MELBA Journal Machine Learning for Biomedical Imaging

Preventing Shortcut Learning in Medical Image Analysis through Intermediate Layer Knowledge Distillation from Specialist Teachers

Christopher **Boland** [1,2], Sotirios A. Tsaftaris [2], Sonia Dahdouh [1],

1 Canon Medical Research Europe, Edinburgh, EH6 5NP, UK
2 School of Engineering, The University of Edinburgh, Edinburgh, EH9 3FG, UK

Abstract

Deep learning models are prone to learning shortcut solutions to problems using spuriously correlated yet irrelevant features of their training data. In high-risk applications such as medical image analysis, this phenomenon may prevent models from using clinically meaningful features when making predictions, potentially leading to poor robustness and harm to patients. We demonstrate that different types of shortcuts—those that are diffuse and spread throughout the image, as well as those that are localized to specific areas—manifest distinctly across network layers and can, therefore, be more effectively targeted through mitigation strategies that target the intermediate layers. We propose a novel knowledge distillation framework that leverages a teacher network fine-tuned on a small subset of task-relevant data to mitigate shortcut learning in a student network trained on a large dataset corrupted with a bias feature. Through extensive experiments on CheXpert, ISIC 2017, and SimBA datasets using various architectures (ResNet-18, AlexNet, DenseNet-121, and 3D CNNs), we demonstrate consistent improvements over traditional Empirical Risk Minimization, augmentation-based bias-mitigation, and group-based bias-mitigation approaches. In many cases, we achieve comparable performance with a baseline model trained on bias-free data, even on out-of-distribution test data. Our results demonstrate the practical applicability of our approach to real-world medical imaging scenarios where bias annotations are limited and shortcut features are difficult to identify a priori.

Keywords

Algorithmic Bias, Shortcut Learning, Knowledge Distillation, Spurious Correlations

Article informations

https://doi.org/10.59275/j.melba.2025-8888

Volume 3, Received: 2025/03, Published 2025/11

Corresponding author: christopher.boland@mre.medical.canon
Special issue: Fairness of AI in Medical Imaging (FAIMI) 2025
Guest editors: Veronika Cheplygina, Aasa Feragen, Andrew King, Ben Glocker, Enzo Ferrante, Eike Petersen, Esther Puyol-Antón, Melanie Ganz-Benjaminsen

1. Introduction

Neural networks frequently demonstrate a preference for the path of least resistance during training, a phenomenon termed "simplicity bias" (Shah et al., 2020). This tendency can lead these models to rely on features that, while strongly correlated with class labels in their training datasets, are irrelevant to the task. Such features, often referred to as "shortcuts" or "spurious correlations", yield an effective decision rule-set within the distribution of the training dataset but one which fails to generalize to data beyond this distribution (Geirhos et al., 2020). For example, a model trained to identify cows in images may learn to detect grassy backgrounds rather than learning to understand what a cow looks like, if most training images show cows in grassy pastures. When presented with images of cows in novel contexts, such as on a beach, the model's prediction accuracy declines significantly (Beery et al., 2018). Because the decision rules learned by these systems prioritize such spurious features over robust, task-relevant ones, they fail to generalize to data where the spurious features are not available. In contrast, a system that is trained to leverage reliable and task-relevant visual features should exhibit consistent performance even amidst shifts in data distribution.

In high-risk applications such as disease diagnosis, where clinical decisions rely on the accurate identification of subtle

