A Comprehensive Survey on Data Distillation: Techniques, Frameworks, and Future Directions

Qaiser Razi, Somya Singh, Riya Priyadarshini, Vikas Hassija, and GSS Chalapathi, *Senior Member, IEEE,*

***Abstract*—The increased adoption of machine learning techniques has led to exponential growth in data generation and utilization. This growth has necessitated efficient storage, processing, and utilization of this data, which presents critical challenges, particularly in resource-constrained environments such as the Internet of Things (IoT) and edge devices. Data distillation has emerged as a promising solution that reduces dataset size while preserving essential information and optimizing computational resources. This survey provides a comprehensive analysis of data reduction techniques, covering methodologies such as knowledge distillation, coreset selection, hyperparameter optimization, and generative modeling. We further explore various data distillation learning frameworks, including performance, gradient, parameter, and distribution matching, highlighting their effectiveness in different data modalities such as images, graphs, and text. Furthermore, we examine the implications of data distillation in key areas such as continual and federated learning, privacy preservation, security, healthcare, IoT applications, and edge computing. By enabling lightweight models with minimal computational overhead, data distillation facilitates real-time inference and decision-making on edge devices, making it highly relevant for low-power, bandwidth-limited environments. Data distillation offers numerous advantages in improving model efficiency, reducing training costs, and enhancing privacy. However, data distillation faces numerous challenges related to scalability, computational complexity, and information retention. This survey identifies these challenges and outlines potential future research directions, providing insights for researchers seeking to leverage data distillation for scalable and efficient machine-learning applications.**

***Index Terms*—Data Distillation, Computational Resource, Knowledge Distillation, Core-Set Selection, Generative Modeling, Parameter Matching, and Kernel Ridge Regression.**

I. INTRODUCTION

The rapid surge in global data generation has prompted the International Data Corporation (IDC) to estimate that the total data sphere will reach 175 zettabytes by 2025 [1]. With groundbreaking advancements in Artificial Intelligence (AI), deep learning has made remarkable strides, driving significant progress in natural language processing (NLP) [2], computer vision [3], and speech recognition [4]. These innovations have enabled the creation of highly complex models capable of performing intricate tasks with exceptional precision. However, the ever-growing data volumes required for training such models have introduced major challenges related to storage, computational efficiency, and scalability.

Dataset distillation has emerged as a key solution to these challenges. This technique compresses large datasets into smaller, information-dense representations, retaining the most essential patterns for model training [5]. By reducing dataset size without compromising critical information, distillation enhances computational efficiency and minimizes storage requirements. This method is gaining significant traction across both academic research and industrial applications, transforming how machine-learning models are developed and deployed. With this approach, AI systems can achieve high performance while consuming fewer resources.

Despite these improvements, the exponential growth of data continues to present obstacles. Managing massive datasets demands extensive storage and substantial processing power, resulting in higher computational costs. As AI models become more complex, their training requirements increase dramatically. For instance, GPT-3, with its 175 billion parameters [6], relies on enormous datasets and consumes vast amounts of energy during training. In such cases, dataset distillation becomes essential, enabling the creation of highly compressed yet effective datasets, which reduces both storage needs and training time while preserving model accuracy [7]. The "scale-is-everything" concept emphasizes that enhancing AI performance depends on training larger models with more parameters on massive datasets, backed by increasingly powerful computing resources [8]. This strategy has led to the success of well-known neural networks like AlexNet [3], ResNet [9], BERT [2], and DALLE [10]. These models have unlocked new AI capabilities by achieving good accuracy. However, these models require large computational resources, so scalability is the major issue.

Handling these large datasets demands tremendous effort in data gathering, storage, and computation. It also depends upon the hardware, such as GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units), to train state-of-the-art AI systems, which is very expensive. It is very difficult to maintain the proper balance between accuracy and computational resources, as day by day, the volume of data is increasing. One of the major issues any model faces is that it loses acquired knowledge when it is exposed to new information. This is the major issue continual learning faces, where the model must learn new knowledge without losing the older one [11]. Without effective data handling strategies, models are unable to preserve important past information, which reduces their overall reliability.

Dataset distillation provides a practical alternative,

Qaiser Razi and GSS Chalapathi are with the Department of Electrical and Electronics Engineering, Birla Institute of Technology and Science, Pilani, Pilani Campus, Vidya Vihar, Pilani, Rajasthan 333031, India (e-mail: p20210070@pilani.bits-pilani.ac.in, gssc@pilani.bits-pilani.ac.in).

Somya Singh, Riya Priyadarshini, and Vikas Hassija are with the School of Computer Engineering, KIIT University, Bhubaneswar, Odisha, India -751024, (email: 2205510@kiit.ac.in, 2205664@kiit.ac.in, vikas.hassijafcs@kiit.ac.in).

particularly for devices with constrained resources such as smartphones, IoT systems, and sensors. Since these devices have limited memory and processing capacity, running large AI models directly is often not feasible. Distillation creates smaller but informative datasets, allowing efficient on-device training while reducing storage and communication overheads [12]. This is especially useful in federated learning, where training takes place across multiple devices without sending raw data to a central server, improving both privacy and system efficiency [13]. By keeping essential information while reducing dataset size, distillation makes AI training more cost-effective, faster, and easier to scale. As AI continues to expand, dataset distillation will play an important role in enabling accessible and efficient learning, particularly in resource-limited environments. The acronyms used are listed in Table I.

TABLE I: Acronyms and Definitions

Acronyms	Definitions
DD	Dataset Distillation
DC	Dataset Condensation
KD	Knowledge Distillation
RANSAC	Random Sampling and Consensus
VAEs	Variational Auto-encoders
GANs	Generative Adversarial Networks
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GMM	Gaussian Mixture Model
SVM	Support Vector Machine
KRR	Kernel Ridge Regression
KIP	Kernel Inducing Points
FRePo	Feature Reproduction
NNGP	Neural Network Gaussian Process
TBPTT	Truncated Backpropagation Through Time
IDC	Instance Discrimination Contrastive
PSG	Prediction, Similarity, and Grouping Model
Gcond	Gradient Matching Condensation
DM	Distribution Matching
CAFE	Counterfactual Adversarial Feature Encoding
IT-GAN	Information Theoretic Gen-Adversarial Network
FEDAVG	Federated Averaging
MAML	Model-Agnostic Meta-Learning
PGD	Projected Gradient Descent
LLM	Large Language Models

A. Motivation

The exponential growth of data has fueled remarkable progress in AI but has also exposed practical limitations in terms of scalability, efficiency, and privacy. While large-scale datasets enable the development of powerful machine-learning models, their sheer size often hampers accessibility and efficiency. Researchers studied several techniques to reduce dataset sizes without sacrificing performance to balance efficiency with effectiveness. One popular technique, core-set or instance selection [14], tries to minimize dataset sizes by preserving indicative samples; they are still bounded by raw-data usage, not only restricting flexibility with hard budgets on data but also with possible leakage of sensitive information. These weaknesses thereby call for more sophisticated techniques that move beyond sample selection.

Data Distillation can fill this gap by creating a smaller synthetic but useful dataset that retains the original data's key characteristics while decreasing the computational and storage requirements [15]. The synthetic samples created by data distillation can give comparable efficiency to the original datasets and provide privacy guarantees. Optimization-driven, generative, and privacy-sensitive developments over the last few years brought about the area's rapid growth, while it remains highly fragmented. This motivated us to conduct a comprehensive survey that not only explains the different methods and techniques used for distillation but also explores its various applications in the real world, its various challenges, and future research directions. This type of survey is essential for guiding researchers working in the field of data distillation to build scalable, privacy-preserving, and resource-efficient AI systems across various domains like edge computing, IoT, healthcare, and autonomous systems.

B. Our Contributions

Dataset distillation has the potential to redefine the way machine learning models are trained, offering an efficient means of data compression without compromising performance. Despite the rapid progress in this field, a structured and in-depth study of existing techniques and their applications remains lacking. This survey aims to fill this gap by providing a comprehensive review of dataset distillation methods, their advancements, and their real-world implications. We systematically analyze key algorithms, compare their performance, and discuss their applications, limitations, and future research directions.

Our key contributions are as follows:

1) We have presented a coherent and organized outline of data distillation.
2) We have done a comprehensive performance comparison of the various data distillation algorithms.
3) The current survey gives a detailed account of various applications of dataset distillation.
4) It also provides an in-depth analysis of limitations, including the potential future scope of data distillation.

C. Structure of the paper

The paper is structured as follows: Section II introduces the background work done in this field. Section III discusses different data reduction techniques like knowledge distillation, core-set or instance selection, hyper-parameter optimization, and generative modeling. Section IV covers data distillation learning frameworks like performance matching, which includes meta-model matching and kernel-ridge regression. The other learning frameworks discussed are gradient matching, parameter matching, and distribution matching. Different data topologies where data distillation techniques

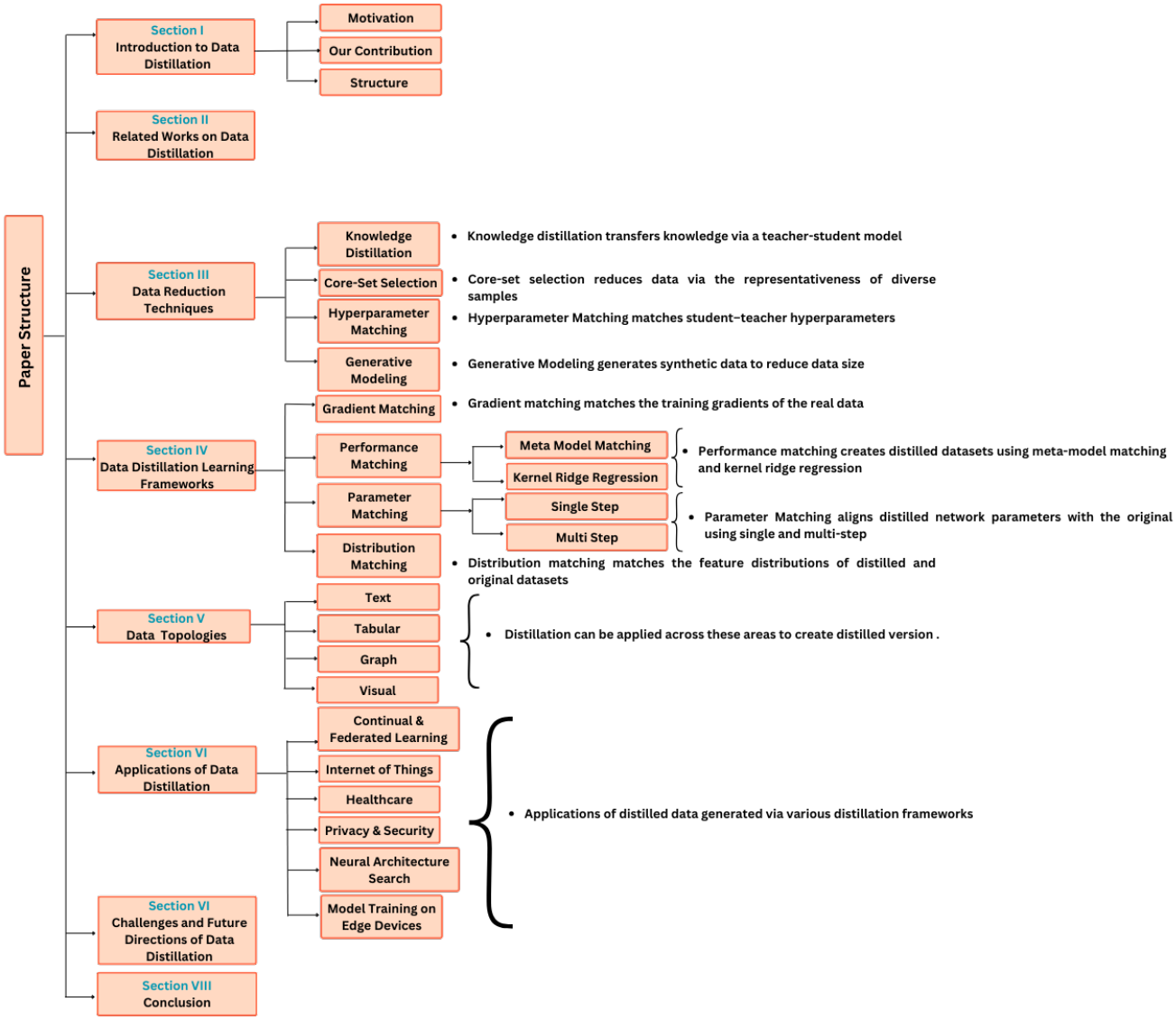

Fig. 1: SURVEY OVERVIEW

can be applied are investigated in Section V. Section VI explores various applications of data distillation techniques, highlighting their crucial role in preserving privacy while also reducing computational time and resource consumption. Section VII discusses various limitations of data distillation, its future aspects, and emerging trends in privacy engineering technologies. Finally, Section VIII summarizes the lessons learned and concludes the paper. The organizational overview of this survey is also shown in Fig. 1.

