Computers and Electrical Engineering 129 (2026) 110822

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Reconstruction resilient privacy-aware dataset distillation using feature distribution matching technique

Qaiser Razi [a], G.S.S. Chalapathi [a,*], Vikas Hassija [b]

[a] Department of Electrical and Electronics Engineering, Birla Institute of Technology and Science, Pilani, 333031, Rajasthan, India
[b] School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, 751024, Odisha, India

ARTICLE INFO

Keywords:
Data distillation
Distribution matching
Reconstruction attack
Privacy preservation
Diffusion model
Medical data sharing
Medical image classification

ABSTRACT

Dataset distillation has emerged as a powerful technique for reducing the size of training data while preserving model performance. This offers significant advantages in various domains, particularly in medical imaging, where data annotation is expensive, storage is limited, and computational resources are constrained. In this paper, we propose a novel distillation technique that combines diffusion models with distribution matching to generate distilled data for three medical image datasets, i.e., Pneumonia, COVID-19, and Brain Tumor detection. The diffusion component enables the generation of high-quality and diverse synthetic samples, while distribution matching ensures alignment with the underlying data distribution, thereby preserving discriminative features. We further design an autoencoder-based reconstruction framework to analyze and compare the vulnerability of original, conventional distilled, and our proposed distilled datasets to reconstruction attacks. In addition, we incorporate privacy risk evaluations using membership inference attacks (MIA) and attribute inference attacks (AIA). Experimental results show that our method achieves better classification accuracy and stronger privacy preservation compared to existing distillation approaches. These findings suggest that dataset distillation, particularly with our proposed framework, not only improves computational efficiency but also acts as an effective privacy-enhancing mechanism, making it a promising approach for secure and scalable medical artificial intelligence (AI) applications.

1. Introduction

In a machine learning (ML) model, data plays a crucial role. High-quality datasets are essential to ensure the accuracy, robustness, and generalization of ML models across various domains. In the medical domain, to facilitate the development of ML models for innovative medical treatment, to enhance patient care, and for medical research, a vast amount of medical data is shared between different hospitals and organizations [1]. Since medical data contains sensitive or personal information of the patient, the privacy and security of this data should be maintained before sharing [2]. However, the ever-growing data volumes required for training such models have introduced significant storage, computational efficiency, and scalability challenges.

Dataset distillation plays a significant role in addressing these challenges. The process distills large datasets into smaller yet information-rich representations, preserving the important patterns for model training [3]. By minimizing the dataset size without affecting the important information, distillation improves computational efficiency and reduces storage needs. Distillation is increasingly finding relevant attention in both academic research and industrial settings, revolutionizing the way ML models are developed and deployed [4]. Dataset distillation offers an innovative solution, particularly for resource-constrained devices such as

* Corresponding author.
E-mail addresses: p20210070@pilani.bits-pilani.ac.in (Q. Razi), gssc@pilani.bits-pilani.ac.in (G.S.S. Chalapathi), vikas.hassijafcs@kiit.ac.in (V. Hassija).

https://doi.org/10.1016/j.compeleceng.2025.110822
Received 3 July 2025; Received in revised form 18 September 2025; Accepted 5 November 2025

Available online 10 November 2025

smartphones, sensors, and IoT systems [5]. These devices are low in memory and processing power, and it is hard to implement huge artificial intelligence (AI) models. Distillation reduces large datasets into small equivalent datasets. These small datasets can be used to train ML models on such resource-limited devices with minimal storage and communication needs [6]. By compressing dataset sizes without compromising on the essential information, dataset distillation speeds up AI training with reduced cost and improved scalability. With accelerating growth in AI, distillation will be essential to achieve effective and affordable models even in resource-constrained environments [7].

Apart from efficiency, privacy preservation is an essential issue, particularly for sensitive domains such as medical imaging. Datasets for medical images tend to involve personally identifiable information (PII) that can lead to serious ethical and legal consequences if it is leaked. Medical images shared for research or developing diagnostic models can result in unwanted information leakage. Hence, protecting data privacy while preserving model performance is an important research objective [8]. Dataset distillation offers a promising approach not just to reduce the computational burden but also to enhance privacy through the creation of synthetic data, abstracting away patient-identifiable information.

Diffusion models have recently emerged as a powerful generative framework for producing high-quality synthetic data. These models are more stable to train and can better capture the global data distribution, thereby reducing the risk of overfitting to specific samples [9]. Their iterative denoising process allows fine-grained control over the fidelity and diversity of generated samples, making them particularly suitable for medical imaging tasks where both realism and privacy are critical [10]. By leveraging diffusion models within dataset distillation, it becomes possible to create synthetic datasets that preserve essential task-related features while reducing the leakage of sensitive, patient-specific details. To evaluate the privacy implications, reconstruction techniques based on autoencoders [11] can be used to measure how much sensitive visual information can be reconstructed from the original as well as distilled datasets. If reconstructions from distilled data have poor similarity with original inputs, that means there is a lower risk of privacy leakage.

In this paper, we have created a distilled version of the original medical image datasets using a novel distillation technique that combines diffusion models and distribution matching, named DDM. The diffusion model contributes by generating high-quality and diverse synthetic samples through its iterative denoising process, which ensures stability and better coverage of the underlying data distribution. On the other hand, distribution matching ensures that the generated synthetic data aligns closely with the real dataset at a class-consistent level, preserving discriminative features necessary for classification tasks. Together, this hybrid approach leverages the generative strength of diffusion models and the alignment capability of distribution matching to produce distilled datasets that are both informative and privacy-preserving. We have evaluated the performance of ML models trained on the original and distilled datasets, and we employed an autoencoder-based reconstruction process across three medical image classification tasks. By comparing the reconstructed images with the originals using various image similarity metrics, we quantify the extent of information retained in each dataset. Finally, we evaluate the trade-off between classification accuracy and privacy preservation of the original and distilled data.

The primary contributions of our work can be summarized as follows:

- Introduced a novel (DDM), a hybrid dataset distillation technique that combines diffusion models and distribution matching, leveraging diffusion for generating diverse and high-quality synthetic samples and distribution matching for preserving class-consistent discriminative features.
- Evaluated privacy preservation using an autoencoder-based technique to reconstruct images and assess reconstruction quality with multiple similarity metrics (SSIM, PSNR, LPIPS, NCC, FSIM). Additionally, evaluated membership inference attacks (MIA) and attribute inference attacks (AIA) to enable more rigorous privacy-oriented evaluations.
- Conducted extensive experiments on three medical imaging datasets (Pneumonia, COVID-19, and Brain Tumor) to demonstrate the effectiveness of the proposed approach in achieving a better trade-off between classification accuracy and privacy compared to existing data distillation approaches.
- Showed that our proposed framework yields distilled datasets that retain high classification accuracy while significantly reducing the risk of privacy leakage compared to existing data distillation techniques.

2. Related works

Dataset Distillation methods have been proposed as a promising solution to create compact but informative summaries from large datasets to ensure that only the most critical knowledge is retained from the larger datasets for training machine learning models. Keeping the size of the dataset small and retaining the important information of the larger dataset, data distillation offers an effective solution for addressing data complexity. Sucholutsky and Schonlau [12] discussed data distillation as a way of sample reduction. They emphasized the role of distillation in decreasing the memory footprint of datasets without sacrificing their learning potential, especially in image and text data handling. Wang et al. [4] discussed backpropagation-based distillation, showing its potential in resisting adversarial attacks and dealing with computationally costly tasks like data poisoning defense and adversarial robustness. Their research highlighted how the optimization of data representation using distillation can significantly cut down on training time and improve model generalization.

Nguyen et al. [13] 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. Guang Li et al. [14] introduced a novel data distillation approach for secure medical data sharing. By compressing and anonymizing medical image datasets, their

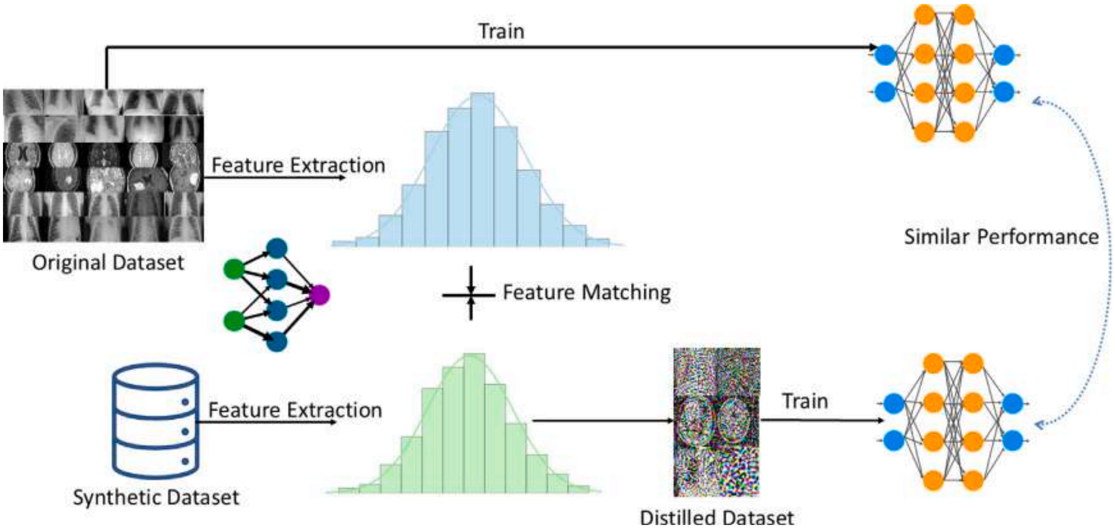

Fig. 1. Data distillation using distribution matching.

method facilitates efficient and privacy-preserving data exchange between hospitals. Experimental evaluations on COVID-19 chest X-ray images demonstrated its advantages over existing techniques. Beyond traditional data distillation, researchers have extended its applications to graph learning, continual learning, and domain adaptation. Ze et al. [15] proposed kernel ridge regression-based graph dataset distillation (KIDD), which exhibited strong performance in both forward and backward propagation processes. Their experiments on seven different datasets revealed that KIDD outperformed other data distillation methods and, in some cases, even surpassed models trained on the full dataset. Jin et al. [16] 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. The authors in [17] introduce a method that transfers knowledge from multiple teacher feature maps to a low-resolution student model, improving recognition accuracy over baseline distillation but with the added computational cost of using multiple feature sources during training. Chen et al. [18] proposed Adaptive Synthetic Data Distillation, a technique that generates distilled datasets optimized for specific learning objectives using reinforcement learning-based strategies. Their work highlights how adaptive distillation can enhance model performance by dynamically selecting the most informative data points. Zhao and Bilen [19] explored DM (Dataset Matching) for domain adaptation, ensuring distilled datasets generalize well across various tasks by matching the feature distributions between different domains.

3. Background

3.1. Data distillation

Dataset distillation is a novel method intended to distill large sets of data into compact but very informative subsets, allowing for efficient training while maintaining model performance [20]. As machine learning increases, dealing with massive amounts of data becomes a problem because it involves enormous storage and computing expenses. Training on complete datasets tends to result in inefficiencies, and thus, it is essential to find and keep only the most important data points. Dataset distillation does this by combining or sampling representative samples that capture the insight of the original dataset [21]. Distilled data enables model deployment on devices with limited resources so that ML can be applied to a wider domain. Yet, finding the right balance between data reduction and model effectiveness entails strategic approaches to guarantee that distilled datasets still contain learning signals.

