Continual-GEN: Continual Group Ensembling for Domain-agnostic Skin Lesion Classification

Nourhan Bayasi[1[0000−0003−4653−6081]], Siyi Du[1[0000−0002−9961−4533]], Ghassan Hamarneh[2[0000−0001−5040−7448]], and Rafeef Garbi[1[0000−0001−6224−0876]]

[1] BiSICL, University of British Columbia, Vancouver, BC, Canada
{nourhanb,siyi,rafeef}@ece.ubc.ca
[2] Medical Image Analysis Lab, Simon Fraser University, Burnaby, BC, Canada
hamarneh@sfu.ca

Abstract. Designing deep learning (DL) models that adapt to new data without forgetting previously acquired knowledge is important in the medical field where data is generated daily, posing a challenge for model adaptation and knowledge retention. Continual learning (CL) enables models to learn continuously without forgetting, typically on a sequence of domains with known domain identities (e.g. source of data). In this work, we address a more challenging and practical CL scenario where information about the domain identity during training and inference is unavailable. We propose Continual-GEN, a novel forget-free, replay-free, and domain-agnostic subnetwork-based CL method for medical imaging with a focus on skin lesion classification. Continual-GEN proposes an ensemble of groups approach that decomposes the training data for each domain into groups of semantically similar clusters. Given two domains, Continual-GEN assesses the similarity between them based on the distance between their ensembles and assigns a separate subnetwork if the similarity score is low, otherwise updating the same subnetwork to learn both domains. At inference, Continual-GEN selects the best subnetwork using a distance-based metric for each test data, which is directly used for classification. Our quantitative experiments on four skin lesion datasets demonstrate the superiority of Continual-GEN over state-of-the-art CL methods, highlighting its potential for practical applications in medical imaging. Our code: https://github.com/nourhanb/Continual-GEN.

Keywords: Continual Learning · Domain-agnostic · Out-of-Distribution Detection · Skin Lesion Classification · Dermatology.

1 Introduction

Deep learning (DL) models have emerged as powerful tools, surpassing human experts in certain cases, particularly in skin lesion classification [7]. However, the conventional clinical practice of training DL models only once falls short of addressing the steady stream of medical data, where data is generated daily and often exhibits a domain shift arising from various factors such as diverse clinical practices, variations in clinical devices or diagnostic workflows, or differences

in data populations [23,8]. Thus, there is a pressing need to design DL models that can effectively learn a stream of heterogeneous data and adeptly adapt to the substantial domain shift encountered across the different straining domains. The straightforward approach of fine-tuning DL models with either new lesions or heterogeneous data, without access to the initial training data, easily leads to overwriting of previously learned knowledge, resulting in *catastrophic forgetting*.

Continual learning (CL) [22] aims to enable DL models to adapt to changing environments and learn from new data while retaining previous knowledge. Replay-based methods [18,17] store a subset of data samples and replay them periodically to retain past domain information. However, these methods face challenges in medical domains due to data privacy policies that restrict unregulated data storage and transfer [24]. Regularization-based methods [16,2] impose restrictions on parameter updates to preserve prior knowledge while learning new domains. However, with the complex and heterogeneous medical data, the performance of these methods is significantly limited. Architecture-based methods [9] assign specialized architectural components for each domain, but encounter increased memory usage as new domains emerge. A promising recent approach has been developed that utilizes different subnetworks within a *fixed-size* dense network to learn the different domains [3,15,6]. Taking advantage of the over-parameterization of DL models, this subnetwork-based approach effectively addresses memory usage limitations in architecture-based methods by pruning unimportant weights, leading to optimized memory footprint and comparable or superior performance. However, existing CL methods face a crucial limitation in their practical deployment in dynamic real-world environments, particularly healthcare, due to the assumption of known data domain identities, such as the source of data or the specific device used for data generation. In practice, the anonymization process may erase domain identity information, making it infeasible to rely on such information during training or inference. As a consequence, current CL methods often underperform when evaluated in a domain-agnostic setup [19].