II. Related Works

Dataset Distillation techniques have emerged as a promising approach. These methods create concise yet informative summaries from extensive datasets, ensuring that only the most essential knowledge is retained for training machine learning models. Reducing the dataset size while preserving critical information provides a viable solution for handling data complexity. Several survey articles have explored data distillation from different perspectives, emphasizing techniques, applications, and challenges.

Lee *et al*. [16] reviewed synthetic dataset generation, discussing dataset distillation in relation to data augmentation, self-supervised learning, and meta-learning. Wang, Kai *et al*. [17] analyzed how dataset distillation intersects with generative modeling, particularly GAN-based methods. Their research explores the potential of synthetic data generation in improving dataset quality and efficiency. Geng *et al*. [18] provided a comprehensive survey on dataset distillation, proposing a taxonomy that categorizes methods based on learning frameworks, enhancement strategies, and data modalities. Their study systematically reviews existing dataset distillation approaches and highlights their role in continual learning, privacy preservation, and neural architecture search. The survey also discusses optimization challenges and possible improvements in distillation strategies. Ruonan Yu *et al*. [19] explored dataset distillation techniques, focusing on privacy-aware distillation and resource-efficient learning.

TABLE II: Related Surveys on Dataset Distillation

Reference	Description	Contributions	Limitations
Lee *et al*. [16]	Reviews synthetic dataset generation, discussing dataset distillation in relation to data augmentation, self-supervised learning, and meta-learning.	Highlights the intersection of dataset distillation with other learning paradigms, providing insights into how these techniques can be integrated.	Does not provide a unified framework for integrating these techniques.
Wang, Kai *et al.* [17]	Analyzes how dataset distillation intersects with generative modeling, particularly GAN-based methods. Explores the potential of synthetic data generation in improving dataset quality and efficiency.	Investigates the synergy between generative models and dataset distillation, emphasizing their role in improving the fidelity and diversity of synthetic data.	Lack of systematic benchmarking, making it difficult to compare GAN-based distillation with other techniques.
Geng *et al*. [18]	Introduces Dataset Distillation, a method to compress large datasets into a small set of synthetic images that retain the original dataset's learning properties.	Proposes dataset distillation to synthesize compact training datasets.	This paper provides fewer real-world experimental comparisons to validate the effectiveness of techniques.
Ruonan Yu *et al.* [19]	Presents a taxonomy of DD methods, including performance matching, parameter matching, and distribution matching, and discusses theoretical interconnections between these techniques.	Establishes connections between performance, parameter, and distribution matching methods.	The paper is dense and technical, making it difficult for non-experts to grasp key insights quickly, and it lacks a clear discussion of application and data topologies.
Noveen Sachdeva *et al.* [20]	Presents a formal framework, categorizes different data distillation approaches, and explores its applications across various data modalities.	Reviews various techniques such as meta-model matching, gradient matching, trajectory matching, and distribution matching. Highlights applications in differential privacy, federated learning, continual learning, and neural architecture search.	The paper primarily categorizes data distillation methods but lacks a deep quantitative comparison of different techniques.
Lei *et al.* [21]	Classifies existing dataset distillation methods into meta-learning, data matching, and factorized approaches while exploring challenges related to scalability, multimodal applications, and label complexity.	Reviews factorized dataset distillation techniques that use latent representations instead of direct data compression.	Limited exploration of interdisciplinary applications, such as dataset distillation for natural language processing or structured data.
Our Work	Provided the most recent and comprehensive survey on dataset distillation, covering emerging methods (e.g., gradient matching, distribution matching) and practical applications across diverse domains.	Presented a coherent and organized outline of dataset distillation, along with a comprehensive comparison of existing distillation techniques. Covers a wide range of applications spanning healthcare, IoT, and edge AI. Delivers an in-depth analysis of limitations and explicitly highlights future research directions and opportunities.	Focuses specifically on dataset distillation, without extending detailed comparisons to other data reduction techniques.

They analyzed dataset distillation's applications in federated learning and model compression, emphasizing how synthetic datasets can be optimized for performance while preserving user data confidentiality. Their review also compares dataset condensation with related data selection and knowledge distillation techniques. Sachdeva *et al.* [20] explored data distillation beyond standard datasets, applying distillation techniques to images, graphs, and recommender systems. Their work identifies challenges in dataset summarization across different modalities, emphasizing how distillation can enhance data efficiency in domain-specific learning tasks. They also differentiate dataset distillation from related fields such as data pruning and knowledge distillation, providing an extensive comparison of existing methods.

Lei *et al.* [21] presented a detailed survey on dataset distillation, classifying existing approaches into meta-learning and data-matching frameworks. Their work highlights the computational challenges of distillation methods, particularly when dealing with high-resolution datasets and complex label spaces. They also discuss optimization techniques to improve dataset quality and generalization in distilled datasets. Wang *et al.* [15] focused on backpropagation-based distillation, demonstrating its effectiveness in mitigating adversarial attacks and handling computationally expensive tasks such as data poisoning defense and adversarial robustness. Their study emphasized how optimizing data representation through distillation can significantly reduce training time and enhance model generalization. Nguyen *et al.* [22] investigated Kernel Inducing Points (KIP), a method that distills datasets by optimizing kernel-based representations. Their research showcased the effectiveness of KIP in reducing training costs while incorporating privacy-preserving mechanisms like ρ-corruption, a technique designed to prevent data leakage in federated learning. Beyond traditional data distillation, researchers have extended its applications to graph learning, continual learning, and domain adaptation. Jin *et al.* [23] proposed GCond, a distillation technique for graph learning, addressing the challenge of condensing structured data while optimizing for neural architecture search (NAS). Their research demonstrated that graph distillation could improve model efficiency without requiring access to the full dataset. Zhao and Bilen [24] explored DM (Dataset Matching) for domain adaptation, ensuring distilled datasets generalize well across various tasks by matching the feature distributions between different domains. These studies collectively provide significant insights into data distillation, covering its applications in privacy, adversarial robustness, continual learning, and efficiency improvements. While each work presents unique methodologies, as shown in Table II, our survey aims to consolidate these perspectives, emphasizing the

intersection of data efficiency, security, and domain-specific optimizations in modern distillation methodologies.

III. Data Reduction Techniques

As machine learning models grow increasingly, training on large datasets becomes computationally expensive and resource-intensive. Dataset Distillation (DD) addresses this challenge by condensing a large dataset into a much smaller, synthetic subset while retaining the model's predictive power. This enables efficient training and deployment, particularly in resource-constrained environments like edge computing and real-time applications.

Dataset distillation builds upon several well-established data reduction techniques, each contributing unique strategies for compressing and preserving essential information. One such approach is Knowledge Distillation (KD), introduced by Hinton *et al.* [25], where knowledge from a complex teacher model is transferred to a simpler student model. By using soft probability distributions instead of hard labels, KD preserves rich class relationships and improves training efficiency. Another key technique is Hyperparameter Optimization (HPO), which is widely used in automated machine learning (AutoML) to fine-tune parameters such as learning rate and batch size. Methods like Grid Search, Bayesian Optimization, and Genetic Algorithms optimize hyperparameters to enhance model performance with minimal data. Core-set selection, originating from active learning and computational geometry, identifies a small but representative subset of data that approximates the full dataset's distribution, reducing training costs while maintaining accuracy. Meanwhile, Generative Modeling, which gained prominence through deep generative networks such as GANs [26] and VAEs [27], enables the generation of synthetic datasets that retain the statistical properties of real data. Notably, dataset distillation shares conceptual similarities with knowledge distillation, where a large, complex model (teacher) transfers knowledge to a smaller model (student) while preserving its predictive ability. While knowledge distillation focuses on model compression, dataset distillation reduces data volume while retaining essential learning signals. These techniques collectively enhance dataset distillation, optimizing storage, training efficiency, and overall model performance. Some of the data reduction techniques are shown in Table III and are discussed below:

A. Knowledge Distillation

Knowledge Distillation (KD) transfers knowledge from a larger teacher model to a smaller student model, enabling comparable performance with reduced computational cost [28], [29] as shown in Fig. 2. Since its introduction by Hinton *et al.* [25], research has refined teacher–student architectures, distillation strategies, and knowledge representations to improve transfer effectiveness [30]. KD has been widely applied in model compression [31], [32] and ensemble learning [33]. While initially proposed as a model compression technique, KD's broader importance lies in how it formalizes the idea of capturing and reusing essential knowledge. This perspective directly inspires dataset distillation [15] instead of compressing parameters into a lightweight student; the goal becomes compressing data itself into a smaller yet equally informative set. In this way, KD provides both the conceptual and methodological foundation and preserves task-relevant structure, which makes dataset distillation feasible.

B. Core-set or Instance Selection

One of the traditional selection-based techniques for reducing the training set volume is core-set or instance selection. To ensure that the models using this methodology show results similar to those using the full dataset, a portion of the primary data containing the key samples is retained. As it is NP-hard (Nondeterministic Polynomial time-hard), to find the subgroup with the best performance, existing core-set methods primarily choose samples using heuristic procedures (trial-and-error methods). Several early approaches typically foresee a consistent distribution of data between core sets and primary data [34], [35], [36], [37]. For example, Random Sampling and Consensus (RANSAC) [38] repeatedly chooses random samples from the data and counts the number of data points that, within a tolerance, fit the model. The model with the highest consensus set becomes the core set. Any legitimate goal for core-set selection on a functional level can also be used for DD. DD and core-set, however, differ significantly from one another. In contrast to core sets, distilled datasets aren't always subsets of the original data. Instead, these are frequently brand-new, artificial data points refined to resemble the original dataset's training effectiveness. In addition, because core-set approaches are NP-hard, they usually rely on greedy tactics or heuristic criteria to strike a balance of efficiency and performance.

C. Hyper-parameter Optimization

Hyperparameter optimization automatically adjusts machine learning model parameters to achieve the best performance, reducing the need for manual fine-tuning while improving efficiency and accuracy, as depicted in Fig. 3. Hyperparameters, such as learning rate, batch size, and regularization strength, are crucial in determining how well a model learns from data. Traditional methods like grid search and random search systematically explore different hyperparameter values, while more advanced techniques such as Bayesian optimization and genetic algorithms aim to find optimal configurations more efficiently. Early studies, such as [39] and [40], used non-gradient model-based techniques to fine-tune models by evaluating errors after complete training. If each example in an artificial dataset is treated as a high-dimensional hyperparameter, Dataset Distillation (DD) can be seen as a type of hyperparameter optimization. However, DD and hyperparameter optimization have different goals. Hyperparameter optimization aims to improve model performance by adjusting parameters, while DD focuses on creating smaller, more efficient datasets that speed up training. Because of these differences, other DD techniques like distribution matching [22] and parameter matching [20] have little connection to hyperparameter optimization.

