis does not yield identical rankings for each phase, it consistently identifies the UGS as the most critical parameter. The UGS, which is closely related to the conditioning temperature, plays a central role in class-conditioned production, which makes this result logical. Contrary to initial expectations, the second most important parameter is not the number of denoising steps, which is typically pre-

Table 3.6: Comparative performance analysis. Top-1 Accuracy achieved on identical real test sets by ResNet20 models trained on authentic data juxtaposed with synthetic variants of increasing cardinality. The overall optimal score is emphasised in **bold**, whilst the most favourable score derived from synthetic training sets is denoted by underline.

Dataset	Real	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
CIFAR10	83.26	79.09	81.49	83.67	83.82	83.67	84.52	**85.82**	85.47	85.57	85.79
CIFAR100	**60.55**	45.19	47.68	48.93	49.50	51.51	51.46	51.16	52.54	52.73	52.93
PathMNIST	**88.70**	85.19	84.52	81.36	84.90	81.33	84.03	83.90	82.18	81.84	86.45
DermaMNIST	**75.71**	66.98	67.03	66.93	68.57	68.47	67.38	69.42	67.38	68.23	67.73
BloodMNIST	**96.22**	85.56	84.83	84.80	84.89	86.58	85.85	86.29	82.58	85.32	83.02
RetinaMNIST	47.75	40.50	43.25	47.00	**52.00**	47.00	43.00	51.25	46.25	49.00	46.50

sented as one of the most important SD parameters, but rather the adaptation epoch.

The comprehensive CAS results for the full range of synthetic datasets generated at the end of the pipeline are reported in Table 3.6 for top-1 accuracy. Synthetic datasets with cardinality equivalent to the training sets do not achieve the CAS of their real counterparts. Nevertheless, the CAS achieved is sufficient to claim that SD has been effectively adapted to generate synthetic versions of each target dataset. As the cardinality of the generated datasets increases, significant improvements are observed. Indeed, the best results are obtained from synthetic datasets with cardinality between 4 and 10 times that of their real counterparts, accompanied by significant improvements in terms of CAS.

The most remarkable results are those for the CIFAR10 and RetinaMNIST datasets, where models trained on synthetic datasets outperform those trained on real data. While for the latter dataset this improvement can be attributed to data scarcity, for the former it can be generalised to the hypothesis that the SD model is indeed able to compete with a comprehensive and well-structured dataset. Regarding the other results, it is noteworthy that there is often potential for further improvement, as the optimal results often correspond to the

maximum cardinality used, suggesting the scalability of the presented approach.

3.5. Conclusions

This chapter presented two methodologies for improving the performance of deep learning models when trained solely on synthetic data: the Gap Filler (GaFi) pipeline and an adaptation of the Stable Diffusion 2.0 model.

The GaFi pipeline, a set of post-processing techniques for generative models, improves the Classification Accuracy Score (CAS). This approach introduces three methods: Dynamic Sample Filtering, Dynamic Dataset Recycle, and Expansion Trick. When properly implemented, these techniques produce measurable benefits across the three datasets examined. In parallel, the study explores an adaptation of the pre-trained Stable Diffusion 2.0 model, using ImageNet-1K for synthetic dataset generation. This process encompasses Transfer Learning, Fine Tuning, and optimisation of generation parameters, aimed at enhancing the efficacy of downstream classifiers. The results show that in one-third of cases, models trained on this synthetic data outperformed those trained on real data.

It is worth noting that the overall accuracy achieved with synthetic data remains slightly below that attained with real data. This difference raises questions about the nature of generative modelling and the feasibility of perfectly replicating real data distributions. The outcomes of both approaches highlight the potential of generative models and synthetic data in research and industry. Whilst acknowledging the existing challenges, the research team maintains a positive outlook on further reducing the performance gap. Overcoming these obstacles could lead to advancements in various fields. In conclusion, this research establishes new benchmarks in synthetic data generation and classification accuracy, paving the way for further innovations in generative modelling. The results obtained underscore the growing importance of synthetic data and generative models in academic and industrial spheres, indicating new directions for machine learning and artificial intelligence.

4 | Knowledge Recycling

Generative artificial intelligence has revolutionised the production of synthetic data, offering novel solutions to challenges such as data scarcity and privacy, which are particularly pressing in fields like medicine. Despite these advancements, a persistent challenge remains in effectively utilising synthetic data to train high-performance models. This paper proposes a solution to this challenge by introducing the Knowledge Recycling (KR) pipeline, designed to enhance both the generation and application of synthetic data in training downstream classifiers. Central to this pipeline is the technique of Generative Knowledge Distillation, which markedly improves the quality and relevance of information provided to classifiers. This is achieved through a mechanism that regenerates synthetic datasets and applies soft labelling. The KR pipeline has been rigorously evaluated across a diverse set of datasets, with particular attention given to six highly heterogeneous medical image datasets, ranging from retinal images to organ scans. The results demonstrate a substantial reduction in the performance disparity between models trained on real data and those trained on synthetic data. In certain cases, models trained on synthetic data even surpassed the performance of those trained on real data. Additionally, the models produced exhibit near-total resistance to Membership Inference Attacks, thereby providing privacy properties that are notably absent in models trained using traditional methods.

4.1. Introduction

The emergence of generative deep learning represents a pivotal technological advancement that is swiftly infiltrating various aspects of society, significantly influencing the everyday lives of individuals. This technology facilitates the effortless creation and interaction with high-quality synthetic data, encompassing images, text, audio, and video. Consequently, the ease with which such artificially generated content can be produced increasingly blurs the line between human-made and algorithmically generated outputs.

Simultaneously, the rapid expansion of generative models is instigating a transformation across numerous sectors, leading to profound implications. On one hand, it unlocks new opportunities, while on the other, it presents ethical and social challenges regarding the utilisation and potential misuse of these technologies.

Presently, the advancements in this field bring forth challenges related to the dissemination of algorithmically generated content, often falsely attributed to human creation. Nevertheless, these developments also unveil opportunities for beneficial applications, particularly in scenarios where it becomes difficult to discern between human and algorithmic production. This indistinguishability has catalysed the use of generative models to enhance real datasets and, more recently, the ambitious goal of generating entirely synthetic datasets.

However, generating complete synthetic datasets is a complex endeavour requiring models capable of producing substantial amounts of data in a reasonable timeframe, while ensuring a careful balance between data quality and diversity. It is well-documented that models trained exclusively on synthetic data tend to underperform when compared to those trained on real data. Additionally, data privacy, a concern of escalating importance, must be meticulously addressed both before and after model training. This consideration is especially critical within the context of medical data, where safeguarding privacy is paramount to maintaining the trust between healthcare professionals and patients. In this context, generative technology presents untapped potential for the secure utilisation of

medical data, paving the way for innovations and advancements in healthcare research.

4.1.1. Main Contributions

This chapter introduces the Knowledge Recycling (KR) strategy, a comprehensive pipeline designed to enhance the generation of synthetic datasets and improve the training of downstream classifiers using solely synthetic images. The process begins with the training of the generator alongside an auxiliary classifier, referred to as the Teacher Classifier. The optimal checkpoint for the Generator is identified by training a Student Classifier at each checkpoint, employing the proposed Generative Knowledge Distillation technique. In this approach, the Teacher Classifier generates soft labels for the synthetic images, enabling the Student Classifier to learn about uncertainties and class correlations, thus improving its prediction accuracy on both synthetic and real images.

Upon determining the optimal checkpoint, the generation parameters are fine-tuned by adjusting the size of the synthetic dataset, the frequency of dataset regeneration during Student Classifier training, and the Generator's standard deviation. Following the completion of the optimal Student Classifier training, its robustness against Membership Inference Attacks is evaluated and compared to that of the Teacher Classifier [24].

The objective of this work is to demonstrate that classifiers trained on synthetic data can achieve performance levels comparable to those trained on real data, while offering enhanced resistance to Membership Inference Attacks. The primary contributions of this research are as follows:

- The introduction of Knowledge Recycling, a novel pipeline that optimises the generation and application of synthetic data within the context of classifier training.
- The development of Generative Knowledge Distillation, a technique aimed at enhancing the quality of information transferred from synthetic data to classifiers, thereby narrowing the performance gap between models trained

on synthetic versus real data.

- Validation of the proposed pipeline's effectiveness in producing models with significant resistance to Membership Inference Attacks, thereby achieving a favourable balance between performance and privacy protection.

4.2. Related Works

The landscape of image generation is currently dominated by two principal model families: Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPMs) [10, 12]. Despite their differing operational mechanisms, both model types demonstrate a robust capacity to generate and manipulate high-resolution images when conditioned in various ways [147, 148]. While this research primarily focuses on generating high-quality individual images, a parallel research trajectory has emerged, leveraging these generative capabilities to produce fully synthetic datasets or to augment existing ones.

Initial efforts in this domain targeted scenarios where data collection and annotation are notably challenging and time-intensive, such as in medical imaging. For instance, Frid-Adar et al. utilised GANs to synthesise images of liver lesions, revealing that supplementing the original dataset with these synthetic images significantly enhanced the diagnostic accuracy of classification models for liver lesions [149]. Building on this, Sedigh et al. and Islam et al. applied GAN models to generate synthetic images for skin cancer and brain PET scans, respectively, with both studies reporting improved classification performance when real images were augmented with synthetic ones [150, 151].

Furthermore, research has advanced beyond merely enriching existing datasets to focus on generating entirely new datasets and evaluating their efficacy through downstream machine learning tasks [137, 138, 152]. Early studies indicated that synthetic data alone often lacks sufficient semantic depth for models trained on such data to perform effectively on real-world inference tasks [153]. To counter this, various strategies have been proposed to optimise the utility of the vast image sets generated by these models. For example, recycling synthetic

datasets during training and generating synthetic datasets with greater cardinality than the original training set have both shown to significantly improve performance [133]. Additionally, filtering techniques have been developed to exclude synthetic images classified with low confidence or inaccurately by auxiliary classifiers, thus enabling sampling from sparser distributions and further enriching the synthetic datasets [136].

As the development of novel models and techniques progresses, the potential attack surfaces of these models expand, necessitating increased attention to privacy protection. Among the most prevalent attack vectors are Membership Inference Attacks (MIAs), Model Inversion, Model Extraction, and Data Poisoning, each applicable depending on the attacker's access level and interaction capabilities with the targeted model [24, 45, 154, 155]. MIAs, in particular, are notable for their simplicity, as they can be executed even in black-box scenarios, relying solely on model logits. These attacks aim to determine whether a given input sample was part of the model's training set. Upon successful identification, MIAs can facilitate more sophisticated attacks, such as Model Inversion, Model Extraction, or further inference attacks designed to extract more detailed information.

Numerous defensive strategies have been proposed to mitigate these threats, many of which leverage relaxed forms of Differential Privacy [25]. Although these methods are effective at thwarting MIAs, they often introduce significant training overheads and can degrade the performance of the protected model, as demonstrated in Chapter 2. Recently, however, alternative approaches have emerged that offer a more favourable trade-off between model performance and privacy protection. These include adversarial training techniques and private training at individual steps, rather than over the entire training process [156]. Empirical metrics have also been developed to better quantify this trade-off, further advancing the field.

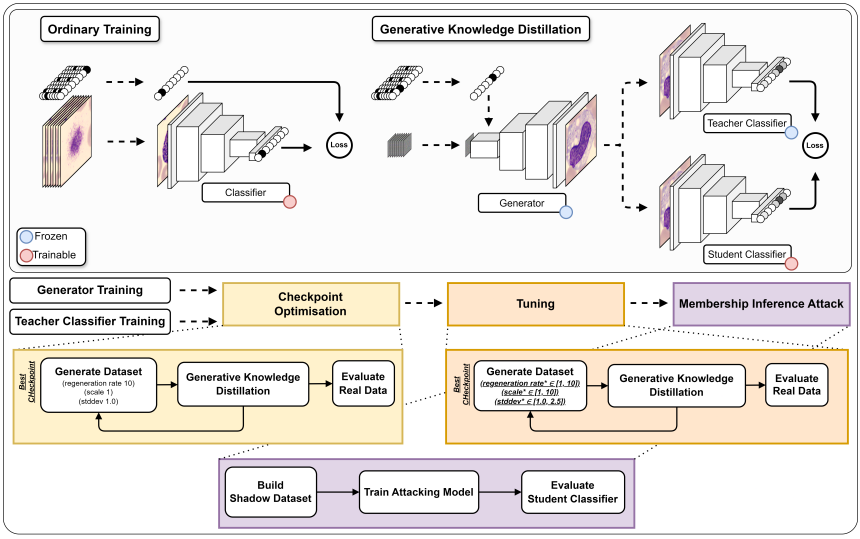

Figure 4.1: The Knowledge Recycling pipeline is illustrated, showcasing its key components. A comparison between the proposed Generative Knowledge Distillation technique and Ordinary Training is presented. This figure highlights the innovative approach to synthetic dataset creation and subsequent classifier training.

4.3. Method

This section delineates the Knowledge Recycling (KR) pipeline, which is designed for the creation of synthetic datasets and their subsequent application in training downstream classifiers. The process initiates with a preparatory phase wherein both an auxiliary classifier, referred to as the **Teacher Classifier**, and a data generator, termed the **Generator**, are trained on an identical real dataset.

The first core stage of the pipeline, known as **Checkpoint Optimisation**, focuses on determining the most effective checkpoint of the Generator. During this phase, classifiers—designated as **Student Classifiers**—are trained for each

checkpoint of the Generator. These Student Classifiers share the same architecture as the Teacher Classifier and undergo training via identical techniques. The synthetic datasets for each checkpoint are generated using the method termed Generative Knowledge Distillation (GKD), which is elaborated in Subsection 4.3.4.

Following the identification of the optimal checkpoint, the pipeline progresses to the **Tuning** stage. In this stage, the generation parameters are refined, and the final Student Classifier is trained. The concluding phase, termed the **Membership Inference Attack**, evaluates the robustness of the Student Classifier against the corresponding privacy attack. A visual overview of this pipeline is provided in Figure 4.1.

4.3.1. Teacher Classifier

The Teacher Classifier is integral to the KR pipeline, serving as the cornerstone of the GKD technique and providing the benchmark against which the Student Classifiers are evaluated for both accuracy and resistance to privacy attacks. To ensure a fair and robust comparison, the architecture and training methodology of the Teacher Classifier are meticulously replicated in the Student Classifiers.

Given the need to balance performance with training efficiency—considering that a complete training process is required for each checkpoint—the ResNet14 architecture was selected for its effectiveness [143]. Training is conducted in Mixed Precision over 500 epochs using the SGD optimiser, with an initial learning rate of 0.5. The learning rate is adjusted via a cosine annealing scheduler, and data augmentation is implemented using TrivialAugment and MixUp techniques [157–159]. Further specifics can be found in Appendix D.

4.3.2. Generator

Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPMs) currently represent the pinnacle of image generation techniques [10, 12]. Although these approaches differ fundamentally in their op-

erational mechanisms, both demonstrate high and comparable performance in producing various forms of media content, with the flexibility to condition the generation process in different ways. GANs are renowned for their rapid inference capabilities, though they are often plagued by instability during training. In contrast, DDPMs provide more stable training processes but are hindered by prolonged generation times. Despite advancements that have significantly reduced the number of denoising steps required by DDPMs, their generation speed remains insufficient to rival GANs in large-scale data generation tasks [160].

For this study, a GAN-based approach was selected, prioritising inference speed. Specifically, a modified version of the BigGAN-Deep model was employed, a model recognised as a significant milestone in GAN development [139]. BigGAN introduced several key innovations, including conditional batch normalisation, the truncation trick to balance quality and diversity in generation, and advanced optimisation techniques like spectral normalisation to manage large networks [161].

In the proposed implementation, the original BigGAN-Deep model underwent several modifications. The hinge loss function was replaced with a logistic loss, and the tanh activation function was substituted with a sigmoid. Furthermore, regularisation techniques such as label smoothing were incorporated to enhance the discriminator's performance, and the AdamW optimiser with a weight decay of 0.0005 was adopted. These alterations were designed to bolster the stability of training and the quality of the generated images.

The model was trained over 500 epochs, maintaining a 4:1 update ratio between the discriminator and generator. To ensure model robustness, checkpoints were systematically saved at intervals of 5 epochs. A comprehensive description of the implementation, alongside a comparative analysis between the vanilla model and the modified version, is available in Appendix D.

4.3.3. Evaluation Metric

In the assessment of image generators, the most commonly employed metrics in the literature are the Inception Score (IS) and the Fréchet Inception Distance (FID) [118, 119]. The IS evaluates the quality of the generated distribution by considering both the clarity and diversity of the produced images. Conversely, the FID offers a more comprehensive measure by comparing the generated distribution to the actual distribution used to train the Generator, thereby capturing both the quality and fidelity of the synthetic images.

However, recent studies have exposed limitations in these metrics, particularly when evaluating the utility of generated images in downstream learning tasks. It has been observed that IS and FID often lack correlation with the effectiveness of generated data in subsequent classification tasks. Additionally, a trade-off between the quality of individual images and the diversity of the generated distribution has been identified [162]. Maximising IS and FID often prioritises the quality of generated images at the expense of diversity, which is essential for creating synthetic datasets that enhance the generalisation capabilities of models trained on them.

In this study, the Classification Accuracy Score (CAS) is adopted as the primary metric. CAS measures the validation accuracy on real data of a classifier trained on synthetic datasets [116]. This metric is crucial for identifying the training epoch that yields the most effective synthetic dataset. Additionally, similar to IS and FID, CAS helps prevent mode collapse in the Generator.

4.3.4. Checkpoint Optimisation

With both the Teacher Classifier and the Generator defined, trained on real data, and subsequently frozen, the next phase in the pipeline is the Checkpoint Optimisation step. The objective of this initial phase is to identify the optimal checkpoint that maximises the performance of the downstream Student Classifier models.

For each Generator checkpoint, a Student Classifier is trained using a methodol-

ogy akin to that of the Teacher Classifier, though the number of training epochs is reduced to 100 (as opposed to 500) to enhance efficiency. At the start of each training session, a synthetic dataset, equal in size to the real dataset, is generated using the current checkpoint. The input noise is drawn from a multivariate Gaussian distribution with a standard deviation of 1.0. To preserve data diversity, the synthetic dataset is fully regenerated every 10 epochs during training.

Previous research has underscored the benefits of filtering generated data to improve the Classification Accuracy Score (CAS). For instance, Dat et al. employed a model similar to the Teacher Classifier to exclude images that produced inconsistent predictions [136].

