Medical Image Analysis (2025)

Contents lists available at ScienceDirect

Medical Image Analysis

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

BiasPruner: Mitigating Bias Transfer in Continual Learning for Fair Medical Image Analysis

Nourhan Bayasi[a,*], Jamil Fayyad[b], Alceu Bissoto[c], Ghassan Hamarneh[d], Rafeef Garbi[a]

[a] University of British Columbia, Vancouver, BC, Canada
[b] University of Victoria, Victoria, BC, Canada
[c] University of Bern, Bern, Switzerland
[d] Simon Fraser University, Burnaby, BC, Canada

ARTICLE INFO

Keywords:
Bias transfer
Catastrophic forgetting
Continual learning
Debiased representations
Debiased subnetwork
Dynamic subnetwork pruning
Spurious correlations

ABSTRACT

Continual Learning (CL) enables neural networks to learn new tasks while retaining previous knowledge. However, most CL methods fail to address bias transfer, where spurious correlations propagate to future tasks or influence past knowledge. This bidirectional bias transfer negatively impacts model performance and fairness, especially in medical imaging, where it can lead to misdiagnoses and unequal treatment. In this work, we show that conventional CL methods amplify these biases, posing risks for diverse patient cohorts. To address this, we propose BiasPruner, a framework that mitigates bias propagation through debiased subnetworks, while preserving sequential learning and avoiding catastrophic forgetting. BiasPruner computes a bias attribution score to identify and prune network units responsible for spurious correlations, creating task-specific subnetworks that learn unbiased representations. As new tasks are learned, the framework integrates non-biased units from previous subnetworks to preserve transferable knowledge and prevent bias transfer. During inference, a task-agnostic gating mechanism selects the optimal subnetwork for robust predictions. We evaluate BiasPruner on medical imaging benchmarks, including skin lesion and chest X-ray classification tasks, where biased data (e.g., spurious skin tone correlations) can exacerbate disparities. Our experiments show that BiasPruner outperforms state-of-the-art CL methods in both accuracy and fairness. Code is available at: BiasPruner.

1. Introduction

In domains such as medical imaging, data evolves continuously due to the emergence of new disease classes, shifting imaging protocols, and population variability, leading to distributional shifts that challenge model generalization (Fayyad et al., 2024b; Fayyad, 2023; Bayasi et al., 2022). Recent efforts have explored data synthesis and model adaptation to mitigate these effects (Fayyad et al., 2025; Du et al., 2023). Still, deep neural networks (DNNs) face the critical challenge of catastrophic forgetting (Lewandowsky and Li, 1995), where new learning can overwrite prior knowledge. Continual Learning (CL) addresses this by enabling models to balance adaptation (plasticity) and retention (stability) (Wang et al., 2023a; Bayasi, 2025). Towards this, a variety of CL techniques have been proposed in the literature, including replay-based techniques, which retain and replay subsets of past data to reinforce prior knowledge (Perkonigg et al., 2021a; Kiyasseh et al., 2021); regularization-based approaches, which constrain parameter updates to preserve previously learned representations (Lenga et al., 2020; Bayasi et al., 2024b); and architecture-based methods, which dynamically modify network structures by expanding or specializing components to accommodate new

*Corresponding author: nourhanbayasi92@gmail.com

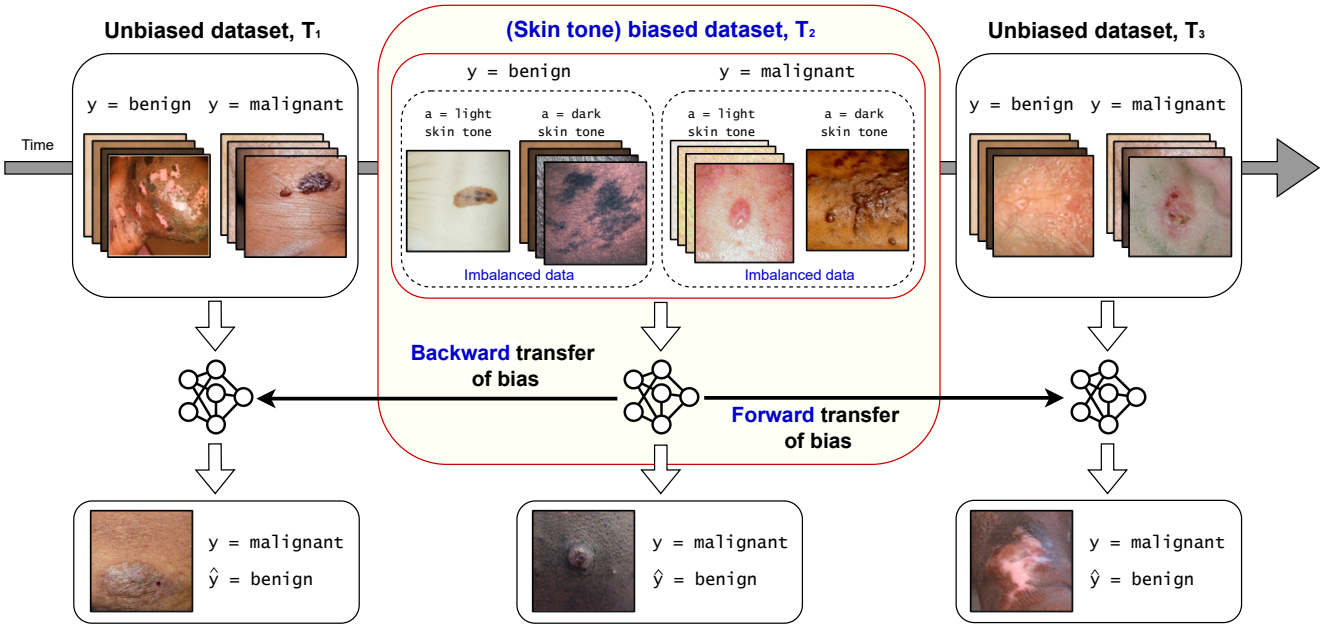

Fig. 1. Illustration of forward and backward bias transfer in CL across three tasks (T=3). The central panel depicts a biased dataset where skin tone serves as a spurious attribute a (i.e., lighter skin tones are correlated with malignant conditions, while darker skin tones are associated with benign lesions). In this case, the true label is represented as y, and the predicted label is represented as $\hat{y}$. The bias propagates forward to subsequent unbiased datasets (forward bias transfer), affecting predictions ($\hat{y}$) even in the absence of explicit bias in the data. Furthermore, bias can retroactively influence predictions in previously learned unbiased tasks (backward bias transfer), altering the model's output. These dual effects of bias transfer highlight how spurious correlations undermine fairness and reliability in continual learning models.

tasks (Elkhayat et al., 2025; González et al., 2022; Bayasi et al., 2023, 2021).

Although previous CL methods have shown success, they often fail to address a more realistic and challenging scenario: *the presence of dataset bias*. This bias arises from the imbalanced distributions of sensitive attributes, such as ethnicity, age or gender that are inadvertently correlated with disease classes in the training data. In such cases, models may unintentionally rely on spurious correlations tied to these attributes (Larrazabal et al., 2020; Luo et al., 2022), performing well only when these biases coincidentally match true disease prevalence in the target population (Larrazabal et al., 2020; Luo et al., 2022; Brown et al., 2023). However, relying on such biased information can have detrimental consequences, leading to unfair and biased decisions. For example, models trained on predominantly lighter skin tones may struggle to detect melanoma in individuals with darker skin tones, significantly compromising both performance and generalization on test data that lack these correlations (Brown et al., 2023).

In CL, the challenge of spurious correlation learning is intensified by bias transfer, where biases learned in one task (a specific subset of data with its own objectives and characteristics) persist and affect subsequent tasks, even when later tasks are based on unbiased data (Salman et al., 2022). Due to the sequential nature of CL, this effect accumulates over time, as recent mathematical analyses highlight (Busch et al., 2024). Standard CL methods exacerbate this issue by retaining unintended biases (e.g., skin tone bias), leading to forward bias transfer (where early task biases influence future tasks) and backward bias transfer (where new task training amplifies prior biases). As illustrated in Fig. 1, this bidirectional bias transfer reinforces spurious correlations, compromising both fairness and reliability. Moreover, naive debiasing strategies applied to the current task can inadvertently erase essential knowledge from previous tasks, causing catastrophic forgetting and further complicating bias mitigation in CL.

In our recent work (Bayasi et al., 2024a), we introduced `BiasPruner`, the first CL framework that mitigates both catastrophic forgetting and bias propagation by identifying and pruning biased network units. By constructing task-specific, debiased subnetworks, `BiasPruner` prevents spurious correlations from carrying over to future tasks while preserving essential knowledge for continual learning. In this extended journal version, we take a necessary step back to establish, for the first time, direct empirical evidence of bias transfer in CL for medical imaging classification tasks. While prior work has focused predominantly on catastrophic forgetting and overall knowledge retention, the specific challenge of bias accumulation across sequential tasks has been largely overlooked. Through carefully controlled experiments, we reveal how spurious correlations not only persist from earlier tasks but can intensify in subsequent training, even when later tasks themselves are unbiased. This analysis provides clear scientific motivation for debiasing continual learners and underscores why methods like `BiasPruner` are critical for achieving both fairness and reliability in real-world medical applications.

We evaluate `BiasPruner` on three medical imaging classification benchmarks, including datasets for skin lesion and chest

X-ray classification, each with a unique bias attribute. The results demonstrate that `BiasPruner` consistently outperforms state-of-the-art CL methods in both classification performance and fairness. Notably, `BiasPruner` achieves these results without relying on explicit bias annotations, addressing a critical challenge in medical contexts where identifying dataset biases is both costly and constrained by privacy considerations (Luo et al., 2022). To the best of our knowledge, this is the first work in the medical domain to systematically explore bias transfer in CL and propose a practical CL framework that effectively balances performance and fairness across sequential tasks.

2. Related Work

2.1. Continual Learning in the Medical Domain

The primary goal of CL is to develop methods that allow a network to learn new tasks while retaining knowledge of previous tasks. In other words, it aims to enable the network to continually adapt to new information without completely erasing or degrading its performance on earlier tasks; i.e., without catastrophic forgetting (Wu et al., 2024a; Kumari et al., 2023). Generally speaking, CL methods in the medical domain can be grouped into three categories: replay-based, regularization-based, and architecture-based methods.

