# L3DMC: Lifelong Learning Using Distillation via Mixed-Curvature Space

Kaushik Roy[1,2(✉)], Peyman Moghadam[2], and Mehrtash Harandi[1]

[1] Department of Electrical and Computer Systems Engineering, Faculty of Engineering, Monash University, Melbourne, Australia
{Kaushik.Roy,Mehrtash.Harandi}@monash.edu
[2] Data61, CSIRO, Brisbane, QLD, Australia
{Kaushik.Roy,Peyman.Moghadam}@csiro.au

**Abstract.** The performance of a lifelong learning (L3) model degrades when it is trained on a series of tasks, as the geometrical formation of the embedding space changes while learning novel concepts sequentially. The majority of existing L3 approaches operate on a fixed-curvature (*e.g.*, zero-curvature Euclidean) space that is not necessarily suitable for modeling the complex geometric structure of data. Furthermore, the distillation strategies apply constraints directly on low-dimensional embeddings, discouraging the L3 model from learning new concepts by making the model highly stable. To address the problem, we propose a distillation strategy named L3DMC that operates on mixed-curvature spaces to preserve the already-learned knowledge by modeling and maintaining complex geometrical structures. We propose to embed the projected low dimensional embedding of fixed-curvature spaces (Euclidean and hyperbolic) to higher-dimensional Reproducing Kernel Hilbert Space (RKHS) using a positive-definite kernel function to attain rich representation. Afterward, we optimize the L3 model by minimizing the discrepancies between the new sample representation and the subspace constructed using the old representation in RKHS. L3DMC is capable of adapting new knowledge better without forgetting old knowledge as it combines the representation power of multiple fixed-curvature spaces and is performed on higher-dimensional RKHS. Thorough experiments on three benchmarks demonstrate the effectiveness of our proposed distillation strategy for medical image classification in L3 settings. Our code implementation is publicly available at https://github.com/csiro-robotics/L3DMC.

**Keywords:** Lifelong Learning · Class-incremental Learning · Catastrophic Forgetting · Mixed-Curvature · Knowledge Distillation · Feature Distillation

## 1 Introduction

Lifelong learning [31,34] is the process of sequential learning from a series of non-stationary data distributions through acquiring novel concepts while preserving already-learned knowledge. However, D̲eep N̲eural N̲etworks (DNNs) exhibit a significant drop in performance on previously seen tasks when trained in continual learning settings.

This phenomenon is often called catastrophic forgetting [27,28,35]. Furthermore, the unavailability of sufficient training data in medical imaging poses an additional challenge in tackling catastrophic forgetting of a DNN model.

In a lifelong learning scenario, maintaining a robust embedding space and preserving geometrical structure is crucial to mitigate performance degradation and catastrophic forgetting of old tasks[23]. However, the absence of samples from prior tasks has been identified as one of the chief reasons for catastrophic forgetting. Therefore, to address the problem, a small memory buffer has been used in the literature to store a subset of samples from already seen tasks and replayed together with new samples [18,32]. Nevertheless, imbalances in data (*e.g.*, between current tasks and the classes stored in the memory) make the model biased towards the current task [18].

Knowledge distillation [4,9,16,31] has been widely used in the literature to preserve the previous knowledge while training on a novel data distribution. This approach applies constraints on updating the weight of the current model by mimicking the prediction of the old model. To maintain the old embedding structure
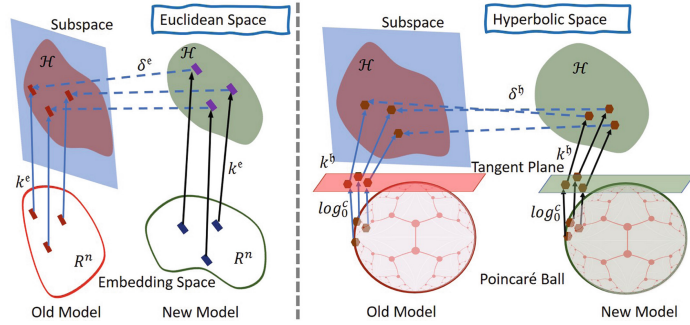


**Fig. 1.** Geometrical interpretation of the proposed distillation strategy, L3DMC via mixed-curvature space that optimizes the model by combining the distillation loss from Euclidean with zero-curvature (left) and hyperbolic with negative-curvature (right) space. L3DMC preserves the complex geometrical structure by minimizing the distance between new data representation and subspace induced by old representation in RKHS.