In the KR pipeline, the proposed and adopted technique is **Generative Knowledge Distillation** (GKD). Unlike filtering methods, GKD leverages the Teacher Classifier to evaluate the generated images and produce soft labels for the Student Classifier. These probability labels are more informative than binary labels, as they encapsulate uncertainties and correlations between classes, leading to a substantial improvement in CAS, as elaborated in Appendix D. This approach optimises both the quality of the information transferred to the Student Classifier and the efficiency of synthetic dataset generation, enabling the desired dataset size to be achieved more rapidly than with filtering-based techniques.

4.3.5. Tuning

Following the identification of the optimal checkpoint based on the CAS metric, the pipeline advances to the Tuning step, where the generation parameters are optimised to further enhance the performance of the Student Classifiers. These parameters, which were held constant during the Checkpoint Optimisation step, are now recalibrated to maximise the effectiveness of the synthetic datasets.

The parameters targeted for optimisation are as follows:

- The regeneration rate of the synthetic dataset, which was previously fixed at 10 epochs, is now adjusted to vary between 1 and 10 epochs.

- The scale of the dataset cardinality, initially set at 1, is now varied between 1 and 10.
- The standard deviation for sampling from the multivariate Gaussian distribution, which was originally set to 1.0, is now adjusted within the range of 1.0 to 2.5.

Previous studies have demonstrated that more frequent regeneration of the synthetic dataset and the creation of larger datasets contribute positively to the improvement of CAS. Regarding the standard deviation, this approach contrasts with the Truncation Trick used in the vanilla BigGAN-Deep model. Here, the goal is to encourage more varied generation, even if it comes at the cost of the perceptual quality of the generated images [139].

The Tuning step is conducted using a Tree-structured Parzen Estimator (TPE) coupled with a Hyperband pruning mechanism [163, 164]. The optimisation process runs for 50 iterations, with each iteration involving the training of a Student Classifier using the same procedure as in Checkpoint Optimisation. However, the synthetic data in this phase is generated based on the current parameter configuration, with the objective of maximising CAS.

Upon completion of the search, the optimal parameter configuration is utilised to train the final Student Classifier over 500 epochs.

4.3.6. Membership Inference Attack

The final stage of the KR pipeline is dedicated to evaluating the robustness of the Student Classifier against a Membership Inference Attack (MIA). This type of attack targets the privacy of the model by attempting to identify whether specific data points were part of the model's training dataset. In this study, it is important to note that sensitive training data is never directly exposed to the Student Classifier; rather, it is solely utilised in training the Generator and Teacher Classifier.

The purpose of this step is to assess the effectiveness of MIA in discerning the data used to train the Generator, as inferred through the Student Classifier,

and to compare this effectiveness with that of a similar attack on the Teacher Classifier.

To conduct the MIA, the shadow models technique, as proposed by Shokri et al. [24], is employed. The implementation consists of the following steps:

1. Creation of 10 shadow models, each identical in architecture and training technique to the Teacher Classifier.
2. The validation dataset is used to train each shadow model, with a 45/10/45 split to simulate the training, validation, and test sets. Data is shuffled before splitting to ensure different splits for each shadow model.
3. Generation of the shadow dataset: training and test data are fed into each shadow model. The resulting logits serve as features for the shadow dataset, while a binary label indicates whether a given logit corresponds to a training set instance or an external (test) set instance.
4. The shadow dataset is divided according to the classes of the original dataset used to train the shadow models.
5. Three models—Logistic Regression, Support Vector Classifier with RBF kernel, and Random Forest—are trained for each class in the dataset.
6. For each class, the best-performing model is selected based on the Area Under the Receiver Operating Characteristic Curve (AUC) metric.

The final attack model is composed of the best classifier for each class in the dataset and is applied to both the Teacher Classifier and the Student Classifier.

The resistance to MIAs is evaluated using two primary metrics: the AUC, which is a standard metric for assessing the effectiveness of such attacks, and the Accuracy Over Privacy (AOP). The AOP metric provides an estimate of the trade-off between model performance, measured by test accuracy, and resistance to MIAs, as described in Chapter 2.

The primary objective of this evaluation is to determine whether the proposed training on synthetic data could offer an additional layer of privacy, thereby reducing the effectiveness of the MIA.

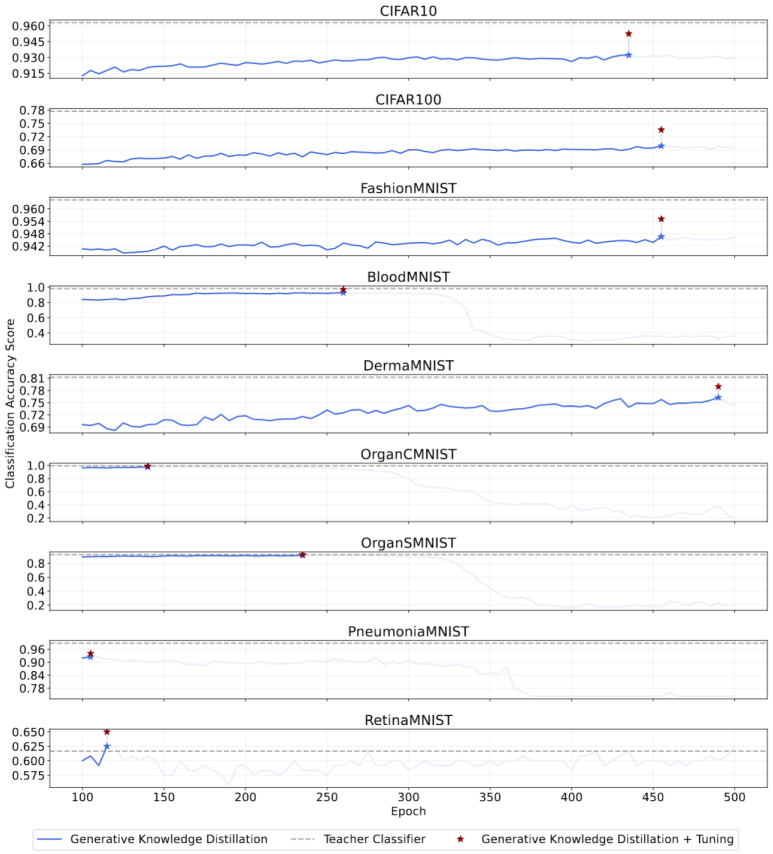

Figure 4.2: The optimal checkpoint's Classification Accuracy Score (CAS) after Generative Knowledge Distillation (GKD) training, using Tuning step parameters, is denoted by a red star. Continuous blue lines represent CAS during Checkpoint Optimisation via GKD, with blue stars marking optimal checkpoints. Dashed grey lines indicate the Teacher Classifier's best validation accuracy. The figure displays Validation CAS across Generator checkpoints for various datasets.

Table 4.1: The Tuning step identified optimal generation parameters for each dataset under consideration. Improvements in validation Classification Accuracy Score (Δ CAS) compared to default generation parameters are presented. The table displays Standard Deviation, Regeneration Rate, and Cardinality Scale for each dataset.

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

4.4. Experiments and Results

The experiments were conducted on nine image datasets, all rescaled to 32x32 pixels. CIFAR10, CIFAR100, and FashionMNIST were employed both for the final comparisons and to calibrate and test the Knowledge Recycling (KR) pipeline, as detailed in D and D. The six medical datasets—BloodMNIST, DermaMNIST, OrganCMNIST, OrganSMNIST, PneumoniaMNIST, and RetinaMNIST—consist of real images from the MedMNIST v2 benchmark [165]. These medical datasets represent the primary domain for applying the proposed technique. The KR pipeline, initially calibrated on the three aforementioned datasets, is subsequently applied to these medical datasets without further specific adaptations. This strategy allows for an evaluation of the technique's effectiveness and robustness in a specialised and complex context distinct from the initial calibration datasets. The experiments were executed on 4 NVIDIA Quadro RTX 6000 GPUs.

Figure 4.2 displays the results from the Checkpoint Optimisation and Tuning

phases, presented as Classification Accuracy Score (CAS) on the respective validation sets, and compared with the optimal accuracy performance of the Teacher Classifier on the same set. The significance of selecting the optimal checkpoint and evaluating it via CAS is evident for Generators with more stable checkpoints (e.g., DermaMNIST, RetinaMNIST) and for those prone to mode collapse and consequent performance drops (e.g., BloodMNIST, OrganSMNIST).

The application of the Generative Knowledge Distillation (GKD) technique alone proves sufficient to achieve results close to those of the Teacher Classifier. Notably, in the case of the RetinaMNIST dataset, a more accurate model is obtained from synthetic data alone. The Tuning step further enhances performance, increasing the validation CAS by a minimum of 0.85% for FashionMNIST and a maximum of 4.03% for BloodMNIST, as shown in Table 4.1.

This improvement stems from two main factors. First, the increased availability of information due to the higher cardinality of the generated datasets and their more frequent regeneration. Second, the enhanced data diversity resulting from sampling with a larger standard deviation, which, when combined with the GKD technique, allows for the utilisation of images that might otherwise be uninformative if associated with hard labels. Such images would likely be filtered out and discarded under alternative synthetic data training methods.

Table 4.2 presents the final comparison between the Teacher Classifier and Student Classifier. The testing CAS of the Student Classifiers approaches, and in some cases exceeds, the testing accuracy of the Teacher Classifiers, particularly in the cases of PneumoniaMNIST and RetinaMNIST.

Regarding resilience to Membership Inference Attacks (MIA), the Student Classifiers demonstrate significantly greater resistance, with attacker performance nearing random guessing levels. The Accuracy Over Privacy (AOP) metric, which evaluates the trade-off between performance and MIA resilience, indicates that Student Classifiers consistently outperform Teacher Classifiers. This suggests that the slight reduction in CAS on the test set is positively offset by the substantial increase in MIA resilience.

Table 4.2: A comprehensive comparison between Teacher and Student Classifiers is presented, focusing on test set performance. Metrics include Accuracy (↑), AUC_{MIA} (↓), and AOP (↑). The table highlights the best scores in **bold**, showcasing improvements in minimum, average, and maximum performance across datasets.

	Accuracy (↑)		AUC_{MIA} (↓)		AOP (↑)	
Model	Teacher Classifier	Student Classifier	Teacher Classifier	Student Classifier	Teacher Classifier	Student Classifier
CIFAR10	**96.24**	95.83	56.22	**51.21**	76.13	**91.34**
CIFAR100	**77.65**	74.92	63.75	**52.31**	47.77	**68.44**
FashionMNIST	**95.94**	95.21	52.69	**50.50**	86.38	**93.34**
BloodMNIST	**97.49**	96.26	51.20	**50.38**	92.96	**94.81**
DermaMNIST	**79.50**	76.46	59.41	**50.43**	56.31	**75.15**
OrganCMNIST	**93.16**	90.23	55.73	**53.07**	74.98	**80.11**
OrganSMNIST	**79.78**	78.76	54.29	**52.38**	67.66	**71.78**
PneumoniaMNIST	86.54	**86.70**	50.39	**50.00**	85.19	**86.70**
RetinaMNIST	54.25	**55.00**	52.36	**50.58**	49.48	**53.74**
Min Imp	-	-3.04	-	-0.39	-	1.51
Avg Imp	-	-1.24	-	-3.90	-	7.72
Max Imp	-	0.75	-	-11.40	-	20.66

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

The average performance gap between Teacher Classifiers and Student Classifiers was reduced to -1.24% in terms of Classification Accuracy Score (CAS) on the test sets, a significant improvement over previous results. This progress is particularly noteworthy considering the use of a single Generator, in contrast to earlier works. Dat et al. achieved an average gap of -10.08% with a single

Generator and -5.81% with six, while the experiments proposed in Chapter 4 achieved -3.87% with a single Generator and -2.63% with six. The approach proposed in this study surpasses these results, indicating potential for further improvement through the use of multiple Generators in parallel.

The inclusion of a metric to empirically assess privacy-related aspects, such as resistance to MIAs, proved essential for a more comprehensive evaluation of the proposed method, especially when dealing with medical images that may be sensitive to privacy violations. Teacher Classifiers, trained with regularisation and augmentation techniques, demonstrated partial resistance to MIAs, confirming the privacy-enhancing properties associated with such techniques. However, Student Classifiers exhibited almost complete resistance to these attacks, underscoring the significant privacy advantage of the proposed approach.

The primary limitations of this study involve the relatively small size of the images used (32x32 pixels) and the selection of models that, while efficient, do not match the performance of the current state-of-the-art models for their respective tasks. These choices were driven by considerations of computational efficiency, given the intensive nature of the KR pipeline. The use of higher-resolution images and more sophisticated models, both for the Classifier (ResNet14) and the Generator (BigGAN-Deep), could lead to further performance enhancements. Specifically, upgrading the Generator model could further narrow the performance gap between the Teacher and Student Classifiers, potentially enabling the Student Classifier to surpass the Teacher in performance.

The scalability of the proposed approach, in terms of both the number of Generators and the cardinality and frequency of data generation, presents promising opportunities for future developments. With ongoing advancements in hardware, it is conceivable that this technique could soon be applied to more complex models and larger datasets, opening new avenues in the fields of private learning and high-quality synthetic data generation.

4.5. Conclusions

This chapter introduced the Knowledge Recycling (KR) pipeline, illustrating the process of generating synthetic data and leveraging it to train downstream classifiers. The application of the Generative Knowledge Distillation (GKD) technique within this pipeline has been shown to enhance the quality of information transferred to downstream classifiers, surpassing the effectiveness of previously proposed methods in the literature. Through this approach, it was possible to significantly narrow the performance gap between models trained on real data and those trained on generated data, thereby setting a new benchmark in the field. Moreover, the synthetic data-driven models exhibited robust privacy characteristics, rendering Membership Inference Attacks largely ineffective. The effectiveness of the KR pipeline was validated on real medical image datasets, demonstrating that it is feasible to maintain high performance while simultaneously reducing the vulnerability to privacy attacks.

5 | Federated Knowledge Recycling

Recent advancements in collaborative model development have positioned federated learning as a crucial paradigm, enabling the construction of robust models without centralising sensitive data. However, conventional federated learning approaches remain vulnerable to privacy and security breaches, as the exposure of models, parameters, or updates may serve as potential attack vectors. This study presents Federated Knowledge Recycling (FedKR), an novel cross-silo federated learning methodology that utilises locally generated synthetic data to enhance inter-institutional collaboration. FedKR integrates advanced data generation techniques with a dynamic aggregation mechanism, thereby significantly reducing the risk of privacy attacks. This approach offers a more secure alternative to existing methods by minimising the attack surface. Empirical evaluations conducted on both generic and medical datasets demonstrate that FedKR not only maintains competitive performance compared with federated learning-based alternatives, but also achieves an average accuracy improvement of 4.24% compared to models trained exclusively on local data. The methodology proves particularly beneficial in scenarios characterised by data scarcity. The implementation of FedKR represents a substantial advancement in the field of federated learning, addressing critical privacy concerns whilst maintaining high performance standards. This research contributes to the ongoing efforts to develop secure and efficient collaborative learning methodologies in data-sensitive domains.

5.1. Introduction

Deep learning has seen remarkable advancements in recent years, with widespread applications across various sectors of society. Among its most promising branches, generative deep learning stands out, capable of producing high-quality synthetic data across diverse types and conditioned forms. Concurrently, federated learning is gaining prominence as a paradigm that fosters collaboration among multiple entities, striking a balance between the growing demand for data to fuel increasingly powerful models and the critical need to protect privacy while providing robust guarantees to users.

The relevance of federated learning is particularly pronounced in the healthcare sector, where it is often challenging to amass datasets of sufficient size within a single institution. In this scenario, the approach is termed cross-silo federated learning, as participating entities typically serve as aggregation centres for end-user data, which, in this context, pertains to individual patient information. The principal objective is to maintain the utmost privacy for such highly sensitive data by employing techniques that enable information sharing without directly exposing the data to the public domain. However, while conventional federated learning techniques address the challenge of direct data exchange, they inadvertently expose other forms of information, such as model weights, gradients, or logits, thereby introducing new potential attack surfaces.

To counter these emerging vulnerabilities, various protection strategies have been proposed, each tailored to mitigate specific security risks. Although these techniques often succeed in preserving privacy, they frequently do so at the cost of the federation's overall performance, resulting in either excessively inefficient models or impractically long computation times.

5.1.1. Main Contributions

In response to these challenges, this chapter introduces Federated Knowledge Recycling (FedKR), a cross-silo federated learning technique designed to address the privacy risks inherent in current federated systems. By leveraging genera-

tive models and restricting data exchanges to synthetic data only, FedKR aims to safeguard the privacy of both the data and models involved in the federation process while ensuring that the performance remains advantageous for all participants.

In the FedKR framework, participation in the federation is contingent upon each member contributing a set of synthetic data corresponding to a specific problem type or data category to be addressed. This contribution serves as an entry requirement to the federation and grants access to the synthetic data generated by other participants, which is stored on a Central Server. This process creates a shared pool of knowledge without compromising the confidentiality of the original, sensitive data.

Furthermore, FedKR introduces a technique known as Dynamic Dataset Aggregation, which allows each participant to locally compile an aggregated dataset by selecting the most effective synthetic datasets. This enables the training of an optimal model tailored to their specific data, all while avoiding the public exposure of private data or models. As a result, FedKR offers enhanced privacy protection and superior properties compared to conventional federated learning methods.

The efficacy of FedKR is demonstrated through simulations on medical datasets of varying types and sizes, highlighting its potential in healthcare applications where privacy is paramount. The approach not only safeguards sensitive data but also facilitates the development of high-performance models that surpass those achievable through non-collaborative efforts.

The principal contributions of this research are as follows:

- The development of Federated Knowledge Recycling (FedKR), an innovative approach to federated learning. This method employs exclusively synthetic data for information exchange amongst participants, thereby substantially enhancing data privacy and markedly reducing vulnerability to privacy breaches in comparison to conventional techniques.
- The conception of the Dynamic Dataset Aggregation strategy within the

FedKR framework. This approach enables the dynamic optimisation of shared synthetic datasets, allowing each federation participant to construct a bespoke training dataset, thus further minimising the potential for privacy infringements.

- The empirical assessment of FedKR in the context of medical imaging. This evaluation demonstrates the methodology's capability to effectively balance the safeguarding of sensitive information with the imperative to develop high-performance models, particularly in scenarios characterised by limited data availability.