and often hard-to-detect disease features, shortcut learning represents a risk to patient safety. Consider pneumothorax detection: popular chest X-ray datasets often contain images acquired post-treatment, once patients have been fitted with treatment devices like chest drains, which are visible in X-ray images. Naturally, this creates a correlation between the presence of the treatment device and the disease label, which a network can learn, incorrectly, to use as a predictive feature of disease presence. Consequently, the model is less accurate at detecting disease in patients who have not yet been treated (Murali et al., 2022). Similarly, models trained to detect atelectasis in chest X-rays can learn to leverage the presence of ECG cables as a predictive feature (Olesen et al., 2024). Disease detection models often rely inappropriately on such confounding features in addition to more subtle features including image acquisition protocol or even demographic characteristics (Souza et al., 2024; Konz and Mazurowski, 2024; Seyyed-Kalantari et al., 2021). Sources of shortcut features in medical data are numerous, and their interactions are complex - exacerbating the challenge of monitoring and accounting for bias. This is magnified by the inconsistencies in metadata collection and labeling practices across datasets and healthcare institutions, making it impractical to track and account for all potential spurious features of the data.

Emerging regulatory frameworks underscore the importance of these challenges. The European Union AI Act, due to come into effect in 2026, establishes comprehensive requirements for AI systems in high-risk domains like healthcare. The act mandates rigorous testing and monitoring of AI systems to identify and mitigate potential biases. Similarly, the World Health Organization's guidelines for AI in healthcare emphasize the need to safeguard patient safety and guarantee equitable treatment outcomes. The FDA's guidelines for AI and machine learning systems in healthcare applications necessitate detailed information regarding the metrics employed and how they ensure patient safety. Additionally, the guidelines request clarity on how to address any new or previously unidentified sources of bias that may arise, as well as details on how to disclose potential biases that could impact the model's effectiveness to users (FDA et al., 2023). FDA approval for many AI systems in healthcare often requires a demonstration of "practical equivalence" to existing systems performing the same task, including evidence that the system's safety is on par with that of current processes (Petrick et al., 2023). These frameworks highlight the risks of deploying systems that may perpetuate or amplify existing healthcare disparities through learned biases. This regulatory landscape creates an urgent need for systematic approaches to identify and mitigate shortcut learning in medical AI systems.

Current approaches to shortcut mitigation can be categorized according to their intervention point in the model development pipeline. Data-centric techniques address bias during pre-processing, where training data distributions are modified through resampling, reweighting, or augmentation to reduce imbalances with respect to bias features (Wu et al., 2023; Li and Vasconcelos, 2019; Ahmed et al., 2022; Zhang et al., 2022; Liu et al., 2021; Wang et al., 2024; Yun et al., 2019). Model-centric techniques include (1) in-processing methods, which incorporate additional loss terms or penalties during training to discourage reliance on spurious features (Sagawa et al., 2019; Müller et al., 2023; Zhang et al., 2022; Boland et al., 2024a) and (2) post-processing approaches, which attempt to remove learned biases from already trained models through fine-tuning or pruning (Xue et al., 2024; Wu et al., 2022; Ghadiri et al., 2024; Bayasi et al., 2024). A critical limitation across many of these methods is their dependency on accurate bias annotations for all training data. The assumption of access to comprehensive and reliable bias labels presents significant practical challenges in medical contexts, where the sources of bias are often numerous, interrelated, and difficult to identify a priori. Even when bias sources are known, obtaining accurate labels across diverse healthcare institutions with inconsistent metadata collection practices is prohibitively resource-intensive, limiting the real-world applicability of these approaches (Banerjee et al., 2023). Consequently, there is a need for methods that can address or reduce this burden of bias annotation while maintaining mitigation efficacy.

Recently, knowledge distillation (KD) has shown potential as a promising in-processing approach for preventing bias learning (Boland et al., 2024a; Cha et al., 2022; Bassi et al., 2024; Kenfack et al., 2024). KD was originally proposed as a model compression technique where a smaller student network learns to mimic the predictions of a larger teacher network through an additional loss term that minimizes the divergence between the student model's outputs and those of the teacher model (Hinton, 2015). In the context of shortcut learning, a teacher trained on carefully curated data might help guide a student away from spurious correlations present in larger, potentially biased datasets. Traditional knowledge distillation approaches typically focus only on matching the final layer outputs. However, several works have demonstrated that learned biases can be detected in the intermediate layers of neural networks and can even be localized to specific network layers (Boland et al., 2024b; Glocker et al., 2023; Stanley et al., 2025). Distillation approaches for debiasing that target the intermediate layers of the network may be able to mitigate biases more effectively.