intact in the new model, feature distillation strategies have been introduced in the literature [9,18]. For instance, LUCIR [18] emphasizes on maximizing the similarity between the orientation of old and new embedding by minimizing the cosine distance between old and new embedding. While effective, feature distillation applies strong constraints directly on the lower-dimensional embedding extracted from the old and new models, reducing the plasticity of the model. This is not ideal for adopting novel concepts while preserving old knowledge. Lower-dimensional embedding spaces, typically used for distillation, may not preserve all the latent information in the input data [19]. As a result, they may not be ideal for distillation in lifelong learning scenarios. Furthermore, DNNs often operate on zero-curvature (*i.e.*, Euclidean) spaces, which may not be suitable for modeling and distilling complex geometrical structures in non-stationary biomedical image distributions with various modalities and discrepancies in imaging protocols and medical equipment. On the contrary, hyperbolic spaces have been successfully used to model hierarchical structure in input data for different vision tasks [13,21].

In this paper, we propose to perform distillation in a Reproducing Kernel Hilbert Space (RKHS), constructed from the embedding space of multiple fixed-curvature

spacess. This approach is inspired by the ability of kernel methods to yield rich representations in higher-dimensional RKHS [10, 19]. Specifically, we employ a Radial Basis Function (RBF) kernel on a mixed-curvature space that combines embeddings from hyperbolic (negative curvature), and Euclidean (zero curvature), using a decomposable Riemannian distance function as illustrated in Fig. 1. This mixed-curvature space is robust and can maintain a higher quality geometrical formation. This makes the space more suitable for knowledge distillation and tackling catastrophic forgetting in lifelong learning scenarios for medical image classification. Finally, to ensure a similar geometric structure between the old and new models in L3, we propose minimizing the distance between the new embedding and the subspace constructed using the old embedding in RKHS. Overall, our contributions in this paper are as follows:

- To the best of our knowledge, this is the first attempt to study mixed-curvature space for the continual medical image classification task.
- We propose a novel knowledge distillation strategy to maintain a similar geometric structure for continual learning by minimizing the distance between new embedding and subspace constructed using old embedding in RKHS.
- Quantitative analysis shows that our proposed distillation strategy is capable of preserving complex geometrical structure in embedding space resulting in significantly less degradation of the performance of continual learning and superior performance compared to state-of-the-art baseline methods on BloodMNIST, PathMNIST, and OrganaMNIST datasets.

## 2 Preliminaries

**Lifelong Learning (L3).** L3 consists of a series of $T$ tasks $\mathcal{T}_t \in \{\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_T\}$, where each task $\mathcal{T}_t$ has it's own dataset $\mathcal{D}_t = \{\mathcal{X}_t, \mathcal{Y}_t\}$. In our experiments, $\mathbf{X}_i \in \mathcal{X}_t \subset \mathcal{X}$ denotes a medical image of size $W \times H$ and $y_i \in \mathcal{Y}_t \subset \mathcal{Y}$ is its associated disease category at task $t$. In class-incremental L3, label space of two tasks is disjoint, hence $\mathcal{Y}_t \cap \mathcal{Y}_{t'} = \emptyset; t \neq t'$. The aim of L3 is to train a model $f : \mathcal{X} \to \mathcal{Y}$ incrementally for each task $t$ to map the input space $\mathcal{X}_t$ to the corresponding target space $\mathcal{Y}_t$ without forgetting all previously learned tasks (*i.e.*, $1, 2, \cdots, t-1$). We assume that a fixed-size memory $\mathcal{M}$ is available to store a subset of previously seen samples to mitigate catastrophic forgetting in L3.

**Mixed-Curvature Space.** Mixed-Curvature space is formulated as the Cartesian product of fixed-curvature spaces and represented as $M = \times_{i=1}^{C} M_i^{d_i}$. Here, $M_i$ can be a Euclidean (zero curvature), hyperbolic (constant negative curvature), or spherical (constant positive curvature) space. Furthermore, $\times$ denotes the Cartesian product, and $d_i$ is the dimensionality of fixed-curvature space $M_i$ with curvature $c_i$. The distance in the mixed-curvature space can be decomposed as $d_M(\mathbf{x}, \mathbf{y}) := \sum_{i=1}^{C} d_{M_i}(\mathbf{x}^i, \mathbf{y}^i)$.

