

Dataset Distillation: Techniques, Advances, and Strategic Relevance for AIKosh

Prateek Khanna

SolFinder Research

Abstract

Dataset distillation is an emerging technique that compresses large-scale training datasets into small, informative synthetic or selected subsets while preserving the essential knowledge required for training high-performing machine learning models. This paper analyses the techniques in dataset distillation, with a particular focus on selection-based methods, and examines how these techniques can address AIKosh's core challenges related to efficient storage, reduced computational costs, cross-architecture generalization, and democratized access to high-quality training data. The research proposes that integrating dataset distillation into AIKosh's ecosystem can help accelerate India's sovereign AI mission by enabling resource-constrained researchers and startups to train competitive models on compressed, representative data.

Introduction

Dataset Distillation: Definition and Motivation

The explosive growth of deep learning has been fueled by the availability of massive datasets and computational resources. However, training on large datasets is computationally expensive, storage-intensive, and often inaccessible to resource-constrained researchers and organizations. Dataset distillation addresses this challenge by generating a compact dataset - typically 1% to 10% of the original size - that retains the essential information needed to train models with comparable or superior performance. Dataset distillation aims to create an optimized dataset specifically designed to maximize training efficiency and model performance. The distilled dataset can take two forms:

Synthetic data: Artificially generated images, text, or features optimized to match training dynamics.

Selected real data: A carefully chosen subset of the original data that best represents the overall distribution.

AIKosh: India's National AI Infrastructure

AIKosh (AIKosha: IndiaAI Datasets Platform) was officially launched on March 6, 2025, by the Ministry of Electronics and Information Technology (MeitY) as a core pillar of the

IndiaAI Mission. The platform serves as a centralized repository for India-specific datasets, pre-trained models, and development tools, with the explicit goal of reducing India's dependence on foreign-origin training data and fostering sovereign AI development.

As of early 2026, AIKosh hosts over 10,000 datasets across 20 sectors along with over 300 AI models. AIKosh constitutes one of seven pillars of the IndiaAI Mission, alongside IndiaAI Compute, Application Development, FutureSkills, Research, Startups, and Safe and Trusted AI.

Why Dataset Distillation Matters for AIKosh

AIKosh faces several structural challenges that dataset distillation can directly address:

Storage and bandwidth constraints: With 10,000+ datasets and growing, efficient storage and transfer of data is critical.

Computational accessibility: GPU resources are limited. Distilled datasets reduce training time and GPU usage.

Cross-architecture generalization: AIKosh hosts diverse models (LLMs, SLMs, vision models, speech models). Distilled datasets that generalize across architectures maximize platform utility.

Data quality and curation: Automated distillation can serve as an intelligent curation mechanism, identifying the most informative samples.

Democratization: Smaller, high-quality datasets lower the barrier to entry for startups and academic researchers.

Literature Review

Dataset distillation has evolved significantly since Wang, T. et al. (2018) first formulated the problem of synthesizing compact training sets that preserve full-dataset performance. Early optimization-based methods, including gradient matching (Zhao, B. et al., 2020) and trajectory matching (Cazenavette, G. et al., 2022), established the paradigm of treating synthetic data as learnable parameters updated through bilevel optimization. However, these approaches suffer from architecture dependency and high computational cost.

Distribution matching methods (Zhao, B., & Bilen, H., 2023) offered a computationally efficient alternative by aligning statistical properties using Maximum Mean Discrepancy. The integration of diffusion models marked a significant breakthrough: Su, D et al. (2024) introduced D4M, using latent diffusion with prototype conditioning to generate high-fidelity images while improving cross-architecture generalization. Moser, B. et al. (2026) extended this with LD3M, learning distilled latents end-to-end through pre-trained diffusion models and achieving up to 4.8% accuracy gains on ImageNet subsets. Selection-based methods represent a distinct paradigm. Liu, Y. et al. (2023) proposed DREAM, using K-Means clustering to select representative sub-cluster centers, reducing distillation iterations by over $8\times$. Sun P. et al. (2024) proposed RDED, an optimization-free approach selecting informative real patches with soft labels. Zhou, X. et al. (2022) formulated coreset selection as probabilistic bilevel optimization, demonstrating architecture-agnostic properties. Recent work has shown that combining selection with soft labels yields performance competitive with complex heuristics, suggesting selection-based approaches may be undervalued.