In our previous work (Boland et al., 2024a), we introduced an oracle-guided training approach to mitigate shortcut learning using a "specialized teacher", a model trained specifically on task-relevant, bias-free data. While

demonstrating promising results, this approach relied on matching batch-wise class probability distributions for knowledge transfer—a technique sensitive to batch composition. Here, we significantly extend this foundation through several methodological improvements. We replace batch-wise probability matching with sample-level Kullback-Leibler (KL) divergence between teacher and student predictions. This more principled approach provides direct guidance for each sample. We also extend the original framework to incorporate knowledge distillation in the final classification layer, complementing the intermediate layer guidance. Furthermore, we significantly expand the experimental validation with extensive evaluation on out-of-distribution (OOD) test sets, systematic analysis of partial-layer distillation, utilization of compact teacher architectures to guide larger student networks, and evaluation of our method's efficacy when training data is corrupted with multiple, simultaneous shortcuts. All of these extensions serve to demonstrate the enhanced generalizability and practical applicability of our approach. Experiments across several network architecture designs, such as AlexNet, ResNet-18, DenseNet-121, and a lightweight 3D CNN, and over multiple medical image analysis tasks in different modalities, demonstrate that our approach is not modality-, task-, or architecture-specific. To further strengthen the viability of the approach in real-world scenarios, we validate our method on a recently released synthetic brain MRI dataset featuring subtle structural bias features, which are hard to detect upon simple inspection (Stanley et al., 2023).

Our proposed approach utilizes a teacher model trained on a small, carefully curated dataset to guide a student network's learning on larger, potentially biased datasets. By distilling knowledge at intermediate network layers, we encourage the student to learn robust, task-relevant features rather than relying on spurious correlations. Our **contributions** are as follows:

1. We demonstrate that intermediate-layer knowledge distillation from a teacher fine-tuned on a small amount of unbiased, task-relevant data effectively mitigates shortcut learning of a student trained on bias-corrupted data and leads to improved generalization as demonstrated through validation on out-of-distribution (OOD) test data.

2. We provide empirical evidence that distillation at intermediate network layers significantly improves bias mitigation compared to final-layer distillation alone.

3. We show that fine-tuning the teacher on task-relevant data leads to performance gains and reductions in bias compared to alternative distillation approaches, such as using a teacher pre-trained on ImageNet data or through confidence regularization of the intermediate layers.

4. We establish that compact teacher architectures (e.g., AlexNet) can effectively guide larger student networks with different architectures (e.g., ResNet-18), critical for real-world deployment where much larger models, which would overfit to small, bias-free training subsets, are likely to be used.

2. Related Works

This section reviews relevant literature in two key areas related to our work: approaches that address shortcut learning in deep neural networks and knowledge distillation techniques that can be leveraged for bias mitigation. We first explore various shortcut mitigation strategies and then examine how knowledge distillation can be adapted to address this challenge.

2.1 Shortcut mitigation

Shortcut mitigation techniques are grouped according to whether they modify the training data or the model's learning process. We review both data-centric and model-centric approaches, highlighting their respective strengths and limitations.

2.1.1 Data-centric techniques

Data-centric approaches address bias at the source by modifying training data distributions. Common approaches include up-sampling and down-sampling the dataset to remove the imbalance in the data with respect to the bias features, or re-weighting the loss to reduce the influence of the bias (Wang et al., 2020; Sagawa et al., 2019). Such approaches require bias labels and sufficient data diversity after resampling or augmentation. In contrast, methods like Just Train Twice (JTT) (Liu et al., 2021) and Discover and Cure (Wu et al., 2023) assign pseudo-labels of the bias feature to identify potentially biased samples before up-sampling or reweighting, avoiding the need for explicit bias annotations. These approaches estimate bias through model accuracy patterns or feature space representations.