In this work, we introduce `Continual-GEN`, the first subnetwork-based CL approach for skin lesion classification that is not only forget-free and replay-free, but also domain-agnostic during training and inference. Specifically, we introduce a continual OOD detection method that is triggered when a domain shift occurs, allowing us to initialize a new subnetwork for learning the new domain during training. Our approach involves decomposing the semantic space for each training domain into distinct clusters with similar semantics, enabling the detection of new domains based on their distance to the clusters of previous domains. However, selecting an optimal number of clusters is challenging due to the complex heterogeneity of skin data. To this end, we introduce the novel ensemble of groups technique, which partitions the features into different groups, each with a different number of clusters. This approach enhances OOD detection reliability without the need for determining an optimal number of clusters. During inference, `Continual-GEN` utilizes a distance-based metric to select the most appropriate subnetwork for each test data, which is directly used for classification.

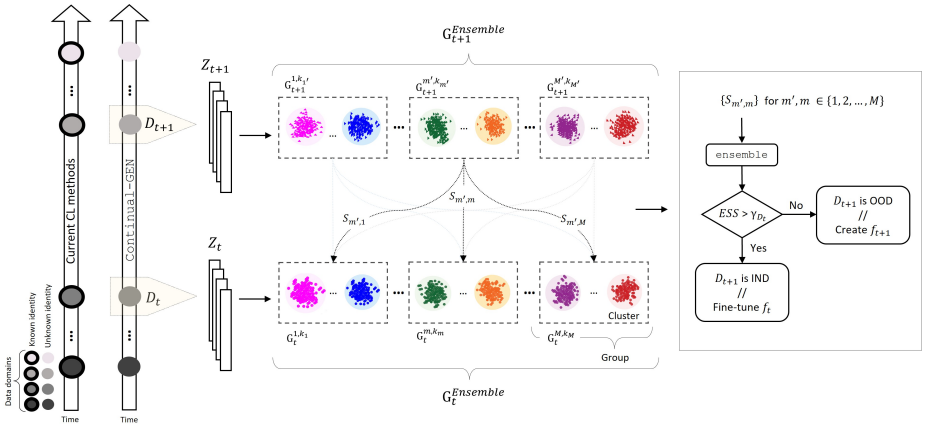

Fig. 1. `Continual-GEN` decomposes data into ensemble of M groups in the feature space, where the m-th group contains k_m clusters. To identify the similarity between two domains D_{t+1} and D_t, an ensemble similarity score ESS is calculated from the `ensemble` of all the minimum distances of each pair of groups, $\{(\mathbf{G}_{t+1}^{m',k_{m'}}, \mathbf{G}_t^{m,k_m})\}_{m',m=1}^{M}$. A large ESS score indicates higher similarity between the two domains, i.e., D_{t+1} is IND with respect to D_t and f_t is updated with D_{t+1}. Else, D_{t+1} is OOD and a new subnetwork f_{t+1} is initialized to learn D_{t+1}.

Experimental results on four diverse skin image datasets provide strong evidence supporting the superiority of our method compared to others.

2 `Continual-GEN`

Preliminaries. We propose a CL framework where a network f, of a fixed size, learns T domains $D = \{D_1, \ldots, D_t, \ldots, D_T\}$ sequentially over time while retaining previously acquired knowledge. The t-th domain $D_t = \left\{\left(x_t^i, y_t^i\right)\right\}_{i=1}^{N_t}$ contains N_t tuples of input samples $x_i^t \in \mathcal{X}$ and their corresponding labels $y_i^t \in \mathcal{C}$. When encountering the t-th domain with unknown identity, the data from previous domains $\{D_i\}_{i=1}^{t-1}$ is either unavailable or restricted. Our objective is to identify an optimal domain-specific subnetwork f_t for D_t, which is only updated when encountering a new, in-distribution (IND) domain. Else, f_t remains frozen and a new subnetwork is created to learn the OOD domain. The network f should be deployable at any time and capable of extracting predictions using the best subnetwork without knowledge of the test image's specific identity.

Domain-specific Subnetwork Formation. After training f on a specific domain D_t, we utilize a culpability-based pruning technique [3] to identify the optimal subnetwork f_t. This technique involves pruning units with high culpability scores, effectively removing them as they are considered unimportant. Through this process, we ensure that the subnetwork f_t maintains performance

comparable to the full network, while simultaneously creating room, i.e., preserving capacity, within the network f to effectively learn knowledge encountered in future domains. The pruning is performed based on a predefined pruning percentage p, which is set by the user.