At present, limited research exists on incorporating developments in data distillation techniques within sovereign AI infrastructure such as AIKosh. This work attempts to address this gap.

Dataset Distillation Methods

Dataset distillation methods can be broadly categorized into three paradigms:

Optimization-Based Methods

These methods treat synthetic data as learnable parameters and optimize them through iterative training loops.

Gradient Matching (Dataset Condensation): Minimizes the difference between gradients computed on real and synthetic data.

Trajectory Matching (MTT): Matches the training trajectory over multiple steps rather than a single gradient step. This provides better long-term consistency but is computationally expensive.

Distribution Matching (DM): Aligns the statistical distribution of synthetic data with real data using metrics like Maximum Mean Discrepancy (MMD) or KL divergence. This avoids the bilevel optimization problem entirely.

Generative and Diffusion-Based Methods

Recent advances leverage diffusion models to generate high-fidelity synthetic data.

D4M (Dataset Distillation via Disentangled Diffusion Model): Uses a Latent Diffusion Model (LDM) to generate images up to 512×512 resolution. It first extracts prototypes via clustering, then conditions the diffusion process on label text embeddings. This was the first method to overcome the pronounced architecture dependency of traditional distillation.

LD3M (Latent Dataset Distillation with Diffusion Models): Learns gradient-based distilled latents and class embeddings end-to-end through a pre-trained LDM. It solves the vanishing gradient problem in long denoising chains via a linearly decaying skip connection. On ImageNet subsets, LD3M improves accuracy by up to 4.8 percentage points over prior state-of-the-art.

Diversity-Driven Diffusion Distillation: Introduces self-adaptive memory to align distilled and real dataset distributions, using the degree of alignment to guide the diffusion model toward more diverse generation.

Selection-Based Methods

Selection-based methods choose a subset of real data points rather than synthesizing new ones. Selection-based methods are particularly relevant for AIKosh because they operate on authentic data samples that naturally generalize across model architectures. Their computational efficiency eliminates the need for expensive bilevel optimization, while their reliance on real data improves interpretability and facilitates the incorporation of soft labels to enhance downstream performance.

Selection-Based Methods: A Deep Dive

Core Principle

Selection-based dataset distillation operates by scoring and ranking real data samples using various importance metrics, then selecting the top-ranked subset to form a compact, informative dataset. Unlike optimization-based methods, the data itself is not altered — only the selection is optimized.

The general pipeline is:

Scoring: Assign an importance score to each sample using a teacher model or statistical metric.

Ranking: Sort samples by their scores.

Selection: Choose the top- k samples (or representative cluster centers) under a budget constraint (e.g., Images Per Class — IPC).

Optional Enhancement: Augment selected data with soft labels or generative models.

Key Methods

DREAM: Representative Matching (ICCV 2023)

DREAM addresses the inefficiency of random sampling in gradient matching by using K-Means clustering within each class to generate sub-clusters. The sub-cluster centers are selected as representative samples because they: - Are representative of surrounding samples - Evenly cover the whole class distribution - Avoid over-representing boundary samples with large gradients

DREAM reduces distillation iterations by more than $8\times$ without performance degradation. It also uses clustering-based initialization for synthetic images, selecting center samples from each sub-cluster to accelerate convergence.

Algorithm (DREAM):

Input: Original dataset T , number of sub-clusters K , budget IPC

For each class c in T :

1. Extract features $f(x)$ for all samples x in class c
2. Apply K-Means clustering on $f(x)$ to obtain K sub-clusters
3. Select the center sample from each sub-cluster
4. If $|\text{centers}| > \text{IPC}$, further subsample using diversity criteria
5. If $|\text{centers}| < \text{IPC}$, augment with nearest neighbors

Output: Selected representative set S

RDED: Realistic, Diverse, and Efficient Dataset Distillation

RDED is an optimization-free approach that: - Selects a diverse set of informative patches directly cropped from original data - Combines these patches into new images - Uses a teacher model trained on the full dataset to identify informative patches and assign soft labels

This method emphasizes realism - since patches come from real data, they avoid the spurious features that optimization-based methods often produce.