Beyond basic resampling and re-weighting approaches, advanced data augmentation techniques have emerged as powerful tools for disrupting potential shortcuts. Cutout (Zhong et al., 2020) introduces random occlusions by masking image regions, forcing models to learn more distributed representations. Mixup (Zhang et al., 2017) creates synthetic training examples by interpolating between image pairs and their labels, reducing overfitting to training artifacts. CutMix (Yun et al., 2019) combines these approaches by replacing removed regions with patches from other training images.

While effective for natural images, these augmentation strategies face limitations in medical contexts. Disease

features in medical images are often subtle and localized, unlike the prominent objects in natural image datasets. Random augmentations risk occluding critical diagnostic features, and they fail to target specific shortcut features systematically.

Ahmed et al. (2022) propose the use of a comprehensive pre-processing pipeline for pneumonia detection in chest X-rays involving normalization, region-of-interest (ROI) cropping, rotations, etc. Through evaluation with both IID and OOD test data, they validate that the influence of biases in the training data is significantly reduced compared to a model trained without applying this pre-processing. While this is relatively straightforward to implement, such an approach requires domain-specific tuning, knowledge of possible shortcut sources, and task-specific domain knowledge to inform some of the augmentation strategies, such as ROI cropping.

2.1.2 Model-centric techniques

Model-centric techniques mitigate learned biases by adjusting the model's weights and learning process, rather than targeting the training data itself. These can be further broken down into in-process techniques, which are applied at training time, and post-process techniques, which are applied after training is complete.

Adversarial training methods (Correa et al., 2024; Zhang et al., 2018) introduce competing objectives to discourage reliance on biased features. However, these often require explicit labels for the bias sources and the competing objectives can introduce instability. Feature disentanglement techniques (Müller et al., 2023) attempt to separate task-relevant from spurious features but also often require explicit bias labels and rely on the assumption that such features are entirely task irrelevant, which may not always hold in practice.

Post-processing methods like pruning (Wu et al., 2022) and fine-tuning (Xue et al., 2024) attempt to remove shortcuts after training. Such approaches are particularly useful when it is not possible to re-train the model, for example, when the full, original training data is not available.

2.2 Knowledge Distillation

While not originally developed for bias mitigation, knowledge distillation-inspired approaches to bias mitigation have shown promise in recent years.

2.2.1 Traditional Knowledge Distillation

Knowledge distillation, originally proposed as a method for model compression (Hinton, 2015), has recently shown effectiveness in mitigating bias learning (Boland et al., 2024a; Cha et al., 2022; Bassi et al., 2024; Kenfack et al., 2024). Adopting a student-teacher training regime, knowledge from a large, well-trained teacher network is "distilled" into a smaller student network. Typically, this process incorporates an additional loss term that quantifies the divergence in predicted class probabilities between the two models (Buciluǎ et al., 2006).

2.2.2 Distillation for bias mitigation

Tian et al. (2024) demonstrate that distillation from a teacher trained on a balanced subset of training data can effectively mitigate learned biases in a student network trained on the full, biased dataset. However, they assume access to labels that accurately portray the source of bias in all of the teacher's training data, and they focus exclusively on the alignment of features in the final network layer, which may allow for more effective bias mitigation. Chai et al. (2022) propose training a teacher model to overfit on its training data, and using its softened logits as training labels for a student model. The authors demonstrate that this soft labeling approach effectively functions as an error-based re-weighting mechanism that can improve fairness metrics without explicit demographic data. However, such an approach may not effectively capture all bias sources in the training data.

While traditional knowledge distillation focuses on final layer outputs, recent work has explored distillation at the intermediate layers. Cha et al. (2022) propose MIRO. Utilizing a large pre-trained network as an "oracle" network, the authors formulate domain generalization as maximizing mutual information between the "oracle" model's representations and a target model's. Similarly, Bassi et al. (2024) propose "explanation distillation" as a technique to prevent shortcut learning in deep neural networks. Their approach distills explanations from a teacher model pre-trained on a massive, diverse dataset, but not necessarily one with task-specific knowledge. This lack of knowledge pertaining to the specific task of the student in the teacher network limits its ability to guide the student network to robust, task-relevant features.