5.2. Related Works

Federated learning is a collaborative framework designed to centralise knowledge and enhance the development of models that are more effective than those generated in isolation, while simultaneously safeguarding the privacy of the underlying data utilised during training. Over time, numerous strategies have been proposed to optimise this approach.

The most widely adopted model in federated learning is Federated Averaging (FedAvg). This technique aggregates local model updates, specifically the parameters or gradients, through an averaging process to construct a global model. The workflow involves local training by clients, the transmission of updates to a central server, and the subsequent redistribution of the aggregated model [73]. Building upon FedAvg, FedProx introduces a proximity term to address client heterogeneity, thereby mitigating divergence between local and global models [72]. To counter client drift, SCAFFOLD employs control variables to guide and correct the update flow, enhancing convergence in environments with heterogeneous data [166]. In scenarios characterised by non-IID data, FedNova implements a normalisation mechanism that balances each client's contribution by adjusting local updates according to variations in data quantity and iteration frequency [167].

Recently, the integration of synthetic data into federated learning has garnered

attention as a promising alternative to traditional methods, aimed at bolstering privacy while maintaining efficient learning. FedGAN exemplifies this approach, where each participant locally trains a Generative Adversarial Network (GAN) on its own dataset. The resulting local datasets, which are enriched by a mix of real and generated data, are then employed in a process analogous to FedAvg. This method seeks to enhance privacy by introducing ambiguity into the real data [168]. Variants of this approach, which incorporate differential privacy guarantees into the generators, focus on transmitting only synthetic data to the central server, rather than combining real and synthetic data, for the global model training [169].

Another noteworthy approach is FedMatch, which utilises a semi-supervised technique involving a pseudo-label generator. Clients generate pseudo-labels for unlabelled data locally, which are then utilised in the federated learning process. This method leverages synthetic pseudo-labelled data to boost performance, particularly in scenarios with limited labelled data [170]. Conversely, FedSyn employs a shared generator to create synthetic data that represents the global distribution. Clients use this synthetic data to regularise their local models [171]. SGDE, on the other hand, is a federated system where each user locally trains Variational Autoencoder (VAE)-type generators with differential privacy guarantees using their own private data. These generators are then uploaded to a central server, enabling access to all shared generators for the local generation of synthetic datasets, which are subsequently used for training individual models. For more details, refer to Chapter 3.

5.2.1. Threats and Defences in Federated Learning

While federated learning was originally conceived as a paradigm to maintain privacy while sharing information to develop more effective models, the very act of sharing such information or models inherently expands the attack surface. Various types of attacks have been studied extensively, each targeting different aspects of the federated learning process:

- **Model Inversion**. This attack focuses on reconstructing the training

data by exploiting common model parameters. By meticulously analysing these parameters, an attacker can potentially deduce sensitive details about the original training data [45].

- **Membership Inference**. The objective of this attack is to ascertain whether a specific data sample was part of the model's training set. By observing the model's responses to certain inputs, an attacker can infer, with some probability, the inclusion of particular data in the training process [24].
- **Data Poisoning**. Data poisoning attacks aim to degrade the global model's performance by injecting malicious or tampered data during the local training process. Such attacks can introduce biases or compromise the model's overall accuracy [172].
- **Byzantine**. These attacks target the aggregation phase by sending arbitrary or malicious updates from compromised clients, thereby disrupting the training convergence and diminishing the quality of the final model [173].
- **Sybil**. In a Sybil attack, the adversary manipulates the aggregation process by creating multiple fraudulent identities, thereby amplifying their influence over the global model. This tactic can result in disproportionate control over the training outcomes [174].
- **Evasion**. Evasion attacks are designed to deceive the model during inference by introducing adversarial inputs specifically crafted to trigger misclassifications. Such attacks are particularly effective in circumventing security mechanisms like malware detection, influencing decisions in autonomous systems, or compromising medical diagnosis systems [175].
- **Gradients/Weights Leakage**. Similar to model inversion, this attack attempts to reconstruct the training data by analysing the gradients or weights exchanged during the federated model update process [176].
- **Free-Riding**. In a free-riding attack, some participants exploit the benefits of the global model without contributing meaningful updates. They may send false or low-quality updates, leading to a degradation of the overall system performance [177].

- **Property Inference**. Property inference attacks aim to deduce statistical properties of the training dataset by analysing how the model behaves under various inputs. The goal is to extract aggregate information about the data without directly accessing it [178].
- **Temporal**. Temporal attacks leverage the timing differences in responses or updates to glean information about the hardware or datasets of the participants. This information can be exploited to identify vulnerabilities or specific characteristics of the participating devices [179].

The evolving landscape of these threats has spurred the development of sophisticated countermeasures, which, although effective, often come with trade-offs in performance.

Differential privacy is a prominent defence against membership inference and model inversion attacks. By adding carefully calibrated noise to data or model parameters, it effectively protects individual information. However, this added noise can also reduce the accuracy of the model, with the extent of the trade-off being proportional to the desired privacy level [25, 53]. Alternative strategies, such as regularisation or adversarial training, have been explored to mitigate the same attacks, aiming to achieve a more favourable balance between accuracy and resilience to membership inference and model inversion. Fore more details, refer to Chapter 2.

Homomorphic encryption offers robust protection against data leakage during computation by allowing operations on encrypted data. While this method provides strong security guarantees, it introduces considerable computational overhead, which can significantly slow down the training process [180].

Secure aggregation techniques safeguard privacy during the model aggregation phase, effectively countering gradients/weights leakage attacks. These techniques strive to maintain a balance between security and efficiency, although they may introduce delays in the global model update process [181].

Techniques such as model pruning and quantisation reduce the attack surface for model inversion by limiting the amount of information shared. However, these

methods can compromise model performance, especially in complex tasks, due to the reduction in model capacity [182].

These advanced countermeasures highlight the ongoing effort to enhance the security of federated learning systems while carefully managing the inherent trade-offs in performance and efficiency.

5.3. Method

This section introduces Federated Knowledge Recycling (FedKR), a novel approach in federated learning that aims to enhance privacy guarantees through the utilisation of generated data. FedKR is specifically applied and evaluated within a federated environment involving the sharing of medical images for classification tasks.

The core principle of FedKR revolves around the exclusive sharing and aggregation of synthetic data amongst federation participants. To engage in the federation, each member must generate a synthetic dataset using a locally trained generator based on their private data. This synthetic dataset is subsequently uploaded to a central server, which functions as a shared repository and distribution hub for all federation members.

By contributing with their synthetic dataset, participants gain reciprocal access to synthetic datasets provided by other federation members. These datasets can be downloaded and utilised locally, enabling each participant to construct a bespoke aggregated synthetic dataset tailored to their specific requirements. Participants can then independently and locally train their models using this aggregated dataset, thereby preserving the autonomy and privacy of their data.

This chapter also presents Dynamic Dataset Aggregation, a technique designed to manage and optimise synthetic datasets within the FedKR framework. This approach allows for the dynamic modulation of each dataset's contribution to the final aggregated dataset. By adjusting the weight assigned to each synthetic dataset, it becomes possible to optimise model performance, prioritising the most relevant datasets whilst excluding less pertinent ones. This adaptive

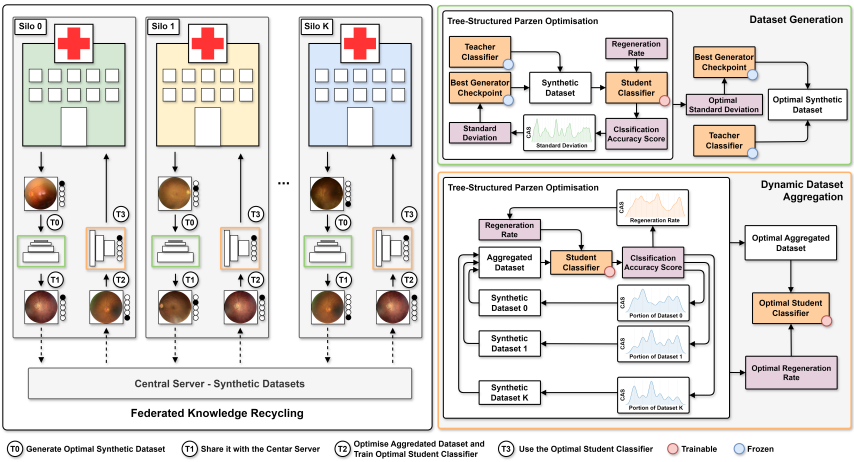

Figure 5.1: Graphical representation of the Federated Knowledge Recycling technique.

approach enhances the overall efficacy of the model training. Figure 5.1 provides a graphical representation of this process.

5.3.1. Knowledge Recycling

To facilitate the exchange of synthetic data that holds significant utility for deep learning tasks amongst federation members, specific generation and post-processing methods are crucial. The literature increasingly explores various techniques and strategies for generating synthetic datasets capable of effectively replacing real ones [136].

In this study, each federation member employs the Knowledge Recycling pipeline, a comprehensive method for generating synthetic datasets and subsequently utilising them for classifier training. This pipeline is specifically designed to maximise the utility of synthetic data and to develop models that are highly resilient to membership inference attacks. For further details, refer to Chapter 4. In line

with the reference work, the process involves using private data solely to train a Generator and a Teacher Classifier. The ultimate objective is to produce the most effective Student Classifier, a model structurally identical to the Teacher Classifier but trained exclusively on synthetic data generated by the Generator. For this purpose, a modified version of the BigGAN-Deep model serves as the Generator, whilst both the Teacher and Student Classifiers utilise a ResNet14 architecture. Detailed implementation and training procedures can be found in Chapter 4.

Throughout the entire process, from the selection of the optimal checkpoint to the final dataset generation, the Generative Knowledge Distillation technique, as elucidated in Chapter 4, is employed. In this technique, the Teacher Classifier evaluates the synthetic images and assigns soft labels, which constitute probability distributions over classes rather than binary labels. These soft labels encapsulate subtle nuances and uncertainties, thereby enriching the synthetic dataset with more profound knowledge.

The process commences with the identification of the optimal checkpoint of the Generator, utilising the Classification Accuracy Score (CAS) metric, which quantifies the accuracy of a classifier trained on generated data when evaluated against real data [116]. Subsequently, a 50-step tuning phase is conducted to determine the optimal standard deviation for data generation, with the objective of maximising the CAS of the Student Classifier. This optimisation employs a Tree-structured Parzen Estimator, in conjunction with a Hyperband pruning mechanism [163, 164], with permissible values for the standard deviation ranging from 0.5 to 2.5. Upon finalisation of these parameters, the Generator is utilised to create the definitive synthetic dataset, generating images that closely approximate the real data distribution. Ultimately, a synthetic dataset five times larger than the original real dataset is produced and uploaded to the Central Server for utilisation by all federation members.

5.3.2. Dynamic Datasets Aggregation

The final stage of the FedKR technique, termed Dynamic Datasets Aggregation (DDA), involves leveraging the synthetic datasets available across the entire federation for local model training. This process commences with the local download of the selected synthetic datasets, followed by the optimisation of their aggregation through a 50-step procedure. This optimisation utilises a Tree-structured Parzen Estimator in conjunction with a Hyperband pruning mechanism, mirroring the approach employed in earlier stages.

During this optimisation phase, two critical parameters undergo fine-tuning. The first parameter pertains to the percentage contribution of each synthetic dataset relative to its original size. This allows for precise modulation of each dataset's contribution, ranging from 0% to 100% of its data, with adjustments made in increments of 20%. The second parameter concerns the regeneration rate of the aggregated dataset during the classifier's training process.

The regeneration rate determines the frequency at which the aggregated dataset is regenerated during training. This can range from regenerating the dataset at every epoch to regenerating it only once at the commencement of training. If the rate is set to regenerate only at the outset, the entire aggregated dataset is utilised in one iteration. Conversely, if multiple regenerations are specified, the dataset is divided into equivalent portions, thus fully exploiting the abundance of synthetic data. Both parameters are optimised with the objective of maximising the local CAS.

The classifier employed in this training process is the same ResNet14 model used in previous steps. Regardless of the chosen regeneration rate, the training process is conducted over 100 epochs. This method ensures that the optimisation of synthetic data aggregation and usage contributes to the creation of a robust and high-performing model.

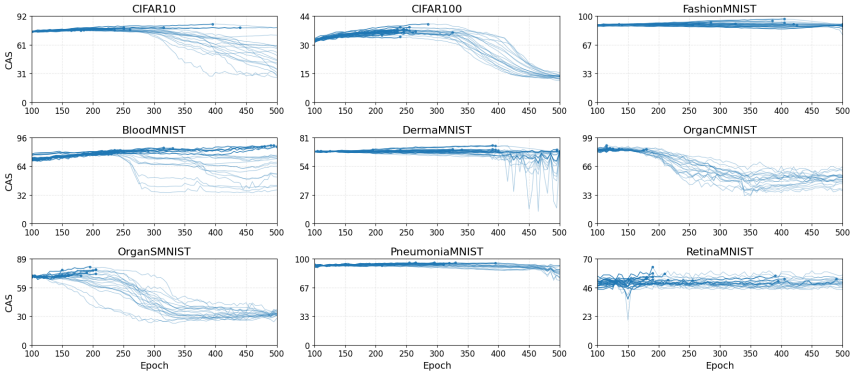

Figure 5.2: The performance evaluation of federation members in successfully identifying and selecting the Generator's most effective checkpoint for network-wide distribution.

5.4. Experiments and Results

This section presents the experimental findings and discusses the principal performance and privacy aspects of different information-sharing protocols. The objective is to evaluate how various federated methods influence classification accuracy and the potential for privacy breaches. The following four approaches are examined:

1. Ordinary Training, in which each participant trains a local model on private data and shares the trained model with the federation;
2. Centralised Training, a scenario where all participants contribute their private data to build a single global model (generally unfeasible under strict privacy requirements);
3. FedAvg, where participants periodically communicate model updates (weights or gradients) to a central server, which aggregates them into a global model;
4. FedKR (Federated Knowledge Recycling), the proposed technique, in which

only synthetic datasets are exchanged, accompanied by a Dynamic Dataset Aggregation (DDA) step. In FedKR, once local models are trained, they are not shared.

Nine image datasets, each rescaled to 32×32 pixels, were used in the experiments. The first three (CIFAR10, CIFAR100, and FashionMNIST) are generic datasets employed both to calibrate the FedKR approach and for final comparisons. The subsequent six (BloodMNIST, DermaMNIST, OrganCMNIST, OrganSMNIST, PneumoniaMNIST, and RetinaMNIST) are medical datasets with real images taken from the MedMNIST v2 benchmark, which constitute the principal area of interest once the technique has been calibrated. Full descriptions of all datasets can be found in Appendix A.

A cross-silo setting was simulated with 20 federation members, dividing the training, validation, and test partitions of each dataset into 20 equal parts without class stratification. The experiments were performed on four NVIDIA Quadro RTX 6000 GPUs.

Under FedKR, each participant trains a local Generator, selecting the best checkpoint and standard deviation from its own validation set. A synthetic dataset is then produced and shared with the Central Server. Figure 5.2 shows the Classification Accuracy on each member's validation data, highlighting consistently high performance and the variability caused by the non-stratified allocation. Once the synthetic dataset is created, no local models or parameters are transferred: every member solely uploads the synthetic data to the Central Server and downloads the synthetic datasets provided by all other participants. A Dynamic Dataset Aggregation phase follows, during which each member augments its local training with the newly collected synthetic data.

The proposed FedKR is compared with FedAvg (implemented without additional privacy techniques), Ordinary Training (a lower bound where participants do not effectively collaborate), and Centralised Training (an upper bound obtained by sharing all private data). In all cases, ResNet14 classifiers were trained for 100 epochs using the same methodology (further details are provided in Appendix D and in Chapter 4).

Table 5.1 reports the test accuracy results, indicating that participating in the federation using FedKR yields an average improvement of 4.24% compared with Ordinary Training, varying from 0.72% on BloodMNIST to 10.33% on PneumoniaMNIST. This shows the combined effect of synthetic data generation and DDA. Although FedAvg achieves a larger mean improvement (10.10%) over Ordinary Training, it involves exchanging gradients or weights at each communication round, which creates considerable privacy and security risks. Centralised Training generally provides the highest performance but is impractical due to confidentiality constraints. In the PneumoniaMNIST dataset, each member has only 235 training samples, yet FedKR surpasses all other techniques, including Centralised Training, demonstrating that synthetic data can be highly advantageous when data are scarce, as in certain medical contexts.

A further observation is that FedKR typically shows a lower standard deviation in performance across the federation compared to FedAvg. Thanks to DDA, synthetic datasets can be adapted to each participant's needs, leading to a more equitable distribution of benefits. By contrast, FedAvg, focusing on optimising the global average performance, may yield significant disparities between different nodes.

Table 5.2 compares the privacy-attack resilience of each approach, focusing on the risks arising when sharing entire trained models (Ordinary Training), full private datasets (Centralised Training), aggregated model parameters (FedAvg), or only synthetic data (FedKR). No additional defences such as differential privacy or homomorphic encryption are considered here. A thorough formal analysis of why certain attacks are mitigated or inapplicable can be found in Appendix E.

Ordinary Training, which exposes locally trained models, and Centralised Training, which shares the entire dataset, both carry significant vulnerabilities. Ordinary Training is susceptible to attacks such as Model Inversion and Membership Inference. Centralised Training entails public exposure of all private data, making the risk even greater. FedAvg aggregates participants' gradient or weight updates into a global model, which can be inverted or examined by adversaries to infer private information or determine membership. This single global model

Table 5.1: Comparison of test accuracy obtained by different approaches. Accuracy for FedKR is evaluated as a Classification Accuracy Score.

Model	Ordinary Training	Centralised Training	FedAvg	FedKR (ours)
CIFAR10	72.20 ± 3.06	94.95 ± 0.52	86.95 ± 0.55	80.16 ± 0.81
CIFAR100	27.71 ± 3.19	75.70 ± 1.07	55.75 ± 2.16	36.60 ± 1.26
FashionMNIST	89.24 ± 2.38	95.54 ± 0.49	93.43 ± 1.30	91.68 ± 0.71
BloodMNIST	90.19 ± 4.26	97.93 ± 0.47	95.79 ± 0.77	90.91 ± 1.03
DermaMNIST	67.92 ± 5.26	79.40 ± 1.98	73.46 ± 2.87	70.47 ± 2.42
OrganCMNIST	80.68 ± 3.63	93.29 ± 0.59	88.95 ± 0.80	83.19 ± 0.62
OrganSMNIST	69.45 ± 3.31	82.64 ± 0.71	78.05 ± 0.88	70.19 ± 0.99
PneumoniaMNIST	79.25 ± 9.30	83.65 ± 4.09	88.72 ± 6.04	89.59 ± 2.47
RetinaMNIST	49.00 ± 15.45	53.75 ± 4.55	55.99 ± 11.47	50.99 ± 6.64
Min Imp	-	4.39	4.19	0.72
Mean Imp	-	14.50	10.10	4.24
Max Imp	-	47.70	28.04	10.33

is also more vulnerable to Data Poisoning, Byzantine, and Sybil attacks.