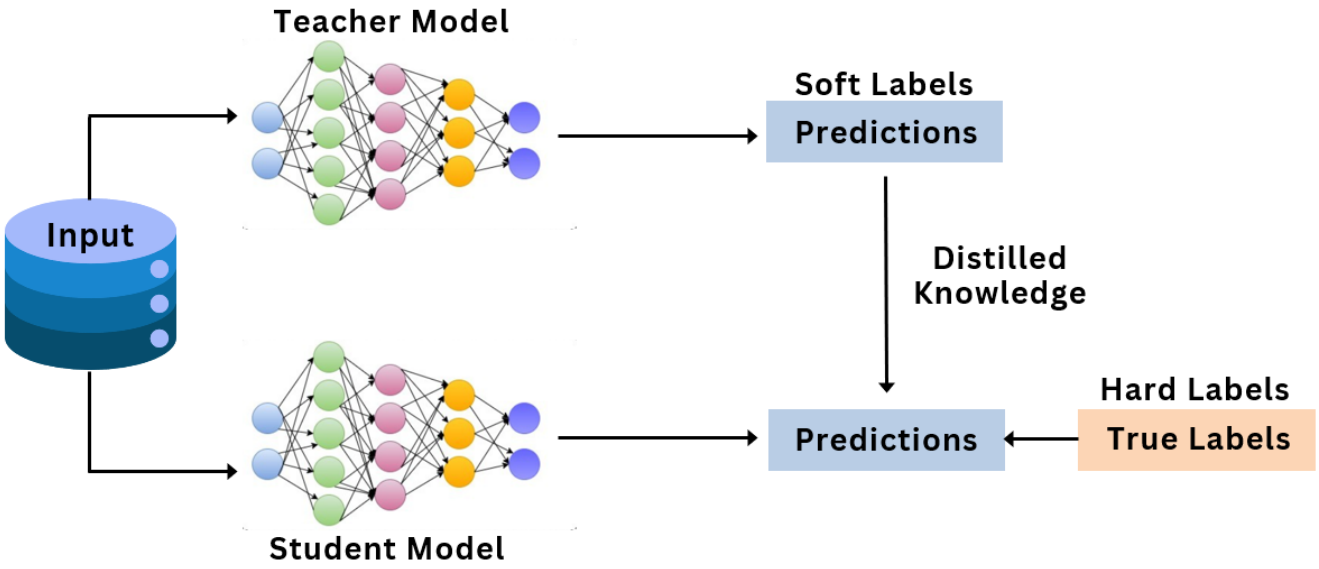

Fig. 2: Knowledge Distillation

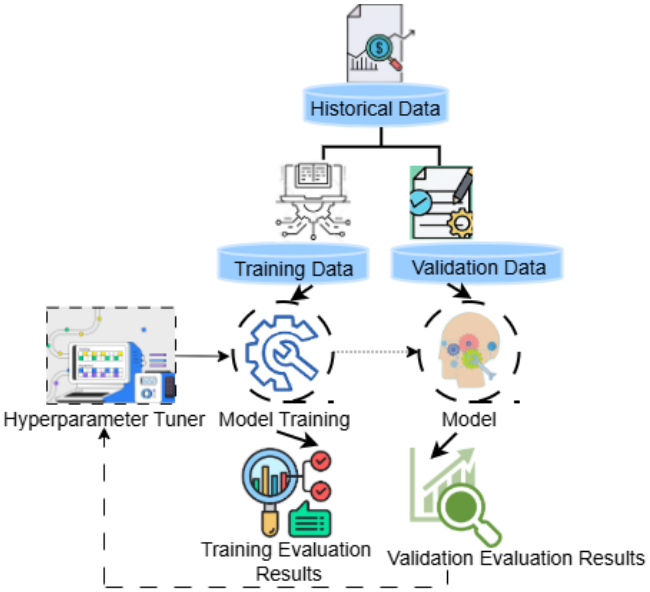

Fig. 3: Hyper-Parameter Optimization

D. Generative Modelling

Generative models such as Generative Adversarial Networks (GANs), auto-regressive models [26], [41], and Variational Auto-encoder (VAEs) [27] can synthesize data that reflects the underlying distribution of large datasets, as shown in Fig. 4. In the context of dataset distillation, this capability is particularly useful as it can generate small, representative synthetic datasets that either replace real data or serve as augmentations to improve robustness [42]. Beyond sample generation, these models compress information into low-dimensional latent spaces, enabling efficient storage and faster training. Unlike meta-learning-based distillation, where optimization explicitly constructs condensed samples, generative models produce them implicitly through learned distributions. Generative models, therefore, act as a bridge between raw data and distilled representations. By learning compact distributions, they can capture the essential statistical structure of large datasets while discarding redundancy. This makes them a promising tool for scaling dataset distillation, especially in scenarios where direct access to the full dataset is costly or impractical.

IV. Data Distillation Learning Frameworks

Dataset distillation is an emerging technique that aims to condense large datasets into smaller yet highly informative subsets, as shown in Fig. 5, enabling efficient training while preserving model performance [15]. As machine learning continues to scale, handling vast amounts of data becomes increasingly challenging due to high storage and computational costs. Training on full datasets often leads to inefficiencies, making it crucial to identify and retain only the most essential data points. Dataset distillation achieves this by synthesizing or selecting representative samples that encapsulate the knowledge of the original dataset. These condensed datasets not only accelerate training but also enhance generalization by filtering out noise and redundant information. Additionally, distilled data facilitates model deployment on resource-constrained devices, making machine learning more accessible across various domains. However, balancing data reduction with model efficacy requires strategic methodologies that ensure distilled datasets retain meaningful learning signals.

Data learning frameworks provide structured approaches to optimize dataset distillation. These frameworks define how distilled data is generated, selected, and refined to maintain essential learning signals while significantly reducing data volume. Techniques such as meta-model matching, parameter matching [22], and distribution matching [43] enable models to retain critical training dynamics as shown in Fig. 6, ensuring effective generalization despite working with distilled data. The following sections delve deeper into these frameworks,

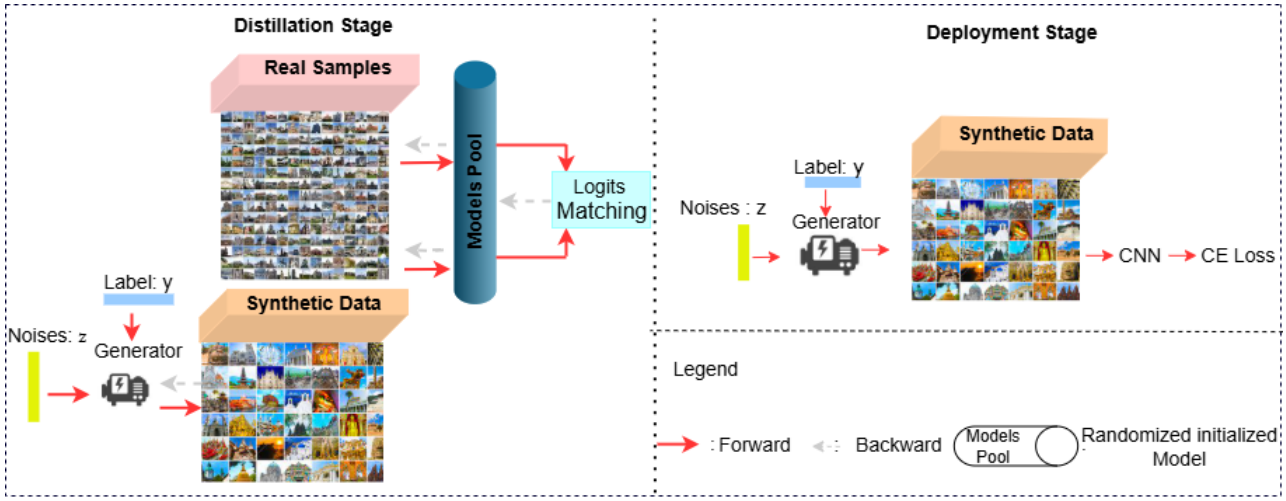

Fig. 4: Generative Modelling

TABLE III: Summary of Reduction Techniques

Technique	Description	Advantage	Disadvantage
Knowledge Distillation [25]	A smaller model (student) learns to replicate a larger model (teacher) to achieve similar performance.	Reduces model size and complexity and improves inference speed.	Loss of accuracy requires more computational resources.
Core-set Selection [34]	Selecting a representative subset of the original dataset while preserving essential patterns.	Reduces computational costs and maintains model performance.	Difficulty in selecting accurate representative subsets, and potential loss of important data.
Hyperparameter Optimization [39]	Tuning parameters such as learning rates or batch sizes to enhance model performance and generalization.	Improves model accuracy and convergence speed, and reduces manual tuning effort.	Computationally expensive, prone to overfitting.
Generative Modelling [42]	Models learn to generate new data samples resembling a given dataset's distribution.	Useful for data augmentation, anomaly detection, and generating synthetic data.	Struggles with realistic output representation, hindering quality.

exploring their methodologies and contributions to advancing dataset distillation.

A. Performance matching

Performance matching-based dataset distillation techniques focus on creating distilled datasets that achieve comparable model performance to training on the full dataset. These methods aim to maximize the generalization ability of models trained on the distilled data while maintaining efficiency. Some of the performance-matching techniques are discussed below:

1) Meta-model matching: Meta-model matching-based data distillation techniques optimize the transferability of models learned on the condensed datasets at the time of application to the primary dataset.

$$\arg\min_{D_{\text{syn}}} \mathcal{L}_{D_{\text{syn}}}\left(\theta^{D_{\text{syn}}}\right) \quad \text{s.t.} \quad \theta^{D_{\text{syn}}} \triangleq \arg\min_{\theta} \mathcal{L}_{D_{\text{syn}}}(\theta) \quad (1)$$

The outer loop systematically streamlines the data summary process, ensuring the application of an optimized learning algorithm to the primary dataset. Meanwhile, the inner loop iteratively refines a representative learning algorithm until it converges. However, Eq. 1 is computationally expensive regarding memory and processing time, necessitating additional assumptions in related methods to improve efficiency.

The data distillation was first proposed by Wang *et al.* [15], which optimizes the system using the meta-model matching framework. DD uses two methods to make the optimization in Eq. 1; one is local optimization in the inner loop using stochastic gradient descent (SGD), and the second is streamlining the loop outside using Truncated Back-Propagation Through Time (TBPTT), which involves unrolling a restricted number of loops that remain inside.

$$\arg\min_{D_{\text{syn}}} \mathop{\mathbb{E}}_{\theta_0 \sim \mathbf{P}_\theta} \left[\mathcal{L}_{D_{\text{syn}}}\left(\theta_T\right)\right] \quad \text{s.t.} \quad \theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla_\theta \mathcal{L}_{D_{\text{syn}}}\left(\theta_t\right) \quad (2)$$

Here, η represents a configurable learning rate, P symbolizes a parameter initialization distribution of choice, and T stands for truncation in Truncated Backpropagation Through Time (TBPTT). D_{syn} represents the synthetic (distilled) dataset we want to optimize. $L_{D_{\text{syn}}}$ represents the loss function evaluated on the real dataset D, using model parameters $\theta_{D_{\text{syn}}}$. Notably,

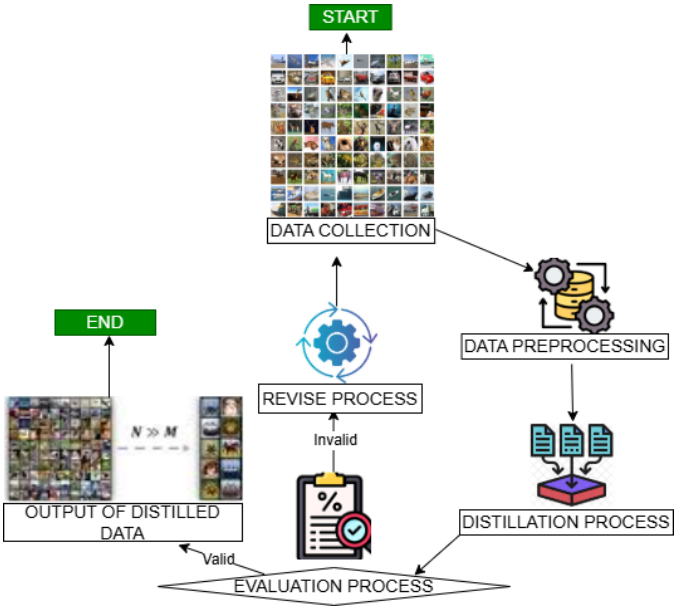

Fig. 5: Data Distillation

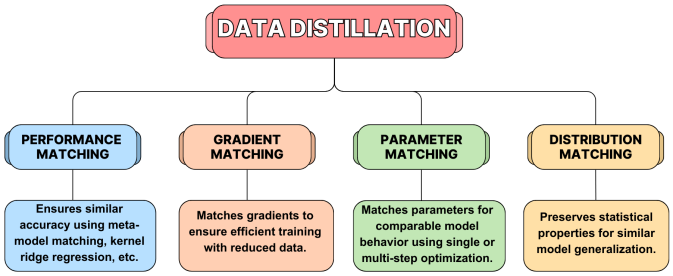

Fig. 6: Data Distillation Learning Frameworks

TBPTT has several limitations, such as:

1) Inner-loop unrolling is computationally expensive.
2) Truncated unrolling introduces bias [44].
3) Loss landscapes can be poorly conditioned, especially during lengthy unrolls [45].