**Hyperbolic Poincaré Ball.** Hyperbolic space is a Riemannian manifold with negative curvature. The Poincare ball with curvature $-c$, $c > 0$, $\mathbb{D}_c^n = \{\mathbf{x} \in \mathbb{R}^n : c\|\mathbf{x}\| < 1\}$ is a model of $n$-dimensional hyperbolic geometry. To perform vector operations on $\mathbb{H}^n$, Möbius Gyrovector space is widely used. Möbius addition between $x \in \mathbb{D}_c^n$ and $y \in \mathbb{D}_c^n$ is defined as follows

$$\mathbf{x} \oplus_c \mathbf{y} = \frac{(1 + 2c\langle\mathbf{x},\mathbf{y}\rangle + c\|\mathbf{y}\|_2^2)\mathbf{x} + (1 - c\|\mathbf{x}\|_2^2)\mathbf{y}}{1 + 2c\langle\mathbf{x},\mathbf{y}\rangle + c^2\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2} \tag{1}$$

Using Möbius addition, geodesic distance between two input data points, $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{D}_c^n$ is computed using the following formula.

$$d_c(\mathbf{x},\mathbf{y}) = \frac{2}{\sqrt{c}}\tanh^{-1}(\sqrt{c}\|(-\mathbf{x}) \oplus_c \mathbf{y}\|_2) \tag{2}$$

Tangent space of data point $\mathbf{x} \in \mathbb{D}_c^n$ is the inner product space and is defined as $T_x\mathbb{D}_c^n$ which comprises the tangent vector of all directions at $\mathbf{x}$. Mapping hyperbolic embedding to Euclidean space and vice-versa is crucial for performing operations on $\mathbb{D}^n$. Consequently, a vector $\mathbf{x} \in T_x\mathbb{D}_c^n$ is embedded onto the Poincaré ball $\mathbb{D}_c^n$ with anchor $\mathbf{x}$ using the exponential mapping function and the inverse process is done using the logarithmic mapping function $\log_{\mathbf{v}}^c$ that maps $\mathbf{x} \in \mathbb{D}_c^n$ to the tangent space of $\mathbf{v}$ as follows

$$\log_{\mathbf{v}}^c(\mathbf{x}) = \frac{2}{\sqrt{c}\lambda_{\mathbf{v}}^c}\tanh^{-1}(\sqrt{c}\| - \mathbf{v} \oplus_c \mathbf{x}\|_2)\frac{-\mathbf{v} \oplus_c \mathbf{x}}{\| - \mathbf{v} \oplus_c \mathbf{x}\|_2} \tag{3}$$

where $\lambda_{\mathbf{v}}^c$ is conformal factor that is defined as $\lambda_{\mathbf{v}}^c = \frac{2}{1-c\|v\|^2}$. In practice, anchor $\mathbf{v}$ is set to the origin. Therefore, the exponential mapping is expressed as $\exp_0^c(\mathbf{x}) = \tanh(\sqrt{c}\|\mathbf{x}\|)\frac{\mathbf{x}}{\sqrt{c}\|\mathbf{x}\|}$.

## 3   Proposed Method

In our approach, we emphasize on modeling complex latent structure of medical data by combining embedding representation of zero-curvature Euclidean and negative-curvature hyperbolic space. To attain richer representational power of RKHS [17], we embed the low-dimensional fixed-curvature embedding onto higher-dimensional RKHS using the kernel method.

**Definition 1.** (*Positive Definite Kernel*) A function $f : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is positive definite (pd) if and only if **1.** $k(\mathbf{x}_i,\mathbf{x}_j) = k(\mathbf{x}_j,\mathbf{x}_i)$ for any $\mathbf{x}_i,\mathbf{x}_j \in \mathcal{X}$, and **2.** for any given $n \in \mathbb{N}$, we have $\sum_i^n\sum_j^n c_ic_jk(\mathbf{x}_j,\mathbf{x}_i) \geq 0$ for any $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n \in \mathcal{X}$ and $c_1,c_2,\cdots,c_n \in \mathbb{R}$. Equivalently, the Gram matrix $K_{ij} = k(\mathbf{x}_i,\mathbf{x}_j) > 0$ for any set of $n$ samples $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n \in \mathcal{X}$ should be Symmetric and Positive Definite (SPD).

Popular kernel functions (*e.g.*, the Gaussian RBF) operate on flat-curvature Euclidean spaces. In $\mathbb{R}^n$, the Gaussian RBF kernel method is defined as

$$k^{\mathfrak{e}}(\mathbf{z}_i,\mathbf{z}_j) := \exp\big(-\lambda\|\mathbf{z}_i - \mathbf{z}_j\|^2\big); \lambda > 0. \tag{4}$$