Data Distillation 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 [13], and distribution matching [22] enable models to retain critical training dynamics, ensuring effective generalization despite working with distilled data. The various data distillation frameworks are performance matching, gradient matching, parameter matching, and distribution matching. Performance matching aims to create distilled datasets using methods like meta-model matching [20] and kernel ridge regression [23]. Gradient matching tries to generate synthetic data so that the training gradients of the synthetic match those of the real data [24]. The parameter matching technique tries to align the neural network parameters trained on the distilled data with those trained on the original data through single-step matching or multi-step matching [25]. On the other hand distribution matching method aligns the feature distributions of distilled and original datasets by minimizing the distance between their mean feature representations, offering a scalable and memory-efficient alternative to parameter-based approaches without requiring bilevel optimization distillation. We have used a distribution matching technique to generate distilled data, which is explained in the following subsection.

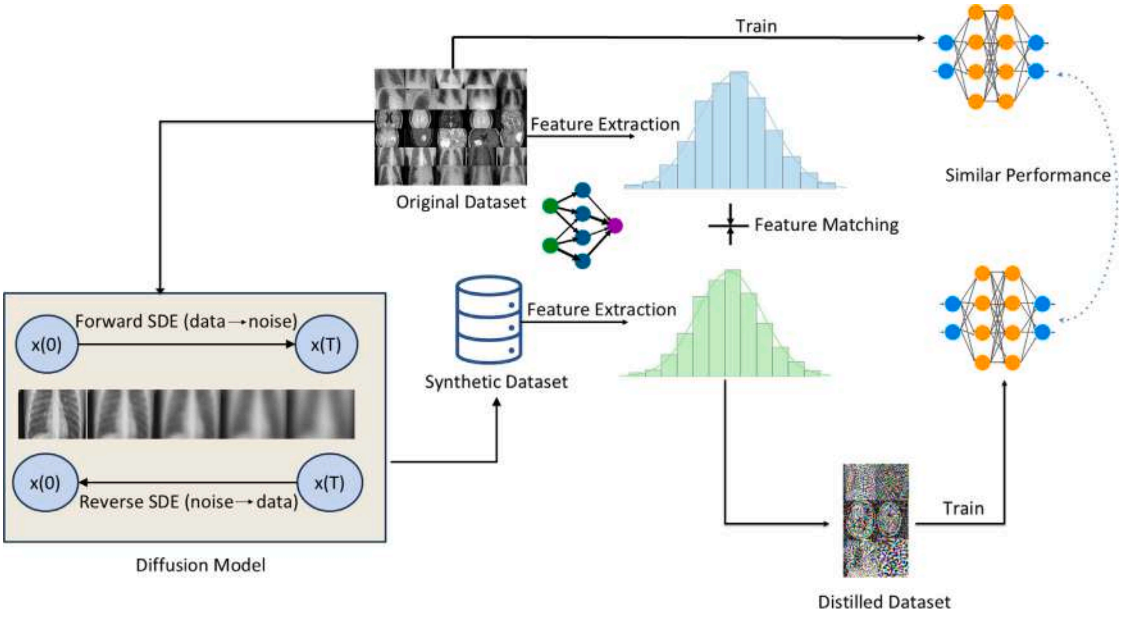

Fig. 2. Data distillation Using DDM technique (proposed).

3.1.1. *Distribution matching*

Although the parameter-wise evaluation demonstrated promising performance, Zhao and Bilen [19] conducted a visualization of the distilled data in a two-dimensional space, revealing a significant distributional gap between the distilled and target datasets. This indicates that the distilled data does not fully capture the underlying feature space of the target distribution. To address this limitation, they introduced a method that emphasizes distribution-level alignment between synthetic and target data in the dataset distillation process, as shown in Fig. 1.

The objective function used to optimize the synthetic data is:

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

where f_θ is parameterized by θ, and θ is sampled from a random distribution P_θ. x_i and $\hat{x}_i$ are the samples of the original and distilled datasets, respectively. $|\mathcal{S}|$ and $|\mathcal{T}|$ are the cardinality of dataset $\mathcal{S}$ and $\mathcal{T}$, respectively.

As illustrated in Eq. (1), distribution matching operates independently of model parameters and avoids the need for bilevel optimization, resulting in reduced memory consumption, making it more scalable for large datasets and high-dimensional data. The limitation of distribution matching is that it primarily aligns feature representations at the dataset level rather than enforcing fine-grained instance-level alignment.

3.1.2. *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. [24] in their work on Dataset Condensation (DC). The optimization objective, as shown in Eq. (2), 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:

$$\arg\min_{D_{\text{syn}}} \mathbb{E}_{\theta_0\sim P_\theta, c\sim C}\left[\sum_{t=0}^{T} D\left(\nabla_\theta \mathcal{L}_D^c(\theta_t), \nabla_\theta \mathcal{L}_{D_{\text{syn}}}^c(\theta_t)\right)\right] \qquad \text{s.t.} \quad \theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla_\theta \mathcal{L}_{D_{\text{syn}}}(\theta_t) \tag{2}$$

3.2. *Diffusion model*

Diffusion models are a class of deep generative models that learn to create realistic data by reversing a gradual noising process. The core idea is to start with real data and progressively add Gaussian noise over several steps until the data is completely destroyed into random noise. A neural network is then trained to learn the reverse process, i.e., how to remove noise step by step and recover the original data distribution. This formulation makes diffusion models both stable to train and capable of producing high-quality synthetic samples. Ho et al. introduced the Denoising Diffusion Probabilistic Model (DDPM) in which the training objective is simplified into a noise-prediction problem [9]. Given a clean data sample x_0, noise $\epsilon \sim \mathcal{N}(0, I)$ is added to produce a noisy version $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$, where the timestep t is sampled uniformly from $\{1, \dots, T\}$. The model $\epsilon_\theta(x_t, t)$ is then trained to predict the added noise using the following mean squared error (MSE) loss:

$$L(\theta) = \mathbb{E}_{x_0,\epsilon,t} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right]. \tag{3}$$

This simplified formulation is the standard objective function for diffusion models, as it is computationally efficient, easy to implement, and highly effective in enabling the generation of high-fidelity synthetic data. In our work, we leverage this property to generate realistic synthetic images from the original dataset, which can be used for tasks such as synthetic data generation, privacy preservation, and improving model generalization.

3.3. Autoencoder for image reconstruction

Autoencoders have emerged as a popular unsupervised learning alternative to traditional neural networks [26,27]. Their primary goal is to learn compact, informative representations of input data while preserving essential features. A basic autoencoder operates using backpropagation, reconstructing the original input at the output by passing it through an encoder with a reduced number of hidden neurons. This compressed representation captures the underlying structure of the data and is expected to enable accurate reconstruction of the original input.

Given a set of training samples $\mathbf{x} = [x_1, x_2, \dots, x_M]$, where each $x_k \in \mathbb{R}^m$ for $k = 1, \dots, M$, the goal of an autoencoder is to minimize the reconstruction loss between the original image and the reconstructed image:

$$\min \sum_{k=1}^{M} \left\| x_k - \hat{x}_k \right\|^2 \tag{4}$$

where x_k and $\hat{x}_k$ are the input and the reconstructed output, respectively.

A basic autoencoder learning process involves iteratively updating the network's weights or parameters to minimize the error between the input data and its reconstructed output. This process is continued till the reconstruction error goes below a predefined threshold [28].

3.4. Evaluation metrics

3.4.1. Structural Similarity Index Measure (SSIM)

SSIM is a metric that measures image quality loss due to processing like compression, transmission distortion, and reconstruction. In this approach, image loss is viewed as the alteration of perception in structural information, which comes closer to human vision perception. It includes major perception-oriented factors like structure, luminance, and contrast. The term structural information highlights strongly interdependent pixels or spatially closed pixels. Luminance denotes differences in light that are less apparent along the borders of an image, whereas contrast denotes differences that are less apparent in the texture of the image. SSIM is applied to assess perceived image quality by quantifying the similarity between the original and reconstructed images [29]. The SSIM between two image x and y is defined as:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

where:

- μ_x: mean intensity of image x
- μ_y: mean intensity of image y
- σ_x^2: variance of image x
- σ_y^2: variance of image y
- σ_{xy}: covariance between image x and y
- C_1 and C_2 are small constants to stabilize the division when denominators are close to zero.

The SSIM score ranges from −1 to 1, where 1 indicates perfect structural similarity between the original and reconstructed, and values closer to 0 or negative indicate increasing levels of dissimilarity.

3.4.2. Peak Signal-to-Noise Ratio (PSNR)

PSNR is a widely used metric for assessing the quality of reconstruction in image processing applications like compression, denoising, and reconstruction. PSNR relies on pixel-wise Mean Squared Error (MSE) between the reconstructed and original image. PSNR is the ratio of the maximum value of a signal to the power of noise that distorts the quality of its representation [30]. PSNR is mathematically defined as follows:

$$\mathrm{PSNR} = 10 \cdot \log_{10} \left(\frac{L^2}{\mathrm{MSE}} \right) \tag{6}$$

where:

- L represents the highest possible pixel value in the image,
- MSE denotes the Mean Squared Error between the original and the reconstructed image.

The Mean Squared Error (MSE) is given by:

$$\text{MSE} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [I(i,j) - K(i,j)]^2 \tag{7}$$

where:

- $I(i,j)$: pixel value of the original image at position (i,j),
- $K(i,j)$: pixel value of the reconstructed image at position (i,j),
- $m \times n$: dimensions of the image.

Higher PSNR indicates better image quality (i.e., the reconstructed image is closer to the original), while a lower PSNR indicates the reconstructed image differs from the original image.

3.4.3. *Learned Perceptual Image Patch Similarity (LPIPS)*

The LPIPS metric evaluates the perceptual similarity between two images using deep neural network feature maps. It measures the distance between feature activations extracted from networks like AlexNet or VGG [31]. LPIPS is defined as:

$$\text{LPIPS}(x,y) = \sum_{l} w_l \cdot \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \left\| \hat{\phi}_l^x(h,w) - \hat{\phi}_l^y(h,w) \right\|_2^2 \tag{8}$$

where:

- ϕ_l is the activation from layer l of a pretrained network.
- $\hat{\phi}_l$ represents normalized feature vectors.
- H_l, W_l denote the height and width of the lth feature map.
- w_l is the learned weight for each layer.
- x, y are the original and reconstructed images.

Lower LPIPS values indicate higher perceptual similarity between the original and the reconstructed image.

3.4.4. *Normalized Cross-Correlation (NCC)*

NCC quantifies the linear correlation between two images and is robust to brightness and contrast changes [32]. NCC is given as:

$$\text{NCC}(x,y) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}} \tag{9}$$

where:

- x_i, y_i are pixel values from the original and reconstructed images.
- $\bar{x}, \bar{y}$ are the mean values of x and y respectively.
- N is the total number of pixels.

NCC ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation). Higher values indicate better reconstruction.

3.4.5. *Feature Similarity Index Measure (FSIM)*

FSIM evaluates perceptual similarity based on the human visual system, incorporating phase congruency and gradient magnitude [33]. FSIM is defined as:

$$\text{FSIM}(x,y) = \frac{\sum_{i \in \Omega} PC_m(i) \cdot S_L(i)}{\sum_{i \in \Omega} PC_m(i)} \tag{10}$$

where:

- Ω is the spatial domain of the image.
- $PC_m(i) = \max(PC_x(i), PC_y(i))$ is the maximum phase congruency at pixel i.
- $S_L(i)$ is the similarity measure at pixel i considering phase congruency and gradient magnitude.

FSIM values lie between 0 and 1, with higher values indicating better perceptual quality between the original and reconstructed image.

3.4.6. Earth Mover's Distance (EMD)

EMD measures the minimum cost of transforming one distribution into another, often applied to image histograms [34]. EMD is defined as:

$$\mathrm{EMD}(P,Q) = \frac{\sum_{i,j} f_{ij} \cdot d_{ij}}{\sum_{i,j} f_{ij}} \tag{11}$$

where:

- f_{ij} is the optimal flow from p_i to q_j.
- d_{ij} is the ground distance between bins i and j.

Lower EMD indicates that the original and the reconstructed images are more similar.