2) Kernel-ridge regression: The method above uses meta-learning and bi-level optimization to calculate validation loss gradients. This process requires computationally expensive outer optimization steps and a significant amount of GPU memory for inner loops. The inadequate number of inner loops hinders optimization and limits results, making it unsuitable for scaling to bigger frameworks. To address this challenge, a set of techniques based on Kernel Ridge Regression (KRR) has been introduced [46], [22], [47]. KRR employs convex optimization, offering an exact solution for the linear framework and eliminating the need for resource-intensive inner loop training. For the regression model, $f(\boldsymbol{x}) = \boldsymbol{w}^\top \psi(\boldsymbol{x})$, where $\psi(\cdot)$ is a nonlinear mapping and the respective kernel is $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \psi(\boldsymbol{x}), \psi(\boldsymbol{x}') \rangle$, there exists a closed form solution for w when the regression model is made to learn on $\mathcal{S}$ with KRR:

$$\boldsymbol{w} = \psi(X_s)^\top \left(\mathbf{K}_{X_s X_s} + \lambda I\right)^{-1} y_s \tag{3}$$

where $\mathbf{K}_{X_s X_s} = \left[K(\boldsymbol{s}_i, \boldsymbol{s}_j)\right]_{ij} \in \mathbb{R}^{n \times n}$ is called the kernel matrix or Gram matrix associated with K and the dataset S and $\lambda > 0$ is a fixed regularization parameter. Therefore, the mean square error (MSE) of predicting T with the model trained on S is

$$\mathcal{L}(\mathcal{S}) = \frac{1}{2} \left\| y_t - \mathbf{K}_{X_t X_s} \left(\mathbf{K}_{X_s X_s} + \lambda I\right)^{-1} y_s \right\|^2 \tag{4}$$

, where $\mathbf{K}_{X_t X_s} = \left[K(\boldsymbol{x}_i, \boldsymbol{s}_j)\right]_{ij} \in \mathbb{R}^{m \times n}$. The meta-gradient of the loss is subsequently utilized to amend the condensed sample. Because of the exact solution in KRR, θ does not require a repetitive update, and the gradient's backward pass avoids the recursive calculation graph. Nguyen *et al.* [22], present a novel technique called neural feature regression with pooling (FRePo) [46], which replaces the neural tangent network (NTK) in the knowledge inference pipeline (KIP) by using a more flexible conjugate kernel with neural features [48], [49], [50], [51]. Meanwhile, Random Feature Approximation for Distillation (RFAD) was proposed by Loo *et al.* [47], employing the Neural Network Gaussian Process (NNGP) [52], [53] kernel to replace the NTK used in KIP. This approach minimizes the computation of the Gram matrix to $\mathcal{O}(|\mathcal{S}|)$, which is linear with a number of elements of the synthetic set, as compared to $\mathcal{O}\left(|\mathcal{S}|^2\right)$, the complexity of accurately calculating the NTK kernel matrix.

B. Gradient matching

Gradient matching-based data distillation optimizes a synthetic dataset D_{syn} such that its training gradients approximate those of the original dataset D. This approach was first introduced by Zhao *et al.* [54] in their work on Dataset Condensation (DC). The optimization objective, as shown in Eq. 5, minimizes the distance between the gradients of a model trained on D and D_{syn} over T-steps of training. The process assumes T-step inner-loop optimization for computational traceability, local smoothness of the parameter space, and first-order approximation of the model trajectory using D_{syn}. By avoiding the unrolling of the inner loop, DC achieves a significant reduction in computational overhead compared to meta-model matching frameworks:

$$\arg\min_{D_{\text{syn}}} \mathbb{E}_{\theta_0 \sim P_\theta, c \sim C} \left[\sum_{t=0}^{T} D\left(\nabla_\theta \mathcal{L}^c_D(\theta_t), \nabla_\theta \mathcal{L}^c_{D_{\text{syn}}}(\theta_t) \right) \right] \quad \text{s.t.} \quad \theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla_\theta \mathcal{L}_{D_{\text{syn}}}(\theta_t). \tag{5}$$

Lee *et al.* [49] proposed a significant extension of DC by incorporating class-contrastive signals to improve stability and performance. Their approach, called DCC (Dataset Condensation with Contrastive signals) refines the gradient-matching process by matching class-averaged gradients across the original dataset D and the synthetic dataset D_{syn}. The optimization objective for DCC, as shown in Eq. 6, ensures that gradients from all classes $c \in C$ are aligned effectively, resulting in better representation of class-specific features and more stable optimization:

$$\arg\min_{D_{\text{syn}}} \mathbb{E}_{\theta_0 \sim P_\theta} \left[\sum_{t=0}^{T} D\Big(\mathbb{E}_{c \in C}[\nabla_\theta \mathcal{L}^c_D(\theta_t)], \mathbb{E}_{c \in C}[\nabla_\theta \mathcal{L}^c_{D_{\text{syn}}}(\theta_t)] \Big) \right] \tag{6}$$

Kim *et al.* [55] further advanced gradient matching through their Instance Discrimination Contrastive (IDC) framework, addressing key limitations in prior work. Their contributions include a multi-formation strategy, where D_{syn} is stored at a lower resolution to reduce memory overhead and upscaled during usage, and matching gradients over the entire dataset D rather than just D_{syn}. This approach overcomes challenges such as strong coupling between inner and outer loop optimizations and vanishing gradients due to the small size of D_{syn}. The optimization objective for IDC, shown in Eq. 7, ensures robust and scalable optimization:

$$\arg\min_{D_{\text{syn}}} \mathbb{E}_{\theta_0\sim P_\theta, c\sim C}\left[\sum_{t=0}^{T} D\left(\nabla_\theta \mathcal{L}_D^c(\theta_t), \nabla_\theta \mathcal{L}_{f(D_{\text{syn}}^c)}(\theta_t)\right)\right] \quad \text{s.t.} \quad \theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla_\theta \mathcal{L}_D(\theta_t). \tag{7}$$

These equations represent a progressive refinement of gradient-matching techniques for data distillation. DC established a computationally efficient framework for matching gradients between D and D_{syn}. DCC improved upon this by incorporating class-contrastive signals for better class representation and optimization stability. IDC further enhanced the framework by addressing scalability and robustness through multi-formation and full-dataset gradient matching, providing a state-of-the-art approach to data distillation.

C. Parameter matching

Zhao *et al.* [54] initially introduced the matching parameters of neural networks in the data distillation method, which have since been further explored in several subsequent studies [47], [56]. The fundamental principle behind parameter matching is to train the same network using various parameters, unlike performance matching, which maximizes the performance of networks trained on synthetic datasets. They encourage the consistency of their trained neural parameters using synthetic and original datasets for specific steps. Two streams can be distinguished within parameter matching approaches based on the number of training steps using $\mathcal{S}$ and $\mathcal{T}$: single-step and multi-step parameter matching.

1) Single Step Parameter Matching: This approach focuses on updating a neural network in a single step using synthetic data $\mathcal{S}$ and real training data $\mathcal{T}$, ensuring that the gradients computed during optimization remain consistent with respect to the model parameters Θ. Specifically, after each update using synthetic data, the network is further trained on real data for multiple steps to refine the gradients. The objective function in this instance is expressed as:

$$\begin{aligned} \mathcal{L}(\mathcal{S},\mathcal{T}) &= \mathbb{E}_{\theta^{(0)}\sim\Theta}\left[\sum_{t=0}^{T} \mathcal{D}\left(\mathcal{S},\mathcal{T};\theta^{(t)}\right)\right] \\ \theta^{(t)} &= \theta^{(t-1)} - \eta\nabla l\left(\mathcal{S};\theta^{(t-1)}\right) \end{aligned} \tag{8}$$

where metric $\mathcal{D}$ measures the distance between gradients $\nabla l(\mathcal{S};\theta^{(t)})$ and $\nabla l(\mathcal{T};\theta^{(t)})$.

Since only a single-step gradient is necessary, and updates of synthetic data and networks are decomposed, this approach is memory-efficient compared with meta-learning-based performance matching. However, this method has notable limitations. The distance metric evaluates gradients for each class independently, failing to capture the interconnections between different classes. As a result, the approach does not effectively preserve class-discriminative features, which can affect the quality of the generated synthetic dataset. Wang *et al.* [15] introduced modifications to address these limitations by utilizing early-stage models to construct a candidate model pool. Instead of relying on entirely random models, they apply weight perturbations to a subset of early-stage models within this pool to enhance model diversity during dataset distillation. This improvement speeds up the process by reducing the need for excessive random networks and optimization iterations while still producing synthetic datasets with equivalent performance. By ensuring a more diverse selection of models in the early stages, the method improves the efficiency of dataset distillation and enhances the representational quality of the synthetic data.

2) Multi-Step Parameter Matching: When synthetic data is used to update models across multiple steps during an evaluation, error accumulation can occur due to the reliance on single-step gradient matching in parameter updates. To address this issue, Cazenavette *et al.* [61] introduced a multi-step parameter matching method known as Matching Training Trajectory (MTT). This approach selects and initializes θ from training trajectory checkpoints obtained from the primary dataset, ensuring a more stable and representative parameter selection. By extending beyond single-step updates, MTT improves the alignment between synthetic and real training data over multiple iterations, leading to more accurate synthetic dataset generation.

$$\begin{aligned} \mathcal{L}(\mathcal{S},\mathcal{T}) &= \mathbb{E}_{\theta^{(0)}\sim\Theta}\left[\mathcal{D}\left(\theta_{\mathcal{S}}^{(T_s)}, \theta_{\mathcal{T}}^{(T_t)}\right)\right] \\ \theta_{\mathcal{S}}^{(t)} &= \theta_{\mathcal{S}}^{(t-1)} - \eta\nabla l\left(\mathcal{S};\theta_{\mathcal{S}}^{(t-1)}\right) \\ \theta_{\mathcal{T}}^{(t)} &= \theta_{\mathcal{T}}^{(t-1)} - \eta\nabla l\left(\mathcal{T};\theta_{\mathcal{T}}^{(t-1)}\right) \end{aligned} \tag{9}$$

In this context, $\mathcal{D}$ measures the difference between the optimized parameters obtained using the synthetic dataset $\mathcal{S}$ and those obtained using the real dataset $\mathcal{T}$. Here, l represents the loss function of the learning model, and Θ denotes the distribution of stored initialization states. This formulation ensures that the synthetic dataset leads to parameter updates that closely match those from the full dataset, improving model efficiency and generalization.

$$\mathcal{D}\left(\theta_{\mathcal{S}}^{(T_s)}, \theta_{\mathcal{T}}^{(T_t)}\right) = \frac{\left\|\theta_{\mathcal{S}}^{(T_s)} - \theta_{\mathcal{T}}^{(T_t)}\right\|^2}{\left\|\theta_{\mathcal{T}}^{(T_t)} - \theta^{(0)}\right\|^2} \tag{10}$$

An important aspect of this approach is the normalization of the loss using the distance between the expert endpoint $\theta_{\mathcal{T}}^{(T_t)}$ and the initial parameter state $\theta^{(0)}$. This normalization helps to adjust for magnitude differences across neurons and enhances the signal's robustness, particularly in later training epochs when expert updates become smaller. By leveraging this multi-step parameter matching technique, the

TABLE IV: Summary of Distillation Frameworks and Applications

References	Distillation Framework	Data Modalities	Computationally Demanding Task	Privacy	Robustness	Applications
Wang *et al.* [15]	Back Propagation Through Time	Image	—	—	—	Data Poisoning
Sucholutsky e*t al.* [57]	Back Propagation Through Time	Image, Text	—	—	—	Data Compression
Nguyen *et al.* [22]	Kernel Ridge Regression	Image	—	ρ-corruption ion	—	Efficient Model Training
Zhou *et al.* [46]	Kernel Ridge Regression	Image	Continual Learning	Membership Inference Attack	Noise-less regression	Model Compression
Kim *et al.* [55]	Parameter Matching	Image	Continual Learning	—	—	Dataset Condensation
Jin *et al.* [23]	Parameter Matching	Graph	Neural Architecture Search	—	—	Graph Learning
Zhao and Bilen [24]	Distribution Matching	Image	Continual Learning , Neural Architecture Search	—	—	Domain Adaptation
K. Wang *et al.* [58]	Distribution Matching	Image	—	—	Detecting distribution shift	Robust Learning
H. Brendan *et al.* [59]	Gradient Matching	Image	Federated Learning	—	—	Federated Learning
Chelsea Finn *et al.* [60]	Gradient Matching	Image	Reinforcement Learning	—	—	Meta Learning

model significantly outperforms its single-step counterpart, resulting in a more refined and effective synthetic dataset distillation process.