However, using the geodesic distance in a hyperbolic space along with an RBF function similar to Eq. (4) (*i.e.*, replacing $\|\mathbf{z}_i - \mathbf{z}_j\|^2$ with the geodesic distance) does not lead to a valid positive definite kernel. Theoretically, a valid RBF kernel is impossible to obtain for hyperbolic space using geodesic distance [11,12]. Therefore, we use the tangent plane of hyperbolic space and employ $\log_0^c$

$$\log_0^c(\mathbf{z}) = \frac{\mathbf{z}}{\sqrt{c}\|\mathbf{z}\|}\tanh^{-1}\big(\sqrt{c}\|\mathbf{z}\|\big) , \tag{5}$$

to embed hyperbolic data to RKHS via the following valid pd kernel (see [10] for the proof of positive definiteness):

$$k^{\natural}(\mathbf{z}_i, \mathbf{z}_j) = \exp\big(-\lambda\|\log_0^c(\mathbf{z}_i) - \log_0^c(\mathbf{z}_j)\|^2\big). \tag{6}$$

Now, in L3 setting, we have two models $h^t$ and $h^{t-1}$ at our hand at time $t$. We aim to improve $h^t$ while ensuring the past knowledge incorporated in $h^{t-1}$ is kept within $h^t$. Assume $\mathbf{Z}_t$ and $\mathbf{Z}_{t-1}$ are the extracted feature vectors for input $\mathbf{X}$ using current and old feature extractor, $h_{feat}^t$ and $h_{feat}^{t-1}$, respectively. Unlike other existing distillation methods, we employ an independent 2-layer MLP for each fixed-curvature space to project extracted features to a new lower-dimensional embedding space on which we perform further operations. This has two benefits, **(i)** it relaxes the strong constraint directly applied on $\mathbf{Z}_t$ and $\mathbf{Z}_{t-1}$ and **(ii)** reduce the computation cost of performing kernel method. Since we are interested in modeling embedding structure in zero-curvature Euclidean and negative-curvature hyperbolic spaces, we have two MLP as projection modules attached to feature extractors, namely $g_e$ and $g_h$.

***Our Idea.*** Our main idea is that, for a rich and overparameterized representation, the data manifold is low-dimensional. Our algorithm makes use of RKHS, which can be intuitively thought of as a neural network with infinite width. Hence, we assume that the data manifold for the model at time $t-1$ is well-approximated by a low-dimensional hyperplane (our data manifold assumption). Let $\mathbf{Z}_{t-1}^{\mathfrak{e}} = \{\mathbf{z}_{t-1,1}^{\mathfrak{e}}, \mathbf{z}_{t-1,2}^{\mathfrak{e}}, \cdots, \mathbf{z}_{t-1,m}^{\mathfrak{e}}\}$ be the output of the Euclidean projection module for $m$ samples at time $t$ (*i.e.*, current model). Consider $\mathbf{z}_t^{\mathfrak{e}}$, a sample at time $t$ from the Euclidean projection head. We propose to minimize the following distance

$$\delta^{\mathfrak{e}}(\mathbf{z}_t^{\mathfrak{e}}, \mathbf{Z}_{t-1}^{\mathfrak{e}}) := \left\|\phi(\mathbf{z}_t^{\mathfrak{e}}) - \text{span}\{\phi(\mathbf{z}_{t-1,i}^{\mathfrak{e}})\}_{i=1}^m\right\|^2 \tag{7}$$
$$= \min_{\alpha \in \mathbb{R}^m}\left\|\phi(\mathbf{z}_t^{\mathfrak{e}}) - \sum_{i=1}^m \alpha_i\phi(\mathbf{z}_{t-1,i}^{\mathfrak{e}})\right\|^2 .$$

In Eq. (7), $\phi$ is the implicit mapping to the RKHS defined by the Gaussian RBF kernel, *i.e.* $k^{\mathfrak{e}}$. The benefit of formulation Eq. (7) is that it has a closed-form solution as

$$\delta^{\mathfrak{e}}(\mathbf{z}_t^{\mathfrak{e}}, \mathbf{Z}_{t-1}^{\mathfrak{e}}) = k(\mathbf{z}_t^{\mathfrak{e}}, \mathbf{z}_t^{\mathfrak{e}}) - k_{\mathbf{z}\mathbf{Z}}^{\top}K_{\mathbf{Z}\mathbf{Z}}^{-1}k_{\mathbf{z}\mathbf{Z}} . \tag{8}$$