3.4.7. Membership Inference Attack (MIA)

A Membership Inference Attack determines whether a specific sample was included in the training dataset of a target model. Given a machine learning model f_θ, parameterized by θ, and a sample (x,y), where x is the input data and y is the true label, the adversary attempts to infer a binary variable $m \in \{0,1\}$. Here, $m = 1$ indicates that (x,y) was part of the training dataset (a "member"), and $m = 0$ indicates that it was not (a "non-member") [35]. The attacker constructs an inference function as:

$$\hat{m} = g(f_\theta(x), y) \tag{12}$$

where $\hat{m}$ is the predicted membership status, and $g(\cdot)$ is the attack function that uses the model's outputs (e.g., prediction probabilities, losses, or gradients). The performance of MIAs is commonly evaluated using metrics such as attack accuracy, precision, recall, and Area Under the ROC Curve (AUC) [36].

3.4.8. Attribute Inference Attack (AIA)

An Attribute Inference Attack attempts to infer hidden or sensitive features of an individual, given partial information and access to the model. Let a data record be represented as $x = (x_{\mathrm{pub}}, x_{\mathrm{priv}})$, where x_{pub} denotes publicly known or observable attributes and x_{priv} represents the sensitive attribute to be inferred (e.g., medical condition, demographic information). The target model is denoted as f_θ, parameterized by θ. The adversary constructs an inference as:

$$\hat{x}_{\mathrm{priv}} = h(f_\theta(x_{\mathrm{pub}})) \tag{13}$$

where $\hat{x}_{\mathrm{priv}}$ is the predicted sensitive attribute and $h(\cdot)$ is the attack function that uses the model's outputs or intermediate representations. The success of AIAs is measured using metrics such as classification accuracy, F1-score, AUC, or mutual information between the true sensitive attribute x_{priv} and the predicted attribute $\hat{x}_{\mathrm{priv}}$. High values of these metrics indicate greater leakage of private information [37].

4. Evaluation methodology

This section outlines the pipeline followed to evaluate the privacy preservation capabilities of dataset distillation through autoencoder-based reconstruction analysis. The methodology comprises four key components: dataset selection, dataset distillation, image reconstruction using an autoencoder, and quantitative evaluation using similarity metrics.

Algorithm 1 Distribution Matching-Based Dataset Distillation

```
Input: $f_\theta$: feature extractor, $\mathcal{T}$: real dataset; $\mathcal{S}$: distilled data; $\alpha$: step size; $T_{\mathrm{steps}}$: number of optimization steps

Initialize $\mathcal{S} = \{\hat{x}_i\}_{i=1}^{|\mathcal{S}|}$ randomly
Compute real feature mean: $\mu_\mathcal{T} = \frac{1}{|\mathcal{T}|} \sum_{x_i \in \mathcal{T}} f_\theta(x_i)$
for training step $t = 1$ to $T_{\mathrm{steps}}$ do
    Compute synthetic feature mean: $\mu_\mathcal{S} = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} f_\theta(\hat{x}_i)$
    Compute matching loss: $\mathcal{L}_{DM} = \left\| \mu_\mathcal{S} - \mu_\mathcal{T} \right\|^2$
    Update synthetic data: $\mathcal{S} \leftarrow \mathcal{S} - \alpha \nabla_\mathcal{S} \mathcal{L}_{DM}$
end for
Output: Distilled dataset $\mathcal{S}$
```

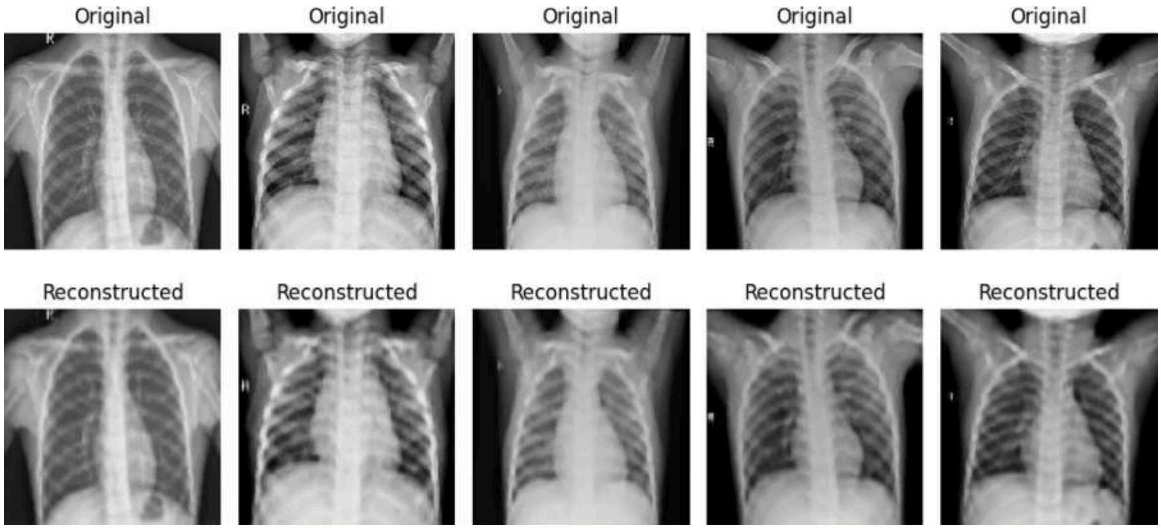

Fig. 3. Original pneumonia and its reconstructed image.

Algorithm 2 Gradient Matching-Based Dataset Distillation

Input: $\mathcal{D}$: real dataset; $\mathcal{D}_{syn}$: distilled dataset; C: number of classes; P_{θ_0}: distribution of model initializations; η: model learning rate; T: number of inner training steps; α: step size for updating $\mathcal{D}_{syn}$; $D(\cdot,\cdot)$: gradient distance function

```
Initialize $\mathcal{D}_{syn}$ randomly
for optimization step = 1 to $N$ do
  Sample model initialization $\theta_0 \sim P_{\theta_0}$
  for training step $t = 0$ to $T - 1$ do
    Compute real gradients: $g^c_{\mathcal{D}}(\theta_t) = \nabla_{\theta_t} \mathcal{L}^c_{\mathcal{D}}(\theta_t)$ for each class $c \in C$
    Compute synthetic gradients: $g^c_{\mathcal{D}_{syn}}(\theta_t) = \nabla_{\theta_t} \mathcal{L}^c_{\mathcal{D}_{syn}}(\theta_t)$
    Compute gradient matching loss: $\mathcal{L}_{GM} = \sum_{c=1}^{C} D\left(g^c_{\mathcal{D}}(\theta_t), g^c_{\mathcal{D}_{syn}}(\theta_t)\right)$
    Update model parameters: $\theta_{t+1} \leftarrow \theta_t - \eta \cdot g^c_{\mathcal{D}_{syn}}(\theta_t)$
  end for
  Update synthetic data: $\mathcal{D}_{syn} \leftarrow \mathcal{D}_{syn} - \alpha \nabla_{\mathcal{D}_{syn}} \mathcal{L}_{GM}$
end for
```

Output: Distilled dataset $\mathcal{D}_{syn}$

4.1. Dataset description

We have used three medical image datasets for our experiment. One is the Chest-Xray-Pneumonia dataset [38], containing chest X-ray images categorized into two classes: ‘Normal’ and ‘Pneumonia’, out of which 8917 are pneumonia cases and 3312 are normal cases. The second dataset used is COVID-19 [39], a chest X-ray image labeled as ‘Normal’ or ‘Infected’ with COVID-19, out of which 3616 images are of the normal class and 3616 images are of the infected class. We have also used Brain Tumor [40], an MRI image of brain tumors. Of these images, 9828 were labeled as having brain tumors, and 9546 were normal images.

4.2. Distillation process

To generate compact, privacy-aware datasets, we applied a dataset distillation technique that compresses the training data into a significantly smaller synthetic dataset. The distillation process preserves task-relevant features needed for classification while eliminating redundant or sensitive information. This distilled data can be used in a resource-limited environment.

4.2.1. Distribution matching based distillation (DM)

The distilled dataset for each domain was generated using a feature distribution matching-based approach having an objective function shown in Eq. (1) that minimizes the statistical discrepancy between the latent representations of real and synthetic data. The algorithm used for creating the distilled images using distribution matching techniques is shown in Algorithm 1. In this algorithm,

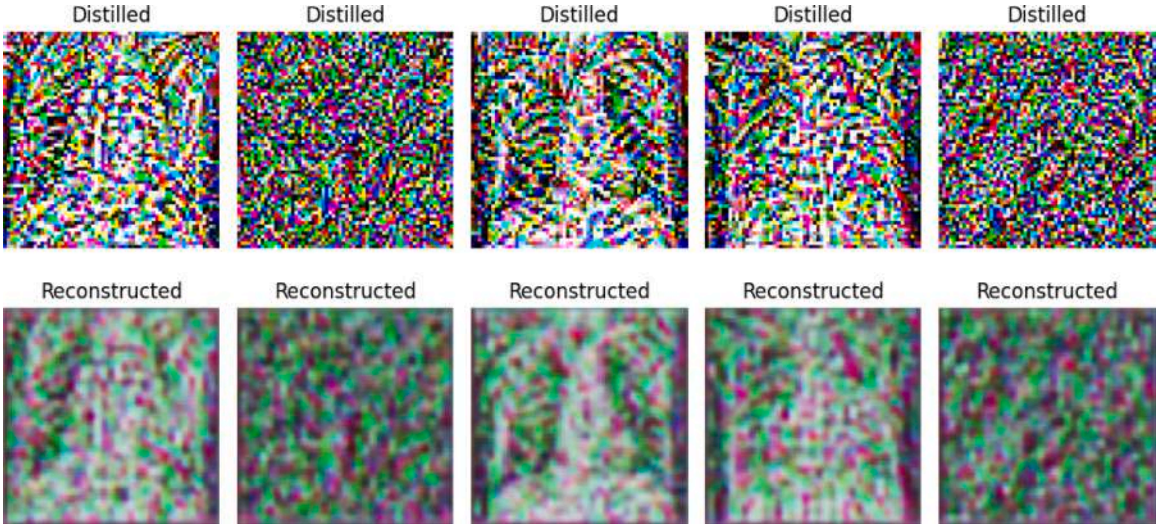

Fig. 4. Distilled pneumonia and its reconstructed image using DM technique.

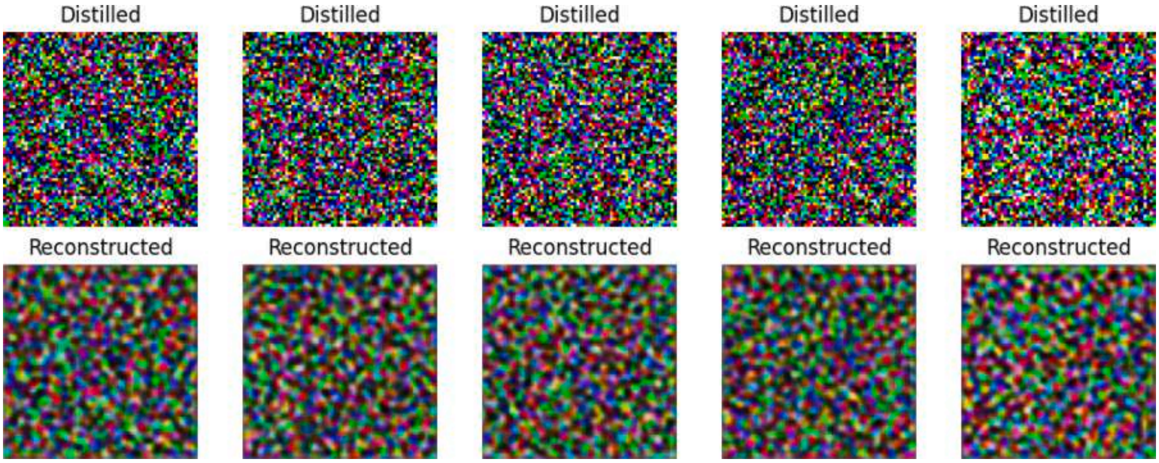

Fig. 5. Distilled pneumonia and its reconstructed image using GM technique.

synthetic data points are iteratively updated to align their representations with those of the original dataset in a high-dimensional feature space. We have used a pre-trained ResNet-18 as the fixed feature extractor to capture semantically meaningful embeddings, which guide the distillation process. Each class in the dataset was distilled into a limited number of images, drastically reducing the dataset size while maintaining high classification performance. This makes the approach particularly suitable for edge devices or resource-constrained environments, where full dataset access is infeasible.