Diffusion-Augmented Coreset Expansion

This two-stage method combines selection with generative expansion:

Stage 1 - Coreset Selection: Select the most informative patches by minimizing cross-entropy loss on random crops. Patches with the lowest cross-entropy loss are chosen. The compression ratio r controls the trade-off between diversity (more patches) and resolution.

Stage 2 - Diffusion-Augmented Expansion: Selected low-resolution patches are used as anchors on the training distribution manifold. They are upsampled, noise-corrupted, and denoised using a Latent Diffusion Model (LDM) to generate high-quality variations.

Each patch can generate multiple variations using different random seeds.

This approach achieves over 10% improvement compared to state-of-the-art on large-scale benchmarks.

Probabilistic Bilevel Coreset Selection (ICML 2022)

This method formulates coreset selection as a probabilistic bilevel optimization problem. A key theoretical insight is that coreset selection is not sensitive to the model since the data itself is not altered - making it more robust across architectures than synthetic distillation.

Soft Labels in Coreset Selection (2025)

Recent research has highlighted the significant role of soft labels, which use probabilistic target distributions instead of conventional one-hot labels, in improving the effectiveness of dataset distillation methods. While soft labels have been widely adopted in several distillation approaches, their potential had not previously been fully explored within coreset selection methods. The study demonstrates that incorporating soft labels consistently enhances the performance of all coreset selection techniques by providing richer supervisory information during training. A simple random selection strategy augmented with soft labels can achieve performance comparable to, and in some cases surpassing, existing sophisticated coreset selection algorithms. These observations suggest that soft labels may contribute more substantially to performance gains than the complexity of the selection strategy itself.

Selection Metrics and Strategies

Metric	Description	Used In
Cross-Entropy Loss	Select patches with lowest teacher	RDED, Diffusion-

Metric	Description	Used In
	model loss	Augmented
Gradient Norm (GraNd)	Measure gradient magnitude as importance signal	Coreset selection baselines
Error L2 Norm (EL2N)	Use prediction error to score sample difficulty	Coreset selection baselines
Forgetting Score	Track how often samples are forgotten during training	Progressive distillation
Clustering (K-Means)	Select sub-cluster centers for even coverage	DREAM
Herding	Iteratively select points near class centers	Traditional coreset selection
Uncertainty (Entropy/Confidence)	Select most uncertain or most confident samples	Object detection coreset
Modular Centrality	Select informative images based on network centrality	InfoDist (medical imaging)

Advantages and Limitations

Selection-based dataset distillation methods offer several advantages. Because these approaches operate on authentic data samples, they are inherently architecture-agnostic, allowing the selected subsets to generalize effectively across diverse machine learning models. They are also computationally efficient, as they do not require the expensive bilevel optimization procedures commonly associated with synthetic data generation methods. Their reliance on real data ensures that the resulting datasets remain realistic, avoiding the artifacts and distortions that can arise during synthetic sample generation. The use of authentic samples also enhances interpretability, since the selected data can be directly examined and understood by human experts. In addition, these methods can support privacy preservation by operating on anonymized datasets without the need to create synthetic artifacts that may inadvertently reveal sensitive information.

Despite these advantages, selection-based approaches also exhibit certain limitations. Their compression ratios are generally lower than those achieved by optimization-based synthetic methods because they cannot encode as much information into a small number of samples. Since they retain actual data instances, they may also require the storage of real images or records rather than compact latent representations, which can impose additional storage requirements. Selection alone may not fully capture the diversity of the original data distribution, particularly for highly heterogeneous datasets. This limitation has motivated the development of hybrid approaches that combine coreset selection with generative augmentation techniques to improve diversity preservation while maintaining computational efficiency.