In Eq. (8), $K_{\mathbf{Z}\mathbf{Z}} \in \mathbb{R}^{m \times m}$ is the Gram matrix of $\mathbf{Z}_{t-1}^{\mathfrak{e}}$, and $k_{\mathbf{z}\mathbf{Z}}$ is an m-dimensional vector storing the kernel values between $\mathbf{z}_t^{\mathfrak{e}}$ and elements of $\mathbf{z}_{t-1}^{\mathfrak{e}}$. We provide the proof of equivalency between Eq. (7) and Eq. (8) in the supplementary material due to the lack of space. Note that we could use the same form for the hyperbolic projection module $g_h$ to distill between the model at time $t$ and $t-1$, albeit this time, we employ the hyperbolic kernel $k^{\natural}$. Putting everything together,

$$\ell_{\text{KD}}(\mathbf{Z}_t) := \mathbb{E}_{\mathbf{z}_t}\delta^{\mathfrak{e}}(\mathbf{z}_t^{\mathfrak{e}}, \mathbf{Z}_{t-1}^{\mathfrak{e}}) + \beta\mathbb{E}_{\mathbf{z}_t}\delta^{\natural}(\mathbf{z}_t^{\natural}, \mathbf{Z}_{t-1}^{\natural}) . \tag{9}$$

Here, $\beta$ is a hyper-parameter that controls the weight of distillation between the Euclidean and hyperbolic spaces. We can employ Eq. (9) at the batch level. Note that in our formulation, computing the inverse of an $m \times m$ matrix is required, which has a complexity of $\mathcal{O}(m^3)$. However, this needs to be done once per batch and manifold (*i.e.*, Euclidean plus hyperbolic). $\ell_{KD}$ is differentiable with respect to $\mathbf{Z}_t$, which enables us to update the model at time $t$. We train our lifelong learning model by combining distillation loss, $\ell_{KD}$, together with standard cross entropy loss. Please refer to the overall steps of training lifelong learning model using our proposed distillation strategy via mixed-curvature space in Algorithm 1.

### 3.1 Classifier and Exemplar Selection

We employ herding based exemplar selection method that selects examples that are closest to the class prototype, following iCARL [32]. At inference time, we use exemplars from memory to compute class template and the nearest template class computed using Euclidean distance is used as the prediction of our L3 model. Assume $\mu_c$ is the class template computed by averaging the extracted features from memory exemplars belonging to class $c$. Then, the prediction $\hat{y}$ for a given input sample $\mathbf{X}$ is determined as $\hat{y} = \underset{c=1,...,t}{\arg\min} \|h_{feat}^t(\mathbf{X}) - \mu_c\|_2$.

## 4   Related Work

In this section, we describe Mixed-curvature space and L3 methods to tackle catastrophic forgetting.

**Constant-Curvature and Mixed-Curvature Space.** Constant-curvature spaces have been successfully used in the literature to realize the intrinsic geometrical orientation of data for various downstream tasks in machine learning. Flat-curvature Euclidean space is suitable to model grid data [37] while positive and negative-curvature space is better suited for capturing cyclical [2] and hierarchical [25] structure respectively. Hyperbolic representation has been used across domains ranging from image classification [26] and natural language processing [29, 30] to graphs [6]. However, a constant-curvature space is limited in modeling the geometrical structure of data embedding as it is designed with a focus on particular structures [14].

---

**Algorithm 1** Lifelong Learning using Distillation via Mixed-Curvature Space

**Input:** Dataset $\mathcal{D}^0, \mathcal{D}^1, ..., \mathcal{D}^T$, and Memory $\mathcal{M}$
**Output:** The new model at time $t$ with parameters $\mathbf{\Theta}^t$

1: Randomly Initialize $\mathbf{\Theta}^0$; $h_{\mathbf{\Theta}}^0 = h_{feat}^0 \circ h_{cls}^0$
2: Train $\mathbf{\Theta}^0$ on $\mathcal{D}^0$ using $\ell_{CE}$
3: **for** $t$ in $\{1, 2, ..., T\}$ **do**
4:     Initialize $\mathbf{\Theta}^t$ with $\mathbf{\Theta}^{t-1}$
5:     **for** iteration 1 to max_iter **do**
6:         Sample a mini batch $(\mathcal{X}_B, \mathcal{Y}_B)$ from $(\mathcal{D}^t \cup \mathcal{M})$
7:         $\mathbf{Z}^t \leftarrow h_{feat}^t(\mathcal{X}_B)$
8:         $\mathbf{Z}^{t-1} \leftarrow h_{feat}^{t-1}(\mathcal{X}_B)$
9:         $\tilde{\mathcal{Y}}_B \leftarrow h_{\mathbf{\Theta}^t}(\mathbf{Z}^t)$
10:         Compute $\ell_{CE}$ between $\mathcal{Y}_B$ and $\tilde{\mathcal{Y}}_B$
11:         Compute $\ell_{KD}$ between $\mathbf{Z}^{t-1}$ and $\mathbf{Z}^t$
12:         Update $\mathbf{\Theta}^t$ by minimizing the combined loss of cross-entropy $\ell_{CE}$ and $\ell_{KD}$ as in Eq. (9)
13:     **end for**
14:     Evaluate Model on test dataset
15:     Update Memory $\mathcal{M}$ with exemplars from $\mathcal{D}^t$
16: **end for**

**Kernel Methods.** A Kernel is a function that measures the similarity between two input samples. The intuition behind the kernel method is to embed the low-dimensional input data into a higher, possibly infinite, dimensional RKHS space. Because of the ability to realize rich representation in RKHS, kernel methods have been studied extensively in machine learning [17].