4.2.2. Gradient matching based distillation (GM)

The distilled dataset for each domain was also generated using a gradient-matching-based approach, whose objective function in Eq. (2) minimizes the discrepancy between the training gradients computed on real data and those computed on synthetic data. The algorithm for gradient matching is provided in Algorithm 2. In this method, for every optimization step, a model is initialized from a distribution of parameters, and class-wise gradients are computed both on the real dataset and the synthetic dataset. The synthetic gradients are then aligned with the real gradients by minimizing a gradient matching loss that aggregates the discrepancy across all classes. During the inner loop, model parameters are updated using synthetic gradients, while in the outer loop, the synthetic data itself is refined by backpropagating through the gradient matching loss. This bidirectional optimization ensures that the distilled dataset induces training dynamics that closely mimic those of the full dataset. Each class is distilled into a small number of synthetic samples, which drastically reduces storage and training time while preserving gradient information critical for generalization. Such an approach is particularly suitable for efficient model training in resource-limited environments, where matching the optimization trajectory of real data is more important than directly matching feature statistics.

Algorithm 3 Diffusion and Distribution Matching (DDM) Based Dataset Distillation

```
Input: $\mathcal{T}$: real dataset with class labels; $\mathcal{S}$: distilled dataset $f_\phi$: feature extractor; $\epsilon_\theta$: diffusion model; $N_{\text{cand}}$: candidates per class; $K$: selected samples per class; $\tau$: diversity threshold

1: Train diffusion model: optimize $\epsilon_\theta$ by minimizing:
       $\mathcal{L}_{\text{Diff}} = \mathbb{E}_{x_0,t,\epsilon}\|\epsilon - \epsilon_\theta(x_t, t)\|^2$
2: Compute real statistics: for each class $c$, compute feature mean:
       $\mu_c = \frac{1}{|\mathcal{T}_c|}\sum_{x\in\mathcal{T}_c} f_\phi(x)$
3: Generate candidates: for each class $c$, sample $N_{\text{cand}}$ synthetic images $\{\tilde{x}\}$ using $\epsilon_\theta$
4: Extract features: compute $f_\phi(\tilde{x})$ for each synthetic candidate
5: Compute Distribution Matching Loss: Refine candidates by minimizing:
       $\mathcal{L}_{\text{DM}} = \sum_c \left\| \frac{1}{|\mathcal{S}_c|} \sum_{\tilde{x}\in\mathcal{S}_c} f_\phi(\tilde{x}) - \mu_c \right\|^2$
       where $\mathcal{S}_c$ denotes the set of current candidates for class $c$.
6: Refine candidates: Update synthetic samples to reduce distribution gap:
       $\tilde{x} \leftarrow \tilde{x} - \eta\nabla_{\tilde{x}}\mathcal{L}_{\text{DM}}$
7: Assign candidates to classes: for each $\tilde{x}$, find nearest class $c$ using
       $d(\tilde{x}, c) = \|f_\phi(\tilde{x}) - \mu_c\|$
8: Diversity-aware top-$K$ selection: for each class $c$, greedily select $K$ candidates with smallest $d(\tilde{x}, c)$, ensuring each new sample has feature distance $\geq \tau$
9: Construct distilled dataset: $\mathcal{S} = \bigcup_c\{\tilde{x}_{\text{selected}}\}$

Output: Distilled dataset $\mathcal{S}$
```

4.2.3. Diffusion and distribution matching based distillation (DDM)

In this work, we propose a novel dataset distillation framework that integrates diffusion-based generative modeling with distribution matching for effective synthetic data selection as shown in Fig. 2 . The distilled dataset for each domain was generated using a novel diffusion–distribution–matching approach (DDM), whose procedure is detailed in Algorithm 3. In this method, a diffusion model is first trained to learn the underlying data distribution by minimizing a denoising objective that reconstructs clean samples from noisy inputs. Once trained, the diffusion model generates a large pool of synthetic candidates for each class, which are subsequently embedded into a feature space using a pretrained feature extractor. To ensure alignment between real and synthetic distributions, class-wise feature means are computed from the real dataset, and a distribution matching loss is applied to refine the synthetic candidates by minimizing their feature discrepancy with respect to these real statistics. This refinement step encourages the generated samples to preserve class-level structure and semantic fidelity. Following refinement, synthetic candidates are assigned to the nearest class mean in feature space, and a diversity-aware selection strategy is employed to retain only the most representative and non-redundant samples per class. The final distilled dataset is constructed from these selected candidates, yielding a compact yet diverse representation of the real data. By unifying diffusion-based generative modeling with distribution-level alignment and diversity-aware selection, this approach effectively balances sample quality, class fidelity, and feature diversity, thereby enhancing the utility of distilled datasets for downstream training.

4.3. Autoencoder-based reconstruction

To evaluate privacy leakage from both the original and distilled datasets, we trained a convolutional autoencoder designed to reconstruct input images. The quality of these reconstructions serves as a measure of how much visual and structural information the dataset retains, which directly correlates with potential privacy risks. The autoencoder architecture consisted of an encoder with convolutional and max-pooling layers that compressed the input into a latent space and a decoder with convolutional and upsampling layers that aimed to reconstruct the original image from this latent representation. The final output layer used a sigmoid activation to ensure the reconstructed pixel values remained within the range [0, 1]. The model was trained using the Mean Squared Error (MSE) loss function and the Adam optimizer over 10 epochs.

The autoencoder was trained separately on the original and distilled datasets to assess the extent of information retained in both datasets. A high-quality reconstruction implies that the dataset still contains substantial visual information, potentially revealing identifiable or sensitive content. Conversely, poor reconstructions suggest reduced fidelity and, therefore, better privacy preservation.

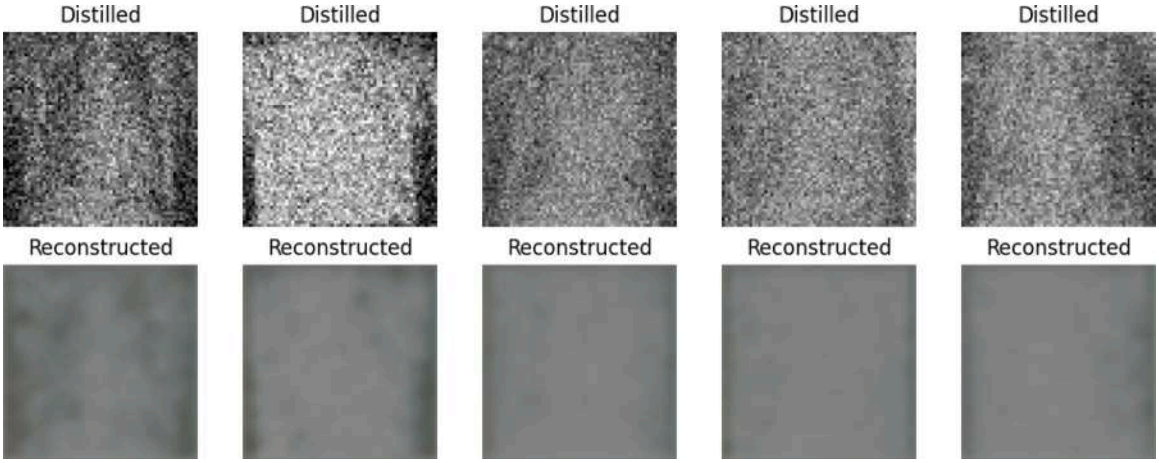

Fig. 6. Distilled pneumonia and its reconstructed image using DDM (proposed) technique.

Table 1
Test accuracy (↑) of ResNet50 classifier trained on original and distilled datasets.

	Pneumonia [38]	COVID-19 [39]	Brain Tumor [40]
Original dataset	96.22%	92.86%	98.92%
Distilled dataset DM	90.48%	85.71%	81.00%
Distilled dataset GM	78.00%	84.00%	80.00%
Distilled dataset DDM (Proposed)	**92.00%**	**88.00%**	**85.00%**

5. Results and discussion

5.1. Classification performance

We evaluated the test accuracy of the ResNet50 [41] classifier trained on both the original and distilled datasets, partitioned into 80% as training and 20% as testing across three medical imaging tasks: Pneumonia detection, COVID-19 classification, and Brain Tumor classification. As expected, models trained on the original datasets consistently achieved the highest accuracy, with 96.22% for Pneumonia, 92.86% for COVID-19, and 98.92% for Brain Tumor. When trained on distilled datasets, the performance dropped due to information compression, yet the results remained competitive. Specifically, distribution matching (DM) achieved 90.48% for the Pneumonia dataset, 85.71% for COVID-19, and 81.00% for the Brain Tumor datasets. While gradient matching (GM) achieved an accuracy of 78.00%, 84.00%, and 80.00% for Pneumonia, COVID-19, and Brain Tumor datasets, respectively. In contrast, our proposed data distillation approach, i.e., diffusion–distribution–matching (DDM) approach, achieved a significantly better accuracy, with 92.00% for Pneumonia, 88.00% for COVID-19, and 85.00% for Brain Tumor, respectively. These results clearly highlight that our proposed method outperforms both DM and GM, narrowing the gap with models trained on full datasets as shown in Table 1.

Overall, this demonstrates that while dataset distillation inevitably introduces a trade-off between data efficiency and model accuracy, our proposed approach more effectively captures essential features for classification. Thus, it offers a promising direction for deploying deep learning models in resource-constrained environments while maintaining robust classification performance.

5.2. Image reconstruction evaluation

To evaluate the privacy-preserving potential of dataset distillation, we employed an autoencoder-based model to reconstruct images from both the original and distilled datasets. The reconstructed image from the original pneumonia dataset is shown in Fig. 3, and the reconstructed image from the distilled pneumonia dataset using the distribution matching technique is shown in Fig. 4, the reconstructed image from the distilled pneumonia dataset using the gradient matching technique is shown in Fig. 5, and the reconstructed image from the distilled pneumonia dataset using our proposed technique i.e., combination of diffusion and distribution matching (DDM) is shown in Fig. 6. It is quite clear from these figures that the autoencoder is able to reconstruct the images that match the original, but the reconstruction of the distilled data differs a lot from the original image visually.