D. Distribution matching

Although the above parameter-wise matching shows satisfying performance, Zhao and Bilen [24] visualized the distilled data in a two-dimensional plane and revealed that there is a large distribution discrepancy between the distilled data and the target data. In other words, the distilled dataset cannot comprehensively cover the data distribution in the feature space. Based on this, they proposed a technique to match the synthetic and target data from a distribution perspective for dataset distillation.

The synthetic data are optimized according to the following objective function:

$$\min_{\mathcal{S}} \mathbb{E}_{\theta\sim P_\theta} \left\| \frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|} f_\theta(\hat{x}_i) - \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|} f_\theta(x_i) \right\|^2 \tag{11}$$

where f_θ is parameterized by θ, and θ is sampled from a random distribution P_θ. $|\mathcal{S}|$ and $|\mathcal{T}|$ are the cardinality of dataset $\mathcal{S}$ and $\mathcal{T}$, respectively.

As shown in Eq. (11), distribution matching does not rely on the model parameters and eliminates bilevel optimization, thereby reducing memory requirements. However, despite its computational efficiency, it empirically underperforms gradient and trajectory-matching approaches. The reason behind this limitation is that distribution matching primarily aligns feature representations at the dataset level rather than enforcing fine-grained instance-level alignment. This could lead to situations where synthetic data fails to accurately replicate the detailed variations of target data, affecting downstream model performance.

To improve distribution matching, researchers have explored various enhancements, such as incorporating multi-layer feature alignment and leveraging additional regularization strategies. By matching feature representations at multiple layers of a neural network, the method can better capture both low-level and high-level feature information, improving the generalization of the distilled dataset. Additionally, combining distribution matching with other distillation strategies, such as gradient matching, can lead to a more comprehensive optimization framework that balances efficiency and accuracy. Data distillation can be applied using these learning frameworks in various data topologies like text, images, tabular, and graphs, as shown in Table IV, which are discussed in the following section.

V. Data Topologies

Dataset distillation offers a robust solution for handling diverse data types, including graphs, images, text, and tabular datasets. By selectively extracting and retaining essential information, it reduces computational expenses while improving model accuracy and efficiency. This technique

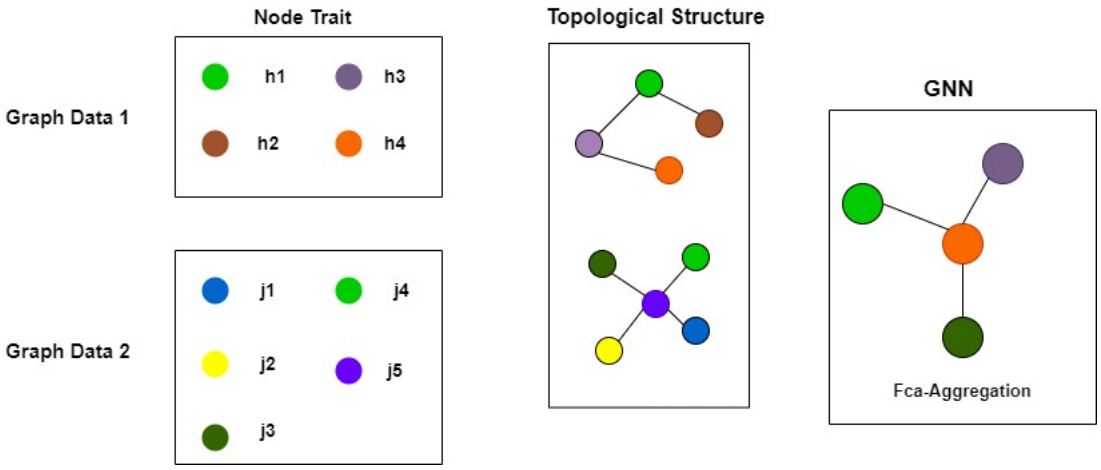

Fig. 7: Depiction of Graph Data

ensures that distilled datasets maintain the key features necessary for effective training, regardless of data format.

Graph-based data presents unique challenges due to its non-Euclidean nature, requiring specialized distillation techniques to preserve the relationships between nodes and edges. On the other hand, tabular data, composed of numerical and categorical values, demands methods that safeguard statistical distributions and feature dependencies, ensuring the distilled dataset retains its predictive power for downstream tasks. In addition to graphs and tabular data, other topologies, such as grid-based image representations and sequential text structures, further expand the applicability of dataset distillation. The following sections explore these topologies in detail, highlighting their distinct characteristics and the role of dataset distillation in optimizing learning across them.

A. Graph

Graphs can be utilized to simulate various data and applications, such as social networks [62], user-item interactions [63], and autonomous driving [64]. Liu *et al.* [65] use distribution matching to improve performance and demonstrate how significantly efficient the data distillation was. For a few datasets, they compressed 99 % of the data to obtain 99 % of the original performance. Graph distillation is a simple solution to most size problems, as shown in Fig. 7; however, producing compact, high-fidelity graphs has the following challenges:

1) Scalability: Graphs follow intrinsic patterns (e.g., spatial correlations [66]) that must be preserved in distilled representations. Graphs often grow to billions of nodes and edges in domains like social networks, e-commerce, or biological data. For example, in fraud detection on platforms like Facebook or payment networks, the full transaction/social graph is far too large to train on directly. A distilled graph must preserve spatial or community structures while dramatically reducing size so that models can still detect anomalies such as fraud rings. If scalability is not handled properly, critical small clusters may disappear during distillation, undermining detection accuracy.
2) Loss of Information: Simplifying a graph can lead to the loss of essential structural details, reducing accuracy in downstream tasks.
3) Preservation of Key Properties: Maintaining essential properties such as community structure and node centrality in the distilled graph presents significant challenges.
4) Heterogeneity: Real-world graphs often contain heterogeneous types of edges and nodes (for example, networking systems of a social kind encompassing different types of relationships). Distilling such graphs while preserving their heterogeneity is challenging because collapsing diverse relations into a simplified structure can distort the task-relevant patterns. For example, in healthcare networks, nodes may represent patients, doctors, and treatments, with edges denoting diagnosis, prescription, or referral relationships. If heterogeneity is not preserved during distillation, the resulting graph may fail to capture critical cross-type interactions, such as how treatment choices influence patient outcomes, thereby limiting the usefulness of the distilled graph for predictive tasks.

B. Visual

Most dataset distillation techniques have been applied to image datasets [15], [18]. Dataset distillation approaches most commonly prefer images due to their complexity, easy visualization, optimized deep learning tools, and significant practical applications in fields like computer vision and medical imaging. The visual nature of images allows for effective compression, as redundant and less informative pixels can be removed while preserving critical features, as shown in Fig. 8. This selective retention reduces dataset size, leading to lower storage requirements and faster model training, ultimately cutting computational costs. Additionally, by distilling high-dimensional image data into representative synthetic subsets, deep learning models require fewer training iterations, minimizing energy consumption and hardware dependency. From MNIST, CIFAR10, and SVHN to more

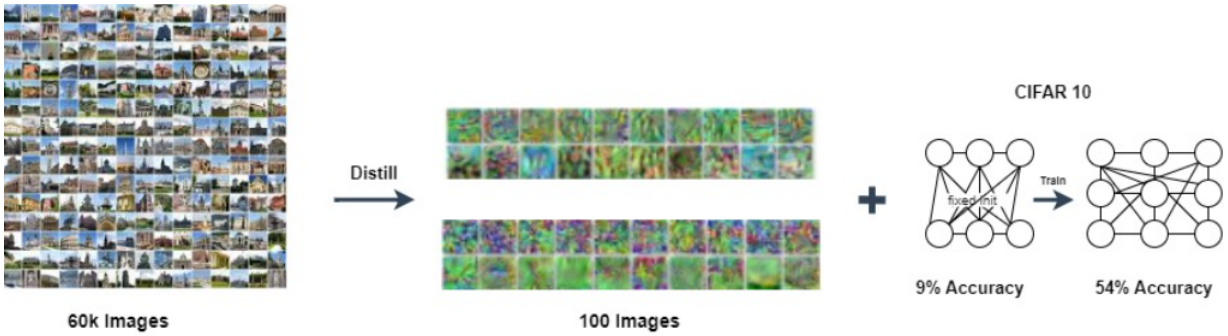

Fig. 8: Visualizing Visual Data

difficult datasets like TinyImageNet and ImageNet, we observe that experimental datasets get more and more complex [61], [67]. As these datasets grow in size and detail, dataset distillation techniques become even more essential in maintaining efficiency. In real-world applications, this has proven highly beneficial for fields such as autonomous driving, where models trained on large-scale image datasets must make split-second decisions, and in medical imaging, where reducing dataset size without losing diagnostic accuracy is crucial for faster and more cost-effective AI-powered analysis.

Visual data poses significant challenges due to its large size and the need for high precision in tasks like object recognition, image segmentation, and classification. As image datasets continue to grow in size and complexity, traditional machine learning methods often struggle with the heavy computational demands they impose. Dataset distillation offers a powerful solution by creating smaller but still representative subsets of the original data. These compact datasets preserve the most important visual features, enabling faster, more efficient model training without compromising accuracy. This is especially critical for real-time applications, where both speed and precision are essential. In computer vision, the impact of distillation is already apparent. In autonomous driving, for example, vehicles must analyze camera feeds instantly to make safe decisions. Training on distilled datasets helps reduce inference latency and improves efficiency, making the models more practical for deployment in low-power, real-world systems. Similarly, in medical imaging, distillation ensures that crucial diagnostic details in X-rays, MRIs, or CT scans are preserved while dramatically reducing dataset size. This not only accelerates model training and development but also supports faster, more accurate diagnoses for healthcare professionals, reducing the burden of data handling and storage.

Beyond performance, dataset distillation also delivers significant cost and sustainability benefits. By shrinking massive image datasets into smaller, information-rich subsets, storage needs, processing time, and memory usage drop substantially. This translates into lower energy consumption, fewer hardware requirements, and ultimately, a smaller carbon footprint. Since distilled datasets also allow models to converge in fewer epochs, researchers can experiment with more advanced architectures at a fraction of the computational cost, a major advantage in a resource-constrained environment. Overall, dataset distillation makes managing large-scale image data more practical, allowing for efficient training, faster inference, and greener AI practices. It is rapidly becoming an indispensable tool for advancing deep learning in domains such as autonomous driving, healthcare, and other vision-based applications.

C. Text

Today, we have access to an overwhelming amount of textual data from sources like websites, news platforms, and academic papers. Large-scale collections, such as the common crawl, capture snapshots of the internet and provide datasets that can reach staggering sizes in some cases, up to 541 terabytes. These massive resources serve as a goldmine for training large language models and advancing research in natural language processing. But while the availability of such data is invaluable, its sheer size and complexity create serious challenges. Storing, processing, and training on hundreds of terabytes of raw text is far from practical for most systems. This is where dataset distillation becomes critical. By condensing these enormous datasets into smaller, information-rich subsets, we can make them much more manageable and computationally efficient for model training without losing the patterns and representations that matter most. As textual data continues to expand at a massive scale, developing effective distillation techniques is becoming essential to make large-scale AI training both feasible and efficient.

Additionally, the innovation of Large Language Models (LLMs) has triggered the training cost escalation of such models on vast datasets, as illustrated by the works of Brown *et al.* [6], Thoppilan *et al.* [68], and Devlin *et al.* [62]. It has not yet been determined whether or not large-scale textual data can be efficiently distilled. The primary barriers to extracting textual data are as follows: The data's intrinsic discreteness means that a token should only appear in a small vocabulary. The underlying structure is complex, with sentences adhering to predetermined grammar patterns. The richness of context leads to varied semantic interpretations for a given text segment depending on various scenarios. Sucholutsky and Schonlau [57] introduce a latent-embedding strategy for textual data distillation. They tackle the discreteness issue in optimization at a high level by conducting distillation within a continuous embedding framework. To create continuous

representations for each word in the condensed text, the authors use a latent space defined using a static text encoder. They then enhance these representations by optimizing them utilizing the TBPTT data distillation framework [69].