**L3 Using Regularization with Distillation.** Regularization-based approaches impose constraints on updating weights of L3 model to maintain the performance on old tasks. LwF mimics the prediction of the old model into the current model but struggles to maintain consistent performance in the absence of a task identifier. Rebuff *et al*. in [32] store a subset of exemplars using a herding-based sampling strategy and apply knowledge distillation on output space like LwF [24]. Distillation strategy on feature spaces has also been studied in the literature of L3. Hou *et al*. in [18] proposes a less-forgetting constraint that controls the update of weight by minimizing the cosine angle between old and new embedding representation.

## 5   Experimental Details

**Datasets.** In our experiments, we use four datasets (*e.g*., BloodMNIST [1], PathMNIST [20], OrganaMNIST [3]) and TissueMNIST [3] from MedMNIST collection [38] for the multi-class disease classification. BloodMNIST, PathMNIST, OrganaMNIST and TissueMNIST have 8, 9, 11, and 8 distinct classes, respectively that are split into 4 tasks with non-overlapping classes between tasks following [8]. For cross-domain continual learning experiments, we present 4 datasets sequentially to the model.

**Implementation Details.** We employ ResNet18 [15] as the backbone for feature extraction and a set of task-specific fully connected layers as the classifier to train all the baseline methods across datasets. To ensure fairness in comparisons, we run each experiment with the same set of hyperparameters as used in [8] for five times with a fixed set of distinct seed values, $1, 2, 3, 4, 5$ and report the average value. Each model is optimized using Stochastic Gradient Decent (SGD) with a batch of 32 images for 200 epochs, having early stopping options in case of overfitting. Furthermore, we use gradient clipping by enforcing the maximum gradient value to 10 to tackle the gradient exploding problem.

**Evaluation Metrics.** We rely on average accuracy and average forgetting to quantitatively examine the performances of lifelong learning methods as used in previous approaches [7,32]. Average accuracy is computed by averaging the accuracy of all the previously observed and current tasks after learning a current task $t$ and defined as: $\text{Acc}_t = \frac{1}{t} \sum_{i=1}^{t} \text{Acc}_{t,i}$, where $\text{Acc}_{t,i}$ is the accuracy of task $i$ after learning task $t$. We measure the forgetting of the previous task at the end of learning the current task $t$ using: $F_t = \frac{1}{t-1} \sum_{i=1}^{t-1} \max_{j \in \{1...t-1\}} \text{Acc}_{j,i} - \text{Acc}_{t,i}$, where at task $t$, forgetting on task $i$ is defined as the maximum difference value previously achieved accuracy and current accuracy on task $i$.

## 6    Results and Discussion

In our comparison, we consider two regularization-based methods (*i.e.*, EWC [22], and LwF [24]) and 5 memory-based methods (*e.g.*, EEIL [5], ER [33], Bic [36], LUCIR [18] and iCARL [32]). We employ the publicly available code[1] of [8] in our experiments to produce results for all baseline methods on BloodMNIST, PathMNIST, and OrganaM-NIST datasets and report the quantitative results in Table 1. The results suggest that the performance of all methods improves with the increase in buffer size (*e.g.*, from 200 to 1000). We observe that our proposed distillation approach outperforms other baseline methods across the settings. The results suggest that the regularization-based methods, *e.g.*, EWC and LwF perform poorly in task-agnostic settings across the datasets as those methods are designed for task-aware class-incremental learning. Our proposed method outperforms experience replay, ER method by a significant margin in both evaluation metrics (*i.e.*, average accuracy and average forgetting) across datasets. For instance, our method shows around 30%, 30%, and 20% improvement in accuracy compared to ER while the second best method, iCARL, performs about 4%, 2%, and 8% worse than our method on BloodMNIST, PathMNIST, and OrganaMNIST respectively with 200

**Table 1.** Experimental results on BloodMNIST, PathMNIST, OrganaMNIST and TissueMNIST datasets for 4-tasks Class-Incremental setting with varying buffer size. Our proposed method outperforms other baseline methods across the settings.