The reconstructed image from the original data for COVID-19 is shown in Fig. 7. The reconstructed images from the distilled data for COVID-19 using distribution matching, gradient matching, and our proposed distillation technique are shown in Figs. 8, 9, and 10, respectively. The figures clearly show that the autoencoder successfully reconstructs images that closely resemble the originals, whereas the reconstructions from the distilled data exhibit significant visual discrepancies. We have also reconstructed the image using an autoencoder for the brain tumor dataset. The reconstructed image from the original data is shown in Fig. 11

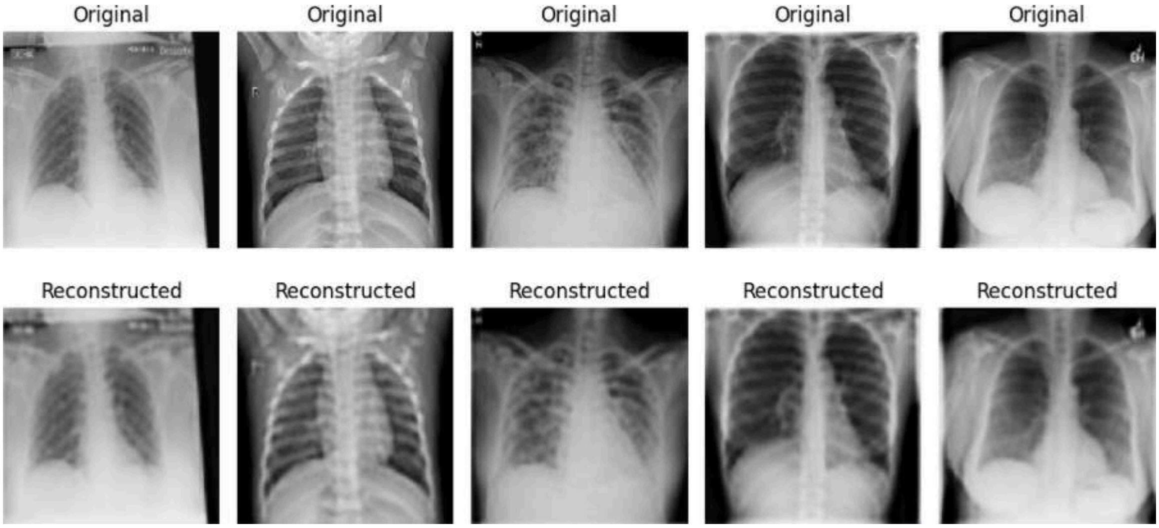

Fig. 7. Original COVID-19 and its reconstructed image.

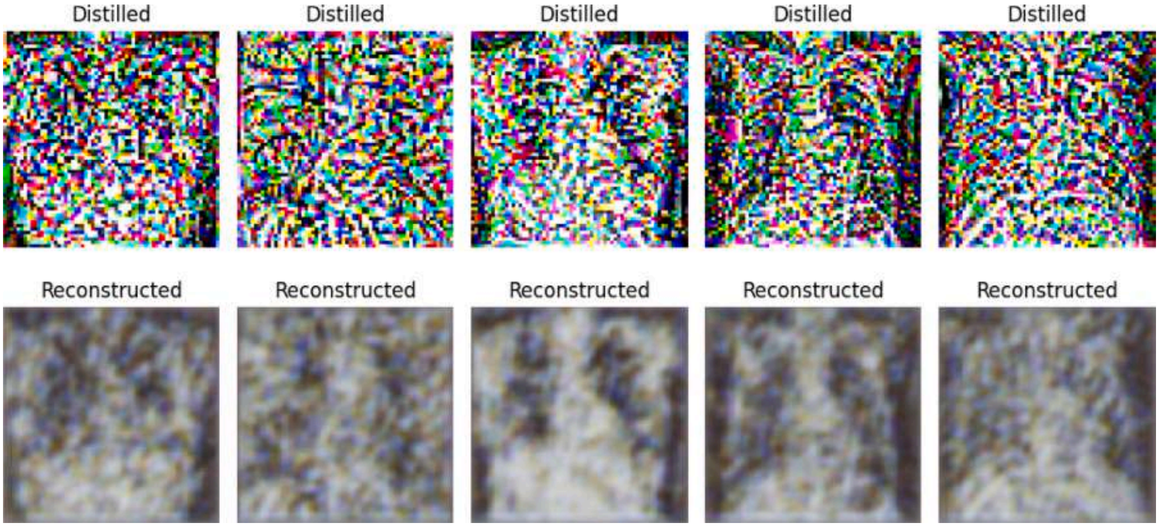

Fig. 8. Distilled COVID-19 and its reconstructed image using DM technique.

and the reconstructed image from the distilled data using the distribution matching approach is shown in Fig. 12. The reconstructed image for the brain tumor dataset using the gradient matching technique is shown in Fig. 13. Fig. 14 depicts the reconstruction done using our proposed distillation technique for the brain tumor dataset. As evident from the figures, the autoencoder effectively reconstructs images that are visually consistent with the originals. On the other hand, reconstructions based on the distilled data display noticeable deviations from the original images. Since no details can be inferred from the distilled reconstructed images, this suggests that the method effectively preserves privacy. So, to evaluate this, we then compared the reconstructions done using various data distillation approaches with the corresponding original images using different image similarity metrics such as SSIM, PSNR, LPIPS, NCC, FSIM, and EMD.

5.2.1. Structural Similarity Index Measure (SSIM)

The results in Fig. 15 presents the SSIM results with 95% confidence interval (CI) across Pneumonia, COVID-19, and Brain Tumor datasets. As expected, reconstructions from the original datasets exhibit the highest similarity to the input data, with SSIM scores of 0.7577, 0.9015, and 0.7042, respectively. These high values confirm that original data reconstructions carry the strongest privacy risk. In contrast, distilled datasets achieve significantly lower SSIM scores, indicating reduced reconstruction fidelity and improved privacy preservation. Distribution matching (DM) yields scores of 0.3694, 0.2155, and 0.2077, while gradient matching (GM) produces moderately higher values of 0.4451, 0.2879, and 0.4139, reflecting weaker privacy protection compared to DM across three datasets.

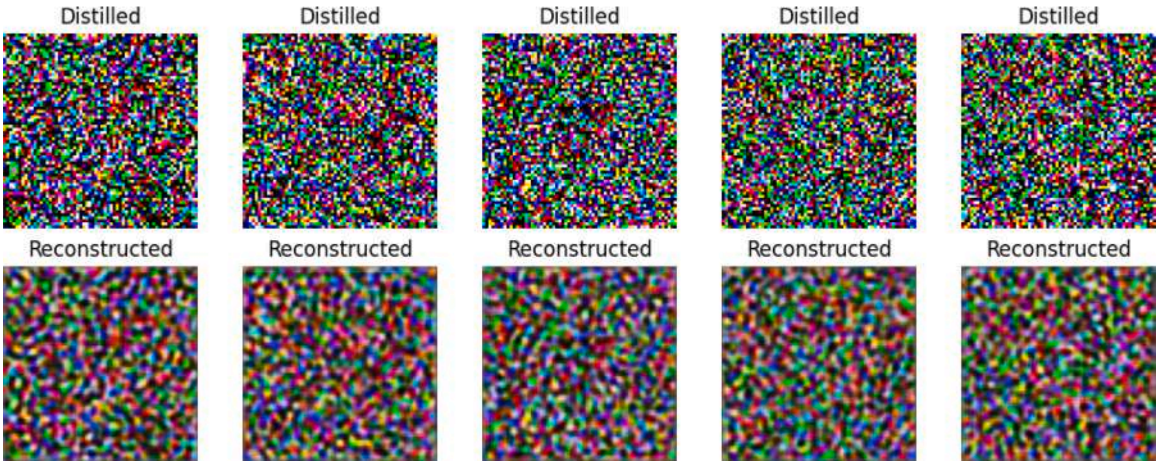

Fig. 9. Distilled COVID-19 and its reconstructed image using GM technique.

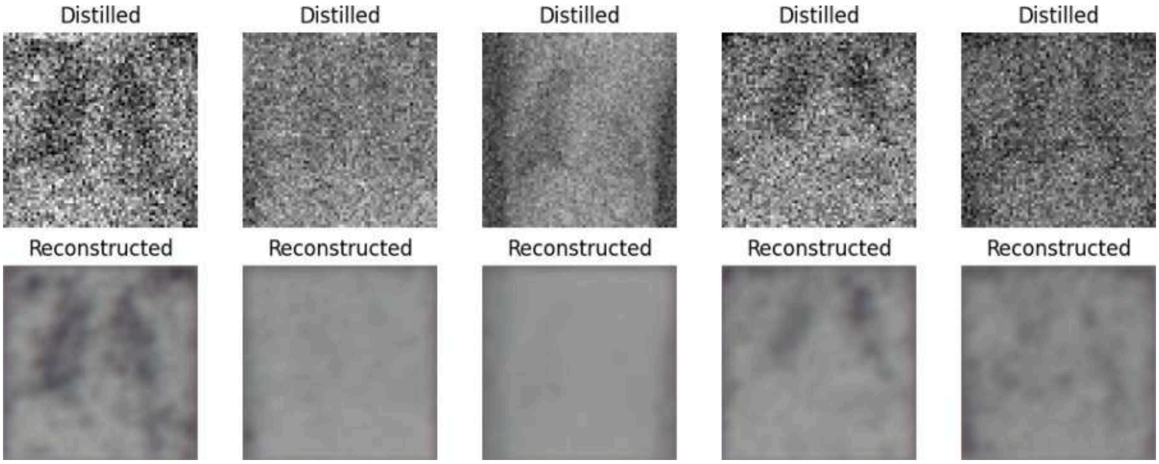

Fig. 10. Distilled COVID-19 and its reconstructed image using DDM (proposed) technique.

Our proposed (DDM) method achieves SSIM values of 0.1773 for Pneumonia, 0.2377 for COVID-19, and 0.2159 for Brain Tumor. These values are consistently lower than those of GM and, in most cases, competitive with or better than DM, striking a favorable balance between utility and privacy. Overall, the findings confirm that dataset distillation reduces reconstructability, thereby protecting privacy, and that our proposed approach provides the strongest privacy advantage while retaining essential features for downstream classification tasks.

5.2.2. Peak Signal-to-Noise Ratio (PSNR)

The PSNR results in Fig. 16 with 95% CI further highlight the disparity in reconstruction quality between original and distilled datasets. Reconstructions from the original datasets achieve the highest PSNR values of 21.54 dB for Pneumonia, 25.52 dB for COVID-19, and 21.20 dB for Brain Tumor, demonstrating strong fidelity to the original data and, consequently, higher privacy risk. In contrast, distilled datasets yield notably lower PSNR scores, highlighting their reduced reconstruction quality and enhanced privacy protection. Specifically, distribution matching (DM) produces values of 14.83 dB, 13.67 dB, and 14.41 dB for Pneumonia, COVID-19, and Brain Tumor, respectively. Gradient matching (GM) results in higher PSNR scores of 19.11 dB, 18.86 dB, and 19.30 dB, reflecting weaker privacy preservation due to closer resemblance to original images.

DDM, our proposed method, achieves PSNR values of 15.45 dB for Pneumonia, 13.52 dB for COVID-19, and 14.60 dB for Brain Tumor. These scores are consistently lower than GM and competitive with DM, demonstrating that our approach maintains a stronger balance between utility and privacy. Overall, the findings confirm that our proposed method reduces reconstruction fidelity more effectively than GM, offering improved privacy while retaining sufficient feature information for downstream tasks.

5.2.3. Learned Perceptual Image Patch Similarity (LPIPS)

The LPIPS results in Fig. 17 with 95% CI further illustrate perceptual differences between original and distilled reconstructions. Reconstructions from the original datasets achieve the lowest LPIPS values, 0.0312 for Pneumonia, 0.0156 for COVID-19, and 0.0179

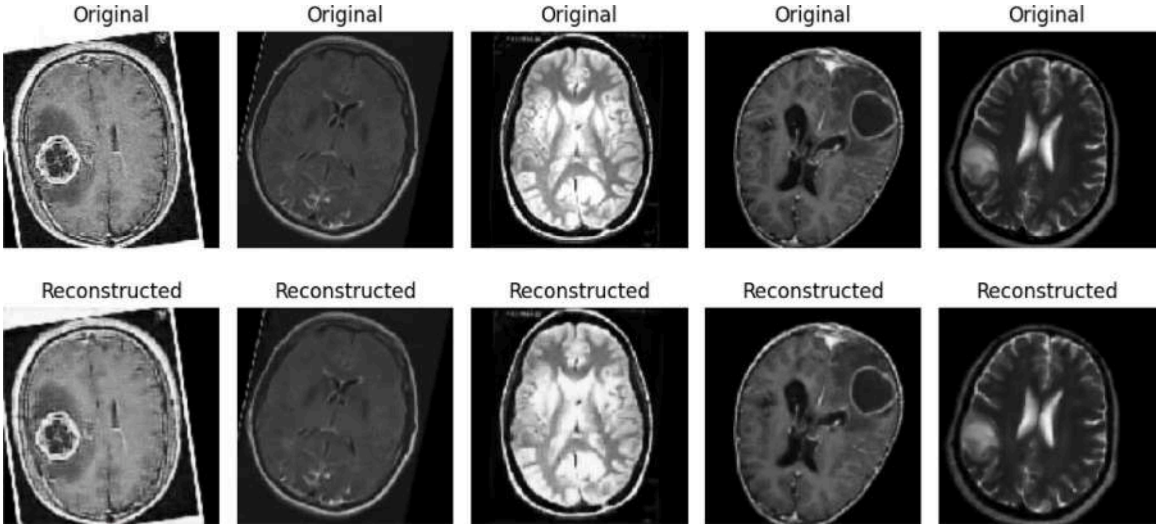

Fig. 11. Original brain tumor and its reconstructed image.