Replay-based CL methods utilize a memory buffer to store samples from old tasks, which are replayed during learning. For instance, iCaRL (Rebuffi et al., 2017) selects representative samples based on their proximity to class prototypes, with successful applications in histopathology tumor classification (Kaustaban et al., 2022) and disease classification (Derakhshani et al., 2022). Advanced sample selection methods include using binary segmentation samples based on gradient contributions (Bera et al., 2023), storing unique samples (Hofmanninger et al., 2020), and leveraging active learning to prioritize informative samples during domain shifts (Perkonigg et al., 2021b). In contrast, generative replay methods address privacy concerns in medical applications by using generative models to synthesize pseudo-representations of past tasks. Examples include CCSI, which employs class-specific synthesis to generate images (Ayromloua et al., 2024); a style-oriented replay module that captures domain-specific adjustments (Li et al., 2022); and GarDA, which integrates appearance transfer and stochastic generators (Chen et al., 2023). Recent advances in diffusion models, such as text-to-image synthesis (Byun et al.), demonstrate remarkable potential for maintaining past performance while adapting to new tasks in medical applications.

Regularization-based CL methods avoid the need for stored examples by introducing regularization terms in the loss function or adjusting learning rates to penalize large parameter updates, thereby mitigating catastrophic forgetting. These methods often utilize a frozen copy of the old model to monitor parameter changes when adapting to new tasks. For instance, Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) uses the Fisher information matrix to assess parameter importance, proving effective in medical imaging (Baweja et al., 2018). In brain MRI segmentation, parameter regularization has facilitated transfer learning from high-quality to lower-quality datasets (van Garderen et al., 2019). Zhang et al. (Zhang et al., 2023a) proposed a joint importance matrix for continual segmentation, selectively regularizing parameters critical to shape and semantic consistency. Liu et al. (Liu et al., 2023) introduced a divergence-aware dual-flow module for incremental learning, incorporating pseudo-label training and self-entropy regularization to balance task rigidity and plasticity. Wu et al. (Wu et al., 2024b) developed MAda, a multi-modal adaptive algorithm for medical image super-resolution, aligning weight spaces to optimize task consistency. Finally, Roy et al. (Roy et al., 2023) proposed L3DMC, a distillation method in mixed-curvature spaces, embedding low-dimensional representations into a higher-dimensional RKHS to preserve prior knowledge during new task learning.

Architecture-based CL methods isolate task-specific knowledge by assigning distinct sets of parameters to each task, preventing knowledge interference and forgetting (Wortsman et al., 2020; Kang et al., 2022; Yan et al., 2021; Abati et al., 2020; Wang et al., 2023b). These methods either maintain a fixed-size architecture (Bayasi et al., 2024c) or expand dynamically to accommodate new tasks, which is more common in the medical literature. For instance, Zhang et al. (Zhang et al., 2023b) proposed adding lightweight, task-specific heads for new organs and tumors, incorporating CLIP embeddings to retain semantic information. Ji et al. (Ji et al., 2023) introduced a unified segmentation model for 143 whole-body organs, employing a fixed encoder and incrementally adding decoders for new tasks. Mousser et al. (Mousser et al., 2022) presented the Incremental Deep Tree framework, which grows hierarchically like a tree by adding branches for new tasks, along with a replay buffer to address catastrophic forgetting. Xie et al. (Xie et al., 2023) proposed a minimalist approach by freezing most of the network and training only batch normalization layers, enabling task adaptation without major architectural changes. While these architecture-based methods primarily aim to mitigate forgetting, they may inadvertently amplify or transfer dataset biases, particularly when subnetworks are constructed through pruning or expansion without considering the underlying data distributions. `BiasPruner` fills this gap by introducing bias-aware pruning, explicitly designed to reduce spurious correlations and prevent bias transfer during continual learning.

2.2. Debiased Representation Learning in Continual Learning

Dataset bias is a long-standing and active area of research in machine learning, with numerous methods proposed for mitigating spurious correlations and improving fairness in static settings (Du et al., 2022; Nam et al., 2020; Bayasi et al., 2025). However, these approaches typically assume access to the entire dataset upfront, rely on explicit bias annotations, or require global rebalancing strategies, all are assumptions that do not hold in CL, where data arrives sequentially, prior data may become inaccessible, and privacy constraints often prohibit replay or annotation-based methods.

In the context of CL, bias and fairness remain largely underexplored. Biased representations and unfair predictions can propagate through tasks, a phenomenon we refer to as bias transfer. Attempts to address this challenge in computer vision are emerging but limited. For example, He et al. (He, 2024;

Bayasi et al., 2013) addressed inter- and intra-task imbalances using gradient reweighting but overlooked biases from inherent or acquired data characteristics (Angwin et al., 2022). Xu et al. (Xu et al., 2024) proposed CLAD, which tackles imbalanced forgetting with class-aware disentanglement to enhance accuracy. FairCL (Truong et al., 2023) focused on class imbalance in semantic segmentation, while FSW (Park et al., 2024) introduced fairness-aware sample weighting to mitigate unfair catastrophic forgetting. However, these approaches are unsuitable for medical applications due to their reliance on replay buffers, which conflict with privacy regulations, or their need for bias annotations, which are costly and often unavailable.

`BiasPruner` takes a fundamentally different approach by directly addressing bias transfer in CL through structural debiasing of the model itself. While LfF (Nam et al., 2020) leverages sample difficulty signals (estimated using the generalized cross-entropy (Zhang and Sabuncu, 2018) loss) to mitigate bias in static learning by distinguishing between easy (potentially biased) and hard (potentially unbiased) examples, `BiasPruner` applies this concept in a novel way: using these difficulty signals to guide the structural pruning of biased network units. This enables the construction of task-specific subnetworks that actively suppress spurious correlations from propagating across tasks. To the best of our knowledge, `BiasPruner` is the first to adapt this idea for bias mitigation in a continual learning setup.

3. Medical Benchmark Datasets

To investigate the effects of bias in CL, we curated three publicly available medical imaging datasets. These datasets were selected based on the following criteria: (1) the presence of potential spurious correlations between bias attributes and disease labels, which enables a controlled study of bias transfer; (2) sufficient class diversity to simulate realistic incremental learning scenarios; and (3) public availability of images and annotations to ensure reproducibility. The selected datasets are:

- **Fitzpatrick17K (FITZ)** (Groh et al., 2021), a dermatology dataset consisting of 16,012 clinical images across 114 class labels. Each image is annotated with a Fitzpatrick skin tone score (I–VI), where I represents the lightest tone and VI the darkest. Previous studies have reported spurious correlations between skin tone and diagnosis, such as the underdiagnosis of melanoma in individuals with darker skin tones (Barros et al., 2023; Du et al., 2022; Bevan and Atapour-Abarghouei, 2022).

- **HAM10000 (HAM)** (Tschandl et al., 2018), another dermatology dataset consisting of 8,678 images across 7 class labels, with accompanying patient metadata. Previous studies have identified age-related biases in skin lesion classification, where older patients are more likely to receive certain diagnoses, leading to disparities in model performance (Khan et al., 2025; Li et al., 2021).

- **NIH ChestX-Ray14 (NIH)** (Wang et al., 2017), a chest X-ray dataset consisting of 19,993 images across 14 class labels, with patient gender annotations (male or female). Prior work has demonstrated gender disparities in chest X-ray interpretation, with models frequently performing worse on female patients due to data imbalances and systemic biases in medical imaging (Glocker et al., 2023; Larrazabal et al., 2020).

In this paper, we utilize the three datasets (FITZ, HAM, and NIH) in two different ways. First, in Section 4, which focuses on the empirical study, we use a controlled subset of each dataset, selecting four classes per dataset that exhibit the strongest bias correlations. This controlled setup allows for a more focused analysis of bias transfer in continual learning. More details on this selection process are provided in Section 4.1.2. In contrast, Section 6 presents the main experiments and results, where we use the full datasets—including all classes and bias attributes—without restrictions. This setup reflects a more realistic continual learning scenario and ensures comprehensive evaluation.

4. Empirical Analysis of Bias Transfer in Continual Learning for Medical Imaging

In this section, we present a comprehensive empirical analysis of bias transfer in CL for medical image classification. Specifically, we examine the forward bias transfer, where biases from earlier tasks propagate to subsequent ones, and the backward bias transfer, where biases from later tasks adversely affect performance on prior tasks. Using the three datasets described in Section 3, our experiments reveal that spurious correlations in the training datasets significantly impair CL performance. Moreover, our analysis establishes a clear relationship between bias transfer and the stability-plasticity trade-off inherent to CL methods: algorithms that prioritize stability tend to exacerbate forward bias transfer, whereas those enhancing plasticity are more prone to backward bias transfer.

4.1. Preliminaries and Experimental Design

4.1.1. Problem Formulation

We consider a realistic CL scenario in which dataset biases are present. In our formulation, a model f is trained sequentially over T tasks (i.e., $t = 1, 2, \ldots$). For each task T_t, the training data consists of samples denoted by the tuple (x_i, y_i, a_i), where x_i is the input (e.g., a medical image), $y_i \in Y_t$ is the associated class label, and $a_i \in A$ represents a bias attribute that may spuriously correlate with the label y_i. Our study is conducted within the class-incremental learning (CIL) paradigm, where class labels are mutually exclusive across tasks (i.e., $Y_i \cap Y_j = \emptyset$ for $i \neq j$). For analytical clarity, we assume a single bias attribute is present across all tasks unless noted otherwise.