Boland et al. (2024a) introduced an oracle-guided training approach for shortcut mitigation that does not require explicit bias labels for the full training dataset of the student, nor does it make assumptions about bias characteristics. Our work builds upon this foundation through methodological improvements for enhanced robustness, exploration of knowledge distillation applied only to subsets of the student network's intermediate layers, and the use of low-capacity teacher networks to guide high-capacity students.

3. Methods

Our proposed approach addresses the challenge of shortcut learning through a novel teacher-student knowledge distillation framework that guides feature learning at multiple network depths (Figure 1). Central to our approach is the observation that the influence of shortcut learning is detectable in a network's intermediate layers, suggesting that effective mitigation strategies should target the entire network rather than just the final output (Boland et al., 2024b).

In this section, we present our method for measuring intermediate-layer model confidence. We then detail our knowledge distillation approach for mitigating shortcut learning, followed by our experimental setup, including datasets, synthetic shortcut designs, and evaluation protocols.

3.1 Model confidence and shortcut learning

Models trained on biased data tend to exhibit overconfidence in their predictions (Utama et al., 2020). Shortcut features in a model's training data provide an easier decision rule-set with which to infer class. These simple features allow the model to achieve high confidence with less effort (Ao et al., 2023). Prior work has demonstrated that these spurious features lead to detectable changes in the internal behavior of the network (Boland et al., 2024b). We are interested in (a) the confidence with which a trained network infers class through the internal layers, (b) how training on shortcut-corrupted data changes this behavior, and (c) if knowledge distillation from an unbiased teacher can mitigate shortcut learning. To understand this, we introduce classification probes (linear classification heads, consisting of an average pooling layer and a single fully connected layer) which are attached to the intermediate layers of both the student and teacher networks. After the network finishes training, these probes are fine-tuned on the downstream task. Once trained, the probes offer insight into a model's ability to infer the true class at different depths of the network, in addition to facilitating knowledge distillation from the teacher network's intermediate layers to the student's.

3.2 Measuring Model Confidence

To quantify a model's confidence over a batch at each layer, we follow prior work and consider the output logits of the classification probes (Taha et al., 2022). The sigmoid of the output logits ranges from 0 to 1, indicating the likelihood that the input belongs to the positive (1) or negative (0) class. We quantify model confidence $C(\mathcal{X})$ as the deviation from maximum uncertainty (0.5), where higher values indicate greater prediction certainty. This is illustrated in Equation 1, where $f(x)$ represents the sigmoid output of the model for input x.

$$C(\mathcal{X}) = \sum_{x \in \mathcal{X}} |f(x) - 0.5| \tag{1}$$

3.3 Mitigation of shortcut learning via knowledge distillation

Our training scheme (Fig. 1) aims to mitigate shortcut learning by preventing the student model from becoming overconfident through the use of shortcut features. The student model is trained to minimize the cross-entropy loss on a biased dataset while matching the teacher network's class probabilities at each layer.

3.3.1 Teacher-Student Architecture

The "specialist teacher" model is defined as a network trained on a small, carefully curated subset of the full training dataset. This subset is manually selected to contain balanced class representation and to be free of the bias features present in the student's training data. Unlike traditional knowledge distillation approaches that use large, general-purpose teachers, our specialized teacher possesses task-specific knowledge while avoiding the spurious correlations that contaminate larger datasets.

Importantly, all samples used to train the teacher model are excluded from the student's training dataset to prevent leakage between the teacher and the student's training data. For the teacher model, we follow a standard training procedure: the network is trained to completion, then frozen before the classification probes are fine-tuned on the downstream task. We use separate optimizers for updating the network parameters and the probe parameters to prevent unintended interactions between their learning objectives.