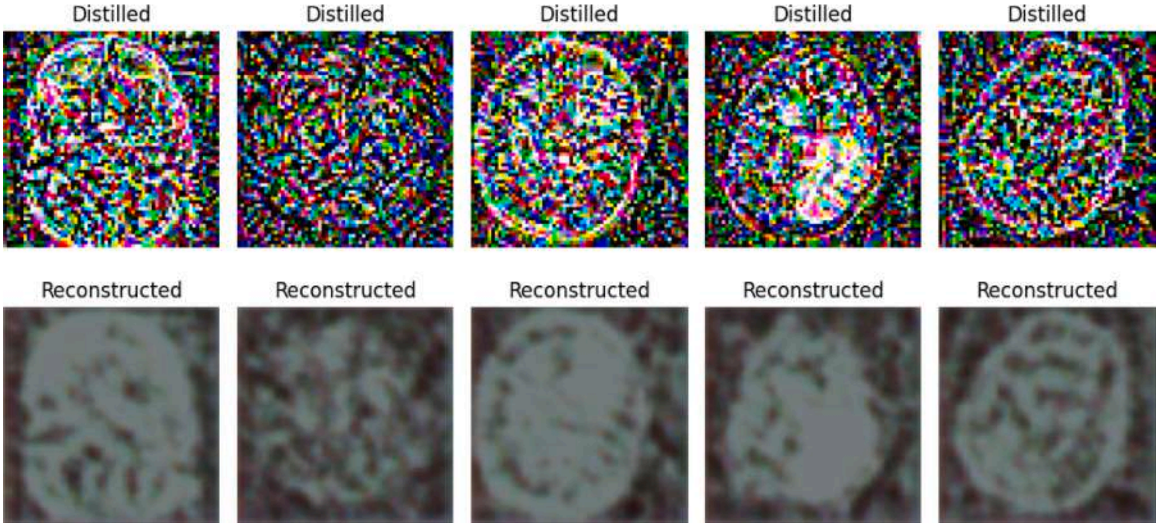

Fig. 12. Distilled brain tumor and its reconstructed image using DM technique.

for Brain Tumor, indicating high visual fidelity to the source images. While this suggests strong data utility, it also reflects weaker privacy preservation. In contrast, distilled datasets yield substantially higher LPIPS scores, highlighting greater perceptual distortion and reduced resemblance to the original inputs. Distribution matching produces 0.5303, 0.7517, and 0.7682 across the three datasets, whereas gradient matching yields comparatively lower scores of 0.3168, 0.2834, and 0.3062 for pneumonia, COVID-19, and brain tumor, respectively, pointing to weaker privacy protection.

DDM data distillation approach achieves 0.6902 for pneumonia, 0.7403 for COVID-19, and 0.7730 for brain tumors. These values are consistently higher than gradient matching and comparable with the distribution matching, demonstrating the ability of the proposed method to enforce stronger perceptual dissimilarity while retaining essential features for classification. These results underline that our method offers an improved trade-off between model performance and privacy, outperforming existing distillation techniques in privacy preservation.

5.2.4. Normalized Cross-Correlation (NCC)

The NCC results in Fig. 18 with 95% CI highlight the reconstruction differences between original and distilled datasets. As expected, the original datasets achieve the strongest correlation with the inputs, with scores of 0.9178 for Pneumonia, 0.9640 for COVID-19, and 0.8540 for Brain Tumor, indicating high pixel-level similarity and, consequently, greater privacy risk. Distillation substantially lowers these correlations. Distribution matching (DM) records values of 0.5484, 0.5861, and 0.5666, while gradient

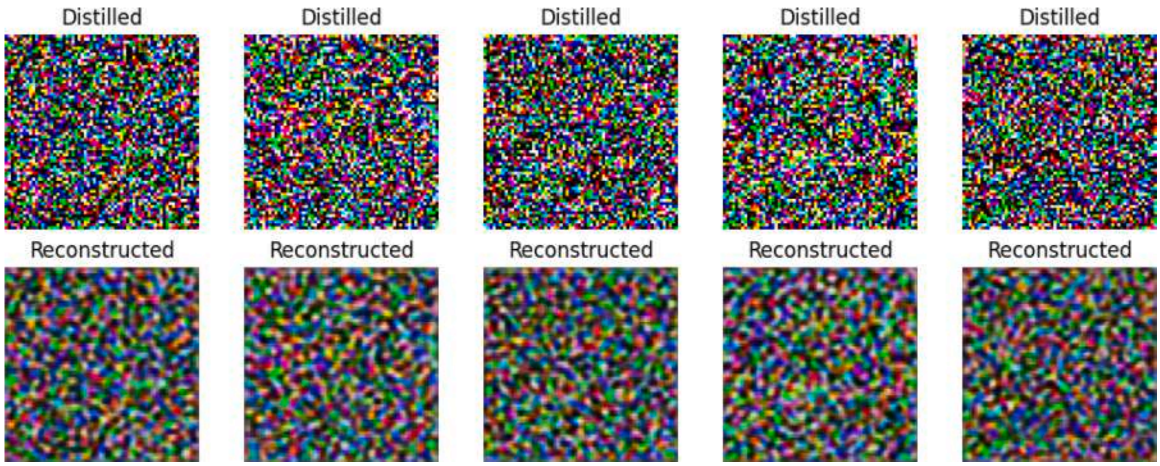

Fig. 13. Distilled brain tumor and its reconstructed image using GM technique.

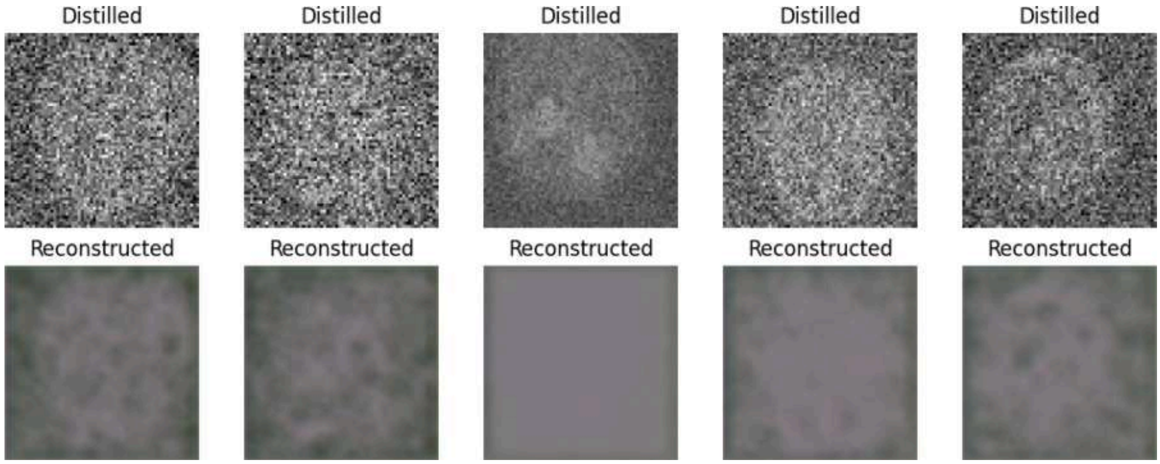

Fig. 14. Distilled brain tumor and its reconstructed image using DDM (proposed) technique.

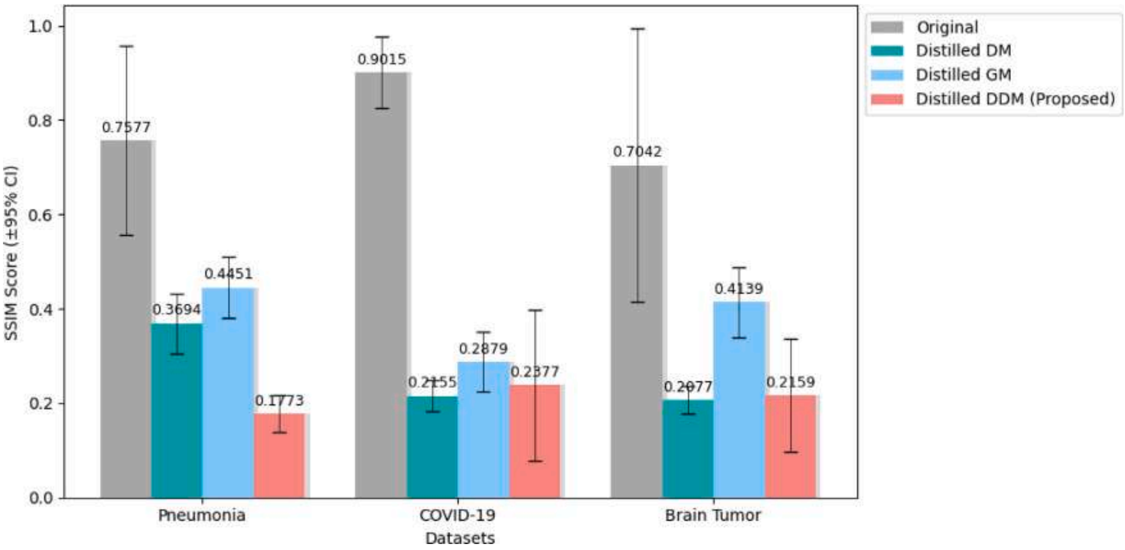

Fig. 15. SSIM score.

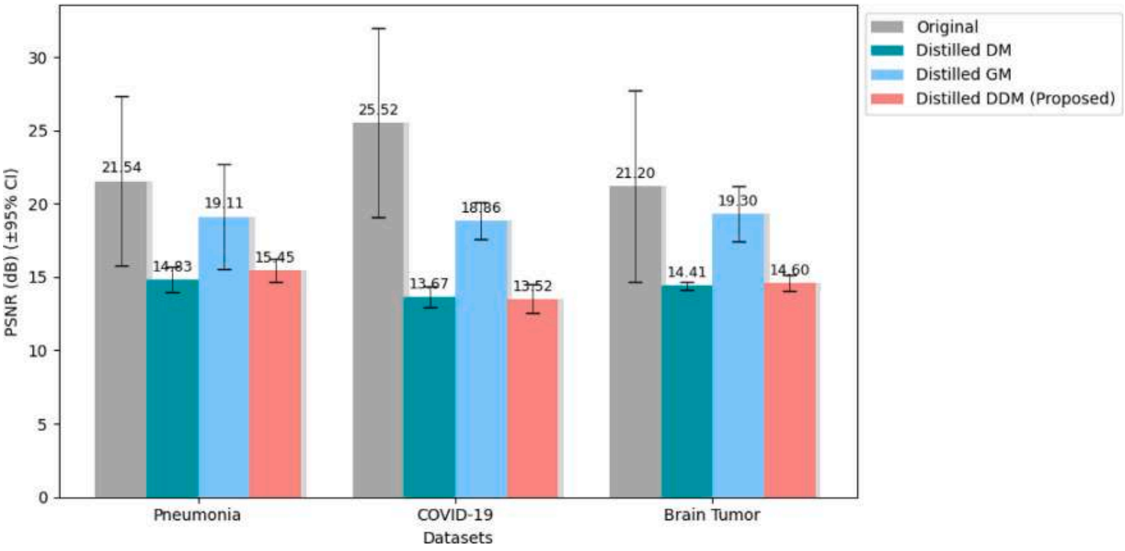

Fig. 16. PSNR score.