4.1.2. Controlled Experimental Setup

We design a controlled experimental framework using the FITZ, HAM, and NIH datasets. First, we binarize the bias attributes in FITZ and HAM. In FITZ, we group skin tones I, II, and III as light skin tones ($a = 1$) and the remaining tones as dark ($a = 0$), similar to (Pundhir et al., 2024). In HAM, patients aged ≥ 60 are assigned $a = 1$, while those aged < 60 are

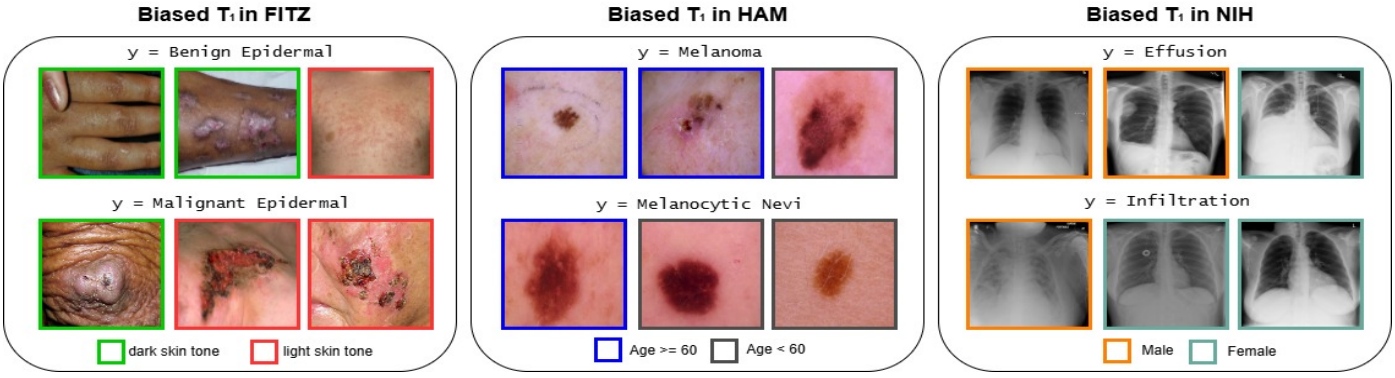

Fig. 2. Samples from Task 1 with a biased setup in FITZ, HAM, and NIH. Each row corresponds to a specific class within the task. Colored boxes highlight images that share the same bias attribute (e.g., red boxes in FITZ indicate light skin tones). In an unbiased setup, each row would display an equal number of images with and without colored boxes, representing a balanced distribution across bias groups.

Table 1. Sample distribution across bias groups ($a = 0, a = 1$) for biased and unbiased tasks. Each cell represents the number of samples per group, where y is the class label and a is the bias attribute. In biased tasks, the imbalance is controlled by β, reducing the sample count in one group while maintaining 200 samples in the other. Unbiased tasks correspond to $\beta = 1.0$, ensuring balanced distributions.

	$y = 0$	$y = 1$
$a = 0$	200	$\beta \times 200$
$a = 1$	$\beta \times 200$	200

assigned $a = 0$, similar to (Fan et al., 2021). In NIH, the gender attribute is already binary, with $a = 0$ for male and $a = 1$ for female. Next, we select a subset of four classes per dataset using Cramér's correlation, which quantifies the association between the bias attribute and class labels. We specifically choose the four classes with the highest Cramér's correlation, as they are more likely to exhibit bias transfer effects. These four classes are then divided across two tasks ($T = 2$) to simulate a two-task CL setup. While this controlled setup simplifies the problem (e.g., binarized bias attributes, two tasks, binary classification per task), Section 6 extends our experiments to more complex settings, validating our findings under realistic continual learning scenarios.

Following (Brown et al., 2023; Banerjee et al., 2023), we create biased and unbiased versions of each task by adjusting the sample distribution across bias groups. In the biased condition, the correlation between the bias attribute and the class label is amplified by removing a fraction β (with $0.1 \leq \beta \leq 0.5$) of samples from the opposing group, while unbiased tasks maintain balanced distributions (i.e., $\beta = 1.0$). Table 1 details the sample distributions, and Fig. 2 shows representative examples from the biased sets. For evaluation, all test sets are curated to be unbiased.

4.1.3. Baselines and Continual Learning Methods

Baselines. We investigate bias transfer by analyzing two common baseline approaches: Joint, where a single model is assumed to have access to all tasks' data simultaneously, serving as an upper bound by eliminating forgetting; and SeqFT (Sequential Fine-Tuning), where a single model is fine-tuned on new tasks sequentially without mechanisms to retain prior knowledge, often leading to catastrophic forgetting and potential bias shifts.

CL Methods. We evaluate four widely recognized CL methods, each representing a different approach to mitigating forgetting. LwF (Learning without Forgetting) (Li and Hoiem, 2017) is a regularization-based method that retains knowledge from previous tasks by distilling predictions from an older model into the new one. EWC (Elastic Weight Consolidation) (Kirkpatrick et al., 2017) is another regularization technique that prevents significant changes to important weights by penalizing updates based on their importance to prior tasks. ER (Experience Replay) (Chaudhry et al., 2019) is a replay-based method that stores a subset of past data and replays it during training to maintain knowledge across tasks. In ER, we use Reservoir Sampling (Vitter, 1985) to efficiently manage the buffer as new tasks arrive. Finally, PackNet (Mallya and Lazebnik, 2018) follows a fixed-size subnetwork-based architecture approach, where a portion of the network is pruned after each task based on weight magnitude, allowing the remaining parameters to be allocated to new tasks while preserving previous knowledge.

Each of these CL methods has its own hyperparameters, such as regularization strength, exemplar memory size, or pruning ratio, which in turn control the stability-plasticity trade-off. This trade-off determines which task's bias is propagated throughout learning. Specifically, when a CL method prioritizes stability (e.g., EWC with high regularization), it retains past knowledge but also transfers bias from earlier tasks to future ones. In contrast, when a method emphasizes plasticity (e.g., ER with larger buffers), it adapts more effectively to new tasks but allows bias from the current task to retroactively influence past tasks. To explicitly indicate the stability-plasticity trade-off of each method in our results, we append (s) for stability-focused variants and (p) for plasticity-focused variants. For example, EWC (s) refers to a version of EWC with a high regularization strength, leading to greater stability, whereas EWC (p) represents a lower regularization strength, allowing for greater plasticity.

4.1.4. Evaluation Metrics

Besides classification accuracy (`ACC`), we utilize two other metrics, discussed below, to evaluate bias transfer in CL. Each

Fig. 3. Analysis of bias transfer in two-task CL on FITZ (top), HAM (middle) and NIH (bottom). We evaluate performance using ACC (higher is better), DCA (lower is better), and DPR (higher is better). In each benchmark, the top row examines forward bias transfer on the unbiased Task 2 (T_2) when Task 1 (T_1) is either unbiased (blue) or biased (red), whereas the bottom row analyzes backward bias transfer on the unbiased T_1 when T_2 is either unbiased (blue) or biased (red). Stability-focused (s) and plasticity-focused (p) variants of CL methods highlight the role of the stability-plasticity trade-off in bias transfer.

is computed per class, and the final value is obtained by averaging across all classes within a task.

Difference in Classification Accuracy Across Bias Attributes (DCA) reflects the degree of bias of a model; i.e., it quantifies the disparity in classification accuracy between the two groups defined by a. For a given class y, it is calculated as:

$$\texttt{DCA} = |Accuracy_{a=0,y} - Accuracy_{a=1,y}|\,, \tag{1}$$

where $Accuracy_{a=i,y}$ represents the accuracy for group $a = i$ for a specific class y. In Eq. 1, DCA closer to 0 indicates a fairer model.

Demographic Parity Ratio (DPR) evaluates whether the likelihood of a positive prediction is independent of the sensitive attribute a. For each class y, DPR is calculated as:

$$\texttt{DPR} = \frac{P(\hat{y} = 1|\, a = 0, y)}{P(\hat{y} = 1|\, a = 1, y)}\,, \tag{2}$$

where $P(\hat{y} = 1|\, a = i, y)$ is the probability of predicting the positive class for group $a = i$ within class y. In Eq. 2, DPR closer to 1 indicates a fairer model.

4.2. Case 1: Forward Transfer of Bias

To assess how bias from T_1 influences T_2 in a CL scenario, we evaluate both baseline and CL methods by systematically varying the bias level in T_1 while maintaining T_2 as unbiased. The results on FITZ in the top row of Fig. 3 demonstrate the significant impact of early-stage bias in T_1 on T_2.

Our findings reveal two key insights. First, bias definitively transfers from T_1 to T_2, evidenced by consistent performance degradation across all metrics. With SeqFT, when bias in T_1 increased from 0% (unbiased) to 50% (biased), T_2's ACC decreased by 15.5 percentage points (72.1% → 56.6%), DPR dropped by 0.29 (0.98 → 0.69), and DCA increased by 0.12 (0.13 → 0.25%). This confirms that biases from earlier tasks systematically propagate to subsequent learning objectives. Second, CL methods prioritizing stability facilitate greater bias transfer than their plasticity-focused variants. EWC (s) with stronger regularization demonstrated a 0.11 higher DCA on T_2 compared to its plasticity-oriented counterpart (EWC (p)) when T_1 contained bias. Similarly, ER (s) with larger memory buffers

showed a DPR value lower than the small-buffer variant (ER (p)) under identical conditions. These findings align with CL theory: stability-focused approaches effectively preserve knowledge from previous tasks to mitigate catastrophic forgetting, but this preservation mechanism operates indiscriminately, retaining both task-relevant knowledge and harmful biases embedded within the learned representations.

4.3. Case 2: Backward Transfer of Bias

Now, we turn our attention to the backward transfer of bias, where we examine how bias from T_2 affects the previously learned T_1. Specifically, we evaluate both baseline and CL methods by varying the bias level in T_2 while maintaining T_1 as unbiased. The results on FITZ (bottom row of the top section in Fig. 3) highlight how the introduction of bias in later tasks (i.e., T_2) impacts earlier performance (i.e., T_1). We omit the results for PackNet, as it freezes the parameters updated in previous tasks, thereby preserving the original predictions for T_1 and preventing any backward transfer effects.

From the results, we notice that the bias in T_2 has a negative influence on T_1. Furthermore, we observe a trend opposite to that of forward bias transfer. For example, plasticity-focused methods facilitate greater backward transfer of bias compared to stability-oriented approaches. When T_2 bias increased from 0% to 50%, T_1's ACC decreased by 20.6 percentage points (from 67.3% to 46.7%) for EWC (p), while EWC (s) showed a 12.7 percentage point reduction. Therefore, CL methods prioritizing plasticity allow greater modification of parameters important for previous tasks, thereby enabling bias from new tasks to retroactively influence earlier knowledge. These findings demonstrate that bias mitigation strategies must consider not only forward but also backward transfer effects. We observe a similar trend on the HAM and NIH benchmarks, shown in the middle and bottom of Fig. 3, respectively.

5. Methodology

Our findings show that conventional CL methods inadvertently transfer biases across tasks, highlighting the need for a debiasing-aware approach. To address this, we introduce BiasPruner[1], a novel CL method that enables sequential learning without forgetting previously acquired knowledge while actively debiasing each task to minimize bias transfer. An overview of the proposed method is presented in Fig. 4.

5.1. Intuition and Overview

Existing CL methods primarily focus on minimizing or eliminating forgetting to retain knowledge across tasks. In contrast, we propose a fundamentally different perspective: rather than viewing forgetting as a limitation, we strategically leverage it to actively 'forget' biases during the learning process. Our proposed method, BiasPruner, achieves this by selectively pruning network units that contribute to biased feature learning.