By contrast, FedKR only shares synthetic datasets, never revealing local models or parameters. As explained in Appendix E, Model Inversion and Membership Inference attacks are inapplicable in FedKR (Theorems E.1 and E.2), since any recovered details would refer solely to data that do not match real samples. Furthermore, Gradient and Weights Leakage attacks are neutralised (Theorem E.6) by the absence of parameter exchange. The Dynamic Dataset Aggregation process provides a mechanism for limiting Data Poisoning, Byzantine, and Sybil threats (Theorem E.3), as any suspicious or low-quality synthetic dataset can be down-weighted. Evasion attacks are also hindered by the restricted attack surface (Theorem E.4). Moreover, Free-Riding and Property Inference become difficult due to the lack of direct access to real data or local models (Theorems E.7 and E.8). Finally, no frequent model synchronisations occur in FedKR, rendering Temporal attacks inapplicable (Theorem E.9).

Table 5.2: The privacy attack resistance properties of the examined approaches compared to the considered attacks. The red dot (●) indicates vulnerability, the yellow dot (●) indicates partial mitigation, the green dot (●) indicates resistance, and the dash (-) indicates that the attack is not applicable. Each rating reflects the ability of an attack to compromise the privacy of real private data using that attack technique.

Attack	Ordinary Training	Centralised Training	FedAvg	FedKR (ours)
Membership Inference	●	●	●	-
Gradients/Weights Leakage	-	●	●	-
Property Inference	●	●	●	●
Model Inversion	●	●	●	-
Temporal	-	●	●	-
Data Poisoning	-	●	●	●
Byzantine	-	-	●	●
Sybil	-	-	●	●
Free-Riding	-	-	●	●
Evasion	●	●	●	-

5.4.1. Discussion and Limitations

FedKR is introduced as a federated learning strategy based on synthetic data sharing, rather than the more common practice of transmitting trained local or global models. The experiments show an average accuracy increase of 4.24% over Ordinary Training. While FedAvg achieves a greater overall accuracy improvement, it necessitates exchanging parameters or gradients, which widens the attack surface. FedKR, by contrast, restricts the information exchanged to synthetic datasets, rendering many standard threats, such as Model Inversion, Membership Inference, and parameter leakage, either entirely inapplicable or significantly less effective. This feature is particularly relevant for sensitive domains like medical imaging.

Progress in generative deep learning may enable FedKR to narrow the performance gap with FedAvg if more sophisticated Generators are employed. Moreover, FedKR typically reduces communication costs by confining data transfers to synthetic datasets, in contrast to FedAvg's iterative sharing of updates.

FedKR can be integrated with additional privacy-preserving mechanisms, such as homomorphic encryption or differential privacy, although computational overhead and performance must be considered. While training local Generators can be demanding, ongoing advances in generative architectures and cloud-based computational services mitigate these challenges.

In conclusion, FedKR represents an appealing alternative for cross-silo federated learning scenarios with strict privacy requirements and uneven or limited data availability. Its application to medical imaging underscores the ability to safeguard private information without compromising adaptability or overall accuracy, paving the way for a more robust and flexible federated framework.

5.5. Conclusions

This chapter presents the Federated Knowledge Recycling (FedKR) technique, a novel approach to federated learning specifically designed for cross-silo environments. Unlike traditional methods, FedKR exclusively relies on the exchange of synthetic data among participants, thereby circumventing the need to share potentially sensitive models, parameters, or metadata. Experimental results show that FedKR offers a favourable trade-off between performance and privacy, with an average accuracy improvement of 4.24% compared to local training, proving particularly effective in data-poor scenarios such as the medical domain. Due to the synthetic nature of the data exchanged, FedKR proved robust against attacks such as Model Inversion, Membership Inference and Gradients or Parameters Leakage, rendering them ineffective. In addition, the system provides significant mitigation against several other types of privacy attacks common to traditional federated learning systems. The Dynamic Datasets Aggregation technique on which FedKR is based has proven effective in optimising the use of shared synthetic data, contributing to the overall performance of the system and providing an additional layer of protection. This methodology allows each participant to benefit from the collective knowledge of the federation while maintaining a high standard of privacy.

Conclusions and Future Directions

The research presented in this thesis was situated at the crossroads of deep learning and data privacy, an area that had emerged as a critical focal point in the evolving landscape of artificial intelligence and data science. As data generation and collection proliferated, accompanied by increasingly sophisticated learning paradigms, unprecedented opportunities for innovation and insight emerged. At the same time, this development has raised significant concerns about privacy and the ethical use of personal data. This research addressed the fundamental trade-off between data utility and privacy in deep learning networks, a challenge that lies at the intersection of competing interests in the field of privacy preserving deep learning. The research sought to reconcile the seemingly contradictory goals of extracting valuable insights from data while protecting individual privacy. This multifaceted challenge remains beyond purely technical considerations and include legal, ethical and societal dimensions.

The motivation for this research stemmed from a confluence of factors, including the implementation of stringent data protection regulations, heightened public awareness of privacy concerns, and high-profile data breaches that underscored the need for robust privacy protections. In this context, the thesis explored the potential of synthetic data generation, federated learning and advanced privacy preserving techniques as a means of bridging the gap between these competing objectives. The research built on a foundation of existing work in differential privacy, federated learning, and generative models. By focusing on the gen-

eration and use of synthetic data, this work sought to fundamentally change the approach to privacy in deep learning, moving from a privacy paradigm to a data synthesis paradigm. This shift in perspective opened up new possibilities for privacy preserving deep learning, potentially facilitating the development of high-performance models without requiring access to sensitive real and private data. At the same time, it raised new questions about the nature of data, the fidelity of synthetic representations, and the ethical implications of creating artificial datasets.

The study began with an examination of existing privacy preservation techniques in deep learning. This initial investigation revealed significant limitations in conventional methods, such as Differential Privacy, particularly in their practical applications. These findings highlighted the need for innovative approaches that could better reconcile the demands of model efficacy and data protection. In response to these challenges, the research presented a series of contributions aimed at advancing the field of privacy-preserving deep learning. The work progressed from exploring regularisation techniques to enhance privacy safeguards whilst maintaining model performance, to developing more sophisticated methodologies. A key development in this progression was the introduction of Discriminative Adversarial Privacy (DAP). This strategy utilised adversarial training to optimise both task performance and privacy protection concurrently. DAP demonstrated improvements over traditional methods, offering a more balanced approach to model accuracy and privacy assurance. As the research deepened, attention turned to federated learning as a framework for collaborative model development. Recognising the vulnerabilities in conventional approaches, the study proposed the Synthetic Generative Data Exchange (SGDE) method. This technique employed generative models to produce synthetic data for exchange within a federated learning context, enhancing privacy protections whilst maintaining model performance. The concept of synthetic data emerged as a central theme, leading to the development of the Gap Filler (GaFi) pipeline. This methodology aimed to narrow the performance gap between models trained on synthetic versus real-world data across various domains. The success of GaFi raised questions about the nature of generative modelling and the potential for

replicating real data distributions. Further exploration in synthetic data generation led to the adaptation of advanced models such as Stable Diffusion 2.0. Through a pipeline incorporating transfer learning, fine-tuning, and generation parameter optimisation, the study demonstrated that models trained on synthetic data could, in some cases, perform comparably to those trained on real data. The research culminated in the development of the Knowledge Recycling (KR) pipeline, which incorporated Generative Knowledge Distillation (GKD) to enhance the quality of information transferred to downstream classifiers. Models trained using the KR pipeline exhibited robust privacy characteristics, showing resilience to Membership Inference Attacks. The final contribution of this thesis was the introduction of Federated Knowledge Recycling (FedKR), an approach to federated learning designed for cross-silo environments. FedKR relied exclusively on synthetic data exchange amongst participants, eliminating the need to share sensitive models, parameters, or metadata. Empirical evidence indicated that FedKR achieved a balance between performance and privacy, showing particular efficacy in data-scarce scenarios.

In reflecting upon the body of work presented in this thesis, several key considerations emerged. The research demonstrated the feasibility of developing artificial intelligence-based systems entirely founded on synthetic data that were both highly performant and respectful of individual privacy. The proposed approaches offered potential solutions to some of the most pressing challenges in balancing data utility and privacy in machine learning applications. Subsequent research undertaken during the publication period of this thesis further confirmed these insights, indicating that even in multimodal contexts conditioned on text and graphs, there exists a substantial margin for the positive impact of synthetic data [183, 184].

This approach can prove exceedingly beneficial in enhancing the quality of existing systems, such as those in industrial or healthcare sectors, by enabling the sharing of vast quantities of information whilst preserving privacy and adhering to current regulations. The potential advantages extend to numerous entities that produce and exploit data to build AI models, including companies, industries, clinics, hospitals, and universities. The ability to securely share synthetic

data creates a dual opportunity: it democratizes access to valuable training resources for smaller organizations that would otherwise face barriers to data availability, and it simultaneously safeguards the privacy of individuals without compromising the construction of efficient and accurate models. This dynamic fosters a virtuous cycle in which privacy and collaboration coexist, laying the groundwork for more open and technologically advanced societies.

The impact of this strategy could be even more pronounced in contexts where certain levels of specialisation or industrialisation are entirely absent, particularly regarding healthcare. Moreover, it could be especially impactful in regions of the world where models trained on synthetic surrogates of existing data could be made available, facilitating a leap from current conditions to the technological frontier. In a geopolitical context such as Europe, where barriers persist among member states, a synthetic data-based sharing system could possess immeasurable economic value. Even when considering a single country, such as Italy, the information held within small- and medium-sized enterprises and clinics represents a resource whose potential remains difficult to quantify. Harnessing synthetic data in full compliance with regulations like the GDPR enables the preservation of individual privacy whilst still granting access to comprehensive aggregated information on populations.

Admittedly, this is not an immediate process, and the work conducted herein aimed to demonstrate the feasibility of such a vision. Thus, whilst a possible direction has been shown, the destination has not yet been reached. At the same time, many open challenges persist. Although current evidence suggests that attacks across federated AI models are relatively ineffective, more sophisticated offensive techniques may emerge. Defining protocols that ensure the safe, traceable, and standardized use of synthetic data at scale remains crucial. These considerations become even more compelling in scenarios where generative models create data to train other generative models, potentially leading to multiple "generations" of synthetic data. The evolution of these data, and the possibility of increasingly autonomous forms of learning, presents both significant technical opportunities and profound ethical questions.

Looking towards the future, this research opened up several promising avenues for further exploration. The potential of synthetic data and generative models in preserving privacy whilst advancing AI capabilities warranted continued investigation. Future work could focus on refining the proposed techniques to further narrow the performance gap between synthetic and real data, exploring the application of these methods in domains beyond those studied in this thesis, and examining the scalability of the proposed approaches to larger and more complex artificial intelligence-based systems. Additionally, addressing remaining challenges in privacy preservation, particularly in the context of evolving privacy regulations and emerging threats, remains a critical area for further research.

The ethical and philosophical implications of synthetic data generation and its potential impact on AI bias, fairness, and transparency also present fertile ground for continued study. As AI systems become increasingly integrated into various aspects of society, ensuring that these technologies remain aligned with human values and respect individual privacy will be of paramount importance. By fostering a culture of collaboration and innovation built upon robust privacy safeguards, the full potential of synthetic data may be realized, ultimately benefiting a diverse array of stakeholders in the broader AI ecosystem.

In conclusion, this research is intended as a contribution towards the ongoing development of technology that is decidedly more exploitable by society and individuals to improve the quality and duration of life whilst preserving privacy. Furthermore, it aims to be situated within a larger context that views these artificial intelligence tools as catalysts for continuously evolving ideas. This evolution must be constantly aligned with the positive ideals that our society has produced and will be able to produce with foresight, aiming at the prosperity of the human species. Whilst substantial progress was made, the field of privacy-preserving deep learning remained ripe for further innovation and discovery, with the potential to profoundly impact the way AI systems are developed and deployed in the years to come.

Bibliography

[1] Eugenio Lomurno and Matteo 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*, pages 223–238. Springer, 2022.

[2] Eugenio Lomurno, Alberto Archetti, Lorenzo Cazzella, Stefano Samele, Leonardo Di Perna, and Matteo Matteucci. Sgde: Secure generative data exchange for cross-silo federated learning. In *Proceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition*, pages 205–214, 2022.

[3] Andrea Lampis, Eugenio Lomurno, and Matteo Matteucci. Bridging the gap: Enhancing the utility of synthetic data via post-processing techniques. *British Machine Vision Conference*, 2023.

[4] Eugenio Lomurno, Alberto Archetti, Francesca Ausonio, and Matteo Matteucci. Discriminative adversarial privacy: Balancing accuracy and membership privacy in neural networks. *British Machine Vision Conference*, 2023.

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

[6] Eugenio Lomurno and Matteo Matteucci. Synthetic image learning: Pre-

serving performance and preventing membership inference attacks. *Pattern Recognition Letters*, 2025.

[7] Eugenio Lomurno and Matteo Matteucci. Federated knowledge recycling: Privacy-preserving synthetic data sharing. *Pattern Recognition Letters*, 191:124–130, 2025.

[8] 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.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[11] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851, 2020.

[13] 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.

[14] Ben Shneiderman. *Human-centered AI*. Oxford University Press, 2022.

[15] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a

survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[17] Maranke Wieringa. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 1–18, 2020.

[18] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

[19] Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679:2016, 2016.

[20] Tambiama Madiega. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 2021.

[21] Huw Roberts, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. The chinese approach to artificial intelligence: An analysis of policy and regulation. *Available at SSRN 3469783*, 2019.

[22] Gleb Papyshev and Masaru Yarime. The limitation of ethics-based approaches to regulating artificial intelligence: regulatory gifting in the context of russia. *AI & SOCIETY*, 39(3):1381–1396, 2024.

[23] Jan Philipp Albrecht. How the gdpr will change the world. *Eur. Data Prot. L. Rev.*, 2:287, 2016.

[24] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[25] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya

Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[26] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 32, 2019.

[27] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 2015.

[28] Liwei Song, Reza Shokri, and Prateek Mittal. Membership inference attacks against adversarially robust deep learning models. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE, 2019.

[29] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 2021.

[30] Mohammad Al-Rubaie and J Morris Chang. Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy*, 17(2):49–58, 2019.

[31] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

[32] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Network and Distributed Systems Security (NDSS) Symposium 2019*, 2019.

[33] Yufei Chen, Chao Shen, Yun Shen, Cong Wang, and Yang Zhang. Amplifying membership exposure via data poisoning. *arXiv preprint arXiv:2211.00463*, 2022.

[34] Xinlei He, Hongbin Liu, Neil Zhenqiang Gong, and Yang Zhang. Semi-leak: Membership inference attacks against semi-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 365–381. Springer, 2022.

[35] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. This person (probably) exists. identity membership attacks against gan generated faces. *arXiv preprint arXiv:2107.06018*, 2021.

[36] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? *arXiv preprint arXiv:2302.01316*, 2023.

[37] Hailong Hu and Jun Pang. Membership inference of diffusion models. *arXiv preprint arXiv:2301.09956*, 2023.

[38] Zihan Wang, Na Huang, Fei Sun, Pengjie Ren, Zhumin Chen, Hengliang Luo, Maarten de Rijke, and Zhaochun Ren. Debiasing learning for membership inference attacks against recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1959–1968, 2022.

[39] Wei Yuan, Chaoqun Yang, Quoc Viet Hung Nguyen, Lizhen Cui, Tieke He, and Hongzhi Yin. Interaction-level membership inference attack against federated recommender systems. *arXiv preprint arXiv:2301.10964*, 2023.

[40] Tomas Chobola, Dmitrii Usynin, and Georgios Kaissis. Membership inference attacks against semantic segmentation models. *arXiv preprint arXiv:2212.01082*, 2022.

[41] Guangsheng Zhang, Bo Liu, Tianqing Zhu, Ming Ding, and Wanlei Zhou. Label-only membership inference attacks and defenses in semantic segmentation models. *IEEE Transactions on Dependable and Secure Computing*, 2022.

[42] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Mem-

bership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*, 2022.

[43] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. Practical blind membership inference attack via differential comparisons. In *ISOC Network and Distributed System Security Symposium (NDSS)*, 2021.

[44] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 17–32, 2014.

[45] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[46] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 225–240, 2019.

[47] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 253–261, 2020.

[48] Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Lim. Exploiting explanations for model inversion attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 682–692, 2021.

[49] Jia Qi Lim and Chee Seng Chan. From gradient leakage to adversarial attacks in federated learning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3602–3606. IEEE, 2021.

[50] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael

Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

[51] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16337–16346, 2021.

[52] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[53] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[54] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

[55] Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

[56] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[57] Prateek Jain, Vivek Kulkarni, Abhradeep Thakurta, and Oliver Williams. To drop or not to drop: Robustness, consistency and differential privacy properties of dropout. *arXiv preprint arXiv:1503.02031*, 2015.

[58] Beyza Ermis and Ali Taylan Cemgil. Differentially private dropout. *arXiv preprint arXiv:1712.01665*, 2017.

[59] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 634–646, 2018.

[60] Ziqi Yang, Bin Shao, Bohan Xuan, Ee-Chien Chang, and Fan Zhang. Defending model inversion and membership inference attacks via prediction purification. *arXiv preprint arXiv:2005.03915*, 2020.

[61] Dingfan Chen, Ning Yu, and Mario Fritz. Relaxloss: defending membership inference attacks without losing utility. *arXiv preprint arXiv:2207.05801*, 2022.

[62] Yigitcan Kaya and Tudor Dumitras. When does data augmentation help with membership inference attacks? In *International conference on machine learning*, pages 5345–5355. PMLR, 2021.

[63] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Toronto University press*, 2009.

[64] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[65] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[66] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[67] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[68] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classifi-

cation over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[69] 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.