| Method | BloodMNIST | | PathMNIST | | OrganaMNIST | |
|---|---|---|---|---|---|---|
| | Accuracy ↑ | Forgetting ↓ | Accuracy ↑ | Forgetting ↓ | Accuracy ↑ | Forgetting ↓ |
| Upper Bound | 97.98 | – | 93.52 | – | 95.22 | – |
| Lower Bound | 46.59 | 68.26 | 32.29 | 77.54 | 41.21 | 54.20 |
| EWC [22] | 47.60 | 66.22 | 33.34 | 76.39 | 37.88 | 67.62 |
| LwF [24] | 43.68 | 66.30 | 35.36 | 67.37 | 41.36 | 51.47 |
| Memory Size: 200 | | | | | | |
| EEIL [5] | 42.17 | 71.25 | 28.42 | 79.39 | 41.03 | 62.47 |
| ER [33] | 42.94 | 71.38 | 33.74 | 80.6 | 52.50 | 52.72 |
| LUCIR [18] | 20.76 | 53.80 | 40.00 | 54.72 | 41.70 | 33.06 |
| BiC [36] | 53.32 | 31.06 | 48.74 | 30.82 | 58.68 | 29.66 |
| iCARL [32] | 67.70 | **14.52** | 58.46 | **-0.70** | 63.02 | **7.75** |
| Ours | **71.98** | 14.62 | **60.60** | 21.18 | **71.01** | 13.88 |
| Memory Size: 1000 | | | | | | |
| EEIL [5] | 64.40 | 40.92 | 34.18 | 75.42 | 66.24 | 34.60 |
| ER [33] | 65.94 | 33.68 | 44.18 | 66.24 | 67.90 | 31.72 |
| LUCIR [18] | 20.92 | 28.42 | 53.84 | 30.92 | 54.22 | 23.64 |
| BiC [36] | 70.04 | 17.98 | – | – | 73.46 | 15.98 |
| iCARL [32] | 73.10 | 13.18 | 61.72 | 14.14 | 74.54 | 10.50 |
| Ours | **77.26** | **10.9** | **67.52** | **12.5** | **76.46** | **9.08** |

---

[1] https://github.com/mmderakhshani/LifeLonger.

**Table 2.** Average accuracy on the cross-domain incremental learning scenario [8] with 200 exemplars. CL3DMC outperforms baseline methods by a significant margin in both task-aware and task-agnostic settings. Best values are in bold.

| Scenario | LwF | EWC | ER | EEIL | BiC | iCARL | LUCIR | L3DMC (Ours) |
|---|---|---|---|---|---|---|---|---|
| Task-Agnostic (Accuracy ↑) | 29.45 | 18.44 | 34.54 | 34.54 | 26.79 | 48.87 | 19.05 | **52.19** |
| Task-Aware (Accuracy ↑) | 31.07 | 29.26 | 37.69 | 33.19 | 33.19 | 49.47 | 27.48 | **52.83** |

exemplars. Similarly, with 1000 exemplars, our proposed method shows consistent performances and outperforms iCARL by 4%, 6%, and 2% accordingly on BloodMNIST, PathMNIST, and OrganaMNIST datasets. We also observe that catastrophic forgetting decreases with the increase of exemplars. Our method shows about 2% less forgetting phenomenon across the datasets with 1000 exemplars compared to the second best method iCARL.

Table 2 presents the experimental results (*e.g.*, average accuracy) on relatively complex cross-domain incremental learning setting where datasets (BloodMNIST, PathMNIST, OrganaMNIST, and TissueMNIST) with varying modalities from different institutions are presented at each novel task. Results show an unmatched gap between regularization-based methods (*e.g.*, Lwf and EWC) and our proposed distillation method. CL3DMC outperforms ER method by around 16% on both task-aware and task-agnostic settings. Similarly, CL3DMC performs around 3% better than the second best method, iCARL.

## 7    Conclusion

In this paper, we propose a novel distillation strategy, L3DMC on mixed-curvature space to preserve the complex geometric structure of medical data while training a DNN model on a sequence of tasks. L3DMC aims to optimize the lifelong learning model by minimizing the distance between new embedding and old subspace generated using current and old models respectively on higher dimensional RKHS. Extensive experiments show that L3DMC outperforms state-of-the-art L3 methods on standard medical image datasets for disease classification. In future, we would like to explore the effectiveness of our proposed distillation strategy on long-task and memory-free L3 setting.