The student model is trained on the biased dataset using both the standard classification loss and the knowledge distillation from the teacher. At each epoch, after the student model's parameters have been updated, the network's encoder and final classification head are frozen, and the probes are fine-tuned on the downstream classification task. This maintains the probes' ability to classify based on the student's currently learned feature embeddings while preventing undesired interaction between the probe training and the student's feature learning.

We encourage alignment between teacher and student by minimizing the KL divergence between the output probability distributions of each model's intermediate layer classification probes. Following other intermediate-layer knowledge distillation literature (Haidar et al., 2021; Bassi et al., 2024), we also apply knowledge distillation loss on the output of the network's final classification head.

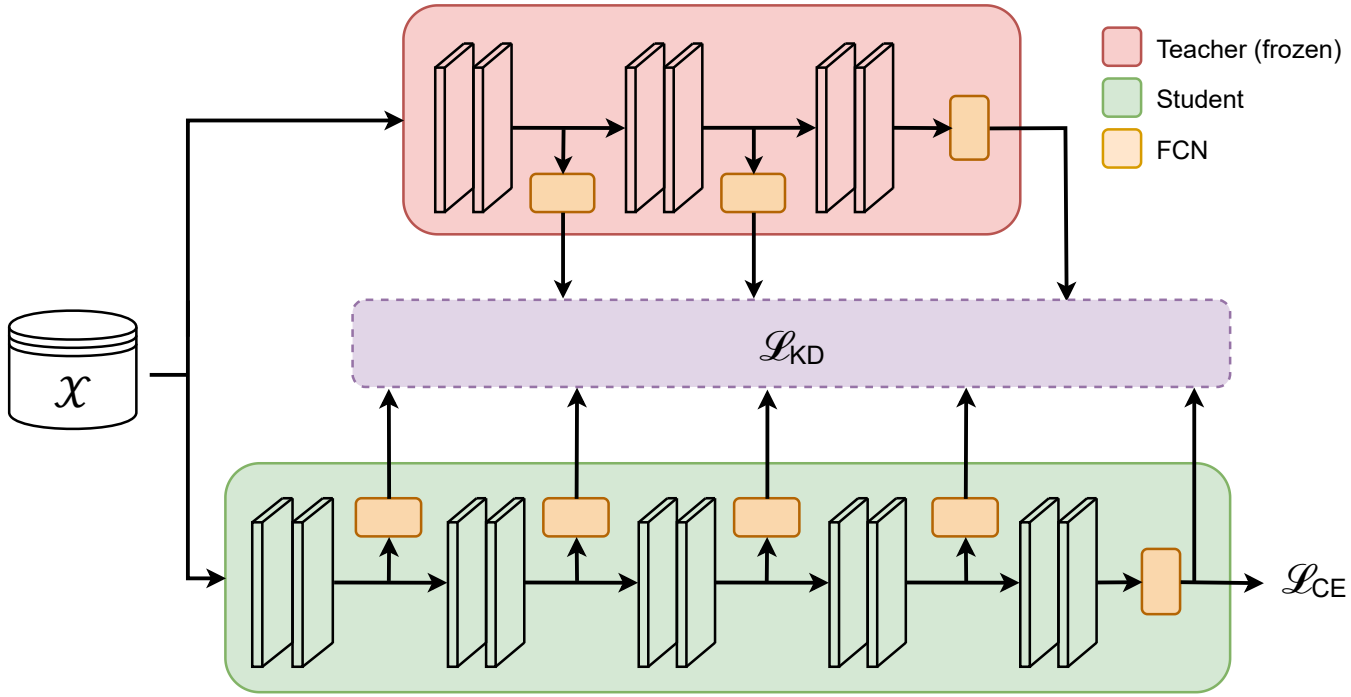

Figure 1: Overview of the proposed student-teacher training method. The teacher network, trained on clean data, guides the student model's learning process through the distillation of task-specific knowledge to the intermediate layers.