Relevance to AIKosh: Strategic Opportunities

Addressing AIKosh's Core Challenges

Efficient Storage and Bandwidth

With 10,000+ datasets and growing, AIKosh faces significant storage and transfer costs. Selection-based distillation can reduce dataset sizes by 90-99% while maintaining training performance. For example, distilling CIFAR-10 from 50,000 images to 500 images (1 IPC) with selection-based methods achieves competitive accuracy, dramatically reducing storage requirements.

Democratizing GPU Access

AIKosh's GPU sandbox is currently limited.

Training on distilled datasets reduces GPU time proportionally. A model that takes 100 epochs on the full dataset may converge in 10-20 epochs on a distilled dataset, making efficient use of limited compute resources.

Cross-Architecture Generalization

AIKosh hosts diverse models: LLMs, SLMs, vision models, speech models, TTS models, and more. Traditional optimization-based distillation is architecture-specific - a dataset distilled for ResNet may not work well for ViT. Selection-based methods, using real data, generalize across architectures naturally, maximizing the utility of each distilled dataset on the platform.

Data Quality and Automated Curation

AIKosh already implements automated pipelines that score datasets on uniqueness, completeness, frequency of updates, and absence of duplication. Integrating selection-based distillation metrics (e.g., gradient norms, forgetting scores, clustering quality) can enhance this scoring system, providing an intelligent curation layer that identifies the most informative samples for each dataset.

Sector-Specific Applications

Healthcare

Medical datasets are often small, sensitive, and expensive to annotate. Selection-based methods can: - Identify the most informative patches in histopathology images (as in InfoDist) - Preserve diagnostic patterns while reducing dataset size - Maintain privacy by selecting from anonymized real data rather than generating synthetic images that may leak patient information

Agriculture

Crop yield optimization and pest detection models benefit from representative selection of diverse field conditions. DREAM's clustering-based approach can ensure coverage of different soil types, weather conditions, and crop varieties.

Language Technologies (Indic Languages)

AIKosh hosts IndicVoices (12,000 hours across 22 languages) and BhasaAnuvaad (44,400 hours across 13 languages). Selection-based distillation can identify the most

informative utterances for ASR and TTS model training, reducing the computational burden of training on massive speech corpora while preserving linguistic diversity.

Governance and Public Services

Datasets for auto-adjudication of medical claims (PMJAY), vendor mapping (MSME), and dispute resolution require high accuracy on edge cases. Selection methods that prioritize uncertain or difficult samples (high EL2N scores) can improve model robustness on critical decision-making tasks.

Integration with AIKosh's Existing Infrastructure

Sandbox Environment

AIKosh's GPU-enabled Jupyter Notebook environment can serve as a practical platform for implementing dataset distillation workflows. Researchers can execute selection-based distillation algorithms directly on uploaded datasets to identify representative samples that preserve the informational content of the original data. The platform can subsequently generate distilled datasets accompanied by comprehensive metadata, including the selection criteria employed, achieved compression ratios, and performance metrics obtained during evaluation. AIKosh can also facilitate systematic benchmarking by enabling comparisons between distilled and full datasets across downstream machine learning tasks. Such benchmarking would allow users to assess trade-offs between dataset size, computational efficiency, and predictive performance, thereby supporting evidence-based decisions regarding the adoption of distilled datasets for diverse AI applications.

Versioning and Collaboration

AIKosh's dataset versioning capabilities can support the continuous refinement and evolution of distilled datasets. Contributors can initially upload original datasets and subsequently generate multiple distilled variants as improved distillation strategies emerge. The versioning framework can maintain a complete record of these iterations, enabling reproducibility and transparent tracking of changes over time. Contributors can also document and share the selection strategies, hyperparameters, and experimental settings employed during the distillation process, thereby facilitating knowledge exchange within the research community. Furthermore, AIKosh can host community-driven benchmarking challenges that encourage researchers to compare and evaluate different distillation approaches under standardized conditions. Such collaborative

initiatives can accelerate methodological innovation, promote best practices, and contribute to the development of robust and efficient dataset distillation techniques for diverse AI applications.