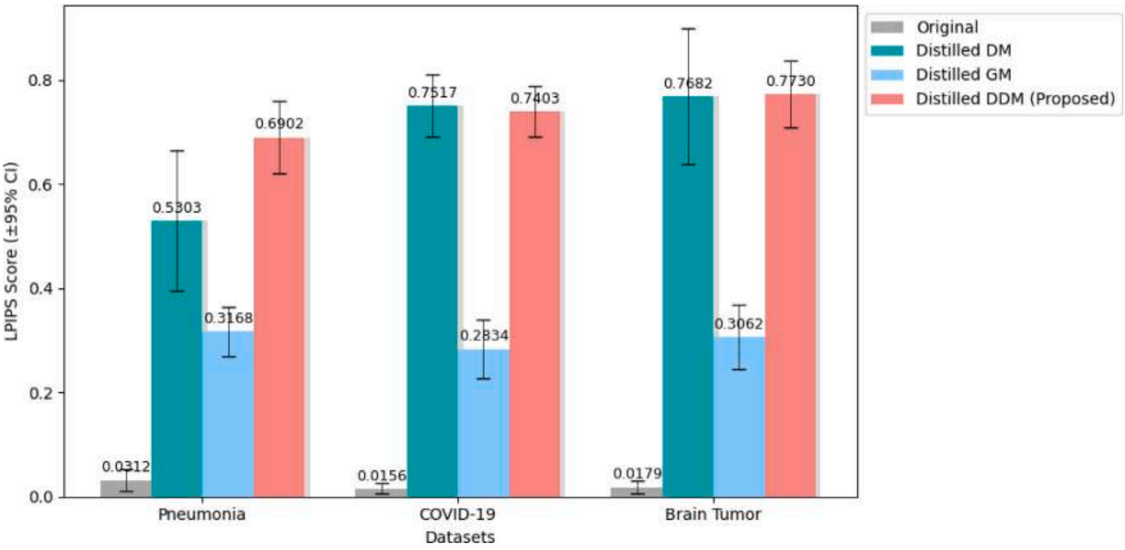

Fig. 17. LPIPS score.

matching (GM) produces 0.5659, 0.6108, and 0.4524 for Pneumonia, COVID-19, and Brain Tumor, respectively. The higher scores observed for GM suggest weaker privacy protection compared to DM.

Our proposed method achieves NCC scores of 0.4654 for Pneumonia, 0.2584 for COVID-19, and 0.4421 for Brain Tumor. These results are consistently lower than GM and DM, demonstrating that our approach effectively reduces pixel-level similarity to the original data. This balance confirms that the proposed method strengthens privacy while retaining sufficient information for classification.

5.2.5. Feature Similarity Index Measure (FSIM)

The FSIM results shown in Fig. 19 with 95% CI highlight clear differences between reconstructions from original and distilled datasets. The original datasets consistently produce the highest FSIM values of 0.7310 for Pneumonia, 0.7829 for COVID-19, and 0.7168 for Brain Tumor, reflecting strong structural feature alignment with the source images and, therefore, greater privacy risks. Among the distilled methods, distribution matching (DM) achieves scores of 0.5785, 0.4435, and 0.4330 across the three datasets, while gradient matching (GM) yields higher values of 0.6752, 0.6694, and 0.6293. The higher FSIM of GM indicates closer resemblance to the original data, suggesting weaker privacy protection.

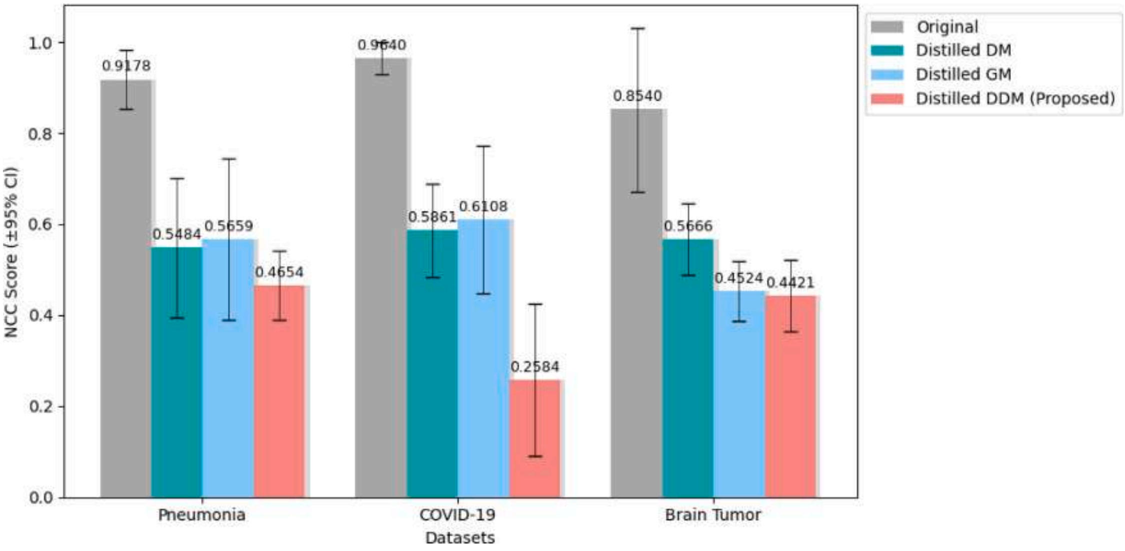

Fig. 18. NCC score.

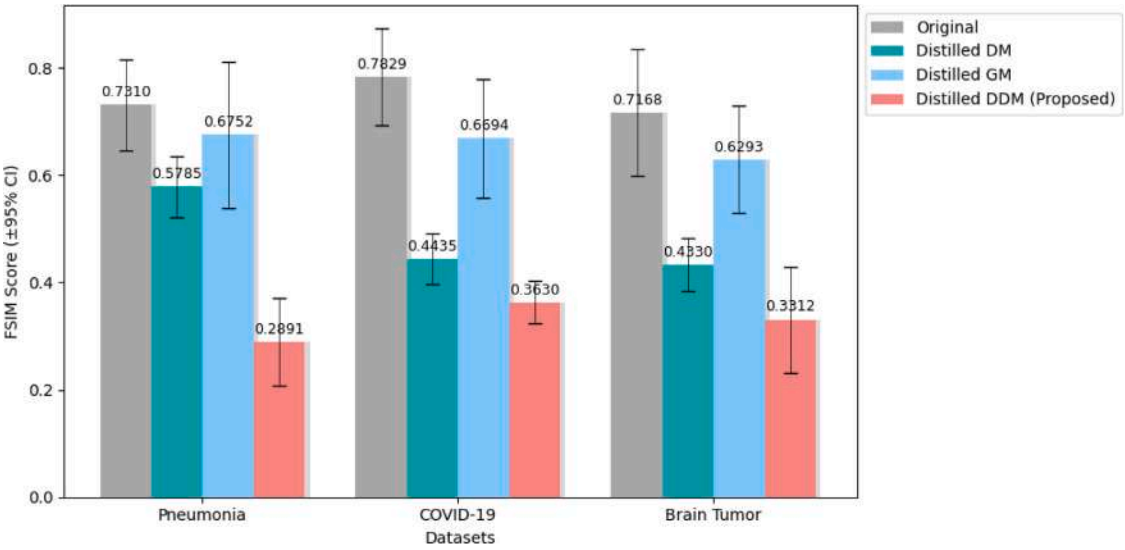

Fig. 19. FSIM score.

DDM distillation technique records FSIM values of 0.2891 for Pneumonia, 0.3630 for COVID-19, and 0.3312 for Brain Tumor. These are consistently lower than GM and DM, demonstrating that the proposed approach effectively reduces structural similarity to original data while retaining the essential discriminative features required for classification. Overall, the FSIM analysis further confirms that our method offers a more favorable balance between privacy preservation and task utility compared to existing distillation techniques.

5.2.6. Earth Mover's Distance (EMD)

The EMD, a metric for assessing distributional divergence, was computed to evaluate the quality of reconstructions. For the Pneumonia dataset, the original reconstructions yielded an EMD score of 2.6784 for Pneumonia, 2.8553 for COVID-19, and 1.5865 for Brain Tumor, indicating greater similarity to the input images. Distilled datasets, on the other hand, produce substantially higher EMD values, confirming reduced reconstruction fidelity and stronger privacy protection. Specifically, distribution matching (DM) yields 23.5902, 27.3656, and 19.6181 for Pneumonia, COVID-19, and Brain Tumor, respectively. Gradient matching (GM), however, achieves lower EMD scores of 8.8439, 10.0002, and 11.3043, compared to the distribution matching technique.

Our proposed method yields EMD scores of 21.7694 for Pneumonia, 24.9758 for COVID-19, and 18.8940 for Brain Tumor, as shown in Fig. 20 with 95% CI. These results are consistently higher than GM and competitive with DM, demonstrating that our

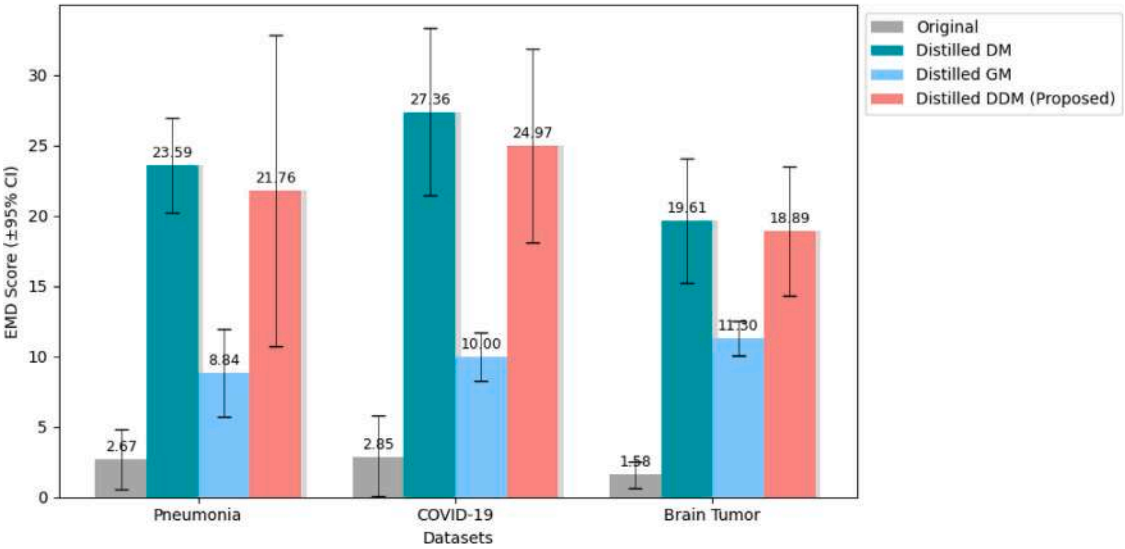

Fig. 20. EMD score.

Table 2
MIA and AIA AUC values for different datasets.