## References

1. Acevedo, A., Merino, A., Alférez, S., Molina, Á., Boldú, L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. In: Data in brief 30 (2020)
2. Bachmann, G., Bécigneul, G., Ganea, O.: Constant curvature graph convolutional networks. In: International Conference on Machine Learning, pp. 486–496. PMLR (2020)
3. Bilic, P., et al.: The liver tumor segmentation benchmark (lits). Med. Image Anal. **84**, 102680 (2023)
4. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. In: NeurIPS (2020)

5. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11216, pp. 241–257. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01258-8_15

6. Chami, I., Ying, Z., Ré, C., Leskovec, J.: Hyperbolic graph convolutional neural networks. In: Advances in Neural Information Processing Systems 32 (2019)

7. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.S.: Riemannian walk for incremental learning: understanding forgetting and intransigence. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 556–572. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_33

8. Derakhshani, M.M., et al.: Lifelonger: A benchmark for continual disease classification. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part II, pp. 314–324. Springer (2022). https://doi.org/10.1007/978-3-031-16434-7_31

9. Douillard, A., Cord, M., Ollion, C., Robert, T., Valle, E.: PODNet: pooled outputs distillation for small-tasks incremental learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12365, pp. 86–102. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58565-5_6

10. Fang, P., Harandi, M., Petersson, L.: Kernel methods in hyperbolic spaces. In: International Conference on Computer Vision, pp. 10665–10674 (2021)

11. Feragen, A., Hauberg, S.: Open problem: kernel methods on manifolds and metric spaces. what is the probability of a positive definite geodesic exponential kernel? In: Conference on Learning Theory, pp. 1647–1650. PMLR (2016)

12. Feragen, A., Lauze, F., Hauberg, S.: Geodesic exponential kernels: when curvature and linearity conflict. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3032–3042 (2015)

13. Ganea, O., Bécigneul, G., Hofmann, T.: Hyperbolic neural networks. In: Advances In Neural Information Processing Systems 31 (2018)

14. Gu, A., Sala, F., Gunel, B., Ré, C.: Learning mixed-curvature representations in product spaces. In: International Conference on Learning Representations (2019)

15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

16. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network, vol. 2(7). arXiv preprint arXiv:1503.02531 (2015)

17. Hofmann, T., Schölkopf, B., Smola, A.J.: Kernel methods in machine learning (2008)

18. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 831–839 (2019)

19. Jayasumana, S., Hartley, R., Salzmann, M., Li, H., Harandi, M.: Kernel methods on riemannian manifolds with gaussian rbf kernels. IEEE Trans. Pattern Anal. Mach. Intell. 37(12), 2464–2477 (2015)

20. Kather, J.N., et al.: Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med. 16(1), e1002730 (2019)

21. Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I., Lempitsky, V.: Hyperbolic image embeddings. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6418–6428 (2020)

22. Kirkpatrick, J., et al.: Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. 114(13), 3521–3526 (2017)

23. Knights, J., Moghadam, P., Ramezani, M., Sridharan, S., Fookes, C.: Incloud: incremental learning for point cloud place recognition. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 8559–8566. IEEE (2022)

24. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 2935–2947 (2017)

25. Liu, Q., Nickel, M., Kiela, D.: Hyperbolic graph neural networks. In: Advances in Neural Information Processing Systems 32 (2019)

26. Mathieu, E., Le Lan, C., Maddison, C.J., Tomioka, R., Teh, Y.W.: Continuous hierarchical representations with poincaré variational auto-encoders. In: Advances in Neural Information Processing Systems 32 (2019)

27. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: the sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989)

28. Nguyen, C.V., Achille, A., Lam, M., Hassner, T., Mahadevan, V., Soatto, S.: Toward understanding catastrophic forgetting in continual learning. arXiv preprint arXiv:1908.01091 (2019)

29. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. In: Advances in Neural Information Processing Systems 30 (2017)

30. Nickel, M., Kiela, D.: Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In: International Conference on Machine Learning, pp. 3779–3788. PMLR (2018)

31. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: a review. Neural Netw. **113**, 54–71 (2019)

32. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: incremental classifier and representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2001–2010 (2017)

33. Riemer, M., et al.: Learning to learn without forgetting by maximizing transfer and minimizing interference. arXiv preprint arXiv:1810.11910 (2018)

34. Ring, M.B.: Child: a first step towards continual learning. Mach. Learn. **28**(1), 77–104 (1997)

35. Robins, A.: Catastrophic forgetting, rehearsal and pseudorehearsal. Connect. Sci. **7**(2), 123–146 (1995)

36. Wu, Y., et al.: Large scale incremental learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 374–382 (2019)

37. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.: A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst. **32**(1), 4–24 (2020)

38. Yang, J., Shi, R., Ni, B.: Medmnist classification decathlon: a lightweight automl benchmark for medical image analysis. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 191–195. IEEE (2021)

# Machine Learning – Explainability, Bias, and Uncertainty I