[1] A preliminary version of this work was presented at MICCAI 2024 (Bayasi et al., 2024a)

This approach offers three key advantages. First, by pruning biased units, BiasPruner preserves valuable network capacity, providing room for future tasks without the risk of catastrophic forgetting. Second, the pruning process naturally results in a subnetwork with reduced bias, enhancing performance not only on the current task but also in subsequent tasks, thanks to the forward transfer of debiased knowledge across subnetworks. Third, backward bias transfer is entirely eliminated, as each task-specific subnetwork is frozen once created, preventing interference with future tasks.

The proposed BiasPruner employs a fixed-size network architecture f, which learns a debiased subnetwork for each task. Notably, BiasPruner does **not** require any prior knowledge of dataset bias during training, i.e., the model processes each task's training data D_t, consisting of pairs (x_t, y_t) with spurious correlations, without explicit bias labels. Furthermore, BiasPruner assumes no access to previous task data, meaning the network is only provided with the training data relevant to the current task, and cannot revisit or rely on data from earlier tasks. To construct a debiased subnetwork, BiasPruner uses a bias score to assess the contribution of each network unit to the learning of biased features. Units with high bias scores are pruned to form a task-specific debiased subnetwork, while the remaining pruned units are made available for learning new tasks. At inference, BiasPruner identifies the optimal subnetwork for predictions on a given data in a task-agnostic setup; i.e., information about the task origin of a test image is unknown or unavailable.

5.2. Forming a Debiased Subnetwork

BiasPruner creates a task-specific, debiased subnetwork, f_t, through a three-step process involving bias-aware pruning, bias scoring, and fine-tuning to selectively remove biased units and improve model fairness.

5.2.1. Bias-Aware Pruning

First, we selectively prune the network by removing the units most responsible for learning the bias encoded in D_t. Specifically, we identify and eliminate the top γ% of units based on their bias scores (discussed next), which include feature maps and their corresponding filters. This leaves $(1-\gamma)$% of the units to form the subnetwork f_t. The result is a pruned subnetwork that eliminates units heavily influenced by spurious features and biases in D_t.

5.2.2. Bias Scoring

The bias score $\mathcal{S}^t_{c,n}$ for each unit n in the biased network is computed based on its contribution to learning biased features for a given class c. To achieve this, we start by intentionally biasing the network using the generalized cross-entropy (GCE) loss (Zhang and Sabuncu, 2018), formulated as:

$$\mathcal{L}_{\text{GCE}}(p(x;\theta), y) = \frac{1 - p_y(x;\theta)^q}{q}\,, \tag{3}$$

where $q \in (0, 1]$ controls the degree of bias amplification. This loss function encourages the network to prioritize easier samples in training dataset D_t by up-weighting the probability of

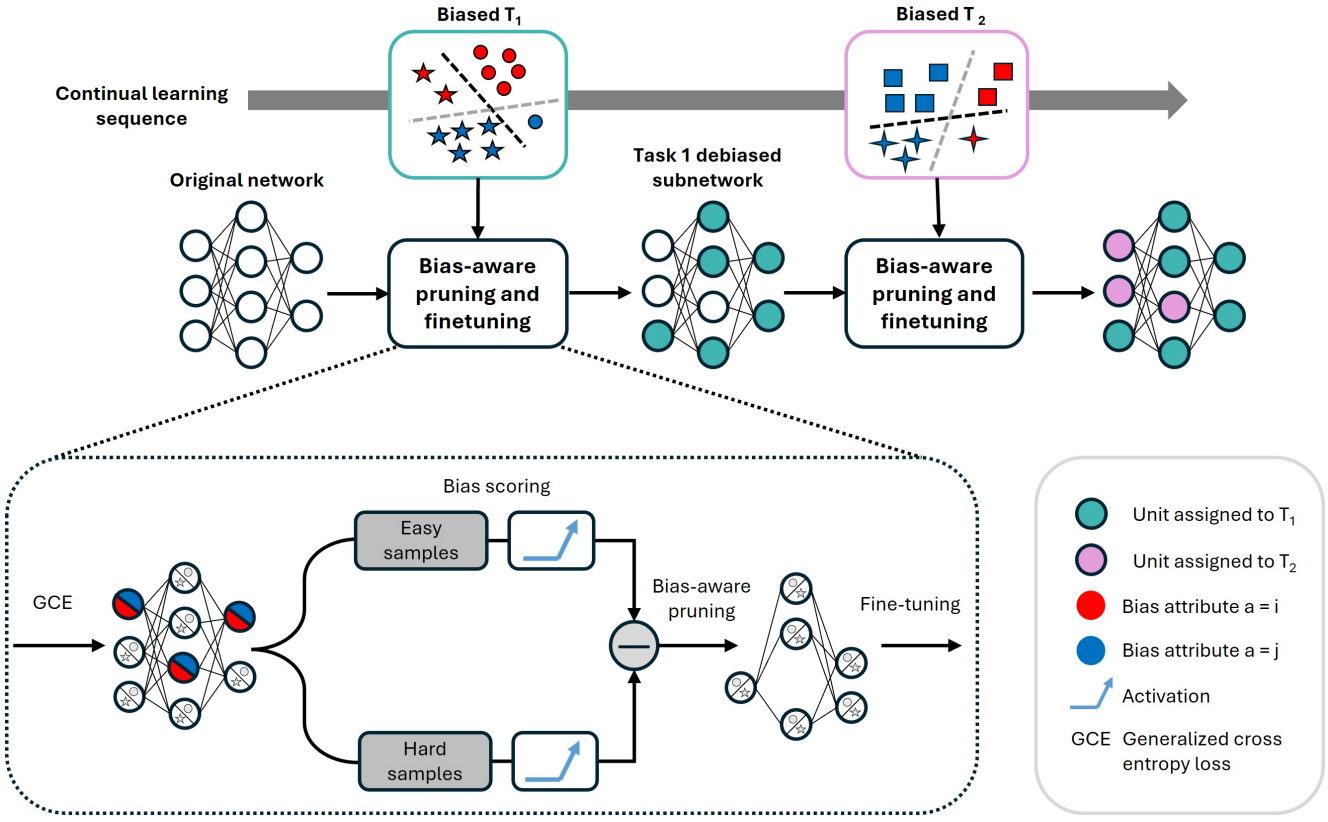

Fig. 4. An overview of the proposed method. (Top) `BiasPruner` **learns sequentially in a continual learning setting, allocating a debiased subnetwork for each task to mitigate bias transfer and avoid forgetting. Each task might be biased, with the same source of bias (e.g., color) affecting all tasks. The debiased subnetwork for Task 1 is shown in green, while the debiased subnetwork for Task 2 is shown in purple. (Bottom)** `BiasPruner` **identifies a debiased subnetwork through bias-aware pruning and fine-tuning. Bias-aware pruning removes units with the highest bias scores, identified by first training the network with generalized cross-entropy loss to amplify bias and then measuring activation differences across two groups (easy vs hard samples). This ensures that the remaining subnetwork retains more robust, unbiased features. The pruned subnetwork is then fine-tuned to enhance generalization.**

correct predictions on these samples. Consequently, the network tends to learn spurious correlations rather than robust features. This deliberate biasing serves a strategic purpose: it highlights which network components overfit to these misleading patterns, allowing us to precisely identify and prune the elements responsible for bias.

After intentionally biasing the network, we partition D_t into two distinct groups for each groundtruth class c to systematically identify which units contribute most to learning biased features for each class:

1. The biased sample set $\mathcal{E}_c^t$ contains samples (x_i, y_i) that the biased network correctly classifies with high confidence (probability $p_{y,i} \geq \tau$). These represent $\mathcal{E}$asier samples that likely contain spurious correlations.
2. The unbiased sample set $\mathcal{H}_c^t$ contains samples (x_i, y_i) that the biased network misclassifies despite high confidence. These represent $\mathcal{H}$arder samples that require more robust feature learning.

Formally, these sets are defined as:

$$\mathcal{E}_c^t = \{i \,|\, y_i = c_i \;\&\; p_{y,i} \geq \tau\} \;\;,\;\; \mathcal{H}_c^t = \{i \,|\, y_i \neq c_i \;\&\; p_{y,i} \geq \tau\} \,. \tag{4}$$

Using these partitions, we calculate a bias score $\mathcal{S}_{c,n}^t$ for each unit n relative to class c by analyzing the unit's ReLU activation patterns:

$$\mathcal{S}_{c,n}^t = \frac{1}{\left|\mathcal{E}_c^t\right|} \sum_{i \in \mathcal{E}_c^t} \mathrm{Var}\left(a_i^n\right) - \frac{1}{\left|\mathcal{H}_c^t\right|} \sum_{i \in \mathcal{H}_c^t} \mathrm{Var}\left(a_i^n\right) \,, \tag{5}$$

Here, $\mathrm{Var}\left(a_i^n\right)$ represents the variance of feature map a_i^n across its spatial dimensions (w, h). The final unit-based bias score $\bar{\mathcal{S}}_n^t$ is computed by averaging across all class-specific scores. Units showing stronger activation responses to biased samples ($\mathcal{E}_c^t$) compared to unbiased samples ($\mathcal{H}_c^t$) receive higher bias scores, effectively identifying them as primary contributors to spurious learning in the network.

5.2.3. *Fine-Tuning the Pruned Subnetwork*

After pruning γ% of the units, we fine-tune the resulting subnetwork f_t to improve its performance on harder-to-learn samples while minimizing performance drops due to pruning. To specifically reduce the risk of reintroducing bias during this phase, we introduce a weighted cross-entropy (WCE) loss function to fine-tune f_t on D_t over a few epochs:

$$\mathcal{L}_{\mathrm{WCE}}(x) = \mathcal{W}(x) \cdot \mathcal{L}_{\mathrm{CE}}\left(f(x), y\right) , \tag{6}$$

$$\text{where} \quad \mathcal{W}(x) = \exp\left(\alpha \cdot \mathcal{L}_{\mathrm{GCE}}(x)\right) \,. \tag{7}$$

Table 2. Summary of the medical imaging benchmarks used for evaluating `BiasPruner`, detailing the number of images, class distribution, task splits, and dataset-specific biases

Dataset	Number of images	Classes	Tasks	Classes per task	Dataset bias
FITZ	16,012	114	6	{19, 19, 19, 19, 19, 19}	Skin tone (I, II, III, IV, V, VI)
HAM	8,678	7	3	{2, 2, 3}	Age (age≥60, age<60)
NIH	19,993	14	3	{4, 5, 5}	Gender (male, female)

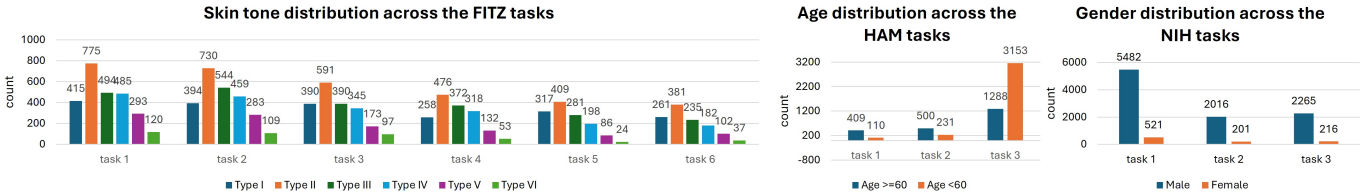

Fig. 5. Bias distribution across the different tasks in FITZ, HAM and NIH.