Quality Scoring Enhancement

AIKosh's existing star-rating framework can be expanded to incorporate specialized evaluation metrics tailored for dataset distillation. A distillation quality score can be introduced to quantify the performance of machine learning models trained on distilled datasets relative to those trained on the original full datasets, thereby assessing the effectiveness of the distillation process. In addition, a diversity score can measure how well the distilled dataset preserves the characteristics and variability of the original data distribution, using clustering-based or distributional similarity metrics. An efficiency score can further evaluate the trade-off between dataset compression and predictive performance by relating the achieved compression ratio to the degree of accuracy retention. Collectively, these metrics would provide users with a transparent and standardized mechanism for assessing the quality, representativeness, and practical utility of distilled datasets, thereby facilitating informed dataset selection and promoting best practices within the AIKosh ecosystem.

Sovereign AI and Data Sovereignty

A critical strategic consideration for AIKosh is the promotion of data sovereignty, which seeks to ensure that national artificial intelligence systems are developed using India-specific datasets without excessive dependence on foreign data sources. Selection-based dataset distillation can contribute significantly to this objective by maximizing the utility of existing domestic datasets and extracting their most informative and representative samples. Such an approach can reduce the necessity of importing external datasets or generating synthetic data that may not accurately reflect India's unique socio-cultural and linguistic contexts. Furthermore, because selection-based methods preserve authentic data instances rather than creating artificial substitutes, they help maintain the cultural specificity and integrity of Indian datasets. This is particularly important for domains such as Indic language technologies, region-specific agricultural practices, and localized healthcare patterns, where contextual nuances play a critical role in model performance and reliability. By enabling efficient utilization of indigenous datasets while preserving their authenticity, selection-based distillation can strengthen AIKosh's role in advancing India's sovereign, inclusive, and context-aware artificial intelligence ecosystem.

Experimental Framework for AIKosh Integration

Proposed Benchmarking Protocol

To evaluate selection-based distillation for AIKosh datasets, we propose the following protocol:

Step 1: Dataset Selection - Choose representative datasets from AIKosh across multiple sectors (healthcare, agriculture, language, governance) - Ensure datasets vary in size (small: <1K samples, medium: 1K-100K, large: >100K)

Step 2: Baseline Establishment - Train reference models on full datasets - Record accuracy, training time, GPU usage, and convergence epochs

Step 3: Distillation - Apply selection-based methods: DREAM, RDED, Coreset with soft labels - Vary compression ratios: 1%, 5%, 10%, 20% - Record selection time and metadata

Step 4: Evaluation - Train identical models on distilled datasets - Compare: accuracy, training time, GPU usage, convergence speed - Test cross-architecture generalization (e.g., train on distilled data with ResNet, evaluate with ViT, EfficientNet)

Step 5: Ablation Studies

- Impact of soft labels vs. hard labels
- Impact of clustering granularity (K-Means K value)
- Impact of generative augmentation (LDM expansion of selected patches)

Challenges and Future Directions

Technical Challenges

Scalability to Large Datasets: Methods like TESLA and TEDDY address this for optimization-based approaches, but selection-based methods need efficient algorithms for datasets with millions of samples (e.g., IndicVoices).

Multimodal Distillation: AIKosh hosts multimodal datasets (image-text, audio-visual). Selection-based methods need to handle cross-modal alignment, as in the bi-trajectory matching framework for vision-language models.

Dynamic Distillation: As datasets on AIKosh are updated (versioning feature), distilled versions need incremental update mechanisms rather than full recomputation.

Fairness and Bias: Selection methods must ensure that compressed datasets preserve demographic and linguistic diversity, particularly critical for India's heterogeneous population.

Policy and Governance Considerations

IP and Prior Art: AIKosh's open-access nature means distilled datasets become part of the public prior art landscape. Contributors should consider patent filing sequences before publishing distilled variants.

Data Licensing: AIKosh datasets have Open or Restricted permissions. Distilled versions must inherit the same licensing terms, and the distillation process itself must be documented for compliance.