[70] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

[71] Max Parasol. The impact of china's 2016 cyber security law on foreign technology firms, and on china's big data and smart city dreams. *Computer law & security review*, 34(1):67–98, 2018.

[72] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[73] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[74] Lingjuan Lyu, Han Yu, and Qiang Yang. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*, 2020.

[75] Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70: 1142–1154, 2022.

[76] Sebastian Caldas, Jakub Konečny, H Brendan McMahan, and Ameet Talwalkar. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018.

[77] Chengliang Zhang, Suyi Li, Junzhe Xia, Wei Wang, Feng Yan, and Yang Liu. {BatchCrypt}: Efficient homomorphic encryption for {Cross-Silo} federated learning. In *2020 USENIX annual technical conference (USENIX ATC 20)*, pages 493–506, 2020.

[78] Nader Bouacida and Prasant Mohapatra. Vulnerabilities in federated learning. *IEEE Access*, 9:63229–63249, 2021.

[79] Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*, 2018.

[80] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14 (1–2):1–210, 2021.

[81] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[82] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31 (9):3400–3413, 2019.

[83] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE communications surveys & tutorials*, 22(3):2031–2063, 2020.

[84] Zhaohui Yang, Mingzhe Chen, Walid Saad, Choong Seon Hong, and Mohammad Shikh-Bahaei. Energy efficient federated learning over wireless communication networks. *IEEE Transactions on Wireless Communications*, 20(3):1935–1949, 2020.

[85] Basak Guler and Aylin Yener. Sustainable federated learning. *arXiv preprint arXiv:2102.11274*, 2021.

[86] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International conference on machine learning*, pages 4615–4625. PMLR, 2019.

[87] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

[88] Qiong Wu, Kaiwen He, and Xu Chen. Personalized federated learning for intelligent iot applications: A cloud-edge based framework. *IEEE Open Journal of the Computer Society*, 1:35–44, 2020.

[89] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[90] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

[91] Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

[92] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7865–7873, 2021.

[93] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[94] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy.

Journal of the Royal Statistical Society: Series B (Statistical Methodology), 84(1):3–37, 2022.

[95] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[96] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2022.

[97] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[98] Qingrong Chen, Chong Xiang, Minhui Xue, Bo Li, Nikita Borisov, Dali Kaarfar, and Haojin Zhu. Differentially private data generative models. *arXiv preprint arXiv:1812.02274*, 2018.

[99] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[100] Vaikkunth Mugunthan, Vignesh Gokul, Lalana Kagal, and Shlomo Dubnov. Dpd-infogan: Differentially private distributed infogan. In *Proceedings of the 1st Workshop on Machine Learning and Systems*, pages 1–6, 2021.

[101] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[102] Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

[103] Uthaipon Tao Tantipongpipat, Chris Waites, Digvijay Boob, Amaresh Ankit Siva, and Rachel Cummings. Differentially private

synthetic mixed-type data generation for unsupervised learning. *Intelligent Decision Technologies*, 15(4):779–807, 2021.

[104] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019.

[105] Gergely Acs, Luca Melis, Claude Castelluccia, and Emiliano De Cristofaro. Differentially private mixture of generative neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1109–1121, 2018.

[106] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019.

[107] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. IEEE, 2019.

[108] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345, 2017.

[109] Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang. Membership inference attack against differentially private deep learning model. *Trans. Data Priv.*, 11(1):61–79, 2018.

[110] Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

[111] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. SPIE, 1993.

[112] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and

José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.

[113] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[114] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *2017 International Conference on Learning Representations (ICLR)*, 2017.

[115] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[116] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019.

[117] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*. IEEE/CVF, 2022.

[118] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[119] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[120] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[121] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[122] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21330–21340, 2022.

[123] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8300–8311, 2021.

[124] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3752–3761, 2018.

[125] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick Van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. *arXiv preprint arXiv:1504.06852*, 2015.

[126] Yo-whan Kim, Samarth Mishra, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogerio Feris. How transferable are video representations based on synthetic data? *Advances in Neural Information Processing Systems*, 35:35710–35723, 2022.

[127] Deqing Sun, Daniel Vlasic, Charles Herrmann, Varun Jampani, Michael Krainin, Huiwen Chang, Ramin Zabih, William T Freeman, and Ce Liu. Autoflow: Learning a better training set for optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10093–10102, 2021.

[128] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017.

[129] Jianxin Ma, Shuai Bai, and Chang Zhou. Pretrained diffusion models for unified human motion synthesis. *arXiv preprint arXiv:2212.02837*, 2022.

[130] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011.

[131] Xi Guo, Wei Wu, Dongliang Wang, Jing Su, Haisheng Su, Weihao Gan, Jian Huang, and Qin Yang. Learning video representations of human motion from synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20197–20207, 2022.

[132] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my gan? In *Proceedings of the European conference on computer vision (ECCV)*, pages 213–229, 2018.

[133] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2020.

[134] Eugenio Lomurno, Samuele Mariani, Matteo Monti, and Matteo Matteucci. Pomonag: Pareto-optimal many-objective neural architecture generator. *arXiv preprint arXiv:2409.20447*, 2024.

[135] Sohyun An, Hayeon Lee, Jaehyeong Jo, Seanie Lee, and Sung Ju Hwang. Diffusionnag: Predictor-guided neural architecture generation with diffu-

sion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[136] Pham Thanh Dat, Anuvabh Dutt, Denis Pellerin, and Georges Quénot. Classifier training from a generative model. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2019.

[137] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis 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.

[138] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[139] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[140] MinGuk Kang, Joonghyuk Shin, and Jaesik Park. StudioGAN: A Taxonomy and Benchmark of GANs for Image Synthesis. *2206.09479 (arXiv)*, 2022.

[141] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[142] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[143] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[144] 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.

[145] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *KDD*. ACM, 2019.

[146] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *ICML*. PMLR, 2014.

[147] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[148] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

[149] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 2018.

[150] Pooyan Sedigh, Rasoul Sadeghian, and Mehdi Tale Masouleh. Generating synthetic medical images by using gan to improve cnn performance in skin cancer classification. In *International Conference on Robotics and Mechatronics*, 2019.

[151] Jyoti Islam and Yanqing Zhang. Gan-based synthetic brain pet image generation. *Brain informatics*, 2020.

[152] Fatemeh Baghdadi, Davide Cirillo, Daniele Lezzi, Francesc Lordan, Fernando Vazquez, Eugenio Lomurno, Alberto Archetti, Danilo Ardagna, and

Matteo Matteucci. Harnessing the computing continuum across personalized healthcare, maintenance and inspection, and farming 4.0. *arXiv preprint arXiv:2403.14650*, 2024.

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

[154] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *USENIX Security Symposium*, 2016.

[155] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the International Conference on Machine Learning*, 2012.

[156] Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems*, 2024.

[157] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2022.

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

[159] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[160] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*, 2024.

[161] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[162] Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models. *arXiv preprint*, 2024.

[163] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, 2011.

[164] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 2018.

[165] 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*, 2023.

[166] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[167] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 2020.

[168] Mohammad Rasouli, Tao Sun, and Ram Rajagopal. Fedgan: Federated generative adversarial networks for distributed data. *arXiv preprint*, 2020.

[169] Sean Augenstein, H Brendan McMahan, Daniel Ramage, Swaroop Ramaswamy, Peter Kairouz, Mingqing Chen, Rajiv Mathews, and Blaise Aguera y Arcas. Generative models for effective ml on private,

decentralized datasets. In *International Conference on Learning Representations*, 2019.

[170] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. In *International Conference on Learning Representations*, 2021.

[171] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *International Conference on Data Engineering*, 2022.

[172] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data poisoning attacks against federated learning systems. In *European Symposium on Research in Computer Security*, 2020.

[173] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, 2018.

[174] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in Neural Information Processing Systems*, 2020.

[175] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases-Volume Part III*, 2013.

[176] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 2019.

[177] Yann Fraboni, Richard Vidal, and Marco Lorenzi. Free-rider attacks on model aggregation in federated learning. In *International Conference on Artificial Intelligence and Statistics*, 2021.

[178] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 25th ACM SIGSAC Conference on Computer and Communications Security*, 2018.

[179] 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.

[180] Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys*, 2018.

[181] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017.

[182] Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. *Advances in neural information processing systems*, 2018.

[183] Giacomo Savazzi, Eugenio Lomurno, Cristian Sbrolli, Agnese Chiatti, and Matteo Matteucci. Neuro-symbolic scene graph conditioning for synthetic image dataset generation. *arXiv preprint arXiv:2503.17224*, 2025.

[184] Nicolo Resmini, Eugenio Lomurno, Cristian Sbrolli, and Matteo Matteucci. Your image generator is your new private dataset. *arXiv preprint arXiv:2504.04582*, 2025.

A | Appendix - Datasets

This Appendix provides the most relevant information about the datasets used in this document. Table A.1 shows the type of each dataset, its number of classes and the size of each of the training, validation and testing sections.

Table A.1: Information about the classification datasets used in this document. Unless otherwise stated, all references are to image datasets.

Dataset	Data Type	# Classes	# Training/Validation/Test
CIFAR10	Various Subjects	10	40000/10000/10000
CIFAR100	Various Subjects	100	40000/10000/10000
FashionMNIST	Clothing	10	50000/10000/10000
MNIST	Handwritten Digits	10	50000/10000/10000
EuroSAT	Satellite Imagery	10	21600/5400/5400
TinyImageNet	Various Subjects	200	100000/10000/10000
OxfordFlowers	Flowers	102	1020/1020/6149
STL10	Various Subjects	10	5000/4000/4000
CINIC10	Various Subjects	10	90000/9000/9000
PathMNIST	Histopathology Images	9	89996/10004/7180
BloodMNIST	Blood Cell Microscope	8	11959/1712/3421
DermaMNIST	Dermatoscope	7	7007/1003/2005
OrganCMNIST	Abdominal CT	11	12975/2392/8216
OrganSMNIST	Abdominal CT	7	13932/2452/8827
PneumoniaMNIST	Chest X-Ray	7	4708/524/624
RetinaMNIST	Fundus Camera	7	1080/120/400
Titanic	Tabular - Demographics	2	713/89/89
Breast Cancer	Tabular - Medical Records	2	457/56/56
Mushrooms	Tabular - Fungi Features	2	6500/812/812
Adult	Tabular - Census Data	2	39074/4884/4884
Wine Quality	Tabular - Chemical Properties	2	5199/649/649

B | Appendix - Discriminative Adversarial Privacy

Analysis of Accuracy Over Privacy

This appendix presents an in-depth examination of the Accuracy Over Privacy (AOP) metric for a range of lambda (λ) values not explored in the main paper. Tables B.1,B.2,B.3,B.4,B.5, and B.6 illustrate the AOP results for λ values of 1, 2, 5, 10, 20, and 50, respectively.

The experiments encompass several datasets: Cifar-10 [63], Cifar-100 [63], FMNIST [65], EuroSAT [66], TinyImagenet [67], OxfordFlowers [68], STL-10 [69], and Cinic-10 [70].

Table B.1 uniquely demonstrates comparable magnitudes of accuracy and Area Under the Curve (AUC) for the Membership Inference Attack (MIA). In this scenario, the trade-off between these metrics clearly favours the Reg model, followed by the DAP models, in comparison to the Baseline model.

As λ increases, AOP scores decrease, indicating a greater penalty proportional to the AUC. Tables B.3 and B.4 demonstrate a significantly faster decline in the average AOP of the Baseline compared to other models. Notably, in Table B.4, DAP models emerge as the optimal choice for managing the trade-off, followed by

Table B.1: Evaluation of various models using the AOP metric ($\lambda = 1$) across multiple datasets. The table presents a comparative analysis of different approaches, including regularization and differential privacy methods. Best results are in **bold**, second-best underlined.

Dataset	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
CIFAR10	60.53	**64.32**	31.02	35.53	40.23	41.54	61.52	60.74
CIFAR100	39.91	**42.83**	3.93	8.03	8.92	7.13	30.53	27.32
FashionMNIST	84.42	82.43	60.32	69.83	73.04	76.63	85.43	**86.14**
EuroSAT	88.03	**90.02**	30.52	58.63	68.12	64.34	89.83	89.13
TinyImageNet	30.32	**31.93**	3.02	3.23	3.14	2.53	25.23	21.32
OxfordFlowers	37.23	**43.13**	2.83	4.73	8.32	13.14	26.93	24.72
STL10	54.23	**57.73**	8.43	13.52	24.72	28.83	47.23	37.92
CINIC10	58.83	57.73	27.93	33.92	38.93	40.23	56.52	**57.83**
Average	56.69	**58.77**	21.00	28.43	33.18	34.30	52.90	50.64

DP models, and lastly, the Reg model, which performs similarly to the Baseline model.

It is noteworthy that DP results tend to deteriorate less than Baseline and Reg results as λ increases, often surpassing them. However, they struggle with datasets containing numerous classes, indicative of poor accuracy despite guaranteed privacy.

Table B.2: Performance comparison of privacy-preserving models using the AOP metric ($\lambda = 2$). This table presents results for various datasets, showcasing the effectiveness of different approaches including regularization and differential privacy techniques. Best results are in **bold**, second-best underlined.

Dataset	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
CIFAR10	46.72	50.93	30.73	33.83	38.72	41.32	**60.73**	60.13
CIFAR100	33.12	**34.52**	3.92	7.83	8.73	7.02	29.63	26.93
FashionMNIST	76.52	73.32	60.03	69.52	72.43	75.92	84.23	**85.02**
EuroSAT	80.92	85.23	30.23	58.32	68.12	64.13	**89.63**	88.93
TinyImageNet	25.12	**27.02**	2.92	3.23	2.93	2.52	24.42	20.92
OxfordFlowers	24.42	**28.12**	2.63	4.42	7.92	12.32	25.02	23.72
STL10	44.92	**51.32**	8.32	12.92	24.52	28.83	46.52	37.52
CINIC10	51.42	47.02	27.92	33.72	38.63	39.92	55.23	**57.02**
Average	47.90	49.69	20.84	27.97	32.75	34.00	**51.93**	50.02

Table B.3: Analysis of model performance using the AOP metric ($\lambda = 5$) for various datasets. The table compares different privacy-preserving approaches, including regularization and differential privacy methods. Best results are in **bold**, second-best underlined.

Dataset	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
CIFAR10	21.42	25.32	29.83	29.02	34.63	40.63	58.23	**58.32**
CIFAR100	18.82	18.02	3.92	7.23	8.42	6.83	**26.92**	26.02
FashionMNIST	56.83	51.63	59.32	68.72	70.72	73.63	80.83	**82.12**
EuroSAT	62.92	72.32	29.32	27.63	68.12	63.32	**89.12**	88.42
TinyImageNet	14.32	16.23	2.72	3.23	2.63	2.42	**22.23**	19.82
OxfordFlowers	6.92	7.92	2.02	3.63	6.72	10.23	20.12	**20.92**
STL10	25.52	35.92	8.23	11.23	23.82	28.63	**44.32**	36.23
CINIC10	34.32	25.42	27.72	33.12	37.92	38.92	51.72	**54.72**
Average	30.13	31.60	20.39	26.73	31.62	33.08	**49.19**	48.32

Table B.4: Comparative evaluation of privacy-preserving models using the AOP metric ($\lambda = 10$) across multiple datasets. This table presents the effectiveness of various approaches, including regularization and differential privacy techniques. Best results are in **bold**, second-best underlined.

Dataset	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
CIFAR10	5.92	7.92	28.32	22.52	28.72	39.42	54.32	**55.52**
CIFAR100	7.42	6.12	3.92	6.23	7.82	6.42	23.02	**24.52**
FashionMNIST	34.63	28.82	58.12	67.42	68.02	70.12	75.42	**77.32**
EuroSAT	41.23	55.12	27.92	56.52	68.12	62.12	88.23	**87.52**
TinyImageNet	5.63	6.92	2.32	3.12	2.12	2.32	**19.02**	18.12
OxfordFlowers	0.82	0.92	1.42	2.52	5.23	7.52	26.92	**24.72**
STL10	9.92	19.82	8.12	8.92	22.63	28.32	**40.92**	34.12
CINIC10	17.52	9.12	27.42	32.12	36.82	37.42	46.42	**51.02**
Average	15.39	16.84	19.70	24.92	29.94	31.71	45.16	**45.61**

Table B.5: Comparative analysis of AOP metric on test sets with $\lambda = 20$. The table presents results for various datasets and methods, including the baseline, regularization (Reg), differential privacy (DP) with different ϵ values, and two versions of DAP. Best results are in **bold**, second-best underlined.

Dataset	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
CIFAR10	0.41	0.83	25.62	13.67	19.84	37.18	47.25	**50.29**
CIFAR100	1.13	0.72	3.78	4.63	6.87	5.78	16.84	**21.71**
FashionMNIST	12.93	8.92	55.96	64.75	62.83	63.41	65.68	**68.62**
EuroSAT	17.73	31.95	25.27	54.38	68.19	59.63	**86.54**	85.82
TinyImageNet	0.92	1.35	1.87	3.17	1.46	2.18	13.84	**15.29**
OxfordFlowers	0.01	0.01	0.67	1.25	3.18	4.06	6.79	**11.37**
STL10	1.52	6.08	7.75	5.68	20.59	27.83	**34.96**	30.27
CINIC10	4.63	1.27	26.94	30.28	34.75	34.58	37.39	**44.46**
Average	4.91	6.39	18.48	22.23	27.21	29.33	38.66	**40.98**

Table B.6: Analysis of AOP metric on test sets with $\lambda = 50$. This table showcases results for various datasets and methods, including the baseline, regularization (Reg), differential privacy (DP) with different ϵ values, and two versions of DAP. The comparative performance across different approaches is presented, with the best results in **bold** and second-best underlined.

Dataset	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
CIFAR10	0.01	0.01	19.07	3.08	6.59	31.04	31.17	**37.35**
CIFAR100	0.01	0.01	3.54	1.93	4.57	4.09	6.58	**15.23**
FashionMNIST	0.74	0.38	49.52	**57.49**	49.46	47.18	43.27	48.02
EuroSAT	1.43	6.29	18.78	48.17	68.19	52.96	**81.42**	80.83
TinyImageNet	0.01	0.01	0.85	2.96	0.47	1.73	5.38	**8.92**
OxfordFlowers	0.01	0.01	0.06	0.13	0.75	0.63	0.75	**3.94**
STL10	0.01	0.27	6.93	1.48	15.26	**26.18**	21.79	21.17
CINIC10	0.09	0.01	25.38	25.39	29.07	27.23	19.46	**29.27**
Average	0.29	0.87	15.52	17.58	21.79	23.88	26.23	**30.59**

Membership Inference Attacks Against Slices

This section explores how the effectiveness of Membership Inference Attacks (MIAs) varies depending on whether they target correctly or incorrectly classified samples from the attacked model. Tables B.7 and B.8 provide concrete evidence for this assertion.