Create a New or Update an Existing Subnetwork. When presented with a new batch of data, D_{t+1}, `Continual-GEN` assesses the similarity between D_{t+1} and each previously encountered domain $\{D_i\}_{i=1}^{t}$. This assessment, which includes three steps (I-III, below), determines whether the new data is IND with respect to any previous domain or OOD. In the case of IND, `Continual-GEN` reuses and updates the corresponding subnetwork, while for OOD, it creates a new subnetwork specifically for D_{t+1}. For notational simplicity, we illustrate the process using only the most recent domain, D_t.

I. Ensemble of Groups. Upon selecting the optimal subnetwork f_t for D_t, we extract the embedding features of D_t in the embedding space, denoted as $Z_t = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\}$. After that, as illustrated in Fig. 1, we partition Z_t into an ensemble of M groups, each with a different number of clusters, i.e.,

$$\mathbf{G}_t^{Ensemble} = [\mathbf{G}_t^{1,k_1}, \ldots, \mathbf{G}_t^{m,k_m}, \ldots, \mathbf{G}_t^{M,k_M}]$$

where $\mathbf{G}_t^{m,k_m}$ is the m-th group with k_m clusters. The mean and covariance of each cluster within each group in the ensemble are computed and stored, occupying only a few KBytes of memory.

II. ESS **Score.** To quantify the similarity between D_t and a new domain D_{t+1}, we form $\mathbf{G}_{t+1}^{Ensemble}$, which mirrors the configurations in $\mathbf{G}_t^{Ensemble}$, by performing a forward pass of D_{t+1} through the trained subnetwork f_t. Then, we measure the Mahalanobis distance between each cluster in a group in $\mathbf{G}_{t+1}^{Ensemble}$ to all the clusters in the mirroring group in $\mathbf{G}_t^{Ensemble}$. Then, for each pair of group configurations, e.g., ($\mathbf{G}_{t+1}^{m',k_{m'}}$, $\mathbf{G}_t^{m,k_m}$), we return the smallest Mahalanobis distance, $S_{m',m}$, representing the similarity score between the m'-th group in $\mathbf{G}_{t+1}^{Ensemble}$ and the m-th group in $\mathbf{G}_t^{Ensemble}$. As demonstrated in Fig. 1 (right), the total $2 \times M$ individual scores are then aggregated using an `ensemble` module, such as averaging in our implementation, yielding the final ensemble similarity score ESS as follows;

$$ESS = \texttt{ensemble}\{S_{m',m}\} \quad \text{for } m', m \in \{1, 2, ...M\}.$$

III. IND vs OOD Decision Making. If ESS exceeds a threshold value γ_{D_t}, indicating a higher degree of similarity between the two domains, f_t is updated with D_{t+1} and the mean and covariance values are recalculated and updated in memory. On the other hand, if ESS falls below γ_{D_t}, suggesting that D_{t+1} is OOD, a new subnetwork is initialized to learn D_{t+1} using the same culpability-based pruning technique. The mean and variance of all clusters and groups in $\mathbf{G}_{t+1}^{Ensemble}$ are calculated from the trained f_{t+1} and stored for future use. If ESS returns IND to multiple domains, we only fine-tune the subnetwork with the corresponding smallest ESS value. We refer the reader to Algorithm-1 in supplementary material for a summary of the training framework of `Continual-GEN`.

Domain-agnostic Inference. For a test image, a forward pass through all subnetworks is performed to calculate ESS with each domain. The subnetwork with the smallest score is selected and directly used to extract a prediction.