$\alpha \in (0, 1)$ is a trainable parameter. This weighted loss adjusts the contribution of each sample during fine-tuning by up-weighting harder-to-learn samples (with larger GCE loss values), which are more likely to correspond to unbiased representations, and down-weighting easier (potentially biased) samples. By doing so, the fine-tuning phase not only restores accuracy but also actively discourages the network from reinforcing spurious correlations.

5.3. Debiased Task-Specific Knowledge Transfer and Adaptation

When learning a new task, `BiasPruner` prunes the full original network f to create a new task-specific debiased subnetwork. This subnetwork includes both free units and pre-assigned subnetworks from previous tasks. To maintain previously acquired knowledge, the subnetworks associated with prior tasks remain frozen during training, while only the free units undergo parameter updates to accommodate the new task requirements. By leveraging the debiased subnetworks from earlier tasks, `BiasPruner` ensures that the transfer of knowledge does not introduce forward bias, thus enabling seamless adaptation without the risk of propagating previously learned biases.

5.4. Task-Agnostic Inference: Handling Unknown Task Identities at Test Time

During inference, `BiasPruner` handles a more challenging yet realistic scenario where the task identity of a test image is unknown. Given a test batch X^{test}, we employ a task prediction strategy based on the maximum output response (Dekhovich et al., 2023). Specifically, the task t^* is predicted as:

$$t^* = \arg\max_{t=1,2,\ldots,T} \sum_{i=1}^{s} \max \varphi_t \left(\theta_t \left(x_i^{\text{test}}\right)\right) , \tag{8}$$

where φ_t represents the fully connected layer of the t-th subnetwork. After selecting the task t^*, the final prediction $\hat{y}$ is made based on the corresponding subnetwork:

$$\hat{y} = f_{t^*} \left(X^{\text{test}}\right) . \tag{9}$$

6. Experiments and Results

6.1. Preliminaries

To evaluate `BiasPruner`, we use three medical imaging benchmarks discussed in Section 3: FITZ, HAM and NIH. Unlike the controlled experimental setup in Section 4.1.2, our evaluation follows a more realistic setting with no constraints. We use all images from each dataset, covering all classes, and divide them into T tasks with non-overlapping classes ($T = 6$ for FITZ, $T = 3$ for HAM, and $T = 3$ for NIH). Additionally, for FITZ, we retain all six Fitzpatrick skin tone levels and report results for each, rather than binarizing them as done in our empirical study. Refer to Table 2 for benchmark details and Fig. 5 for the bias distribution across tasks in each benchmark.

For classification performance, we evaluate `BiasPruner` using both accuracy (`ACC`) and the F1-score (`F`). We report accuracy per sensitive attribute (e.g., male, female) as well as overall accuracy (`overall`). For fairness, we use the Demographic Parity Ratio (`DPR`; Eq. 2), as previously described. In this section, we also introduce the Equal Opportunity Difference (`EOD`), which measures whether the true positive rate is independent of the sensitive attribute a. Specifically, for each class y, `EOD` is calculated as:

$$\texttt{EOD} = P(\hat{y} = 1 | \, a = 0, y = 1) - P(\hat{y} = 1 | \, a = 1, y = 1) \, , \tag{10}$$

where $P(\hat{y} = 1 | \, a = i, y = 1)$ represents the probability of correctly predicting the positive class for group $a = i$. A lower absolute `EOD` value indicates a fairer model, with `EOD` = 0 meaning equal true positive rates across groups. Following standard CL evaluation, all metrics are reported at the end of learning (i.e., after training on all T tasks) and averaged across all tasks.

In addition to the baselines (Joint, SeqFT) and CL methods (EWC, PackNet) discussed in our empirical study (Section 4.1.2), we also compare `BiasPruner` against other architecture-based CL methods: SupSup (Wortsman et al., 2020) learns a unique, sparse subnetwork (or supermask) for each task while keeping a randomly initialized backbone fixed, leveraging parameter superposition to prevent interference; WSN (Kang et al., 2022) similarly isolates tasks by identifying a winning subnetwork for each task during training, but differs

Table 3. Classification performance and fairness on FITZ. Best results marked in bold (excluding Exp. $\mathcal{E}$). Higher is better for all metrics except EOD. Our method is highlighted in gray.

Exp	Method	F	ACC							DPR	EOD
			Type-I	Type-II	Type-III	Type-IV	Type-V	Type-VI	overall		
Comparison against Baselines											
$\mathcal{A}$	JOINT	0.256	0.269	0.304	0.335	0.309	0.365	0.245	0.324	0.137	0.298
	SeqFT	0.188	0.187	0.261	0.299	0.254	0.214	0.192	0.221	0.051	0.721
Comparison against CL Methods											
$\mathcal{B}$	EWC	0.325	0.254	0.356	0.355	0.401	0.412	0.244	0.324	0.212	0.342
	PackNet	0.433	0.366	0.402	0.445	0.447	0.479	0.319	0.414	0.154	0.425
	SupSup	0.451	0.254	0.298	0.441	0.452	0.436	0.410	0.425	0.162	0.431
	WSN	0.462	0.371	0.415	0.458	0.469	0.482	0.328	0.421	0.165	0.437
	DER	0.439	0.282	0.310	0.429	0.440	0.450	0.399	0.416	0.158	0.430
	CCG	0.448	0.360	0.399	0.451	0.460	0.475	0.325	0.418	0.161	0.434
	CBA	0.455	0.267	0.305	0.438	0.445	0.442	0.406	0.428	0.159	0.432
Comparison against CL with Bias Mitigation Methods											
$\mathcal{C}$	EWC+S	0.308	0.264	0.357	0.324	0.411	0.417	0.385	0.341	0.228	0.311
	EWC+W	0.321	0.251	0.356	0.334	0.392	0.401	0.398	0.346	0.216	0.298
	PackNet+S	0.495	0.434	0.485	**0.494**	**0.565**	0.562	**0.584**	0.501	0.184	0.248
	PackNet+W	0.527	0.405	0.477	0.480	0.529	0.546	0.524	0.472	0.144	0.246
	SupSup+S	0.466	0.418	0.467	0.432	0.554	0.561	0.534	0.492	0.182	0.221
	SupSup+W	0.457	0.425	0.451	0.448	0.530	0.561	0.544	0.508	0.178	0.254
	WSN+S	0.503	0.440	0.492	0.488	0.562	**0.568**	0.579	0.506	0.186	0.252
	WSN+W	0.532	0.415	0.481	0.472	0.535	0.552	0.530	0.478	0.148	0.249
	DER+S	0.482	0.423	0.476	0.459	0.550	0.558	0.521	0.497	0.180	0.243
	DER+W	0.475	0.429	0.462	0.443	0.525	0.550	0.510	0.485	0.175	0.250
	CCG+S	0.498	0.437	0.486	0.480	0.559	0.564	0.572	0.503	0.185	0.251
	CCG+W	0.525	0.410	0.473	0.466	0.532	0.549	0.523	0.474	0.147	0.247
	CBA+S	0.489	0.428	0.479	0.468	0.553	0.560	0.529	0.499	0.181	0.246
	CBA+W	0.480	0.432	0.465	0.450	0.528	0.551	0.515	0.483	0.176	0.251
Our Proposed Fair CL Method											
$\mathcal{D}$	BiasPruner	**0.540**	**0.457**	**0.502**	0.435	0.551	0.563	**0.584**	**0.512**	**0.331**	**0.202**
Comparison against a Bias Mitigation Method											
$\mathcal{E}$	FairDisCo	0.542	0.479	0.523	0.468	0.571	0.574	0.615	0.548	0.474	0.192

from SupSup by actively optimizing the subnetwork weights; DER (Yan et al., 2021) takes a different approach by dynamically expanding the network's capacity when new tasks arrive, enabling representational growth to accommodate new knowledge; CCG (Abati et al., 2020) introduces conditional channel gating to selectively activate or deactivate channels for each task, allowing for task-specific execution paths through the network; Finally, CBA (Wang et al., 2023b) focuses on mitigating the accumulation of representation bias over time by learning task-wise feature adaptations, rather than relying on architectural isolation or expansion.

6.2. Implementation Details

We use ResNet-50 (He et al., 2016) as the backbone for feature extraction and a unified classifier for all tasks during inference. Following standard practice, we preprocess all images by resizing, normalizing pixel values using dataset-specific mean and standard deviation, and applying random horizontal flipping for augmentation. For dataset partitioning, we allocate 70% of the data for training, 20% for validation, and 10% for testing. We set $q = 0.7$ in $\mathcal{L}_{\text{GCE}}$, consistent with prior works (Zhang and Sabuncu, 2018; Nam et al., 2020) that recommend this value for effective separation of easy and hard examples. The confidence threshold is set to $\tau = 0.7$, a commonly used value in conformal prediction frameworks (Angelopoulos and Bates, 2021; Fayyad et al., 2024a; Graham-Knight et al., 2024), which performed robustly across our experiments. The pruning ratio is fixed at $\gamma = 0.6$, selected based on preliminary runs to balance fairness and accuracy across tasks. For fine-tuning with $\mathcal{L}_{\text{WCE}}$, we train the pruned subnetwork for 20 epochs and select the checkpoint with the highest average ACC and EOD on the validation set. To ensure robustness, we report average results across three random task orders, mitigating the impact of task ordering on performance. All experiments are conducted on a single NVIDIA TITAN V GPU (24GB).