The experiments utilise the following datasets: Cifar-10 [63], Cifar-100 [63], FMNIST [65], EuroSAT [66], TinyImagenet [67], OxfordFlowers [68], STL-10 [69], and Cinic-10 [70].

Table B.7, which presents MIA results against misclassified samples, reveals particular vulnerability in Baseline and Reg models. Conversely, DP models exhibit high security levels. The DAP technique achieves results approximating random guessing, thus ensuring robust protection.

Table B.8 demonstrates the generally poor performance of MIAs on correctly classified samples. Both Baseline and Reg models allow the attacking model to achieve an AUC close to 0.5, often surpassing the AUC obtained by DP models. This suggests that correctly predicted test set samples share similar characteristics with training data, making them challenging to differentiate.

The most significant finding pertains to DAP models, which not only outperform the Baseline but also emerge as the overall best, surpassing even DP models.

Table B.7: AUC metric for Membership Inference Attacks (MIAs) on misclassified samples. This table presents a comparative analysis of various privacy-preserving techniques across different datasets. The effectiveness of each method in mitigating MIAs is shown, with the best results in **bold** and second-best underlined.

Dataset	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
CIFAR10	75.36	68.97	50.57	**50.38**	**50.38**	50.69	51.06	50.57
CIFAR100	61.07	62.78	**50.15**	50.63	53.83	50.59	51.87	50.94
FashionMNIST	65.97	76.89	50.28	50.67	50.28	**50.09**	52.39	52.75
EuroSAT	69.82	56.38	**50.09**	50.17	52.39	50.47	52.78	52.83
TinyImageNet	58.52	56.59	50.43	50.27	**50.18**	50.56	51.86	50.83
OxfordFlowers	78.53	83.79	52.96	55.58	**51.86**	53.27	54.75	53.78
STL10	63.67	56.75	**50.09**	50.63	50.68	50.49	50.96	50.49
CINIC10	56.04	64.97	**50.19**	50.38	50.35	50.28	50.86	51.09
Average	66.12	65.89	**50.59**	51.09	51.24	50.81	52.07	51.66

Table B.8: Evaluation of AUC metric for Membership Inference Attacks (MIAs) on correctly classified samples. This table presents a comprehensive comparison of various privacy-preserving techniques across different datasets, showcasing their effectiveness in mitigating MIAs. Best results are highlighted in **bold**, while second-best are underlined.

Dataset	Baseline	Reg	$DP_{\varepsilon=0.5}$	$DP_{\varepsilon=1}$	$DP_{\varepsilon=2}$	$DP_{\varepsilon=4}$	DAP_t	DAP_v
CIFAR10	54.59	53.75	**50.09**	50.28	50.35	50.67	50.73	50.49
CIFAR100	51.69	51.78	51.27	52.17	50.85	51.16	**50.72**	50.94
FashionMNIST	51.47	53.28	50.43	50.58	51.18	51.09	50.56	**50.15**
EuroSAT	51.46	50.39	51.63	50.87	50.82	53.28	**50.27**	**50.27**
TinyImageNet	50.96	50.63	50.45	53.02	54.43	52.01	50.28	**50.19**
OxfordFlowers	55.63	59.54	87.75	67.76	58.93	59.06	**52.97**	53.53
STL10	52.26	50.82	52.17	51.36	**50.16**	50.57	50.36	50.42
CINIC10	50.86	53.43	50.58	**50.19**	**50.19**	**50.19**	**50.19**	**50.19**
Average	52.37	52.95	55.55	53.28	52.11	52.25	**50.76**	**50.77**

C | Appendix - Gap Filler

Generative Adversarial Network Architecture

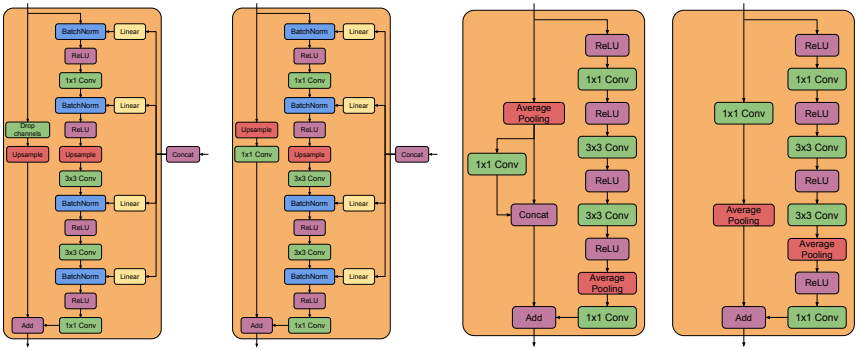

Figure C.1: Architectural comparison of BigGAN Deep blocks. From left to right: conventional Generator block, StudioGAN Generator block (employed in this study), conventional Discriminator block, StudioGAN Discriminator block (employed in this study).

The research utilised the BigGAN Deep architecture as the generative model, drawing upon the implementation provided by the StudioGAN library [140]. This variant introduces subtle modifications to the residual block structure in both the generator and discriminator components. Within the generator's G block, rather than reducing the number of channels, the residual path undergoes upsampling followed by a Conv1x1 operation. This approach ensures consis-

tent channel quantities between the residual and non-residual pathways. Analogously, the discriminator's D block initially processes the residual path through a Conv1x1 layer to achieve the appropriate channel output before applying downsampling. Figure C.1 offers a visual juxtaposition of the original BigGAN Deep blocks and their StudioGAN counterparts.

The training regimen for all models spanned 500 epochs, employing a batch size of 192 and incorporating 3 discriminator (D) steps per generator (G) step. It should be noted that the interpretation of "N D steps per G step" in this context diverges slightly from its original formulation. Empirical observations suggest that this modified approach yields superior outcomes. Specifically, instead of subdividing the batch (size 192) into N sub-batches (size 64) and training the discriminator on each sub-batch individually, N iterations of discriminator training are conducted on the complete batch (size 192) (see Figure C.2). This modification is hypothesised to produce enhanced results on datasets of smaller scale than ImageNet, hence the decision to refrain from adopting the larger batch sizes (2048) proposed by the original BigGAN authors in their ImageNet-based training. The training dataset underwent minimal augmentation, limited to random horizontal flips.

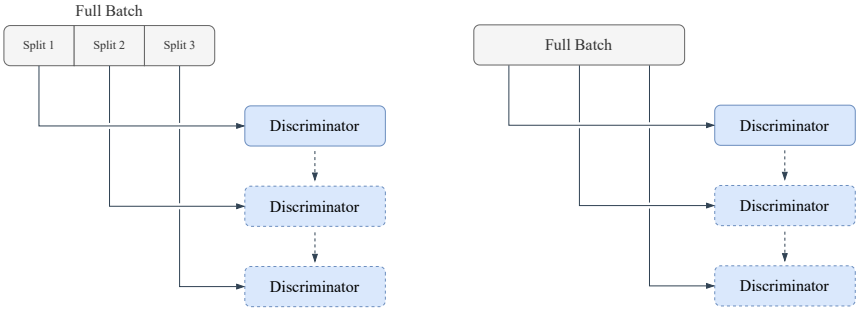

Figure C.2: Comparative illustration of discrimination steps: conventional approach versus the method employed in this study.

Expansion Trick

Figure C.3: Visual representation of the Expansion Technique's effects on generated images. The images correspond to the "Truck" class label, with standard deviations ranging from 1.0 to 2.0 in increments of 0.2 (fixed seed). Upper row: An instance where increased standard deviation diminishes image quality, likely resulting in filtration. Lower row: An example where increased standard deviation enhances diversity without compromising quality.

The Expansion Technique entails an enlargement of the input noise space, as opposed to truncation. This is accomplished by sampling from a normal distribution characterised by a standard deviation exceeding that employed during model training. By broadening the diversity of the input noise space, this approach encourages the generative model to explore underrepresented regions encountered less frequently during the training phase. Consequently, it facilitates the generation of more diverse and novel images, a desirable outcome in scenarios where diversity takes precedence over visual fidelity. As anticipated, the increased standard deviation of the input noise distribution adversely impacts the quality of individual samples, as evidenced in Figure C.3. Thus, the efficacy of the Expansion Technique is enhanced when employed in conjunction with sample filtering methodologies.

D | Appendix - Knowledge Recycling

Image Generators

This appendix section provides a detailed comparison between the original BigGAN-Deep model and the modified version employed in this study, referred to as BigGAN-Deep (ours). The primary differences between the two models are summarised in Table D.1.

Figure D.1 illustrates the Validation Classification Accuracy Scores, computed for each checkpoint, across the CIFAR10, CIFAR100, and FashionMNIST datasets. Notably, both models undergo identical training protocols as those applied during the Checkpoint Optimisation phase of the Knowledge Recycling pipeline, with one key distinction: the synthetic dataset is not regenerated during training, and its labels correspond to the hard labels used to condition the data generation process.

Table D.1: A comparative analysis of key parameters is presented. The comparison is made between the original BigGAN-Deep model, referred to as BigGAN-Deep (vanilla), and its modified version utilized in this study, denoted as BigGAN-Deep (ours).

Model	BigGan-Deep (vanilla)	BigGan-Deep (ours)
Latent dimension	128	128
Shared dimension	128	128
Batch size	64	64
Adversarial loss	Hinge	Logistic
Precision	Full	Mixed
G conditioning	cBN	cBN
D conditioning	PD	PD
G spectral normalisation	True	True
D spectral normalisation	True	True
G optimiser	Adam	Adam
D optimiser	Adam	AdamW
G learning rate	0.0002	0.0002
D learning rate	0.0002	0.0002
G beta1	0.5	0.5
G beta2	0.999	0.999
D beta1	0.5	0.5
D beta2	0.999	0.999
G ema	True	True
G ema decay	0.9999	0.9999
G ema starting step	1000	1000
D basket size	3	1
D label smoothing	-	0.1
D weight decay	-	0.004
D updates per step	2	4
G attention resolution	16	16
D attention resolution	16	16
G convolutional dimension	128	128
D convolutional dimension	128	128
G depth	2	2
D depth	2	2
G model size	41M	41M
D model size	39M	39M

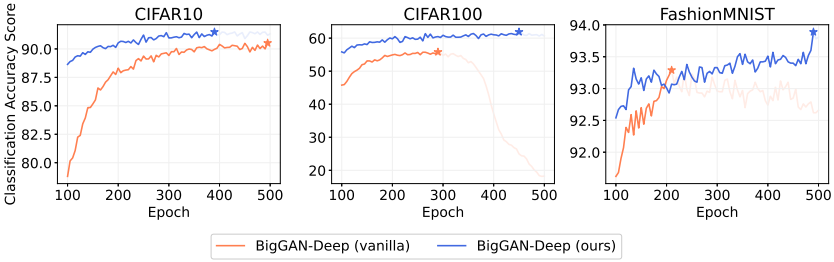

Figure D.1: For the considered datasets, the Classification Accuracy Score (CAS) is presented. This metric is calculated for each checkpoint during validation. The comparison encompasses both the BigGAN-Deep (vanilla) and BigGAN-Deep (ours) generators.

Table D.2: Detailed parameters for configuration, training, and data augmentation are presented. These specifications pertain to each Classifier implemented in this research.

Model	ResNet14
Initial filters	64
Batch size	256
Loss	Categorical cross-entropy
Precision	Mixed
Learning rate	0.5
Learning rate scheduler	Cosine annealing
Warm-up epochs	10
Warm-up learning rate	0.00001
Optimiser	SGD
Momentum	0.9
Weight decay	0.0005
Nesterov	True
Clipnorm	1.0
Label smoothing	0.1
TrivialAugment	True
TrivialAugment interpolation	Bilinear
MixUp	True
MixUp alpha	0.2
Random horizontal flip	True
Padding	2
Random crop	True

Image Classifiers

This appendix section provides a comprehensive overview of the implementation details for the classifiers utilised in this study. Table D.2 presents a detailed breakdown of the configuration, training procedures, and data augmentation parameters.

Training Strategies

This appendix section offers a comparative analysis of the performance outcomes associated with the various training strategies explored during the preliminary phase of this research. Figure D.1 displays the Validation Classification Accuracy Scores for each checkpoint of the BigGAN-Deep (ours) model across the CIFAR10, CIFAR100, and FashionMNIST datasets. The training protocol for each model mirrors the procedure employed during the Checkpoint Optimisation phase of the Knowledge Recycling pipeline, with the sole exception being the generation of the synthetic dataset.

In the Baseline strategy, the synthetic dataset is generated only once at the beginning of training, with a cardinality matching that used for the Generator's training. The Gap Filler strategy implements the filtering technique proposed by Lampis et al., regenerating the synthetic dataset every 10 epochs while maintaining the same dataset size as that of the Generator. The Generative Knowledge Distillation (GKD) strategy, introduced in this work, also regenerates the synthetic dataset every 10 epochs, keeping the dataset size constant.

The analysis of the results underscores the superiority of the GKD strategy, demonstrating that this approach significantly enhances the information content within the generated synthetic datasets compared to the other strategies considered.

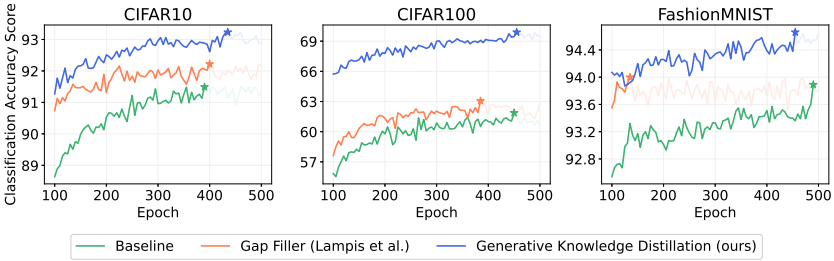

Figure D.2: A comparative analysis of three training strategies for the BigGAN-Deep (ours) generator is presented. The strategies include the Baseline approach with a single dataset generation, the Gap Filler method by Lampis et al., and the proposed Generative Knowledge Distillation (GKD) technique. For each considered dataset, the Classification Accuracy Score (CAS) is calculated at every checkpoint during validation. This comparison demonstrates the enhanced information content in synthetic datasets generated using the GKD approach.

E | Appendix - FedKR Security Analysis

This appendix presents rigorous proofs demonstrating the security properties of the FedKR framework against various attack vectors. The presentation begins with precise definitions and assumptions, followed by detailed proofs for each security claim.

Definitions and Notation:

- $\mathcal{I} = \{1, \ldots, N\}$: Set of federation participants.
- $\mathcal{X} \subseteq \mathbb{R}^{h \times w \times c}$: Input space.
- $\mathcal{Y} = \{1, \ldots, K\}$: Label space for K-class classification.
- $D_i = \{(x_i^u, y_i^u)\}_{u=1}^{U_i} \subset \mathcal{X} \times \mathcal{Y}$: Private dataset of participant $i \in \mathcal{I}$ with cardinality U_i.
- $T_i \colon \mathcal{X} \to \mathcal{Y}$: Teacher classifier of participant $i \in \mathcal{I}$, trained on D_i.
- $\mathcal{Z}_i \subseteq \mathbb{R}^l$: Generator latent space for participant $i \in \mathcal{I}$ with dimension l.
- $G_i \colon \mathcal{Z}_i \to \mathcal{X}$: Generator of participant $i \in \mathcal{I}$, trained on D_i.
- $S_i = \{(G_i(z_i^v),\ T_i(G_i(z_i^v)))\}_{v=1}^{V_i} \subset \mathcal{X} \times \mathcal{Y}$: Public synthetic dataset of participant $i \in \mathcal{I}$ with cardinality V_i.
- $\omega_{ij} \in [0, 1]$: Weight assigned by participant $i \in \mathcal{I}$ to synthetic dataset S_j where $j \in \mathcal{I}$ during their Dynamic Dataset Aggregation (DDA) optimization step.

- $\mathcal{S} = \{S_i\}_{i \in \mathcal{I}}$: Collection of all public synthetic datasets shared within the federation.
- $W_i = \bigcup_{j \in \mathcal{I}} \{s : s \in S_j \text{ with weight } \omega_{ij}\}$: Weighted aggregation of synthetic datasets for participant $i \in \mathcal{I}$.
- $M_i : \mathcal{X} \to \mathcal{Y}$: Student classifier of participant $i \in \mathcal{I}$, trained on W_i.
- θ_i: Parameters of participant i's model M_i.
- $\mathcal{L}_i$: Loss function used during the training of model M_i.
- $\varepsilon \in \mathbb{R}^+$: Negligible positive quantity.
- $\delta \in \mathbb{R}^+$: Non-negligible positive quantity.
- $\mathcal{A}$: Computationally bounded adversary.

Assumptions:

1. **Privacy of Information.** Each participant $i \in \mathcal{I}$ maintains strict privacy of all personal data and internal parameters, including:
   - **Private dataset:** D_i and its cardinality U_i.
   - **Models and their properties:** Generator G_i, Teacher classifier T_i, Student classifier M_i.
   - **Training-related information:** Model architectures, training strategies, data augmentation, and preprocessing techniques.
   - **Operational parameters:** Latent space dimension l, weights $\{\omega_{ij}\}_{j \in \mathcal{I}}$, aggregated dataset W_i.

   The only shared information within the federation is the collection of synthetic datasets $\mathcal{S} = \{S_i\}_{i \in \mathcal{I}}$, where each S_i inherently preserves individual-level privacy through its synthetic nature, while carrying some of the intended population-level statistical properties.
2. **Generator Independence.** Each Generator G_i, where $i \in \mathcal{I}$, is trained independently on its corresponding private dataset D_i, without any form of parameter sharing or collaborative training.