D. Tabular

Tabular data structured in rows and columns used in healthcare, finance, and business analytics poses unique challenges for dataset distillation. Unlike images or text, which often contain redundancy that can be compressed, tabular data is made up of a mix of numerical and categorical variables [70]. This makes distillation more complex because the relationships between features and underlying statistical patterns must be carefully preserved. To tackle this, techniques such as prototype selection, synthetic data generation, and generative modeling are used to produce smaller yet representative datasets while maintaining predictive accuracy [71]. One of the core challenges is ensuring that distilled data reflects feature correlations, missing values, and domain-specific constraints accurately [72]. While image data can be reduced without significant loss, tabular data requires more refined strategies, such as differentiable data selection and gradient-based optimization, to retain its structure and meaning. These approaches not only reduce storage needs and speed up training but also help prevent overfitting by focusing on truly informative samples [73].

Beyond efficiency, dataset distillation for tabular data also improves interpretability and privacy. By eliminating redundancy, models generalize better, and in sensitive domains like medical research, distilled data can enable privacy-preserving machine learning. The integration of distillation with federated learning further strengthens security by allowing decentralized training while simultaneously reducing communication costs. As structured datasets continue to expand across industries, refining distillation techniques will play a crucial role in building scalable, cost-effective, and trustworthy AI applications.

VI. APPLICATIONS

The growing demand for efficient machine learning solutions has driven the need for smarter data management techniques. By creating smaller, highly informative datasets, modern approaches are addressing challenges such as catastrophic forgetting in continual learning, enhancing federated learning efficiency, and optimizing IoT system performance. These innovations are transforming fields like healthcare, privacy, and security while also accelerating model robustness and neural architecture search, as shown in Fig. 9. With its wide-ranging impact, this data-centric strategy is shaping the future of artificial intelligence and its real-world applications. The following sections highlight its key areas of influence.

A. Continual and federated learning

Continual learning, extensively reviewed by Parisi *et al.* [74], frequently confronts the challenge of catastrophic forgetting [75], where previously acquired patterns deteriorate upon exposure to new data or tasks. Data distillation emerges as a potent remedy for this challenge. This technique involves condensing data into succinct summaries, continuously updating them, and storing them in a replay buffer. These summaries serve as crucial representations for subsequent training, as evidenced by studies [76], [77], [78]. Deng and Russakovsky [79] present supplementary proof, highlighting the superiority of a compress-then-recall strategy over the present continual learning methods. Significantly, only data overviews are preserved for respective tasks, and for every new task, a new model undergoes learning from scratch, utilizing previous data overviews.

In contrast, federated or collaborative learning, comprehensively explored in Li *et al.* [80] survey, employs a decentralized training approach. Traditionally, federated learning involves synchronizing local parameter updates with a central server rather than sharing unprocessed data, as suggested in [81]. However, the concept of data distillation introduces an alternative approach, diminishing the requirement of aligning complex frameworks across distributed endpoints and central units. Instead, compactly generated data overviews are transmitted to the central server, training occurring solely on it, minimizing communication overhead. Overall, the integration of data distillation within continual and federated learning presents a promising avenue for more efficient and scalable learning systems. By leveraging compact and informative representations, continual learning can mitigate forgetting, while federated learning can reduce communication overhead and privacy concerns. Future research should explore optimizing these approaches further to enhance adaptability and robustness in real-world applications.

B. Internet Of Things

Integrating IoT with data distillation holds immense potential in fields like agriculture, healthcare, smart cities, and manufacturing. IoT devices generate vast real-time data streams, presenting challenges such as storage constraints, high computational costs, and difficulties in extracting actionable insights. Data distillation addresses these issues by condensing large datasets into essential, manageable forms, enabling efficient analysis and decision-making. Continuous data on medication adherence and vital signs is collected in healthcare by IoT devices such as wearables and remote monitors [82]. Resource pressure and a delay in insights can result from managing this raw data. In order to improve proactive interventions, illness monitoring, and individualized treatment plans, data distillation helps streamline the data and enables healthcare providers to swiftly extract important information [83]. Also, distilled data helps telemedicine platforms by lowering storage and bandwidth needs and improving telemonitoring and remote consultations. IoT sensors monitor environmental variables, crop health, and soil moisture in agriculture [84]. Traditional systems may be overwhelmed by the amount of data, making prompt decision-making more difficult. Farmers can improve animal

Fig. 9: Applications of Data Distillation

management techniques, identify crop illnesses early, and optimize irrigation by using data distillation, which makes the data simpler. This leads to higher yields, improved resource efficiency, and sustainable farming [85]. The computational load of managing raw data is lessened by IoT-driven supply chain systems that use distilled data for product traceability and quality assurance. Additionally, IoT systems in smart cities keep an eye on public safety, energy consumption, and traffic. The massive volume of data may put a burden on the city's infrastructure, causing delays and inefficiency. Data Distillation can help in reducing this burden by reducing the dataset size. Reducing the dataset size can help in faster analysis, which can help in enhancing various services and development. The reduced dataset size can help in maximizing the performance of IoT devices by increasing operational effectiveness, responsiveness, and resource management.

C. Healthcare

Incorporating data distillation into healthcare provides a revolutionary solution to effectively managing and using large medical datasets efficiently. High-volume medical data, such as patient information, imaging data that will be used for diagnosis, and treatment, requires high computational power to process. Data distillation eliminates the requirement for high resources by creating a condensed dataset of the larger datasets [86]. The advanced techniques like machine learning and natural language processing can help the expert in diagnosis and better identify the trends in patient health [87]. To train these models, huge computational resources are required as the size of the dataset is very large. So, the dataset distillation can minimize this computational resource requirement by creating a condensed version of the same data, which can give the same efficiency compared to the larger datasets. Distilling datasets into compact, high-quality versions accelerates data analysis and reduces infrastructure requirements, boosting overall system performance [88]. This is especially crucial in real-time medical decision-making and research settings, where speed and accuracy are paramount.

Additionally, while protecting patient privacy, healthcare informatics makes it easier to create anonymized and de-identified datasets for safe data exchange between researchers and healthcare practitioners. Healthcare informatics facilitates the distillation process and unifies many data sources by increasing the quality and efficiency of data processing and analysis, paving the way for precision healthcare applications and personalized medicine [89]. By providing these summarized datasets, data distillation not only preserves critical information but also enables more effective and faster decision-making, ultimately improving healthcare outcomes while reducing the infrastructure costs required for handling vast amounts of raw medical data.

D. Privacy, Security, and Robustness

Machine learning is defenseless against various privacy attacks like membership inference [90], and property inference attacks [91], where attackers try to extract task-independent private information from the target model and potentially reconstruct the primary training data. Dataset distillation offers an approach that begins with the data alone, preserves privacy, and strengthens the model's robustness. For instance, distributed learning (remote training) [92] transmits generated data collections with condensed annotations rather than genuine datasets to preserve data privacy. Dong *et al.* [93] highlight dataset distillation as a potential method for data privacy protection, highlighting its potential to prevent unintentional data leakage and its theoretical connection to differential privacy. Moreover, Chen *et al.* [94] provide high-dimensional data with differential privacy assurances to share private data privately at a reduced memory and computing expense using DD. As for the robustness of the model, DD enhances robustness by compressing the larger model to a smaller and more informative model that retains essential training information. This process reduces noise and redundancy, promoting better generalization and faster training. It helps the model focus on critical features, implicitly regularizing and reducing overfitting. It combines adversarial training and KIP Kernel Inducing Points (KIP) [22] to enhance the learning data rather than model variables with great effectiveness. Additionally, it offers adequate resilience against Projected Gradient Descent (PGD) attacks [95]. Dataset distillation enhances privacy preservation by significantly reducing the size of the training data, which limits the exposure of sensitive information. By condensing the dataset into essential representations, it becomes more challenging for attackers to extract specific details about the original data, thus mitigating risks associated with privacy

attacks such as membership inference [96]. Additionally, the distillation process aligns with differential privacy principles, offering a secure framework for sharing data with minimal risk of leakage [97]. To ensure that trained networks remain resilient against adversarial attacks, robust dataset learning focuses on optimizing datasets through a min-max tri-level optimization problem, aiming to minimize the robust error of adversarial inputs on models parameterized by robust datasets [24], [98].

E. Neural Architecture Search

Neural Architecture Search (NAS) is a procedure that looks through hundreds of network options to discover the best architecture for a given dataset to improve generalization. To save training time, the NAS procedure often involves training the network candidates on a tiny representation of the primary dataset. Based on these trained network candidates, the generalization ranking can be determined. However, as it involves training multiple models and selecting the best-performing one based on validation, NAS is well known to be costly. Researchers have suggested employing a proxy-distilled dataset as a stand-in for the whole model as a solution to the problem and successfully determining the optimal network. Classical techniques, such as greedy search and random selection, have been developed to create proxy datasets without changing the underlying data [99], [90]. Optimizing a deeply insightful collection of data as a stand-in for model option selection was first put forward by Such *et al.* [42]. Ensuing experiments on the CIFAR-10 dataset have considered NAS as an auxiliary task to test the put forward dataset distillation algorithms DSA [91], and DM [92]. Thus, simulation on the synthetic dataset allows for precise generation ranking for optimal architecture selection, significantly reducing training time.

Additionally, using distilled datasets for simulation allows for precise ranking of generated data, significantly reducing training time and enabling the selection of the optimal architecture. This is particularly beneficial for rapid model identification in resource-limited scenarios. Dataset distillation also supports real-time inference in isolated or disconnected environments by reducing the reliance on large datasets and enabling local predictions. In distributed systems or federated learning, distilling data before distribution helps minimize communication overhead, leading to more efficient model updates and deployments across multiple devices.

F. Efficient Model Training on Edge Devices

As the Internet of Things (IoT) and edge AI continue to grow, one of the biggest challenges is running deep learning models on devices with limited resources [22]. Traditional models rely on massive datasets and heavy computational power, which makes them unsuitable for smartphones, embedded hardware, or industrial IoT systems. This is where dataset distillation comes in. Instead of depending on large collections of raw data, dataset distillation creates smaller but highly informative datasets that capture the essential patterns needed for training. These distilled datasets allow models to be trained more efficiently while consuming far less memory and energy without sacrificing accuracy [100]. Techniques like Kernel Inducing Points (KIP) [101] and gradient-matching distillation [102] have shown that it's possible to keep models lightweight yet reliable. This approach is especially valuable for real-time AI applications such as autonomous drones, smart surveillance, and wearable healthcare devices. In these cases, models need to run directly on the device with minimal lag for tasks like navigation, anomaly detection, or health monitoring. By working with distilled datasets, the models can update quickly and adapt over time, reducing the need for frequent retraining on powerful cloud servers. Minimizing cloud dependence also improves real-time decision-making and helps cut down on communication delays and costs. For embedded AI systems like smart cameras, industrial automation tools, or IoT devices, dataset distillation directly addresses limits on memory, storage, and processing power. By reducing the dataset size, storage requirements shrink, and energy consumption stays low, making it practical to deploy AI in energy-constrained devices such as smartwatches, fitness trackers, or even autonomous vehicles, where efficiency is critical. Another advantage is continual learning. Since edge devices often face changing environments, distilled datasets help models adapt to new data distributions without losing performance. This ensures that AI systems remain effective and relevant in dynamic, real-world conditions. Finally, dataset distillation also strengthens privacy and security. Instead of sending large volumes of sensitive data back to centralized servers, information can be distilled and processed locally. This not only saves bandwidth and reduces reliance on the cloud, but it also ensures sensitive user data stays private, making the approach more secure and scalable for widespread deployment [65].

G. Real World Use Cases

Distillation has found several impactful real-world applications across diverse domains. The authors in [103] proposed a lightweight fault diagnosis in IoT edge computing, where a student model learns from a complex teacher model to achieve high diagnostic accuracy with reduced computational cost, enabling efficient real-time deployment in resource-constrained environments. A multi-agent sensor integration and knowledge distillation system has been introduced to enhance real-time navigation in autonomous vehicles in [104]. In this framework, multiple sensors operate as agents that feed environmental data into an integrated model. A lightweight student network learns from a more complex teacher by distilling knowledge across these sensors, enabling efficient, robust decision-making under stringent latency and computational constraints ideal for embedded vehicle systems requiring real-time autonomy. In [105] authors introduced a soft-label dataset distillation method. This approach compresses tens of thousands of gastric X-ray images into just a few anonymized soft-label images, while also extracting essential weights from Deep Convolutional Neural Networks (DCNNs) to drastically reduce model storage. Despite that extreme

compression, the resulting synthetic dataset achieves high detection performance, demonstrating enhanced efficiency and security in cross-institutional medical image sharing. These applications illustrate how both dataset and model distillation bridge the gap between high-performance machine learning and real-world deployment constraints.