6.3. Evaluation on Non-Binary Bias: FITZ

We begin by reporting the results on the FITZ dataset (Table 3), which presents a more challenging scenario due to the non-binary bias, with six distinct categories of bias levels. In comparison to the baseline methods (Exp. $\mathcal{A}$), BiasPruner demonstrates superior performance in both accuracy and fairness. Specifically, BiasPruner achieves a classification accuracy (overall) of 0.512 and significantly reduces fairness disparities, as indicated by the improved EOD score of 0.202, outperforming all baselines.

In Exp. $\mathcal{B}$, we compare BiasPruner with several CL methods, including both regularization-based (e.g., EWC) and architecture-based (e.g., PackNet, SupSup, WSN, DER, CCG, CBA) approaches. Among these, the architecture-based methods generally achieve better classification accuracy than regularization-based methods, particularly for specific subgroups. However, this improvement often comes at the expense of fairness. This fairness degradation is largely due to how these methods operate: they prune or mask parts of the network to free capacity for new tasks. Such pruning decisions are made agnostic to subgroup-specific features, which risks eliminating units or connections that are critical for certain underrepresented groups, thereby exacerbating disparities in model performance. While some of these architecture-based methods (e.g., SupSup, CCG) show competitive performance on certain metrics, their fairness measures (e.g., DPR, EOD) remain suboptimal. By contrast, BiasPruner achieves a better balance between accuracy and fairness. Although BiasPruner

Table 4. Classification performance and fairness on HAM and NIH. Best results marked in bold (excluding Exp. $\mathcal{J}$). Higher is better for all metrics except EOD. Our method is highlighted in gray.

Exp	Method	HAM						NIH					
		F	ACC			DPR	EOD	F	ACC			DPR	EOD
			<60	≥60	overall				M	F	overall		
Comparison against Baselines													
$\mathcal{F}$	JOINT	0.755	0.781	0.665	0.738	0.239	0.320	0.282	0.306	0.259	0.285	0.706	0.325
	SeqFT	0.431	0.372	0.404	0.416	0.201	0.558	0.219	0.251	0.217	0.231	0.246	0.544
Comparison against CL Methods													
$\mathcal{G}$	EWC	0.788	0.773	0.804	0.772	0.561	0.360	0.398	0.428	0.405	0.417	0.562	0.264
	PackNet	0.824	0.807	0.799	0.808	0.620	0.302	0.434	0.47	0.444	0.458	0.588	0.284
	SupSup	0.831	0.788	0.845	0.822	0.625	0.296	0.448	0.451	0.441	0.445	0.571	0.293
	WSN	0.838	0.799	0.849	0.829	0.628	0.300	0.455	0.463	0.449	0.452	0.582	0.289
	DER	0.819	0.794	0.835	0.812	0.615	0.298	0.442	0.456	0.440	0.448	0.574	0.292
	CCG	0.827	0.802	0.842	0.821	0.623	0.301	0.450	0.468	0.447	0.454	0.585	0.287
	CBA	0.834	0.790	0.847	0.826	0.630	0.299	0.453	0.460	0.445	0.450	0.579	0.290
Comparison against CL with Bias Mitigation Methods													
$\mathcal{H}$	EWC+S	0.834	0.821	0.832	0.827	0.575	0.172	0.412	0.434	0.416	0.421	0.567	0.259
	EWC+W	0.791	0.778	0.784	0.781	0.544	0.168	0.418	0.441	0.423	0.432	0.569	0.251
	PackNet+S	0.839	0.849	0.817	0.829	0.613	0.181	0.419	0.44	0.425	0.434	0.640	0.211
	PackNet+W	0.814	**0.877**	0.819	0.842	0.549	0.189	0.443	0.462	0.456	0.459	0.704	0.192
	SupSup+S	0.849	0.802	0.811	0.817	0.639	0.204	0.432	0.456	0.448	0.451	0.662	0.204
	SupSup+W	0.846	0.797	0.809	0.803	0.536	0.213	0.458	0.481	0.463	0.474	0.731	**0.184**
	WSN+S	0.853	0.812	0.818	0.827	0.641	0.196	0.437	0.448	0.440	0.446	0.656	0.208
	WSN+W	0.844	0.819	0.822	0.834	0.545	0.203	0.462	0.474	0.467	0.470	0.718	0.190
	DER+S	0.847	0.808	0.812	0.819	0.633	0.193	0.429	0.443	0.435	0.439	0.648	0.210
	DER+W	0.840	0.814	0.815	0.824	0.532	0.201	0.452	0.465	0.459	0.463	0.710	0.194
	CCG+S	0.851	0.816	0.820	0.825	0.637	0.198	0.435	0.450	0.443	0.447	0.654	0.207
	CCG+W	0.842	0.822	0.825	0.831	0.540	0.205	0.460	0.472	0.466	0.469	0.722	0.191
	CBA+S	0.849	0.810	0.815	0.821	**0.644**	0.194	0.433	0.445	0.438	0.442	0.650	0.209
	CBA+W	0.838	0.817	0.818	0.828	0.535	0.202	0.455	0.468	0.461	0.465	0.715	0.193
Our Proposed Fair CL Method													
$\mathcal{I}$	BiasPruner	**0.860**	0.851	**0.852**	**0.858**	0.642	**0.127**	**0.488**	**0.525**	**0.484**	**0.507**	**0.821**	0.188
Comparison against a Bias Mitigation Method													
$\mathcal{J}$	FairDisCo	0.873	0.876	0.904	0.893	0.682	0.113	0.486	0.545	0.512	0.538	0.855	0.150

Table 5. Classification (overall) and fairness (DPR) results of BiasPruner from ablation studies. Best results are marked in bold. The results for the default configuration of BiasPruner (from Tables 3 and 4) are highlighted in gray.

Exp	GCE	Bias-aware Pruning	WCE	KT	FITZ		HAM		NIH	
					overall↑	DPR↑	overall↑	DPR↑	overall↑	DPR↑
$\mathcal{D}, \mathcal{I}$	✓	✓	✓	✓	**0.512**	**0.331**	**0.858**	**0.642**	**0.507**	**0.821**
$\mathcal{K}$	×	✓	✓	✓	0.498	0.254	0.834	0.579	0.501	0.779
$\mathcal{L}$	✓	×	✓	✓	0.508	0.328	0.842	0.637	0.498	0.814
$\mathcal{M}$	✓	✓	×	✓	0.481	0.247	0.792	0.576	0.468	0.754
$\mathcal{N}$	✓	✓	✓	×	0.504	0.324	0.851	0.630	0.496	0.803

also uses pruning to manage network capacity across tasks, it does so in a bias-aware manner. Our approach actively preserves those parameters that contribute to fair representation across different subgroups, resulting in consistently strong performance across both accuracy and fairness metrics.

Exp. $\mathcal{C}$ extends this analysis by combining existing CL methods with external bias mitigation techniques; specifically, the widely adopted Resampling (S) and Reweighting (W) algorithms. The Resampling Algorithm adjusts the training distribution by oversampling minority subgroups and undersampling majority subgroups for each combination of skin tone and label, providing the model with more balanced exposure. The Reweighting Algorithm, on the other hand, modifies the contribution of each training example during optimization to counteract the effects of imbalance. These pre-processing strategies help improve fairness metrics to some extent, but their effectiveness is limited because they do not directly address how the internal network structure evolves during continual learning. BiasPruner stands apart because its bias mitigation is integrated within the learning dynamics of continual learning. Instead of relying on data-level adjustments alone, BiasPruner constructs a debiased subnetwork during task learning. This ensures that capacity is preserved for features that are essential to underrepresented groups. As a result, BiasPruner consistently outperforms both CL methods and their bias-mitigation-augmented counterparts in most cases, particularly in balancing high overall accuracy with improved fairness.

Finally, in Exp. $\mathcal{E}$, we compare against FairDisCo (Du et al., 2022), a non-CL bias mitigation technique specifically designed for medical applications, which relies on bias annotations during training. To ensure a fair comparison, we allow FairDisCo to learn each task independently and report the average performance across all tasks (Exp. $\mathcal{E}$). Despite not using bias annotations, BiasPruner demonstrates performance that is slightly lower but still comparable to FairDisCo, reinforcing its ability to mitigate bias effectively without relying on external bias annotations.

6.4. Evaluation on Binary Bias: HAM and NIH

In Table 4, we present the results on the HAM and NIH benchmarks, each associated with a binary bias attribute: age in HAM and gender in NIH. As with the FITZ benchmark, we compare the performance of BiasPruner against baselines, existing CL methods, augmented CL methods, and FairDisCo. BiasPruner (Exp. $\mathcal{I}$) consistently outperforms these methods

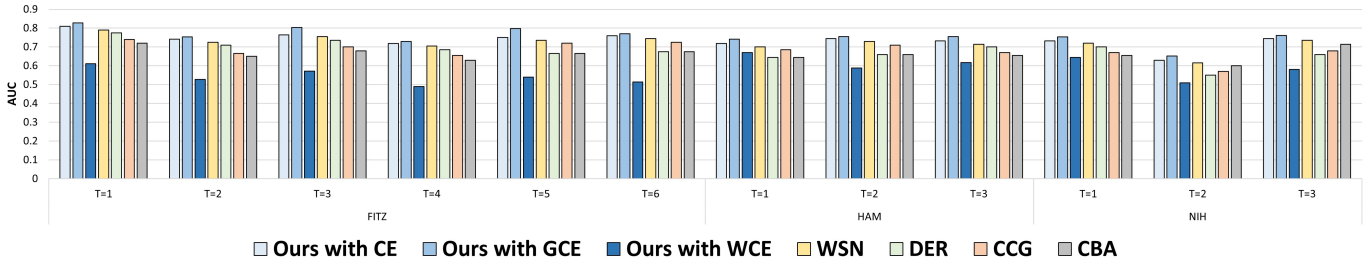

Fig. 6. Bias quantification through AUC for detecting sensitive attributes in frozen models pre-trained for diagnostic tasks. `BiasPruner` demonstrates low AUC values, indicating minimal embedding of sensitive attributes and thus reduced bias in its learned representations.