3. **Training Set Diversity.** For each participant $i \in \mathcal{I}$, the aggregated synthetic dataset W_i incorporates diverse synthetic data from multiple participants $j \in \mathcal{I}$, ensuring rich and varied training samples.
4. **Dynamic Dataset Aggregation.** During the training of M_i, each participant $i \in \mathcal{I}$ assigns weights $\{\omega_{ij}\}_{j\in\mathcal{I}}$ to samples from synthetic datasets $\{S_j\}_{j\in\mathcal{I}}$ in W_i, reflecting the importance of each dataset in their learning process.
5. **Adversarial Access Constraints.** The adversary $\mathcal{A}$ is limited to accessing only the publicly shared synthetic datasets $\mathcal{S} = \{S_i\}_{i\in\mathcal{I}}$.
6. **Computational Limitations.** The adversary $\mathcal{A}$ is restricted to polynomial-time computations, ensuring practical bounds on computational capabilities.

The assessment of security measures employs a chromatic indicator system: (•) in red denotes a security vulnerability, (•) in yellow signifies partial protection, whilst (•) in green demonstrates effective countermeasures. In cases where a specific attack technique is not applicable, a dash (−) is utilised. The rating system reflects the potential effectiveness of each attack methodology in compromising the confidentiality of actual private data.

Theorem E.1 ((-) Model Inversion Inapplicability)**.** *Under the stated assumptions, for any adversary $\mathcal{A}$ and any private sample $x \in D_i$, the probability that $\mathcal{A}$ can reconstruct x by attacking M_i and having access to $\mathcal{S}$ is negligible:*

$$P\left(\mathcal{A}(M_i, \mathcal{S}) = x\right) \leq \varepsilon. \tag{E.1}$$

Proof. For contradiction, suppose there exists an adversary $\mathcal{A}$ capable of reconstructing a private sample $x \in D_i$ with non-negligible probability $\delta > \varepsilon$, such that:

$$P\left(\mathcal{A}(M_i, \mathcal{S}) = x\right) \geq \delta. \tag{E.2}$$

This would imply that $\mathcal{A}(M_i, \mathcal{S})$ approximates $M_i^{-1}\colon \mathcal{Y} \to \mathcal{X}$, thus potentially

exploiting a state-of-the-art method to invert the label and reconstruct $x \in D_i$ [45].

However, since M_i is private and inaccessible (by Assumption 1), $\mathcal{A}$ would need to invert a transformation that cannot be known or estimated. Therefore, equation (E.2) can be rewritten as:

$$P\left(\mathcal{A}(\mathcal{S}) = x\right) \geq \delta. \tag{E.3}$$

Since the adversary lacks access to the target model M_i, performing a model inversion attack becomes infeasible, regardless of any information extracted from $\mathcal{S}$. Thus, the probability of inverting an inaccessible model to reconstruct the input x that generated a prediction y is negligible. Therefore, the probability of successfully reconstructing any $x \in D_i$ is negligible. Therefore:

$$P\left(\mathcal{A}(M_i, \mathcal{S}) = x\right) = P\left(\mathcal{A}(\mathcal{S}) = x\right) \leq \varepsilon. \tag{E.4}$$

□

Attack Surface $\mathcal{S}$. Since $\mathcal{S}$ represents the only available attack surface, an adversary might attempt to exploit the information present in S_i to reconstruct samples $x \in D_i$. Such an attack, if effective with non negligible probabilities, could be formalised as:

$$\begin{aligned} P\left(\mathcal{A}(\mathcal{S}) = x\right) &= P\left(\mathcal{A}(S_i) = x\right) \\ &= P\left(\mathcal{A}\left((G_i(z_i),\ T_i(G_i(z_i)))\right) = x\right) \geq \delta. \end{aligned} \tag{E.5}$$

At most, what could be achieved is a model capable of partially emulating T_i through distillation or surrogate model creation from the pairs $(G_i(z_i),\ T_i(G_i(z_i)))$, and subsequently, at best, approximating T_i^{-1}.

Given that the architectures and all training information related to T_i remain unknown to the adversary $\mathcal{A}$ (by Assumption 1), even if such architectures could be approximated in the worst-case scenario, there would be no means to access

any y associated with $x \in D_i$. Consequently, there would be no non-negligible probability that such an attack could expose confidential information about private data - which constitutes the scope of model inversion attacks. Thus:

$$P\left(\mathcal{A}(\mathcal{S}) = x\right) \leq \varepsilon. \tag{E.6}$$

An adversary might attempt to exploit the attack surface provided by the synthetic datasets $\mathcal{S}$ to reconstruct a private sample $x \in D_i$ [45]. Specifically, the adversary may utilise S_i, the synthetic dataset shared by participant i. The adversary has access to these synthetic samples and their corresponding labels generated by T_i.

The adversary might attempt to approximate classifier T_i through a surrogate model $\tilde{T}_i$, trained on S_i:

$$\tilde{T}_i \approx T_i. \tag{E.7}$$

Similarly, an attempt might be made to approximate generator G_i through a surrogate model $\tilde{G}_i$, using synthetic data in S_i:

$$\tilde{G}_i \approx G_i. \tag{E.8}$$

However, since S_i contains only synthetic samples generated by G_i and labelled by T_i, which are potentially complex models with unknown architectures and parameters (by Assumption 1), the approximations $\tilde{T}_i$ and $\tilde{G}_i$ will necessarily be imprecise:

$$\left\| T_i(x) - \tilde{T}_i(x) \right\| \geq \varepsilon, \quad \left\| G_i(z) - \tilde{G}_i(z) \right\| \geq \varepsilon. \tag{E.9}$$

Furthermore, even if, for the sake of argument, $\tilde{T}_i$ and $\tilde{G}_i$ were available with negligible approximation errors, the adversary could neither reconstruct any $x \in D_i$ exactly nor prove having done so (by Assumption 1).

Assume, for contradiction, that $\mathcal{A}$ could compute approximate inverses $\tilde{T}_i^{-1}$ and $\tilde{G}_i^{-1}$ such that:

$$\tilde{T}_i^{-1}(y) \approx x', \quad \tilde{G}_i^{-1}(x') \approx z', \tag{E.10}$$

where x' represents synthetic data and z' represents an approximated latent vector.

In the case of $\tilde{T}_i^{-1}$, the adversary could at most obtain a model capable of approximately mapping labels y to synthetic data x' not belonging to D_i:

$$x' = \tilde{T}_i^{-1}(y). \tag{E.11}$$

However, since x' is generated through $\tilde{T}_i^{-1}$ which approximates T_i^{-1} using synthetic data, x' cannot be a real sample from D_i.

Similarly, for $\tilde{G}_i^{-1}$, the adversary might attempt to recover the latent vector z' from synthetic data x':

$$z' = \tilde{G}_i^{-1}(x'). \tag{E.12}$$

However, inverting a generative model such as $\tilde{G}_i$ is inherently difficult due to the nature of generative networks [10]. Moreover, even if the adversary obtained z', this would provide no information about $x \in D_i$, as x' is a synthetic sample not belonging to D_i, and vector z' bears no correlation to the real data in D_i, given that G_i maps from $\mathcal{Z}$ to an approximated distribution of $\mathcal{X}$.

Therefore, even in the hypothetical case where the adversary could construct $\tilde{T}_i^{-1}$ and $\tilde{G}_i^{-1}$ with negligible errors, they would be unable to reconstruct any $x \in D_i$ or prove having done so.

This demonstrates that, although the adversary might attempt to exploit the attack surface $\mathcal{S}$, such an attempt provides no advantage in terms of performing

a model inversion attack, and the probability of successfully reconstructing a sample $x \in D_i$ remains negligible.

Theorem E.2 ((-)Membership Inference Inapplicability)**.** *Under the stated assumptions, for any adversary $\mathcal{A}$ and any private sample $x \in D_i$, the adversary's advantage in determining whether $x \in D_i$ by attacking M_i and having access to $\mathcal{S}$ is negligible:*

$$|P\left(\mathcal{A}(M_i, \mathcal{S}, x) = 1 \mid x \in D_i\right) - P\left(\mathcal{A}(M_i, \mathcal{S}, x) = 1 \mid x \notin D_i\right)| \leq \varepsilon. \tag{E.13}$$

Proof. Suppose, for contradiction, that there exists an adversary $\mathcal{A}$ capable of determining the membership of a private sample $x \in D_i$ with non-negligible advantage $\delta > \varepsilon$, such that:

$$|P\left(\mathcal{A}(M_i, \mathcal{S}, x) = 1 \mid x \in D_i\right) - P\left(\mathcal{A}(M_i, \mathcal{S}, x) = 1 \mid x \notin D_i\right)| \geq \delta. \tag{E.14}$$

This would imply that $\mathcal{A}$ can exploit information from M_i and $\mathcal{S}$ to distinguish whether x is a member of D_i with significant advantage. However, since M_i is private and inaccessible (by Assumption 1), the adversary cannot utilize M_i directly in their attack. Therefore, equation (E.14) can be rewritten as:

$$|P\left(\mathcal{A}(\mathcal{S}, x) = 1 \mid x \in D_i\right) - P\left(\mathcal{A}(\mathcal{S}, x) = 1 \mid x \notin D_i\right)| \geq \delta. \tag{E.15}$$

Given the absence of access to the target model M_i, the adversary must rely solely on the shared synthetic datasets $\mathcal{S}$ to perform the membership inference attack [24]. A membership inference attack becomes thus infeasible, given that it necessitates of a target model to be exploited as a target, and this remains true regardless of any information extracted from $\mathcal{S}$. Thus, the probability of guessing the presence of a sample x inside the training set D_i may be at most a

random guess plus some inference noise, thus negligible. Therefore:

$$|P\left(\mathcal{A}(M_i, \mathcal{S}, x) = 1 \mid x \in D_i\right) - P\left(\mathcal{A}(M_i, \mathcal{S}, x) = 1 \mid x \notin D_i\right)| \leq \varepsilon. \tag{E.16}$$

□

Attack Surface $\mathcal{S}$*.* As $\mathcal{S}$ represents the only available attack surface, an adversary might attempt to exploit the information present in $\mathcal{S}$ to determine whether $x \in D_i$. Such an attack, if effective with non-negligible probabilities, could be formalised as:

$$\begin{aligned}
&|P\left(\mathcal{A}(\mathcal{S}, x) = 1 \mid x \in D_i\right) - P\left(\mathcal{A}(\mathcal{S}, x) = 1 \mid x \notin D_i\right)| = \\
&|P\left(\mathcal{A}(S_i, x) = 1 \mid x \in D_i\right) - P\left(\mathcal{A}(S_i, x) = 1 \mid x \notin D_i\right)| = \\
&|P\left(\mathcal{A}((G_i(z_i), T_i(G_i(z_i))), x) = 1 \mid x \in D_i\right) - \\
&P\left(\mathcal{A}((G_i(z_i), T_i(G_i(z_i))), x) = 1 \mid x \notin D_i\right)| \geq \delta.
\end{aligned} \tag{E.17}$$

However, since G_i and T_i are private and unknown to the adversary (by Assumption 1) and thus there are no models to effectively attack, it is no possible to extract membership information about x from $\mathcal{S}_i$.

The adversary could attempt to find exact matches between $x \in D_i$ and samples in $\mathcal{S}_i$. However, since the synthetic data is generated by G_i trained to approximate the distribution of D_i and without the capability of replicating individual training samples, the probability that x exactly matches any sample in $\mathcal{S}$ is negligible:

$$P\left(x = G_i(z)\right) \leq \varepsilon, \quad \forall z \in \mathcal{Z}. \tag{E.18}$$

Therefore, direct comparison does not provide the adversary with significant advantage. The adversary might then attempt to infer statistical properties of D_i from S_i and compare $x \in D_i$ to these properties. However, since synthetic data may at most expose population-level statistics and not individual samples

ones, and given that many data points not in D_i could also conform to these statistics, the adversary cannot reliably infer membership. More formally, for any property ϕ inferred from S_i:

$$P\left(\phi(x) \mid x \in D_i\right) \approx P\left(\phi(x) \mid x \notin D_i\right). \tag{E.19}$$

The adversary could still attempt to train surrogate models $\tilde{G}_i$ and $\tilde{T}_i$ using S_i. However, given that these are complex models with unknown architectures and parameters (by Assumptions 1), and that the adversary is computationally bounded (by Assumption 6), any such attempt would result in significant approximation errors:

$$\left\|G_i(z) - \tilde{G}_i(z)\right\| \geq \varepsilon, \quad \left\|T_i(x) - \tilde{T}_i(x)\right\| \geq \varepsilon. \tag{E.20}$$

Even if the adversary could approximate these models, they would face two fundamental limitations:

1. The surrogate models would be trained solely on synthetic data, never having seen the private data D_i. Therefore, any potential overfitting that might facilitate membership inference would pertain to the synthetic data, not to D_i:

$$\left|P\left(\mathcal{A}(\tilde{T}_i(x)) = 1 \mid x \in D_i\right) - P\left(\mathcal{A}(\tilde{T}_i(x)) = 1 \mid x \notin D_i\right)\right| \leq \varepsilon \tag{E.21}$$

2. The synthetic data generated by $\tilde{G}_i$ would approximate the distribution generated by G_i that is already an approximation of the distribution of D_i, making it not feasible to determine whether a specific sample x belongs to D_i:

$$\left|P\left(x \in D_i \mid x \sim \tilde{G}_i\right) - P\left(x \notin D_i \mid x \sim \tilde{G}_i\right)\right| \leq \varepsilon. \tag{E.22}$$

Therefore, even with access to the attack surface $\mathcal{S}$, the adversary's advantage in determining whether $x \in D_i$ remains negligible:

$$|P\left(\mathcal{A}(\mathcal{S}, x) = 1 \mid x \in D_i\right) - P\left(\mathcal{A}(\mathcal{S}, x) = 1 \mid x \notin D_i\right)| \leq \varepsilon. \tag{E.23}$$

Theorem E.3 ((●) Data Poisoning, Byzantine, and Sybil Attacks Resistance and Detection)**.** *Given the stated assumptions, and considering that there is no central model and only synthetic datasets are shared—causing Data Poisoning, Byzantine, and Sybil attacks to manifest in a similar fashion—in any set of malicious participants $\mathcal{B} \subseteq \mathcal{I}$ (including individual poisoners and Sybil identities), the cumulative impact on any honest participant's model M_i, where $i \notin \mathcal{B}$, is negligible:*

$$\sum_{b \in \mathcal{B}} \omega_{ib} \leq \varepsilon. \tag{E.24}$$

Proof. Consider a set of malicious participants $\mathcal{B} \subseteq \mathcal{I}$, which may comprise both individual poisoners and Sybil identities, who contribute malicious synthetic datasets S_b with $b \in \mathcal{B}$ intended to compromise the performance of other participants' models. Within this framework, as there is no central model and only synthetic datasets are shared, Data Poisoning, Byzantine, and Sybil attacks all manifest through the contribution of malicious synthetic datasets to $\mathcal{S}$. Hence, they can be analysed and mitigated using identical defence mechanisms[172–174]. Participant $i \in \mathcal{I}$, where $i \notin \mathcal{B}$, constructs their aggregated dataset W_i by weighting the synthetic datasets S_j from all participants according to weights ω_{ij} (by Assumption 4). The weights ω_{ij} are selected by participant i to optimise their local model's performance on their private validation set, utilising the Dynamic Dataset Aggregation (DDA) process. Including the malicious datasets S_b with $b \in \mathcal{B}$ in W_i with substantial weights would result in diminished performance on participant i's validation data. Therefore, during the DDA optimisation, participant i will assign negligible weights to all malicious datasets to minimise their validation loss. This yields:

$$\sum_{b \in \mathcal{B}} \omega_{ib} \leq \varepsilon. \tag{E.25}$$

Consequently, the cumulative impact of all malicious synthetic datasets on M_i is negligible, effectively mitigating both individual Data Poisoning and distributed

Byzantine/Sybil attacks. □

Attack Surface $\mathcal{S}$. The sole shared information is the collection of synthetic datasets $\mathcal{S}$. Malicious participants attempt to disrupt the training of honest participants by contributing malicious synthetic datasets to $\mathcal{S}$. In this context, Data Poisoning, Byzantine, and Sybil attacks all manifest similarly through this attack surface.

The Dynamic Dataset Aggregation (DDA) process employs the Tree-structured Parzen Estimator (TPE) to optimally assign weights ω_{ij} to each synthetic dataset S_j. The objective is to minimise the local model M_i's error on participant i's validation set by solving the following optimisation problem:

$$\min_{\omega_{ij}} \ \mathcal{L}(M_i, W_i) \quad \text{subject to} \quad 0 \leq \omega_{ij} \leq 1, \quad \forall j \in \mathcal{I}, \tag{E.26}$$

where $\mathcal{L}(M_i, W_i)$ denotes the loss function of M_i evaluated on the aggregated dataset $W_i = \bigcup_{j \in \mathcal{I}} \omega_{ij} S_j$. In the presence of malicious datasets $S_b b \in \mathcal{B}$, assigning higher weights ω_{ib} would increase the validation loss $\mathcal{L}(M_i, W_i)$. The TPE optimiser models the Expected Improvement (EI):

$$\mathrm{EI}(\omega) = \int_{-\infty}^{l^*} (l^* - l) \cdot p(l \mid \omega)\, dl, \tag{E.27}$$

where l^* is a threshold loss value and $p(l \mid \omega)$ is the probability of observing loss l given the weight configuration ω. Configurations resulting in lower losses are favoured. As the TPE iterates, it identifies that higher weights on malicious datasets lead to increased loss, and consequently reduces ω_{ib} for all $b \in \mathcal{B}$ to minimise $\mathcal{L}(M_i, W_i)$. Over successive iterations, this results in:

$$\sum_{b \in \mathcal{B}} \omega_{ib} \leq \varepsilon, \tag{E.28}$$

where ε is a negligible positive value. This adaptive weighting ensures that the cumulative influence of all malicious datasets on the local model M_i remains minimal. Furthermore, because Data Poisoning, Byzantine, and Sybil attacks

all attempt to influence the training process through malicious datasets in this context, the DDA process with TPE effectively defends against these attacks by assigning negligible weights to harmful datasets based on their impact on validation performance. This demonstrates the unified effectiveness of the DDA with TPE in mitigating these types of attacks within the FedKR framework.