3 Experiments and Results

Datasets and Implementation Details. In our experimental setup, we consider a total of six sequentially presented domains that are constructed using four distinct skin lesion datasets: HAM10000 (HAM)[21] (partitioned into three domains as in [8]), Dermofit (DMF) [1], Derm7pt (D7P) [11], and MSK [10]. We use ResNet-152 as the backbone of network f. For each domain, we train it using the cross-entropy (CE) loss for 150 epochs with a constant learning rate of 1e-5 and a batch size of 16. We partition each domain into three sets: training (60%), validation (20%), and test (20%) sets. We balance all the training domains in PyTorch, and we resize the images to 224×224. To address domain order bias, we averaged the results across all 720 possible domain order combinations. We use p=80% pruning ratio when creating all the subnetworks. Each ensemble consists of M= 8 groups, including one group formed using the Ground Truth (GT) clustering, which cluster features based on the known class labels, i.e., $k = GT$, and seven additional groups created by the Gaussian mixture model (GMM) clustering method, which models the features as a mixture of k Gaussian distributions in the embedding space (k =1, 3, 5, 7, 10, 15, and 20). We use averaging for the `ensemble`, and set γ_{D_t} at twice the mean of all clusters in $\mathbf{G}_t^{Ensemble}$.
Metrics. We evaluate the performance of our `Continual-GEN` using two metrics: 1) the widely-used accuracy of each domain after training all the domains: $ACC = \frac{1}{T}\sum_{t=1}^{T} a_{T,t}$, where $a_{T,t}$ is the test balanced accuracy of t-th domain after a model has learned all the T domains, and 2) the average accuracy computed over all domains, $AVG = \frac{1}{T}\sum_{t=1}^{T} ACC(t)$.
Comparison Against SOTA CL Methods. We compare `Continual-GEN` against several CL methods, including three subnetwork-based methods: CPN [3], PackNet [15] and CP&S [6], and two regularization-based methods: EWC [13] and LwF [14]. All the competitors require the availability of domain identity information, as they were not specifically designed for domain-agnostic scenarios. Additionally, we provide an upper bound performance (JOINT), which is obtained by the usual supervised fine-tuning on the data of all tasks jointly (assuming all available at one time), and a lower bound performance (SeqT), which simply performs sequential training without any countermeasure to forgetting. Our comprehensive evaluation, as summarized in Table 1, demonstrates the performance of `Continual-GEN`, surpassing other CL approaches across all domains. This superiority can be attributed to two key factors. Firstly, we address the potential issue of negative knowledge interference by identifying one HAM domain as OOD and assigning a separate subnetwork for it (the total number of subnetworks in f is 5 in `Continual-GEN` as opposed to 4 in alternative methods). Secondly, we use a culpability-based pruning technique to retain only the most relevant units for each domain, resulting in improved classifica-

Table 1. Performance of `Continual-GEN` against baselines and SOTA CL methods on six skin lesion domains. '# of sub' indicates the total number of subnetworks in f.

Method	Test Sets Performance (ACC) %						Total AVG %	# of sub
	HAM-1	HAM-2	HAM-3	DMF	D7P	MSK		
Baselines								
JOINT	90.72$_{\pm 0.81}$	91.43$_{\pm 0.43}$	89.65$_{\pm 0.15}$	84.07$_{\pm 0.72}$	88.65$_{\pm 0.06}$	84.21$_{\pm 0.99}$	<u>88.12</u>$_{\pm 0.52}$	-
SeqT	40.38$_{\pm 0.28}$	42.06$_{\pm 0.27}$	41.84$_{\pm 0.51}$	44.97$_{\pm 0.08}$	44.52$_{\pm 0.65}$	40.78$_{\pm 0.94}$	42.43$_{\pm 0.45}$	-
Competing CL Methods								
CPN	84.36$_{\pm 0.50}$	83.37$_{\pm 0.12}$	82.63$_{\pm 0.78}$	76.54$_{\pm 0.40}$	80.46$_{\pm 0.57}$	70.11$_{\pm 0.63}$	79.58$_{\pm 0.50}$	4
PackNet	81.04$_{\pm 0.35}$	80.61$_{\pm 0.29}$	79.39$_{\pm 0.81}$	70.05$_{\pm 1.02}$	77.59$_{\pm 0.46}$	64.83$_{\pm 0.59}$	75.59$_{\pm 0.42}$	4
CP&S	80.47$_{\pm 0.68}$	79.51$_{\pm 0.53}$	78.84$_{\pm 0.16}$	71.18$_{\pm 0.31}$	78.55$_{\pm 0.42}$	69.91$_{\pm 0.70}$	76.41$_{\pm 0.47}$	4
EWC	44.15$_{\pm 0.91}$	44.98$_{\pm 0.50}$	43.25$_{\pm 0.83}$	56.34$_{\pm 0.65}$	46.08$_{\pm 0.13}$	43.12$_{\pm 1.12}$	46.32$_{\pm 0.69}$	-
LwF	53.28$_{\pm 0.84}$	54.22$_{\pm 0.30}$	53.01$_{\pm 0.90}$	59.62$_{\pm 0.33}$	47.50$_{\pm 0.46}$	45.14$_{\pm 1.19}$	52.13$_{\pm 0.67}$	-
Proposed Method								
Ours	**85.78**$_{\pm 0.20}$	**84.11**$_{\pm 0.84}$	**85.41**$_{\pm 0.71}$	**77.52**$_{\pm 0.98}$	**81.73**$_{\pm 0.30}$	**71.84**$_{\pm 0.13}$	**81.07**$_{\pm 0.50}$	5