3.3.2 Loss Functions

The training loss of the student is described in Eq. 3, where $\mathcal{L}_{total}$ is the total loss, $\mathcal{L}_{CE}$ is the Cross Entropy (CE) loss, $\mathcal{L}_{KD}$ is the knowledge distillation loss between the teacher and student probes, and λ_i is a weight applied to each loss to allow the trade-off between each objective to be managed. For simplicity, we set all weights equal to 1. KL divergence loss is defined in Eq. 2 where we have two sets of intermediate layer predictions, $S = \{p_1, p_2, ..., p_n\}$ and $T = \{q_1, q_2, ..., q_n\}$ where S represent the set of intermediate layer outputs of the student network, T represents the teacher, and α_i represents the weight of the distillation loss to the i^{th} layer of the student.

$$\mathcal{L}_{KD} = \sum_{i}^{n} \alpha_i D_{KL}(p_i^S || q_i^T) \tag{2}$$

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{CE} + \lambda_2 \cdot \mathcal{L}_{KD} \tag{3}$$

3.4 Experimental Setup

3.4.1 Datasets

We evaluate our proposed method using three medical imaging datasets of different modalities and tasks. For each, we enforce class balancing by downsampling the majority class and combine the original train/validation splits. New splits are generated when we run k-fold cross-validation. Table 1 summarizes the composition of positive/negative class samples across train, validation, and test splits for each dataset.

Table 1: Composition of positive/negative class samples in train, validation, and test splits of our datasets.

Dataset	Train	Valid	Test
CheXpert	1457/1457	365/406	600/300
ISIC	560/560	140/146	207/393
SimBA	1291/1292	323/323	530/544
MIMIC	n/a	n/a	500/500
Fitzpatrick17k	n/a	n/a	69/69

- CheXpert (Irvin et al., 2019): a large-scale chest radiography dataset comprised of $224,316$ chest X-rays from $65,240$ patients and 14 disease labels. In our experiments, we consider the task of pneumothorax detection. We use a subset of the full CheXpert dataset containing an equal number of pneumothorax-positive and no finding-positive images. Our final training dataset consists of $2,914$ images.
- ISIC 2017 (Codella et al., 2018): a popular skin lesion image dataset from the International Skin Imaging Collaboration containing $2,000$ dermoscopic images. We perform binary classification between malignant lesions (melanoma/seborrheic keratosis) and benign lesions. Our training split contains $1,120$ images after class balancing.

- SimBA (Stanley et al., 2023): a fully synthetic brain MRI dataset which allows evaluation on 3D medical data with controlled biases. It contains simulated structural changes associated with the class label alongside artificially introduced morphological deformations as potential shortcuts. We also utilize the original version of the dataset without any bias features added.

To assess the generalization of trained models, we include two out-of-distribution (OOD) datasets for evaluation:

- MIMIC (Goldberger et al., 2000; Johnson et al., 2024, 2019): for CheXpert evaluation, we use a class-balanced evaluation set composed of pneumothorax-positive and no finding samples from the MIMIC dataset, another large-scale chest radiograph dataset acquired from a different institution.
- Fitzpatrick17k (Groh et al., 2021, 2022): for ISIC evaluation, we leverage a second dermatological image dataset.

3.4.2 Synthetic shortcuts

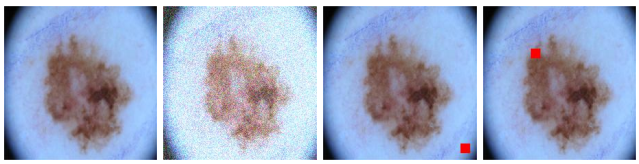

Figure 2: ISIC skin lesion image augmented with synthetic shortcuts: (a) original; (b) noise; (c) square (constant location); (d) square (random location). The noise effect has been amplified here for illustrative purposes.