Ethical AI: The IndiaAI Safe and Trusted AI pillar requires that distillation methods are transparent, explainable, and auditable - selection-based methods naturally satisfy this since the selected data is interpretable.

Future Research Directions

Hybrid Selection-Generation: Combine selection-based anchors with diffusion-based augmentation for optimal compression-diversity trade-off.

Federated Distillation: Enable multiple AIKosh contributors to collaboratively distill datasets without sharing raw data, preserving privacy and sovereignty.

Auto-Distillation: Develop meta-learning approaches that automatically select the best distillation strategy (selection vs. optimization, metric choice, compression ratio) for a given dataset and task.

Cross-Lingual Distillation: For India's 22 official languages, develop selection methods that identify the most informative samples for multilingual model training, maximizing cross-lingual transfer.

Conclusion

Dataset distillation, particularly selection-based approaches, represents a transformative opportunity for AIKosh and India's broader artificial intelligence ecosystem. By compressing large datasets into smaller, representative, and architecture-agnostic subsets, these techniques can substantially reduce storage requirements and computational costs associated with training machine learning models. Such efficiency gains can democratize access to AI development by enabling resource-constrained researchers, educational institutions, and startups to train and experiment with sophisticated models without requiring extensive computational infrastructure. In addition, intelligent and automated data curation can improve overall dataset quality by identifying the most informative and representative samples while minimizing redundancy. Selection-based distillation can also contribute to India's data sovereignty objectives by maximizing the value extracted from domestically curated datasets and reducing dependence on external data sources. Furthermore, because these methods preserve authentic data samples and are largely architecture-agnostic, they can support robust cross-architecture generalization across AIKosh's diverse model ecosystem. Collectively, these capabilities position dataset distillation as a strategic enabler for building an efficient, inclusive, and sustainable AI infrastructure that aligns with India's long-term vision of sovereign and context-aware artificial intelligence.

Methods like DREAM, RDED, and diffusion-augmented coreset expansion offer practical, efficient, and realistic pathways to achieve these goals. As AIKosh continues to grow integrating selection-based distillation into its infrastructure can be a force

multiplier for realizing the IndiaAI Mission's vision of accessible, sovereign, and ethical AI development.

The convergence of dataset distillation research with India's national AI infrastructure creates a unique opportunity to define best practices for efficient, equitable, and sovereign AI infrastructure. As AI systems continue to scale, future competitiveness may increasingly depend on data efficiency rather than computational abundance. In this context, dataset distillation has the potential to become a foundational technology that enables repositories such as AIKosh to transition from passive data storage platforms into active intelligence infrastructures capable of generating optimized, high-quality, and reusable datasets for India's emerging AI ecosystem.

References

Wang, T., Zhu, J. Y., Torralba, A., & Efros, A. A. (2018). Dataset distillation. *arXiv preprint arXiv:1811.10959*.

Zhao, B., Mopuri, K. R., & Bilen, H. (2020). Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*.

Zhao, B., & Bilen, H. (2023). Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 6514-6523).

Cazenavette, G., Wang, T., Torralba, A., Efros, A. A., & Zhu, J. Y. (2022). Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4750-4759).

Su, D., Hou, J., Gao, W., Tian, Y., & Tang, B. (2024). D⁴: Dataset distillation via disentangled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5809-5818).

Moser, B., Raue, F., Palacio, S., Frolov, S., & Dengel, A. (2026). Unlocking dataset distillation with diffusion models. *Advances in Neural Information Processing Systems*, 38, 72581-72614.

Liu, Y., Gu, J., Wang, K., Zhu, Z., Jiang, W., & You, Y. (2023). Dream: Efficient dataset distillation by representative matching. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 17314-17324).

Sun, P., Shi, B., Yu, D., & Lin, T. (2024). 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* (pp. 9390-9399).

Zhou, X., Pi, R., Zhang, W., Lin, Y., Chen, Z., & Zhang, T. (2022, June). Probabilistic bilevel coreset selection. In *International conference on machine learning* (pp. 27287-27302). PMLR.