Theorem E.4 ((-) Evasion Attack Inapplicability)**.** *Under the stated assumptions, the probability that an adversary $\mathcal{A}$ can successfully perform an evasion attack on the private model M_i using only the shared synthetic datasets $\mathcal{S}$ is negligible:*

$$P\left(\exists x_{adv} \in \mathcal{X} \text{ such that } M_i(x_{adv}) \neq y_{true} \text{ and } \mathcal{A}(\mathcal{S}) = x_{adv}\right) \leq \varepsilon. \tag{E.29}$$

Proof. Suppose, for contradiction, that there exists an adversary $\mathcal{A}$ capable of crafting an adversarial input $x_{\text{adv}} \in \mathcal{X}$ such that the private model M_i misclassifies it, i.e., $M_i(x_{\text{adv}}) \neq y_{\text{true}}$, with non-negligible probability $\delta > \varepsilon$, using only the shared synthetic datasets $\mathcal{S}$:

$$P\left(M_i(x_{\text{adv}}) \neq y_{\text{true}} \text{ and } \mathcal{A}(\mathcal{S}) = x_{\text{adv}}\right) \geq \delta. \tag{E.30}$$

However, under Assumption 1, the adversary has no access to the private model M_i, its architecture, parameters, or the private dataset D_i. Additionally, the adversary only has access to the shared synthetic datasets $\mathcal{S}$ (Assumption 5), which contain neither individual data points from D_i nor detailed information about M_i.

Without access to M_i or its gradients, $\mathcal{A}$ cannot compute perturbations δx necessary to craft adversarial examples that exploit specific vulnerabilities of M_i:

$$x_{\text{adv}} = x + \delta x, \quad \text{with } M_i(x_{\text{adv}}) \neq y_{\text{true}}. \tag{E.31}$$

An adversary might attempt to train a surrogate model $\tilde{M}_i$ using the synthetic

datasets $\mathcal{S}$ to approximate M_i:

$$\tilde{M}_i \approx M_i. \tag{E.32}$$

However, owing to Assumptions 2 and 3, each participant's model is trained independently on their private dataset with diverse data, rendering M_i significantly different from any surrogate model $\tilde{M}_i$ trained on $\mathcal{S}$. The approximation error between M_i and $\tilde{M}_i$ remains significant:

$$\forall x \in \mathcal{X}, \quad P\left(\tilde{M}_i(x) = M_i(x)\right) \leq \varepsilon. \tag{E.33}$$

Furthermore, the transferability of adversarial examples relies upon the similarity between the surrogate model and the target model. Given the independent training and diversity of models (from Assumptions 2 and 3), adversarial examples crafted against $\tilde{M}_i$ have a negligible probability of causing M_i to miss-classify:

$$P\left(M_i(x_{\text{adv}}) \neq y_{\text{true}} \mid \tilde{M}_i(x_{\text{adv}}) \neq y_{\text{true}}\right) \leq \varepsilon. \tag{E.34}$$

Therefore, the probability that $\mathcal{A}$ can successfully craft an adversarial input x_{adv} such that $M_i(x_{\text{adv}}) \neq y_{\text{true}}$ is negligible, contradicting equation (E.30). Thus, in conclusion:

$$P\left(\exists x_{\text{adv}} \in \mathcal{X} \text{ such that } M_i(x_{\text{adv}}) \neq y_{\text{true}} \text{ and } \mathcal{A}(\mathcal{S}) = x_{\text{adv}}\right) \leq \varepsilon. \tag{E.35}$$

□

Attack Surface $\mathcal{S}$. The adversary has access solely to the synthetic datasets $S = \{S_j\}_{j \in I}$. They might attempt to exploit this data to craft adversarial examples. Specifically, the adversary could attempt to:

1. Train a surrogate model $\tilde{M}_i$ using the synthetic datasets $\mathcal{S}$ in an attempt to approximate M_i. However, due to Assumptions 2 and 3, the models M_i and $\tilde{M}_i$ are trained on different datasets, and their architectures

and training strategies are private (Assumption 1), leading to significant discrepancies between them.

2. Craft adversarial examples against $\tilde{M}_i$ by finding inputs x_{adv} such that $\tilde{M}_i(x_{\text{adv}}) \neq y_{\text{true}}$. However, owing to the significant differences between $\tilde{M}_i$ and M_i, the likelihood that these adversarial examples will transfer to M_i and cause misclassification is negligible, as shown in equation (E.34).
3. Exploit shared synthetic data to find inherent vulnerabilities. However, the synthetic datasets $\mathcal{S}$ are generated independently by each participant's generator G_i (as per Assumption 2) and lack detailed information about any private model M_i (Assumption 1). Moreover, synthetic data is designed to approximate the overall data distribution without revealing specific weaknesses of participants' models.

Thus, the attack surface provided by $\mathcal{S}$ does not offer the adversary sufficient information or means to perform an effective evasion attack on M_i, confirming that:

$$P\left(\exists x_{\text{adv}} \in \mathcal{X} \text{ such that } M_i(x_{\text{adv}}) \neq y_{\text{true}} \text{ and } \mathcal{A}(\mathcal{S}) = x_{\text{adv}}\right) \leq \varepsilon.$$

Theorem E.5 ((-) Gradient and Parameter Leakage Infeasibility)**.** *Under the stated assumptions, the probability that an adversary $\mathcal{A}$ can extract either the gradients $\nabla_{\theta_i}\mathcal{L}_i$ or parameters θ_i of participant i's private model M_i utilising solely the shared synthetic datasets $\mathcal{S}$ is negligible:*

$$P\left(\mathcal{A}(\mathcal{S}) = \nabla_{\theta_i}\mathcal{L}_i \text{ or } \theta_i\right) \leq \varepsilon. \tag{E.36}$$

Proof. Proceeding by contradiction, suppose there exists an adversary $\mathcal{A}$ capable of extracting either the gradients $\nabla_{\theta_i}\mathcal{L}_i$ or parameters θ_i of participant i's private model M_i utilising solely the shared synthetic datasets $\mathcal{S}$, with non-zero probability.

However, under Assumption 1 (Privacy of Information), participant i maintains

strict privacy of all personal data and internal parameters, including the private model M_i and its parameters θ_i, as well as any computed gradients $\nabla_{\theta_i}\mathcal{L}_i$.

Moreover, Assumption 5 (Adversarial Access Constraints) explicitly states that the adversary $\mathcal{A}$ is restricted to accessing only the publicly shared synthetic datasets $\mathcal{S}$.

Given that participant i does not share any information regarding the gradients or parameters of M_i, and no model updates or gradients are communicated within the federation, the adversary has no direct access to θ_i or $\nabla_{\theta_i}\mathcal{L}_i$.

An adversary might attempt to approximate θ_i or $\nabla_{\theta_i}\mathcal{L}_i$ by training a surrogate model $\tilde{M}_i$ utilising the synthetic datasets $\mathcal{S}$ [176]. However, this approach proves ineffective for the following reasons:

1. According to Assumptions 2 and 3 (Generator Independence and Training Set Diversity), each participant trains their models independently on private datasets D_i that are not shared. Hence, the synthetic datasets $\mathcal{S}$ lack sufficient information to reconstruct or approximate M_i or its parameters θ_i accurately.
2. Any surrogate model $\tilde{M}_i$ trained on $\mathcal{S}$ would differ significantly from M_i due to differences in training data, architectures (which are private per Assumption 1), and training strategies. Therefore, the parameters $\tilde{\theta}_i$ of $\tilde{M}_i$ would not provide meaningful information about θ_i:

$$\left\|\theta_i - \tilde{\theta}_i\right\| \geq \varepsilon. \tag{E.37}$$

3. Since gradients $\nabla_{\theta_i}\mathcal{L}_i$ are computed privately by participant i during the training of M_i and are not shared (Assumption 1), the adversary has no means to access or estimate them.

Therefore, the adversary $\mathcal{A}$ cannot extract the gradients $\nabla_{\theta_i}\mathcal{L}_i$ or parameters θ_i of M_i utilising only $\mathcal{S}$, which contradicts the assumption of non-zero probability. Hence, the probability that $\mathcal{A}$ can extract the gradients or parameters of M_i is zero. □

Attack Surface $\mathcal{S}$. The attack surface $\mathcal{S}$ available to the adversary comprises solely the shared synthetic datasets $\mathcal{S} = \{S_j\}_{j\in\mathcal{I}}$. These synthetic datasets consist of synthetic data-label pairs generated by each participant's generator G_j and labelled by their teacher model T_j:

$$S_j = \left\{ \left(G_j(z_j^v),\ T_j\left(G_j(z_j^v)\right)\right) \right\}_{v=1}^{V_j}. \tag{E.38}$$

The adversary might attempt to approximate the private model M_i by utilising $\mathcal{S}$ to train a surrogate model $\tilde{M}_i$. However, due to Assumptions 2 and 3, the models are trained independently on different private datasets, and their architectures and training details are private (Assumption 1), leading to significant discrepancies between M_i and $\tilde{M}_i$.

Furthermore, the synthetic datasets $\mathcal{S}$ do not contain information about M_i's parameters or gradients, and the adversary cannot compute gradients or perform model inversion attacks without access to M_i or its outputs.

Therefore, the attack surface provided by $\mathcal{S}$ offers no viable means for the adversary to extract gradients or parameters of M_i, thus confirming that the probability of gradient or parameter leakage is zero.

Theorem E.6 ((●) Free-Riding Mitigation and Detection)**.** *Under the stated assumptions, whilst an adversary $f \in \mathcal{I}$ attempting to free-ride by contributing low-quality or irrelevant synthetic data S_f cannot compromise other participants' models due to the Dynamic Dataset Aggregation (DDA) process, they may nonetheless gain access to the benefits provided by the federation. The impact of S_f on any participant's model M_i remains negligible:*

$$\omega_{if} \leq \varepsilon, \quad \textit{for all } i \in \mathcal{I}. \tag{E.39}$$

Proof. Consider that participant $f \in \mathcal{I}$ is a free-rider who contributes a low-quality or irrelevant synthetic dataset S_f whilst attempting to benefit from the synthetic datasets shared by others [177].

Participant $i \in \mathcal{I}$ employs Dynamic Dataset Aggregation (DDA) to assign weights ω_{ij} to synthetic datasets S_j in order to optimise their local model M_i's performance on their private validation set.

Given that S_f is of low quality or irrelevant, including it in W_i would not enhance, or might indeed degrade, the performance of M_i on participant i's validation set.

Consequently, during the DDA optimisation, participant i shall assign a negligible weight to S_f:

$$\omega_{if} \leq \varepsilon. \tag{E.40}$$

As a consequence, the impact of S_f on M_i remains negligible, and participant i is not adversely affected by the free-rider's low-quality contribution.

However, participant f gains access to the shared synthetic datasets $\mathcal{S}$ and may potentially benefit from the high-quality data contributed by other participants.

Whilst the DDA process mitigates the negative impact of free-riding on honest participants, it does not prevent free-riders from accessing the benefits of the federation. Addressing this matter may require additional policies or mechanisms beyond the scope of this work. □

Attack Surface $\mathcal{S}$. The adversary attempts to exploit the federation by contributing low-quality or irrelevant synthetic data S_f to $\mathcal{S}$, with the aim of gaining access to the high-quality synthetic datasets shared by other participants without contributing meaningful data themselves.

Participants employ the Dynamic Dataset Aggregation (DDA) process to assign weights ω_{ij} to each synthetic dataset S_j, optimising their local model M_i's performance on their private validation set. The DDA process resolves the following optimisation problem:

$$\min_{\{\omega_{ij}\}} \mathcal{L}(M_i, W_i) \quad \text{subject to} \quad 0 \leq \omega_{ij} \leq 1, \quad \forall j \in \mathcal{I}, \tag{E.41}$$

where $\mathcal{L}(M_i, W_i)$ denotes the loss function of M_i evaluated on the aggregated dataset $W_i = \bigcup_{j \in \mathcal{I}} \{\omega_{ij} S_j\}$.

The Tree-structured Parzen Estimator (TPE) optimiser is employed to efficiently search the weight space $\{\omega_{ij}\}$ for configurations that minimise the validation loss. The TPE models the Expected Improvement (EI) as:

$$\mathrm{EI}(\omega) = \int_{-\infty}^{l^*} (l^* - l) \cdot p(l \mid \omega)\, dl, \tag{E.42}$$

where l^* is a threshold loss value, and $p(l \mid \omega)$ is the probability of observing loss l given the weight configuration ω.

In the context of free-riding, the low-quality or irrelevant synthetic dataset S_f contributes little to no improvement in M_i's performance on the validation set. Assigning higher weights ω_{if} to S_f would not decrease, and may indeed increase, the validation loss $\mathcal{L}(M_i, W_i)$.

As the TPE optimiser iterates, it identifies that assigning negligible weights to S_f leads to better validation performance. Consequently, the DDA process reduces ω_{if} to minimise $\mathcal{L}(M_i, W_i)$. Over successive iterations, this results in:

$$\omega_{if} \leq \varepsilon, \tag{E.43}$$

where ε is a negligible positive value.

This adaptive weighting ensures that the impact of the free-rider's low-quality dataset S_f on honest participants' models M_i remains minimal.

However, since participants gain access to the shared synthetic datasets $\mathcal{S}$ upon contributing any synthetic dataset, even free-riders like f can access the high-quality synthetic data contributed by others. The DDA process mitigates the negative effects on honest participants but does not prevent free-riders from benefiting from the federation.

To fully address free-riding, additional mechanisms such as contribution-based

access control or incentive schemes may be required to detect and discourage such behaviour, which are beyond the scope of this work.

Theorem E.7 ((●) Property Inference Mitigation)**.** *Under the stated assumptions, for any adversary $\mathcal{A}$ attempting to infer a property ϕ of participant i's private dataset D_i using the shared synthetic datasets $\mathcal{S}$, the probability that $\mathcal{A}$ can accurately estimate $\phi(D_i)$ within a small error δ remains negligible:*

$$P\left(|\mathcal{A}(\mathcal{S}) - \phi(D_i)| \leq \delta\right) \leq \varepsilon, \tag{E.44}$$

for any small $\delta > 0$.

Proof. Let us consider that an adversary $\mathcal{A}$ has access solely to the publicly shared synthetic datasets $\mathcal{S}$ (per Assumption 5) and aims to infer a property ϕ of participant i's private dataset D_i.

The synthetic dataset S_i is generated by the generator G_i, which is trained on D_i. Whilst S_i may reflect certain statistical properties of D_i at the population level, the generative process introduces randomness and abstraction that prevent the disclosure of individual data points.

According to principles analogous to differential privacy, a randomised mechanism provides strong privacy guarantees if the inclusion or exclusion of a single individual's data does not significantly affect the output distribution. Although G_i does not implement differential privacy, the randomness inherent in the generative process serves a similar purpose in protecting individual data.

Moreover, operational parameters such as the latent space dimension l, the architecture of G_i, and training strategies remain private (per Assumption 1). Without access to these parameters and the internal workings of G_i, the adversary cannot accurately reconstruct D_i or deduce individual-level information.

The adversary's estimate $\hat{\phi}_i = \mathcal{A}(\mathcal{S})$ can only approximate $\phi(D_i)$ to the extent that S_i reflects $\phi(D_i)$. Due to the randomness introduced by G_i and the variability inherent in synthetic data generation, there exists significant error in the

adversary's estimation:

$$\left|\hat{\phi}_i - \phi(D_i)\right| \geq \delta.$$

Therefore, the probability that $\mathcal{A}$ can accurately estimate $\phi(D_i)$ within a small error δ remains negligible, bounded by ε.

Thus, the privacy of individual data points in D_i is preserved, and property inference attacks are mitigated. □

Attack Surface $\mathcal{S}$. The adversary attempts to analyse the shared synthetic datasets $\mathcal{S}$ to infer properties ϕ of the private dataset D_i. Whilst S_i may expose approximate statistical information about the population-level distribution of D_i, the individual data points remain protected.

The generative process of G_i introduces randomness and abstraction, rendering it difficult for the adversary to extract precise information about D_i. This is analogous to the concept of differential privacy, where adding noise to the output ensures that individual contributions cannot be discerned. Although differential privacy is not explicitly applied here, the effect of G_i's randomness serves a similar protective function.

Furthermore, since the adversary lacks access to G_i's parameters, the latent space $\mathcal{Z}_i$, and other operational details (as per Assumption 1), they cannot reverse-engineer the generative process to isolate individual data characteristics.

Therefore, even though the synthetic data may reflect some high-level properties of the underlying dataset, the adversary cannot accurately infer specific properties $\phi(D_i)$ of the private data within a negligible error margin. The privacy of individual participants is thus preserved, and property inference attacks are substantially mitigated.

Theorem E.8 ((-) Temporal Attack Inapplicability)**.** *Under the stated assumptions, the probability that an adversary $\mathcal{A}$ can infer private information about*

participant i based on timing information remains negligible:

$$P\left(\mathcal{A} \textit{ infers private information from timing data}\right) \leq \varepsilon. \tag{E.45}$$

Proof. Assume, for contradiction, that there exists an adversary $\mathcal{A}$ capable of inferring private information about participant i's data D_i or model M_i based on timing information, with non-negligible probability $\delta > \varepsilon$.

In the FedKR framework, participants share synthetic datasets $\mathcal{S}$ at discrete intervals or upon joining the federation. There exist no frequent or synchronous updates exchanged between participants, and no model parameters, gradients, or other sensitive information are shared (per Assumption 1 and Assumption 5). The communication between participants and any central server is limited to the transmission of synthetic datasets, which are shared in a manner that remains independent of the participants' private data D_i and models M_i.

Furthermore, operational parameters, such as training times, computational resources, and network communication patterns, remain private (per Assumption 1). The adversary $\mathcal{A}$ has no access to the internal timing of participants' training processes or any hardware-level interactions.

Given that the only information available to $\mathcal{A}$ is the time at which synthetic datasets are shared, and considering that these times are not correlated with the sensitive information in D_i or M_i, the adversary cannot leverage timing information to infer private data.

Moreover, any timing information related to the server where synthetic datasets are collected pertains solely to the network properties of that server and does not provide insights into the participants' private data or models.

Therefore, the probability that $\mathcal{A}$ can successfully perform a temporal attack to infer private information about participant i remains negligible, contradicting the assumption of non-negligible probability δ. Hence, we conclude that:

$$P\left(\mathcal{A} \text{ infers private information from timing data}\right) \leq \varepsilon.$$

□

Attack Surface $\mathcal{S}$*.* The adversary might attempt to utilise the timing of shared synthetic datasets to infer private information about participant i. However, since the sharing of synthetic datasets occurs at discrete and possibly randomised intervals, and remains independent of the participants' private data and models (per Assumption 1), the timing information does not correlate with sensitive information.

Additionally, the adversary cannot interact with the participants' hardware or observe their internal processes. The only timing information available pertains to the server where synthetic datasets are collected, which may reveal certain properties of the server's network but not of the participants' private data.

Therefore, the attack surface provided by timing information remains minimal, and temporal attacks are inapplicable in this setting.