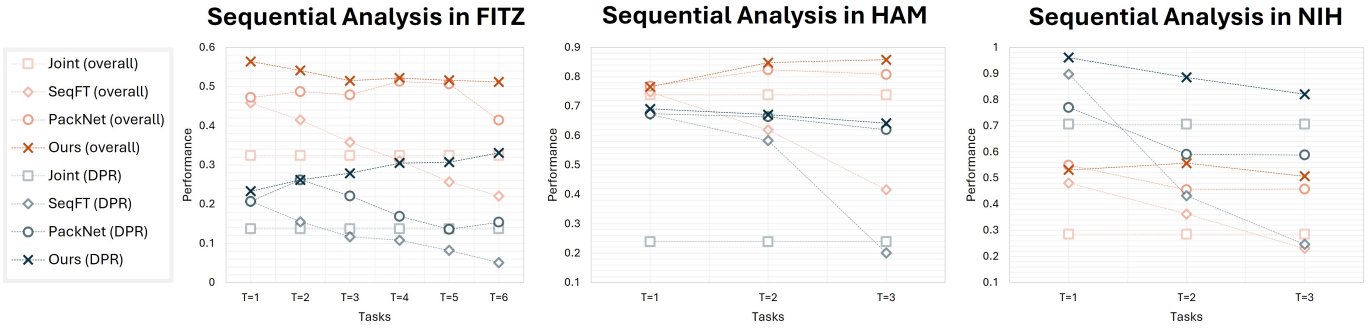

Fig. 7. The `overall` (red-ish) and `DPR` (blue-ish) performance of `BiasPruner` and other methods over all the seen tasks after each training step in the continual learning sequence, where T_i refers to the ith task.

(Exp. $\mathcal{F}$, Exp. $\mathcal{G}$, Exp. $\mathcal{H}$, Exp. $\mathcal{J}$) in both classification accuracy and fairness.

6.5. Ablation Studies

We conduct a series of ablation studies to assess the contribution of each component in `BiasPruner` (Table 5). In Exp. $\mathcal{K}$, we replace the generalized cross-entropy (GCE) loss with standard cross-entropy (CE) loss to evaluate its effect on performance. In Exp. $\mathcal{L}$, we substitute our bias-based pruning strategy with random pruning to isolate the impact of bias-driven pruning. Exp. $\mathcal{M}$ involves fine-tuning the debiased subnetworks using CE loss without weighting, testing the effectiveness of the weighting mechanism in the debiasing process. Finally, in Exp. $\mathcal{N}$, we examine the role of knowledge transfer (KT) by prohibiting any overlap between the subnetworks of the different tasks, effectively simulating the absence of knowledge transfer. Our results reveal that each modification results in a decline in both classification accuracy and fairness compared to `BiasPruner`. Specifically, Exp. $\mathcal{M}$ shows that fine-tuning without weighting (CE loss only) leads to the poorest performance in both metrics, as it increases the likelihood of subnetworks relearning bias. Furthermore, we find that knowledge transfer between debiased subnetworks (Exp. $\mathcal{N}$) significantly improves both performance and fairness in `BiasPruner`.

6.6. Analysis of Model Biases

We conduct a detailed analysis to evaluate how different training strategies impact the degree of bias embedded in learned representations. To do this, we train multiple versions of `BiasPruner` itself: once using standard Cross-Entropy (CE) loss, once using Generalized Cross-Entropy (GCE) loss, and once using our proposed weighted cross-entropy (WCE). For comparison, we also evaluate a set of architecture-based CL methods from prior work, including WSN, DER, CCG, CBA.

First, we train all the models on diagnostic tasks from the FITZ, HAM, and NIH benchmarks. After training, we freeze the feature extractors of each model and train a separate classifier on top to predict sensitive attributes such as skin tone, age, and gender. Higher detection accuracy of these sensitive attributes indicates that the model has encoded more bias in its feature representations. As shown in Fig. 6, the `BiasPruner` trained with CE or GCE loss retain substantial amounts of bias in their learned features, with detection accuracies for sensitive attributes consistently above chance (0.63–0.83). The bias is particularly high for the GCE-trained variant, as GCE encourages the network to rely on shortcut features to improve classification accuracy. The other CL competitors similarly exhibit considerable bias retention. By contrast, our proposed WCE-based `BiasPruner` significantly reduces the presence of sensitive attribute information in the learned features, yielding detection accuracies close to random chance (0.49–0.67). These results demonstrate that combining `BiasPruner` with a bias-aware loss function provides strong protection against spurious correlations, which is important to enhance fairness in CL.

6.7. Sequential Analysis

As depicted in Fig. 7, we conduct a sequential analysis to evaluate the performance of `BiasPruner` over the course of a CL process. This analysis tracks the model's performance step by step across the FITZ, HAM, and NIH benchmarks,

measuring both `overall` and `DPR` at each stage. The results clearly demonstrate that `BiasPruner` consistently outperforms other methods throughout the learning sequence. By observing performance over time, we highlight the advantage of `BiasPruner` in adapting to new tasks while maintaining both high classification accuracy and fairness. Unlike other methods, which may show performance degradation or imbalance between groups as the model progresses, `BiasPruner` manages to maintain a balance, improving task performance without exacerbating fairness disparities.

7. Conclusion and Future Work

In this paper, we introduced `BiasPruner`, a novel continual learning (CL) framework that addresses the challenge of bias transfer in sequential medical image classification. Unlike conventional CL methods that often overlook this issue and inadvertently preserve or propagate dataset biases, `BiasPruner` leverages intentional forgetting to actively mitigate spurious correlations. By identifying and pruning network units that contribute to biased feature representations, `BiasPruner` constructs task-specific debiased subnetworks that retain essential knowledge while discarding spurious associations. While no pruning strategy can fully guarantee the removal of all bias or redundancy, `BiasPruner` effectively prioritizes the removal of the most biased units and complements this with fine-tuning to adapt retained units toward learning robust, unbiased features. Our experiments across diverse medical imaging benchmarks demonstrated that `BiasPruner` consistently achieves superior classification accuracy and fairness, outperforming both recent CL methods and CL methods combined with external bias mitigation strategies. Importantly, these improvements were achieved without requiring explicit bias annotations, addressing practical challenges in real-world medical datasets where such annotations are often unavailable.

Despite these contributions, two important directions remain for future research. First, while `BiasPruner` effectively mitigates individual spurious correlations (e.g., skin tone or gender), real-world data often contain multiple, intersecting sources of bias. Addressing such scenarios remains challenging due to the lack of public datasets with comprehensive and reliable annotations across multiple sensitive attributes. Constructing or curating such datasets, or developing robust methods capable of debiasing in the absence of full bias information, presents an exciting avenue for future work. Approaches like multi-bias attribution scores or generative modeling of missing bias labels may prove promising.

Second, the current implementation of `BiasPruner` incurs an inference overhead proportional to the number of tasks, as it evaluates each task-specific subnetwork to determine the optimal prediction. While this design choice stems from our commitment to handling the more realistic task-agnostic scenario, where test-time task identities are unknown, scaling to larger numbers of tasks may introduce latency challenges in certain deployments. Future work should explore strategies such as lightweight task-routing networks, shared or hierarchical subnetworks to minimize redundant computations, and post-hoc consolidation of subnetworks to improve efficiency.

References

Abati, D., Tomczak, J., Blankevoort, T., Calderara, S., Cucchiara, R., Bejnordi, B.E., 2020. Conditional channel gated networks for task-aware continual learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3931–3940.

Angelopoulos, A.N., Bates, S., 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 .

Angwin, J., Larson, J., Mattu, S., Kirchner, L., 2022. Machine bias, in: Ethics of data and analytics. Auerbach Publications, pp. 254–264.

Ayromloua, S., Tsang, T., Abolmaesumia, P., Li, X., 2024. Ccsi: Continual class-specific impression for data-free class incremental learning. IEEE Transactions on Medical Imaging .

Banerjee, I., Bhattacharjee, K., Burns, J.L., Trivedi, H., Purkayastha, S., Seyyed-Kalantari, L., Patel, B.N., Shiradkar, R., Gichoya, J., 2023. "shortcuts" causing bias in radiology artificial intelligence: causes, evaluation and mitigation. Journal of the American College of Radiology .

Barros, L., Chaves, L., Avila, S., 2023. Assessing the generalizability of deep neural networks-based models for black skin lesions, in: Iberoamerican Congress on Pattern Recognition, Springer. pp. 1–14.

Baweja, C., Glocker, B., Kamnitsas, K., 2018. Towards continual learning in medical imaging, in: Medical Imaging meets NIPS Workshop.

Bayasi, N., 2025. Beyond catastrophic forgetting: advancing continual learning for robust and fair medical image analysis. Ph.D. thesis. University of British Columbia.

Bayasi, N., Du, S., Hamarneh, G., Garbi, R., 2023. Continual-GEN: Continual group ensembling for domain-agnostic skin lesion classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 3–13.

Bayasi, N., Fayyad, J., Bissoto, A., Hamarneh, G., Garbi, R., 2024a. Biaspruner: Debiased continual learning for medical image classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 90–101.

Bayasi, N., Fayyad, J., Hamarneh, G., Garbi, R., Najjaran, H., 2025. Debiasify: Self-distillation for unsupervised bias mitigation, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 3227–3236.

Bayasi, N., Hamarneh, G., Garbi, R., 2021. Culprit-Prune-Net: Efficient continual sequential multi-domain learning with application to skin lesion classification, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), Springer. pp. 165–175.

Bayasi, N., Hamarneh, G., Garbi, R., 2022. Boosternet: Improving domain generalization of deep neural nets using culpability-ranked features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 538–548.

Bayasi, N., Hamarneh, G., Garbi, R., 2024b. Continual-zoo: Leveraging zoo models for continual classification of medical images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4128–4138.

Bayasi, N., Hamarneh, G., Garbi, R., 2024c. GC^2: Generalizable continual classification of medical images. IEEE Transactions on Medical Imaging .

Bayasi, N., Saleh, H., Mohammad, B., Ismail, M., 2013. The revolution of glucose monitoring methods and systems: A survey, in: 2013 IEEE 20th International Conference on Electronics, Circuits, and Systems (ICECS), pp. 92–93.

Bera, S., Ummadi, V., Sen, D., Mandal, S., Biswas, P.K., 2023. Memory replay for continual medical image segmentation through atypical sample selection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 513–522.

Bevan, P.J., Atapour-Abarghouei, A., 2022. Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification, in: MICCAI Workshop on Domain Adaptation and Representation Transfer, Springer. pp. 1–11.

Brown, A., Tomasev, N., Freyberg, J., Liu, Y., Karthikesalingam, A., Schrouff, J., 2023. Detecting shortcut learning for fair medical ai using shortcut testing. Nature Communications 14, 4314.

Busch, F.P., Kamath, R., Mitchell, R., Stammer, W., Kersting, K., Mundt, M., 2024. Where is the truth? the risk of getting confounded in a continual world. arXiv preprint arXiv:2402.06434 .

Byun, Y., Garg, S., Mehta, S.V., Singh, P., Kalpathy-Cramer, J., Wilder, B., Lipton, Z.C., . Conditional diffusion replay for continual learning in medical settings. ICML 2023 Workshop Deployable Generative AI .

Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P., Torr, P., Ranzato, M., 2019. Continual learning with tiny episodic memories, in: Workshop on Multi-Task and Lifelong Reinforcement Learning.

Chen, B., Thandiackal, K., Pati, P., Goksel, O., 2023. Generative appearance replay for continual unsupervised domain adaptation. Medical Image Analysis 89, 102924.

Dekhovich, A., Tax, D.M., Sluiter, M.H., Bessa, M.A., 2023. Continual prune-and-select: class-incremental learning with specialized subnetworks. Applied Intelligence 53, 17849–17864.

Derakhshani, M.M., Najdenkoska, I., van Sonsbeek, T., Zhen, X., Mahapatra, D., Worring, M., Snoek, C.G., 2022. Lifelonger: A benchmark for continual disease classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 314–324.

Du, S., Bayasi, N., Hamarneh, G., Garbi, R., 2023. Avit: Adapting vision transformers for small skin lesion segmentation datasets, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 25–36.

Du, S., Hers, B., Bayasi, N., Hamarneh, G., Garbi, R., 2022. FairDisCo: Fairer ai in dermatology via disentanglement contrastive learning, in: European Conference on Computer Vision, pp. 185–202.

Elkhayat, M., Mahmoud, M., Fayyad, J., Bayasi, N., 2025. Foundation models as class-incremental learners for dermatological image classification. arXiv preprint arXiv:2507.14050 .

Fan, D., Wu, Y., Li, X., 2021. On the fairness of swarm learning in skin lesion classification, in: Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning. Springer, pp. 120–129.

Fayyad, J., 2023. Out-of-distribution detection using inter-level features of deep neural networks. Ph.D. thesis. University of British Columbia.

Fayyad, J., Alijani, S., Najjaran, H., 2024a. Empirical validation of conformal prediction for trustworthy skin lesions classification. Computer Methods and Programs in Biomedicine 253, 108231.

Fayyad, J., Bayasi, N., Yu, Z., Najjaran, H., 2025. Lesiongen: A concept-guided diffusion model for dermatology image synthesis. arXiv preprint arXiv:2507.23001 .

Fayyad, J., Gupta, K., Mahdian, N., Gruyer, D., Najjaran, H., 2024b. Exploiting classifier inter-level features for efficient out-of-distribution detection. Image and Vision Computing 142, 104897.

van Garderen, K., van der Voort, S., Incekara, F., Smits, M., Klein, S., 2019. Towards continuous learning for glioma segmentation with elastic weight consolidation. arXiv preprint arXiv:1909.11479 .

Glocker, B., Jones, C., Roschewitz, M., Winzeck, S., 2023. Risk of bias in chest radiography deep learning foundation models. Radiology: Artificial Intelligence 5, e230060.

González, C., Ranem, A., Othman, A., Mukhopadhyay, A., 2022. Task-agnostic continual hippocampus segmentation for smooth population shifts, in: Domain Adaptation and Representation Transfer MICCAI Workshop, pp. 108–118.

Graham-Knight, J.B., Fayyad, J., Bayasi, N., Lasserre, P., Najjaran, H., 2024. Conformal-in-the-loop for learning with imbalanced noisy data. arXiv preprint arXiv:2411.02281 .

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O., 2021. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1820–1828.

He, J., 2024. Gradient reweighting: Towards imbalanced class-incremental learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16668–16677.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hofmanninger, J., Perkonigg, M., Brink, J.A., Pianykh, O., Herold, C., Langs, G., 2020. Dynamic memory to alleviate catastrophic forgetting in continuous learning settings, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 359–368.

Ji, Z., Guo, D., Wang, P., Yan, K., Lu, L., Xu, M., Wang, Q., Ge, J., Gao, M., Ye, X., et al., 2023. Continual segment: Towards a single, unified and non-forgetting continual segmentation model of 143 whole-body organs in ct scans, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 21140–21151.

Kang, H., Mina, R.J.L., Madjid, S.R.H., Yoon, J., Hasegawa-Johnson, M., Hwang, S.J., Yoo, C.D., 2022. Forget-free continual learning with winning subnetworks, in: International Conference on Machine Learning, PMLR. pp. 10734–10750.

Kaustaban, V., Ba, Q., Bhattacharya, I., Sobh, N., Mukherjee, S., Martin, J., Miri, M.S., Guetter, C., Chaturvedi, A., 2022. Characterizing continual learning scenarios for tumor classification in histopathology images, in: International Workshop on Medical Optical Imaging and Virtual Microscopy Image Analysis, pp. 177–187.

Khan, S.S., Shi, T., Donato-Woodger, S., Chu, C.H., 2025. Mitigating digital ageism in skin lesion detection with adversarial learning. Algorithms 18, 55.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al., 2017. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences 114, 3521–3526.

Kiyasseh, D., Zhu, T., Clifton, D., 2021. A clinical deep learning framework for continually learning from cardiac signals across diseases, time, modalities, and institutions. Nature Communications 12, 4221.

Kumari, P., Chauhan, J., Bozorgpour, A., Azad, R., Merhof, D., 2023. Continual learning in medical imaging analysis: A comprehensive review of recent advancements and future prospects. arXiv preprint arXiv:2312.17004 .

Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E., 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences 117, 12592–12594.

Lenga, M., Schulz, H., Saalbach, A., 2020. Continual learning for domain adaptation in chest x-ray classification, in: Medical Imaging with Deep Learning, pp. 413–423.

Lewandowsky, S., Li, S.C., 1995. Catastrophic interference in neural networks: Causes, solutions, and data, in: Interference and inhibition in cognition, pp. 329–361.

Li, X., Cui, Z., Wu, Y., Gu, L., Harada, T., 2021. Estimating and improving fairness with adversarial learning. arXiv preprint arXiv:2103.04243 .

Li, Z., Hoiem, D., 2017. Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 2935–2947.

Li, Z., Ren, K., Jiang, X., Li, B., Zhang, H., Li, D., 2022. Domain generalization using pretrained models without fine-tuning. arXiv preprint arXiv:2203.04600 .

Liu, X., Shih, H.A., Xing, F., Santarnecchi, E., El Fakhri, G., Woo, J., 2023. Incremental learning for heterogeneous structure segmentation in brain tumor mri, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 46–56.

Luo, L., Xu, D., Chen, H., Wong, T.T., Heng, P.A., 2022. Pseudo bias-balanced learning for debiased chest x-ray classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 621–631.

Mallya, A., Lazebnik, S., 2018. Packnet: Adding multiple tasks to a single network by iterative pruning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7765–7773.

Mousser, W., Ouadfel, S., Taleb-Ahmed, A., Kitouni, I., 2022. Idt: an incremental deep tree framework for biological image classification. Artificial Intelligence in Medicine 134, 102392.

Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J., 2020. Learning from failure: Debiasing classifier from biased classifier. Advances in Neural Information Processing Systems 33, 20673–20684.

Park, J., Kim, M., Whang, S.E., 2024. Fair class-incremental learning using sample weighting. arXiv preprint arXiv:2410.01324 .

Perkonigg, M., Hofmanninger, J., Herold, C.J., Brink, J.A., Pianykh, O., Prosch, H., Langs, G., 2021a. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nature communications 12, 5678.

Perkonigg, M., Hofmanninger, J., Langs, G., 2021b. Continual active learning for efficient adaptation of machine learning models to changing image acquisition, in: Information Processing in Medical Imaging, pp. 649–660.

Pundhir, A., Raman, B., Singh, P., 2024. Biasing & debiasing based approach towards fair knowledge transfer for equitable skin analysis. arXiv preprint arXiv:2405.10256 .

Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H., 2017. iCaRL: Incremental classifier and representation learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2001–2010.

Roy, K., Moghadam, P., Harandi, M., 2023. L3DMC: Lifelong learning using distillation via mixed-curvature space, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp.

123–133.

Salman, H., Jain, S., Ilyas, A., Engstrom, L., Wong, E., Madry, A., 2022. When does bias transfer in transfer learning? arXiv preprint arXiv:2207.02842 .

Truong, T.D., Nguyen, H.Q., Raj, B., Luu, K., 2023. Fairness continual learning approach to semantic scene understanding in open-world environments. Advances in Neural Information Processing Systems (NeurIPS) 36, 65456–65467.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data 5, 1–9.

Vitter, J.S., 1985. Random sampling with a reservoir. ACM Transactions on Mathematical Software (TOMS) 11, 37–57.

Wang, L., Zhang, X., Su, H., Zhu, J., 2023a. A comprehensive survey of continual learning: Theory, method and application. arXiv preprint arXiv:2302.00487 .

Wang, Q., Wang, R., Wu, Y., Jia, X., Meng, D., 2023b. Cba: Improving online continual learning via continual bias adaptor, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 19082–19092.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2097–2106.

Wortsman, M., Ramanujan, V., Liu, R., Kembhavi, A., Rastegari, M., Yosinski, J., Farhadi, A., 2020. Supermasks in superposition. Advances in Neural Information Processing Systems 33, 15173–15184.

Wu, X., Xu, Z., Tong, R.K.y., 2024a. Continual learning in medical image analysis: A survey. Computers in Biology and Medicine 182, 109206.

Wu, Z., Zhu, F., Guo, K., Sheng, R., Chao, L., Fang, H., 2024b. Modal adaptive super-resolution for medical images via continual learning. Signal Processing 217, 109342.

Xie, X., Xu, J., Hu, P., Zhang, W., Huang, Y., Zheng, W., Wang, R., 2023. Task-incremental medical image classification with task-specific batch normalization, in: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pp. 309–320.

Xu, S., Meng, G., Nie, X., Ni, B., Fan, B., Xiang, S., 2024. Defying imbalanced forgetting in class incremental learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 16211–16219.

Yan, S., Xie, J., He, X., 2021. Der: Dynamically expandable representation for class incremental learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3014–3023.

Zhang, J., Gu, R., Xue, P., Liu, M., Zheng, H., Zheng, Y., Ma, L., Wang, G., Gu, L., 2023a. S 3 r: Shape and semantics-based selective regularization for explainable continual segmentation across multiple sites. IEEE Transactions on Medical Imaging .

Zhang, Y., Li, X., Chen, H., Yuille, A.L., Liu, Y., Zhou, Z., 2023b. Continual learning for abdominal multi-organ and tumor segmentation, in: International conference on medical image computing and computer-assisted intervention (MICCAI), pp. 35–45.

Zhang, Z., Sabuncu, M., 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems 31.