Table 2. `Continual-GEN` average performance with different OOD detection methods. '# of sub' indicates the total number of subnetworks in f.

Method	`Continual-GEN`			
	Ours	Method$-\mathcal{A}$	Method$-\mathcal{B}$	Method$-\mathcal{C}$
Total AVG %	81.07$_{\pm 0.50}$	76.51$_{\pm 0.38}$	72.34$_{\pm 0.18}$	73.54$_{\pm 0.62}$
# of sub	5	3	2	6*

$\star$ indicates that the pruning ratio was increased to 85% to accommodate more subnetworks.

tion performance, even with the subnetworks in `Continual-GEN` having fewer parameters than those of other methods.

Comparison Against other Domain-agnostic Methods. To assess the effectiveness of the proposed OOD detection method in `Continual-GEN`, i.e., ensemble of groups, we compare it against alternative domain-agnostic learning techniques. In Method$-\mathcal{A}$, a new subnetwork is initialized when the accuracy on new domain drops below 10%. In Method$-\mathcal{B}$, the Gram distance [17] is used instead of the Mahalanobis distance for both training and inference. In Method$-\mathcal{C}$, domain shifts are detected by computing the Mahalanobis distance between features extracted after the first layer of Batch Normalization (BN) [9]. As demonstrated in Table 2, our proposed method outperforms the alternative approaches. The Gram distance (Method$-\mathcal{B}$) fails to accurately detect distribution shifts in skin datasets, and Methods$-\mathcal{A},\mathcal{C}$ are sensitive to hyperparameter choices, such as the 10% accuracy drop threshold in Method$-\mathcal{A}$ and the selection of the BN layer in Method$-\mathcal{C}$.

Unraveling Cluster Quality for Skin Datasets. The quality of clusters in the embedding space is a fundamental aspect to the success of our method. Therefore, we conduct an extensive analysis to compare the quality of clusters generated by different clustering techniques and training methods, as follows:

Clustering Techniques: In addition to the GT and GMM clustering methods, we explore the use of k-means, which partitions the features into k clusters based on their similarity measured using the Euclidean distance in the embedding space. Although we considered other clustering methods, such as agglomerative clustering and DBSCAN, we found them to be less compatible and requiring