Dataset	Pneumonia [38]		COVID-19 [39]		Brain Tumor [40]	
	MIA (AUC)↓	AIA (AUC)↓	MIA (AUC)↓	AIA (AUC)↓	MIA (AUC)↓	AIA (AUC)↓
Original dataset	0.5149	0.9794	0.5192	0.9172	0.5385	0.9687
Distilled DM	0.4571	0.5000	0.5010	0.6475	0.5186	0.6894
Distilled GM	0.5100	0.8062	0.6500	0.6808	0.5070	0.6818
Distilled DDM (Proposed)	**0.4967**	**0.6164**	**0.4945**	**0.5824**	**0.4911**	**0.5554**

approach enforces stronger distributional divergence while maintaining task-relevant information. The EMD findings reaffirm that our method achieves an improved privacy–utility trade-off, offering enhanced robustness for privacy-sensitive applications such as medical image analysis.

5.3. Attack based evaluation

5.3.1. Membership Inference Attack (MIA)

The Membership Inference Attack (MIA) Area under the ROC Curve (AUC) values across all datasets reveal that distillation consistently offers privacy advantages compared to models trained on original datasets. In the Pneumonia dataset, the original model achieves 0.5149, whereas the distilled DM (0.4571), distilled GM (0.5100), and our DDM distillation approach (0.4967) all show lower or comparable values, indicating reduced vulnerability. For the COVID-19 dataset, the original model yields a value of 0.5192, while the distilled DM (0.5010) and distilled DDM methods (0.4945) achieve lower values. In contrast, distilled GM shows a higher AUC of 0.6500. In the Brain Tumor dataset, the original model obtains 0.5385, whereas distilled DM (0.5186), distilled GM (0.5070), and distilled DDM (0.4911) reduce this vulnerability, as shown in Table 2. These results demonstrate that distillation, and particularly the proposed distillation technique, effectively lowers the success rate of MIAs, thereby enhancing privacy.

5.3.2. Attribute Inference Attack (AIA)

The Attribute Inference Attack (AIA) AUC values show an even stronger privacy benefit from distillation. For the Pneumonia dataset, the original model achieves a very high accuracy of 0.9794, while distilled DM reduces it to 0.5000, distilled GM achieves 0.8062, and the distilled DDM method further lowers it to 0.6164. In the COVID-19 dataset, the original model achieves 0.9172, compared to 0.6475 for distilled DM, 0.6808 for distilled GM, and a much lower 0.5824 with the proposed method. Similarly, in the Brain Tumor dataset, the original model achieves a score of 0.9687, which drops to 0.6894 for distilled DM, 0.6818 for distilled GM, and further to 0.5554 for the distilled DDM, as shown in Table 2. These consistent reductions across all datasets highlight that dataset distillation not only improves resistance against MIAs but also substantially enhances protection against AIAs, making it an effective privacy-preserving strategy.

These results suggest that the autoencoder is less capable of accurately reconstructing images from the distilled datasets, indicating that much of the fine-grained or sensitive visual information is lost during the distillation process. This reduced reconstructability enhances the privacy-preserving properties of the distilled datasets, as it becomes more difficult to regenerate identifiable or diagnostically critical features from them. Thus, while some classification performance is sacrificed, dataset distillation provides a measurable gain in privacy preservation by impeding high-fidelity reconstruction. Importantly, our proposed model further strengthens this trade-off by maintaining stronger classification performance compared to traditional distillation approaches, such as distribution matching and gradient matching, while also achieving improved privacy preservation. Our proposed distillation scheme (DDM) outperforms distribution matching and gradient matching because it explicitly balances utility and privacy by enforcing stronger feature and structure-level dissimilarity while retaining task-relevant information. Unlike GM, which tends to preserve too much pixel and gradient-level similarity, leading to privacy leakage, while DM achieves stronger privacy but does not consistently minimize feature and structure-level similarity, our method integrates structural, perceptual, and distributional cues to suppress reconstructability without discarding discriminative features needed for classification. Moreover, the distilled datasets generated by our method exhibit stronger resilience against adversarial privacy attacks, such as membership inference and attribute inference, further demonstrating their robustness as a privacy-preserving alternative. This makes the distilled data generated by our approach a more efficient alternative, effectively balancing accuracy, privacy, attack resilience, and computational efficiency.

6. Conclusion

In this paper, we explored the dual benefits of dataset distillation in the context of medical imaging: computational efficiency and privacy preservation. By compressing large medical datasets into compact synthetic versions, we demonstrated that high classification performance can still be achieved with significantly fewer data samples, making distillation highly suitable for resource-constrained environments. To further advance this direction, we proposed a novel hybrid distillation framework that combines diffusion models with distribution matching (DDM), leveraging the generative strength of diffusion for producing diverse synthetic samples and the alignment capability of distribution matching for preserving discriminative features. Through experiments on Pneumonia, COVID-19, and Brain Tumor datasets, we demonstrated that while conventional distillation offers privacy advantages, our proposed approach achieves superior trade-offs between accuracy and privacy, as confirmed by both reconstruction-based metrics, such as SSIM, PSNR, and LPIPS, as well as inference attack evaluations, including MIA and AIA. This demonstrates that our approach not only strengthens privacy protection by reducing sensitive detail recoverability but also enhances classification performance compared to existing methods. In summary, dataset distillation, especially with our proposed hybrid approach, offers a practical pathway to balance performance, efficiency, and privacy in medical imaging. Future work can further refine this framework by extending it to multimodal medical data, integrating formal privacy guarantees such as differential privacy, and enabling real-time clinical deployment to maximize its impact in healthcare applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research is funded by Birla Institute of Technology and Science, Pilani. All authors approved the version of the manuscript to be published.

Data availability

No data was used for the research described in the article.

References

[1] Awrahman BJ, Aziz Fatah C, Hamaamin MY. A review of the role and challenges of big data in healthcare informatics and analytics. Comput Intell Neurosci 2022;2022(1):5317760.
[2] Scheibner J, Raisaro JL, Troncoso-Pastoriza JR, Ienca M, Fellay J, Vayena E, Hubaux J-P. Revolutionizing medical data sharing using advanced privacy-enhancing technologies: technical, legal, and ethical synthesis. J Med Internet Research 2021;23(2):e25120.
[3] Maekawa A, Kosugi S, Funakoshi K, Okumura M. DiLM: Distilling dataset into language model for text-level dataset distillation. 2024, p. 2–3, arXiv preprint arXiv:2404.00264.
[4] Wang T, Zhu J-Y, Torralba A, Efros AA. Dataset distillation. 2018, arXiv preprint arXiv:1811.10959.
[5] Guo H, Li C, Zhao W, Peng H, Qiao H, Chen X. Data distillation for sleep stage classification. IEEE Internet Things J 2025.
[6] Sachdeva N, McAuley J. Data distillation: A survey. 2023, arXiv preprint arXiv:2301.04272.
[7] Lei S, Tao D. A comprehensive survey of dataset distillation. IEEE Trans Pattern Anal Mach Intell 2023;46(1):17–32.
[8] Li M, Cui C, Liu Q, Deng R, Yao T, Lionts M, Huo Y. Dataset distillation in medical imaging: A feasibility study. 2024, arXiv preprint arXiv:2407.14429.
[9] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 2020;33:6840–51.
[10] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis. Adv Neural Inf Process Syst 2021;34:8780–94.
[11] Tezcan KC, Baumgartner CF, Luechinger R, Pruessmann KP, Konukoglu E. MR image reconstruction using deep density priors. IEEE Trans Med Imaging 2018;38(7):1633–42.
[12] Sucholutsky I, Schonlau M. Soft-label dataset distillation and text dataset distillation. In: 2021 international joint conference on neural networks. IJCNN, IEEE; 2021, p. 1–8.
[13] Nguyen T, Chen Z, Lee J. Dataset meta-learning from kernel ridge-regression. 2020, arXiv preprint arXiv:2011.00050.

[14] Li G, Togo R, Ogawa T, Haseyama M. Dataset distillation for medical dataset sharing. 2022, arXiv preprint arXiv:2209.14603.

[15] Xu Z, Chen Y, Pan M, Chen H, Das M, Yang H, Tong H. Kernel ridge regression-based graph dataset distillation. In: Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining. 2023, p. 2850–61.

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

[17] Ren J, Zhou Q, Yang S, Yang J. Multi-source feature map distillation for enhanced low-resolution object recognition. Comput Electr Eng 2025;128:110710.

[18] Chen Z, Cao Y, Gu Q, Zhang T. A generalized neural tangent kernel analysis for two-layer neural networks. Adv Neural Inf Process Syst 2020;33:13363–73.

[19] Zhao B, Bilen H. Dataset condensation with distribution matching. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2023, p. 6514–23.

[20] Wang T, Zhu J-Y, Torralba A, Efros AA. Dataset distillation. 2018, p. 1–23, arXiv preprint arXiv:1811.10959.

[21] Xu T-B, Liu C-L. Data-distortion guided self-distillation for deep neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, 2019, p. 5565–72.

[22] Zhao B, Bilen H. Dataset condensation with differentiable siamese augmentation. In: International conference on machine learning. PMLR; 2021, p. 12674–85.

[23] Zhou Y, Nezhadarya E, Ba J. Dataset distillation using neural feature regression. Adv Neural Inf Process Syst 2022;35:9813–27.

[24] Dong T, Zhao B, Lyu L. Privacy for free: How does dataset condensation help privacy? 2022, p. 5378–96.

[25] Cazenavette G, Wang T, Torralba A, Efros AA, Zhu J-Y. Dataset distillation by matching training trajectories. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 4750–9.

[26] Zheng J, Peng L. An autoencoder-based image reconstruction for electrical capacitance tomography. IEEE Sensors J 2018;18(13):5464–74.

[27] Li Y, Yu Z, Chen Y, He T, Zhang J, Zhao R, Xu K. Image reconstruction using pre-trained autoencoder on multimode fiber imaging system. IEEE Photonics Technol Lett 2020;32(13):779–82.

[28] Yang Y, Wu QJ, Wang Y. Autoencoder with invertible functions for dimension reduction and image reconstruction. IEEE Trans Syst Man Cybern: Syst 2016;48(7):1065–79.

[29] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13(4):600–12.

[30] Hore A, Ziou D. Image quality metrics: PSNR vs. SSIM. In: 2010 20th international conference on pattern recognition. IEEE; 2010, p. 2366–9.

[31] Ghazanfari S, Garg S, Krishnamurthy P, Khorrami F, Araujo A. R-LPIPS: An adversarially robust perceptual similarity metric. 2023, arXiv preprint arXiv:2307.15157.

[32] Zhang B, Yang H, Yin Z. A region-based normalized cross correlation algorithm for the vision-based positioning of elongated IC chips. IEEE Trans Semicond Manuf 2015;28(3):345–52.

[33] Zhang L, Zhang L, Mou X, Zhang D. FSIM: A feature similarity index for image quality assessment. IEEE Trans Image Process 2011;20(8):2378–86.

[34] Rubner Y, Tomasi C, Guibas LJ. The earth mover's distance as a metric for image retrieval. Int J Comput Vis 2000;40:99–121.

[35] Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy. SP, IEEE; 2017, p. 3–18.

[36] Hu H, Salcic Z, Sun L, Dobbie G, Yu PS, Zhang X. Membership inference attacks on machine learning: A survey. ACM Comput Surv 2022;54(11s):1–37.

[37] Zhao BZH, Agrawal A, Coburn C, Asghar HJ, Bhaskar R, Kaafar MA, Webb D, Dickinson P. On the (in) feasibility of attribute inference attacks on machine learning models. In: 2021 IEEE European symposium on security and privacy. EuroS&P, IEEE; 2021, p. 232–51.

[38] Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172(5):1122–31.

[39] Wang L, Lin ZQ, Wong A. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Sci Rep 2020;10(1):19549.

[40] Saroja DS, Joshi S. Augmented MR images of brain tumor. 2024, http://dx.doi.org/10.21227/9p7v-ed03.

[41] Mascarenhas S, Agarwal M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In: 2021 international conference on disruptive technologies for multi-disciplinary research and applications. CENTCON, vol. 1, IEEE; 2021, p. 96–9.