VII. Challenges and Future Directions

The research on dataset distillation holds great promise, with numerous algorithms already applied across various fields. Although existing methods have demonstrated significant performance, several obstacles and concerns persist. The present section highlights the limitations of data distillation along with the future direction of the current limitations.

1) **Scalability:** Scalability remains a significant challenge in dataset distillation due to two primary factors. First, many existing methods rely on random-sampling benchmarks when summarizing large datasets, which often leads to suboptimal distilled datasets that fail to retain critical information, as noted by Cui *et al.* [106]. Second, the computational complexity of generating comprehensive data summaries poses a major hurdle, making it difficult to efficiently distill large-scale datasets using currently available techniques. To address these issues, several concrete strategies can be employed. Algorithmic approaches such as kernel-inducing points (KIP) [101] and gradient matching [7] summarize essential data patterns without processing the entire dataset, improving efficiency. Computational enhancements, including distributed computing frameworks, GPU acceleration, and adaptive memory management, can mitigate bottlenecks in processing large datasets [61]. Adaptive or importance-based sampling techniques prioritize high-value data points during distillation, reducing redundant computation while preserving task-relevant information. Implementing these strategies can make dataset distillation more practical for large-scale applications, allowing deep learning models to be trained efficiently while minimizing memory and computational overhead.
2) **Computational Complexity:** A major challenge in dataset distillation is the high computational cost during the optimization process, especially when dealing with large datasets or complex models. Despite reducing memory and storage requirements, distillation still requires significant computational power, particularly with deep learning architectures [7]. The need for efficient hardware accelerators such as GPUs or TPUs becomes crucial, particularly for real-time processing on edge devices. Future research could focus on optimizing distillation techniques to reduce the number of iterations required or exploring hardware-aware methods tailored to resource-constrained environments. Additionally, incorporating distributed computing may alleviate the burden on individual devices and improve scalability [107].
3) **Information Loss:** One of the biggest challenges with data distillation is the risk of losing important information [108]. Since the goal is to create a smaller, more compact version of the original dataset, some details can get left out, leading to a drop in model accuracy. The key, therefore, is finding the right balance in reducing the dataset size while still keeping the information that really matters for the model's performance. To address this, more advanced distillation methods are being explored. For example, techniques that use feature selection or attention mechanisms can help focus on the most relevant patterns in the data, minimizing information loss while boosting accuracy. Beyond that, incorporating domain-specific knowledge into the process adds another layer of refinement [109].
4) **Security and Privacy:** Current research ignores the potential security and privacy of DD in favor of refining techniques to produce more insightful synthetic datasets. Rather than targeting the model after training, a recent study introduces a new backdoor attack technique called DOORPING, which occurs during the dataset distillation process [110]. Future research will address further security and privacy concerns as well as potential defense mechanisms.
5) **Cross Architecture Transferability:** A significant challenge in dataset distillation is its reliance on specific neural network architectures, which limits the generalization of distilled datasets across models. Distilled data often underperforms when applied to unseen architectures, restricting its practical application [22]. To improve cross-architecture transferability, several concrete strategies have been proposed. Latent-space optimization leverages trained generative models to encode datasets into structured, architecture-agnostic representations, allowing distilled data to retain essential features across different model types [111]. Shared feature embedding approaches optimize distilled datasets to match internal representations rather than model-specific outputs, enhancing adaptability across diverse architectures [112]. Meta-learning methods further improve generalization by training distilled datasets on multiple architectures simultaneously, while curriculum-based adaptation gradually exposes the distilled data to increasingly diverse models to enhance robustness [113]. By implementing these strategies, distilled datasets can achieve greater cross-architecture robustness and broader applicability, enabling efficient training across a variety of neural network models in real-world AI applications.
6) **Expanding Frontiers:** Traditional data distillation has focused mainly on image datasets, but new modalities like graphs [10], recommender systems [114], audio [115], and video classification [116] present unique challenges, such as managing long temporal sequences. Furthermore, existing methods focus on classification tasks, limiting their applicability to broader predictive problems. To address these challenges, future research must focus on developing distillation techniques that can handle complex tasks like image generation [117], [118], and representation learning [119], [120], extending

beyond standard classification to improve adaptability and efficiency across diverse machine learning applications.

7) **Class Imbalance in Data Distillation:** One of the major issues of data distillation is to guarantee that the generated data from the distillation process has class balance and diversity comparable to that of the original dataset. Class imbalances indicate that several of these classes are overrepresented while others are fairly underrepresented, something that is common in all real datasets such as medical images, fraud behaviors, and rare event prediction [121]. If over-selection of samples of frequent classes takes place in the distillation stage, it can lead to a generated set of data that causes bias, therefore poor generalization and low model effectiveness, especially for categories of a minority class. This can lead to models misclassifying rare instances due to the absence or low representation of important instances of data. To avoid such an issue from occurring, class-aware sampling, loss-aware selection, and class-weighted loss functions can be applied to guarantee all categories are represented fairly [122].
8) **Application-specific constraints:** Application-specific constraints play a crucial role in shaping data distillation techniques, as different domains impose unique challenges related to privacy, computational efficiency, interpretability, and data availability. In sensitive fields like healthcare and finance, privacy regulations such as GDPR and HIPAA necessitate the development of privacy-preserving distillation methods to prevent data leakage while maintaining model performance. Similarly, applications requiring real-time processing, such as autonomous driving and fraud detection, demand computationally efficient distillation techniques that enable rapid decision-making with minimal latency [123]. Future research should focus on developing flexible and adaptive distillation frameworks that cater to the specific needs of various applications, ensuring both efficiency and effectiveness across different real-world scenarios.

VIII. Conclusion

Data distillation has become a revolutionary technique in managing huge-scale data generated by machine learning models, making it possible to use data efficiently while maintaining model performance. In this survey, several techniques to distill data, like knowledge distillation, core-set or instance selection, hyperparameter tuning, and generative modeling, have been discussed. We further elaborated on different applications of data distillation across multiple domains, i.e., continual learning, privacy, healthcare, federated learning, security, IoT applications, and edge computing. We discussed different topologies of data in which data distillation can be utilized to reduce resource consumption, save computational time, and enhance efficiency across different types of distributed and edge-based environments. The survey of our study finds several key advantages of dataset distillation, like reduced computational overheads, enhanced protection of privacy, and enhanced genericity of models. Conversely, several limitations still persist, i.e., scalability, computational complexity, and inter-architecture transferability between architectures. In order to eliminate these limitations, more adaptable and robust distillation schemes must be studied with the incorporation of advanced optimization algorithms and domain-aware modifications. Data distillation has huge prospects to play its role in developing more efficient, scalable, and privacy-preserving AI models with the rise of machine learning, primarily for IoT and edge-based applications. This progress will pave the way for future advancements in intelligent data management, distributed learning, and resource-efficient AI deployment across various domains.

References

[1] D. R.-J. G.-J. Rydning, J. Reinsel, and J. Gantz, "The digitization of the world from edge to core," *Framingham: International Data Corporation*, vol. 16, pp. 1–28, 2018.
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[4] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," pp. 173–182, 2016.
[5] A. Maekawa, S. Kosugi, K. Funakoshi, and M. Okumura, "Dilm: Distilling dataset into language model for text-level dataset distillation," *arXiv preprint arXiv:2404.00264*, pp. 2–3, 2024.
[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
[7] R. Yu, S. Liu, and X. Wang, "Dataset distillation: A comprehensive review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–2, 2023.
[8] P. Fernandes, B. Ghorbani, X. Garcia, M. Freitag, and O. Firat, "Understanding multi-task scaling in machine translation," 2022.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.
[10] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.
[11] Z. Deng and O. Russakovsky, "Remember the past: Distilling datasets into addressable memories for neural networks," *arXiv preprint arXiv:2206.02916*.
[12] X. Li, Y. Gu, C. T. Dinh, *et al.*, "On-device learning: A survey and outlook," *ACM Computing Surveys*, vol. Volume Number, no. Issue Number, p. Pages, 2020.
[13] J. Konečný, H. B. McMahan, F. X. Yu, *et al.*, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint*, vol. arXiv:1610.05492, 2017. [Online]. Available: https://arxiv.org/abs/1610.05492
[14] J. Nalepa and M. Kawulok, "Selecting training sets for support vector machines: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 857–900, 2019.
[15] T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros, "Dataset distillation," *arXiv preprint arXiv:1811.10959*, pp. 1–23, 2018.
[16] D. B. Lee, S. Lee, J. Ko, K. Kawaguchi, J. Lee, and S. J. Hwang, "Self-supervised dataset distillation for transfer learning," *arXiv preprint arXiv:2310.06511*, 2023.
[17] K. Wang, J. Gu, D. Zhou, Z. Zhu, W. Jiang, and Y. You, "Dim: Distilling dataset into generative model," *arXiv preprint arXiv:2303.04707*, 2023.
[18] J. Geng, Z. Chen, Y. Wang, H. Woisetschlaeger, S. Schimmler, R. Mayer, Z. Zhao, and C. Rong, "A survey on dataset distillation: Approaches, applications and future directions," *arXiv preprint arXiv:2305.01975*, 2023.

[19] R. Yu, S. Liu, and X. Wang, "Dataset distillation: A comprehensive review," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 1, pp. 150–170, 2023.

[20] N. Sachdeva and J. McAuley, "Data distillation: A survey," *arXiv preprint arXiv:2301.04272*, 2023.

[21] S. Lei and D. Tao, "A comprehensive survey of dataset distillation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 17–32, 2023.

[22] T. Nguyen, Z. Chen, and J. Lee, "Dataset meta-learning from kernel ridge-regression," *arXiv preprint arXiv:2011.00050*, 2020.

[23] W. Jin, L. Zhao, S. Zhang, Y. Liu, J. Tang, and N. Shah, "Graph condensation for graph neural networks," *arXiv preprint arXiv:2110.07580*, 2021.

[24] B. Zhao and H. Bilen, "Dataset condensation with distribution matching," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6514–6523.

[25] G. Hinton, O. Vinyals, and J. Dean, "Distilling knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, vol. 27, 2014.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[28] J. A. Romero, R. Sanchis, and E. Arrebola, "Experimental study of event based pid controllers with different sampling strategies. application to brushless dc motor networked control system," in *2015 XXV international conference on information, communication and automation technologies (ICAT)*. IEEE, 2015, pp. 1–6.

[29] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[30] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[31] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 126–136, 2018.

[32] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[33] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4119–4128.

[34] M. Welling, "Herding dynamical weights to learn," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 1121–1128.

[35] Y. Chen, M. Welling, and A. Smola, "Super-samples from kernel herding," *arXiv preprint arXiv:1203.3472*, 2012.

[36] D. Feldman, M. Faulkner, and A. Krause, "Scalable training of mixture models via coresets," *Advances in neural information processing systems*, vol. 24, 2011.

[37] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[38] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[39] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, vol. 25, 2012.

[40] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *International conference on machine learning*. PMLR, 2013, pp. 115–123.

[41] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[42] F. P. Such, A. Rawal, J. Lehman, K. Stanley, and J. Clune, "Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9206–9216.

[43] B. Zhao and H. Bilen, "Dataset condensation with differentiable siamese augmentation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 674–12 685.

[44] Y. Wu, M. Ren, R. Liao, and R. Grosse, "Understanding short-horizon bias in stochastic meta-optimization," *arXiv preprint arXiv:1803.02021*, 2018.

[45] L. Metz, N. Maheswaranathan, J. Nixon, D. Freeman, and J. Sohl-Dickstein, "Understanding and correcting pathologies in the training of learned optimizers," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4556–4565.

[46] Y. Zhou, E. Nezhadarya, and J. Ba, "Dataset distillation using neural feature regression," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9813–9827, 2022.

[47] N. Loo, R. Hasani, A. Amini, and D. Rus, "Efficient dataset distillation using random feature approximation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 877–13 891, 2022.

[48] Z. Chen, Y. Cao, Q. Gu, and T. Zhang, "A generalized neural tangent kernel analysis for two-layer neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 363–13 373, 2020.