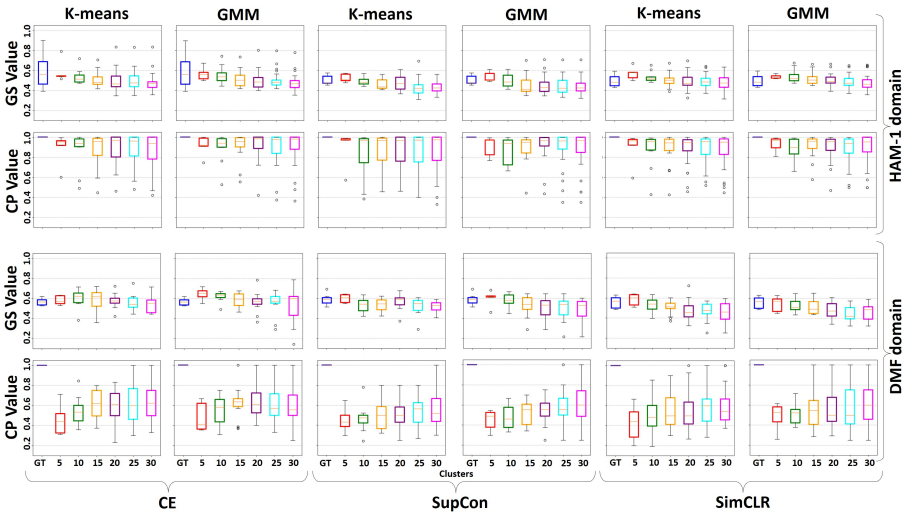

Fig. 2. Comparison of cluster quality for CE, SupCon, and SimCLR based on GS and CP on the HAM-1 and DMF domains. The evaluation process begins with the default GT clusters, followed by k-means or GMM with an increasing number of clusters.

careful hyperparameter tuning, such as selecting appropriate linking strategies for agglomerative clustering or determining the epsilon value for DBSCAN.

Training Methods: Besides the CE loss, we investigate the influence of contrastive learning approaches due to their demonstrated capability in OOD detection [20]. Specifically, we compare two approaches: supervised contrastive learning (SupCon) [12] and the unsupervised approach (SimCLR) [5].

Metrics for Cluster Quality: To evaluate the effectiveness of the different clustering and training approaches, we employ two metrics: Global separation (GS) and cluster purity (CP) [4]. GS quantifies the separability between clusters by evaluating the intra-cluster distances to the enter-cluster distances of the nearest neighboring cluster, whereas CP determines how many samples in a cluster belongs to the same class. Higher values of both metrics indicate higher quality of clusters. We refer the reader to [4] for equations of GS and CP.

Discussion of Results: By analyzing the results of applying the different clustering and training methods on the HAM-1 and DMF domains, as illustrated in Fig. 2, we can derive important observations about the quality of the generated clusters. The following key findings emerge from this analysis: 1) The three learning methods (CE, SupCon, SimCLR) exhibit comparable performance, with CE and SupCon showcasing slightly better results due to their supervised learning nature. 2) The quality of clusters generated by GMM outperforms k-means, particularly in terms of CP values. The higher purity values achieved by GMM reflect its capability to generate more internally homogeneous clusters, predom-

Table 3. `Continual-GEN` average performance with different `ensemble` strategies. '# of sub' indicates the total number of subnetworks in f.

Strategy	Average	Top		Bottom		Trimmed Average	
		q=20	q=40	q=20	q=40	q=20	q=40
Total AVG %	$81.07_{\pm 0.50}$	$77.43_{\pm 0.62}$	$79.3_{\pm 0.48}$	$74.94_{\pm 0.66}$	$81.07_{\pm 0.50}$	$81.07_{\pm 0.50}$	$81.07_{\pm 0.50}$
# of sub	5	3	4	$6^{\star}$	5	5	5

$\star$ indicates that the pruning ratio was increased to 85% to accommodate more subnetworks.

inantly containing samples from the same class, suggesting its ability to capture the underlying data distribution of the skin more effectively. 3) The optimal number of clusters k cannot be easily determined, as the choice of it may not straightforwardly correspond to higher purity and separation. For instance, CE with k =5 of GMM on the DMF dataset exhibits lower purity compared to that of k =10, despite higher values of GS. These results demonstrate the challenge in selecting the ideal clustering technique and k value for skin-related analysis, further emphasizing the unique and effective nature of the proposed ensemble of groups method.

Ablation Study on the Impact of the Ensemble Size. We investigate the impact of the enemsble size (M) on the performance of `Continual-GEN`. Our findings demonstrate that utilizing a substantial number of diverse groups leads to improved average performance. Specifically, `Continual-GEN` achieves a performance of 75.34% and 79.34% for $M \in \{1, 2\}$ and $M \in \{3, 4, 5, 6\}$, respectively. With $M \in \{8, 9, 10, 11\}$, the performance further increases to 81.07%.

Ablation Study on the Impact of the `ensemble` Strategy. We investigate different ensembling strategies to compute the final ESS score: 1) Average (default) averages all distance scores, 2) Top averages the top q sorted scores, 3) Bottom averages the bottom q sorted scores, and 4) Trimmed Average averages remaining scores after removing top and bottom q sorted scores. Notably, the Top method identifies more domains as IND, which potentially led to decreased performance due to negative knowledge interference between domains, resulting in a reduction of 3.64% and 1.77% in performance with q= 20 and 40, respectively, compared to the default method (Average). On the other hand, the Trimmed Average method performs similarly to the default method, indicating that it detects the same IND and OOD domains.