[49] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, "Wide neural networks of any depth evolve as linear models under gradient descent," *Advances in neural information processing systems*, vol. 32, 2019.

[50] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[51] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, "Finite versus infinite neural networks: an empirical study," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 156–15 172, 2020.

[52] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012, vol. 118.

[53] B.-Y. N.-R. S. S. S. P. J. Lee, Jaehoon and J. Sohl-Dickstein, "Deep neural networks as gaussian processes," *arXiv preprint arXiv:1711.00165*, 2017.

[54] T. Dong, B. Zhao, and L. Lyu, "Privacy for free: How does dataset condensation help privacy?" pp. 5378–5396, 2022.

[55] J.-H. Kim, J. Kim, S. J. Oh, S. Yun, H. Song, J. Jeong, J.-W. Ha, and H. O. Song, "Dataset condensation via efficient synthetic-data parameterization," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 11 102–11 118.

[56] S. Lee, S. Chun, S. Jung, S. Yun, and S. Yoon, "Dataset condensation with contrastive signals," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 352–12 364.

[57] I. Sucholutsky and M. Schonlau, "Soft-label dataset distillation and text dataset distillation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[58] K. Wang, B. Zhao, X. Peng, Z. Zhu, S. Yang, S. Wang, G. Huang, H. Bilen, X. Wang, and Y. You, "Cafe: Learning to condense dataset by aligning features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 196–12 205.

[59] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2023. [Online]. Available: https://arxiv.org/abs/1602.05629

[60] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," 2017. [Online]. Available: https://arxiv.org/abs/1703.03400

[61] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu, "Dataset distillation by matching training trajectories," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4750–4759.

[62] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, "Graph neural networks for social recommendation," in *The world wide web conference*, 2019, pp. 417–426.

[63] N. Sachdeva and J. McAuley, "How useful are reviews for recommendation? a critical review and potential improvements," in *proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 2020, pp. 1845–1848.

[64] J. Cui, R. Wang, S. Si, and C.-J. Hsieh, "Scaling up dataset distillation to imagenet-1k with constant memory," in *International Conference on Machine Learning*. PMLR, 2023, pp. 6565–6590.

[65] B. Zhao, K. R. Mopuri, and H. Bilen, "Dataset condensation with gradient matching," *arXiv preprint arXiv:2006.05929*, 2020.

[66] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[67] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu, "Dataset distillation by matching training trajectories," in *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4750–4759.

[68] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.

[69] T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros, "Dataset distillation," *arXiv preprint arXiv:1811.10959*, 2018.

[70] J. Shang, T. Wang, and H. Liu, "Advances in tabular data processing for machine learning," *Journal of AI Research*, 2021.

[71] J. Yoon, D. Jarrett, and M. van der Schaar, "Generating synthetic data for tabular learning," in *NeurIPS*, 2020.

[72] L. Xu and X. Luo, "Handling missing values in tabular data distillation," *IEEE Transactions on Data Science*, 2019.

[73] Z. Wang and Y. Chen, "Gradient-based optimization for efficient dataset distillation," in *ICML*, 2022.

[74] K. R.-P. J. L. K. C. Parisi, German I and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural networks*, vol. 113, pp. 54–71, 2019.

[75] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[76] A. Rosasco, A. Carta, A. Cossu, V. Lomonaco, and D. Bacciu, "Distilled replay: Overcoming forgetting through synthetic samples," in *International Workshop on Continual Semi-Supervised Learning*. Springer, 2021, pp. 104–117.

[77] M. Sangermano, A. Carta, A. Cossu, and D. Bacciu, "Sample condensation in online continual learning," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 01–08.

[78] F. Wiewel and B. Yang, "Condensed composite memory continual learning," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[79] Z. Deng and O. Russakovsky, "Remember the past: Distilling datasets into addressable memories for neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 391–34 404, 2022.

[80] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[81] J. Konečnỳ, R. D. McMahan, H Brendan, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[82] R. Lohiya and A. Thakkar, "Application domains, evaluation data sets, and research challenges of iot: A systematic review," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8774–8798, 2020.

[83] E. Batista, M. A. Moncusi, P. López-Aguilar, A. Martínez-Ballesté, and A. Solanas, "Sensors for context-aware smart healthcare: A security perspective," *Sensors*, vol. 21, no. 20, p. 6886, 2021.

[84] O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, and X. Wang, "Internet of things for the future of smart agriculture: A comprehensive survey of emerging technologies," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 4, pp. 718–752, 2021.

[85] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[86] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, 2017.

[87] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 4–21, 2016.

[88] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[89] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[90] I. Sucholutsky and M. Schonlau, "Secdd: Efficient and secure method for remotely training neural networks (student abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 18, 2021, pp. 15 897–15 898.

[91] T. Dong, B. Zhao, and L. Lyu, "Privacy for free: How does dataset condensation help privacy?" in *International Conference on Machine Learning*. PMLR, 2022, pp. 5378–5396.

[92] D. Chen, R. Kerkouche, and M. Fritz, "Private set generation with discriminative information," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 678–14 690, 2022.

[93] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[94] Y. Wu, X. Li, F. Kerschbaum, H. Huang, and H. Zhang, "Towards robust dataset learning," *arXiv preprint arXiv:2211.10752*, 2022.

[95] C. White, P. Jain, S. Nayak, G. Ramakrishnan, *et al.*, "Speeding up nas with adaptive subset selection," *arXiv preprint arXiv:2211.01454*, 2022.

[96] T. Dong, B. Zhao, and L. Lyu, "Privacy for free: How does dataset condensation help privacy?" in *International Conference on Machine Learning*. PMLR, 2022, pp. 5378–5396.

[97] K. Pan, M. Gong, K. Feng, and H. Li, "Preserving privacy in fine-grained data distillation with sparse answers for efficient edge computing," *IEEE Internet of Things Journal*, 2024.

[98] G. Li, G. Qian, I. C. Delgadillo, M. Muller, A. Thabet, and B. Ghanem, "Sgas: Sequential greedy architecture search," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1620–1630.

[99] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 691–706.

[100] H. Ren, J. Wang, and K. Zhou, "A comprehensive review on dataset distillation techniques for efficient ai model training," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2103–2117, 2021.

[101] T. Aoyama, H. Yang, H. Hanada, S. Akahane, T. Tanaka, Y. Okura, Y. Inatsu, N. Hashimoto, T. Murayama, H. Lee, *et al.*, "Generalized kernel inducing points by duality gap for dataset distillation," *arXiv preprint arXiv:2502.12607*, 2025.

[102] A. Sajedi, S. Khaki, E. Amjadian, L. Z. Liu, Y. A. Lawryshyn, and K. N. Plataniotis, "Datadam: Efficient dataset distillation with attention matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 097–17 107.

[103] Y. Wang, Z. Yu, J. Wu, C. Wang, Q. Zhou, and J. Hu, "Adaptive knowledge distillation-based lightweight intelligent fault diagnosis framework in iot edge computing," *IEEE Internet of Things Journal*, vol. 11, no. 13, pp. 23 156–23 169, 2024.

[104] M. Hijji, K. Ullah, M. Alwakeel, A. Alwakeel, F. Aradah, F. A. Cheikh, M. Sajjad, and K. Muhammad, "Multiagent sensor integration and knowledge distillation system for real-time autonomous vehicle navigation," *IEEE Systems Journal*, 2025.

[105] G. Li, R. Togo, T. Ogawa, and M. Haseyama, "Compressed gastric image generation based on soft-label dataset distillation for medical data sharing," *Computer Methods and Programs in Biomedicine*, vol. 227, p. 107189, 2022.

[106] J. Cui, R. Wang, S. Si, and C.-J. Hsieh, "Dc-bench: Dataset condensation benchmark," *Advances in Neural Information Processing Systems*, vol. 35, pp. 810–822, 2022.

[107] K. Zhang, G. Li, N. Lu, P. Yang, and K. Tang, "Hardware-aware dnn compression for homogeneous edge devices," *arXiv preprint arXiv:2501.15240*, 2025.

[108] P. Sun, B. Shi, X. Shang, and T. Lin, "Information compensation: A fix for any-scale dataset distillation."

[109] S. Khaki, A. Sajedi, K. Wang, L. Z. Liu, Y. A. Lawryshyn, and K. N. Plataniotis, "Atom: attention mixer for efficient dataset distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7692–7702.

[110] Y. Liu, Z. Li, M. Backes, Y. Shen, and Y. Zhang, "Backdoor attacks against dataset distillation," *arXiv preprint arXiv:2301.01197*, 2023.

[111] J. Wang, X. Liu, Y. Zhang, Z. Li, X. He, and T.-S. Chua, "Universal feature matching for cross-architecture dataset distillation," in *International Conference on Machine Learning*, 2023.

[112] H. Zhao, X. Chen, Y. Li, and E. P. Xing, "Towards architecture-agnostic dataset distillation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[113] X. Liu, W. Zhang, L. Chen, and J. Huang, "Enhancing cross-architecture generalization in dataset distillation," *Journal of Machine Learning Research*, vol. 25, pp. 1–20, 2024.

[114] N. Sachdeva, M. Dhaliwal, C.-J. Wu, and J. McAuley, "Infinite recommendation networks: A data-centric approach," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31 292–31 305, 2022.

[115] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[116] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[117] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[118] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[119] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[120] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[121] P. Sun, B. Shi, D. Yu, and T. Lin, "On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9390–9399.

[122] S. Zhang, C. Chen, X. Hu, and S. Peng, "Balanced knowledge distillation for long-tailed learning," *Neurocomputing*, vol. 527, pp. 36–46, 2023.

[123] P. Liang, J. Chen, Y. Wu, B. Pu, H. Huang, Q. Chang, and G. Ran, "Data free knowledge distillation with feature synthesis and spatial consistency for image analysis," *Scientific Reports*, vol. 14, no. 1, p. 27557, 2024.

Qaiser Razi received the M.E degree from Birla Institute of Technology, Mesra, Ranchi, India, in 2019. He is currently a Ph.D. Research Scholar with the Department of Electrical and Electronics Engineering, BITS-Pilani, Pilani Campus, Pilani, India. His research interests include artificial intelligence, privacy and security, IoT, and non-fungible tokens.

Somya Singh is an undergrad student at Kalinga Institute of Industrial Technology (KIIT), Patia, Bhubaneswar. She is currently honing her research expertise through a prestigious internship at Birla Institute of Technology and Science (BITS), Pilani, under the mentorship of Dr. Vikas Hassija. With a strong command of machine learning, federated learning, and data distillation techniques, her research interests extend to reinforcement learning, quantum computing, and deep learning. Committed to pushing the boundaries of innovation, she strives to make impactful contributions to the field of artificial intelligence.

Riya Priyadarshini is currently pursuing B.Tech. degree from Kalinga Institute of Industrial Technology (KIIT), Bhubaneshwar. She is currently pursuing her research internship at the Birla Institute of Technology and Science (BITS), Pilani, under Dr. Vikas Hassija. She is eager to drive advancements in deep learning by exploring innovative techniques to enhance data efficiency. Her research interests include data science, machine learning, and dataset optimization.

Vikas Hassija received the M.E. degree from the Birla Institute of Technology and Science-Pilani, Pilani, India, in 2014. He is an Associate Professor at the School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, India. He was a Post-Doctoral Research Fellow at the National University of Singapore (NUS), Singapore. His current research is in blockchain, non-fungible tokens, the IoT, privacy and security, and distributed networks.

GSS Chalapathi obtained his B.E. in Electrical and Electronics Engineering with distinction from Birla Institute of Technology and Science (BITS), Pilani, in 2009. He obtained an M.E. in Embedded Systems and a Ph.D. from BITS Pilani in 2011 and 2019, respectively. He carried out his postdoctoral research at the University of Melbourne under the supervision of Prof. Rajkumar Buyya, a Distinguished Professor at the University of Melbourne, Australia. During his doctoral studies, he has been a visiting researcher at the National University of Singapore and Johannes Kepler University, Austria. He has published in reputed journals like IEEE Wireless Communication Letters, IEEE Sensors Journal, and Future Generation Computing Systems. He is a reviewer for the IEEE Internet of Things Journal and IEEE Access. His research interests are UAVs, Precision Agriculture, and Embedded Systems. He is a Senior Member of the IEEE and a member of the ACM.