4 Conclusion

We introduced `Continual-GEN`, a subnetwork-based CL approach for skin lesion classification. Our method supports sequential learning without forgetting and does not require domain identity information during training and inference. `Continual-GEN` decomposes the semantic space into groups, detecting domain shifts and assigning domain-specific subnetworks accordingly. Extensive experiments on diverse skin lesion datasets demonstrate its superior performance over SOTA CL methods and domain-agnostic learning techniques. Additionally, `Continual-GEN` ensures memory efficiency by avoiding network expansion and individual sample storage, crucial for maintaining patient privacy.

References

1. Ballerini, L., Fisher, R.B., Aldridge, B., Rees, J.: A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In: Color Medical Image Analysis, pp. 63–86 (2013)
2. Baweja, C., Glocker, B., Kamnitsas, K.: Towards continual learning in medical imaging. arXiv preprint arXiv:1811.02496 (2018)
3. Bayasi, N., Hamarneh, G., Garbi, R.: Culprit-prune-net: Efficient continual sequential multi-domain learning with application to skin lesion classification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24. pp. 165–175 (2021)
4. Bojchevski, A., Matkovic, Y., Günnemann, S.: Robust spectral clustering for noisy data: Modeling sparse corruptions improves latent embeddings. In: Proceedings of International Conference on Knowledge Discovery and Data Mining. pp. 737–746 (2017)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607 (2020)
6. Dekhovich, A., Tax, D.M., Sluiter, M.H., Bessa, M.A.: Continual prune-and-select: class-incremental learning with specialized subnetworks. Applied Intelligence pp. 1–16 (2023)
7. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
8. Fogelberg, K., Chamarthi, S., Maron, R.C., Niebling, J., Brinker, T.J.: Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation. New Biotechnology (2023)
9. González, C., Ranem, A., Othman, A., Mukhopadhyay, A.: Task-agnostic continual hippocampus segmentation for smooth population shifts. In: Domain Adaptation and Representation Transfer MICCAI Workshop. pp. 108–118 (2022)
10. Gutman, D., Codella, N.C.F., Celebi, M.E., Helba, B., Marchetti, M.A., Mishra, N.K., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv **abs/1605.01397** (2016)
11. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. The IEEE journal of biomedical and health informatics **23**(2), 538–546 (2018)
12. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
13. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
14. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence TPAMI **40**(12), 2935–2947 (2017)
15. Mallya, A., Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative pruning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 7765–7773 (2018)

16. Özgün, S., Rickmann, A.M., Roy, A.G., Wachinger, C.: Importance driven continual learning for segmentation across domains. In: Machine Learning in Medical Imaging. pp. 423–433 (2020)
17. Perkonigg, M., Hofmanninger, J., Herold, C.J., Brink, J.A., Pianykh, O., Prosch, H., Langs, G.: Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nature communications **12**(1), 5678 (2021)
18. Perkonigg, M., Hofmanninger, J., Langs, G.: Continual active learning for efficient adaptation of machine learning models to changing image acquisition. In: Information Processing in Medical Imaging. pp. 649–660 (2021)
19. Prabhu, A., Torr, P., Dokania, P.: Gdumb: A simple approach that questions our progress in continual learning. In: The European Conference on Computer Vision (ECCV) (August 2020)
20. Sehwag, V., Chiang, M., Mittal, P.: SSD: A unified framework for self-supervised outlier detection. arXiv preprint arXiv:2103.12051 (2021)
21. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
22. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: Theory, method and application. arXiv preprint arXiv:2302.00487 (2023)
23. Wen, D., Khan, S.M., Xu, A.J., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A.K., Liu, X., et al.: Characteristics of publicly available skin cancer image datasets: a systematic review. The Lancet Digital Health **4**(1), e64–e74 (2022)
24. Yoon, J., Jeong, W., Lee, G., Yang, E., Hwang, S.J.: Federated continual learning with weighted inter-client transfer. In: International Conference on Machine Learning. pp. 12073–12086 (2021)