Inspired by research highlighting common shortcut sources in medical image analysis datasets such as acquisition devices, scanning protocols, hospital tags, and medical devices, we design a controlled environment for the empirical evaluation of our approach. We introduce synthetic bias features into our ISIC and CheXpert training datasets that allow us to assess the generalizability of our method across diverse types of bias. We design several experimental setups featuring a single bias feature and multiple, concurrent bias features. We augment our datasets with one of three unique synthetic bias features (Figure 2):

1. **Diffuse:** leveraging random, uniform noise patterns as a spurious signal spread throughout the image. The noise is generated using a uniform distribution with values between 0 and 0.15 applied to each pixel. Such shortcut features are designed to simulate those that may be caused by acquisition devices and scanning protocols (Ong Ly et al., 2024).

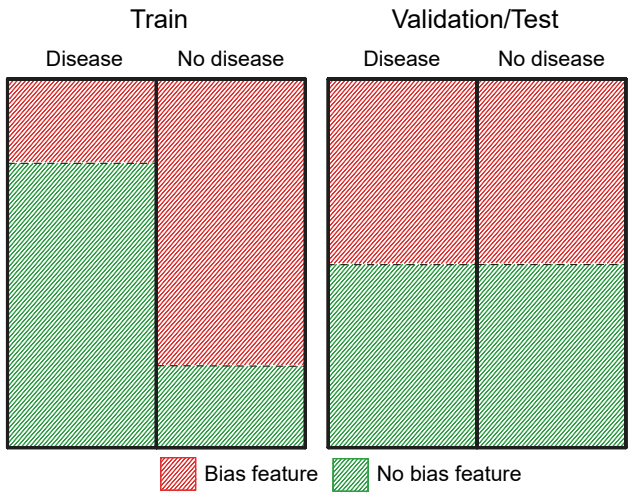

Figure 3: Illustrative representation of the synthetic shortcut feature distribution in our train, validation, and test splits in the CheXpert and ISIC datasets.

2. **Localized:** introducing small square shapes to the image, similar to other work (Dagaev et al., 2023), we aim to simulate more localized shortcut features of various complexities seen in the literature, such as hospital tags and treatment devices (Olesen et al., 2024). We test two variants:
   (a) **Constant location:** the square appears in a fixed spot.
   (b) **Random location:** the location of the square varies among images.

In our training splits, the shortcut features are correlated with the class label (Figure 3). We vary the prevalence of the shortcut feature (the degree of its correlation with the class label in the training data) to assess its influence on training and mitigation efforts. In the validation and test splits, shortcut features are balanced across both classes. In cases with two simultaneous shortcut features, each is correlated with a different class label.

Notably, in the case of the SimBA dataset, all data splits exhibit the same bias prevalence. Experiments on SimBA allow us to validate the efficacy of our method when it is not possible to access an unbiased validation set.

3.5 Evaluation metrics and statistical analysis

We evaluate our experiments by considering both overall performance metrics and bias-specific metrics. For classification performance, we report the Area Under the Receiver Operating Characteristic Curve (AUC), providing a threshold-free measure of discriminative ability across all datasets. Similar to recent research investigating bias and

shortcut learning in medical image analysis, we quantify the impact of shortcut features on model predictions through True Positive Rate disparity (ΔTPR) (Glocker et al., 2023; Stanley et al., 2025). ΔTPR directly measures the model's ability to maintain consistent sensitivity across bias-aligned and bias-contrasting groups, a critical requirement for clinical deployment where missed diagnoses (false negatives) carry severe consequences. We used a 0.5 classification threshold in all cases as the natural decision boundary for binary classification.

We define bias-aligned samples as those whose combination of class label and shortcut feature presence matches the class-bias correlation established in the training data. Meanwhile, bias-contrasting samples represent those whose combination of class label and shortcut feature presence opposes the class-bias correlation in the training data. Statistical significance between performance differences is assessed using paired t-tests with Bonferroni correction to account for multiple comparisons where appropriate.

3.6 Benchmark methods

We compare our approach with several established shortcut learning mitigation methods. The network trained on the original, clean dataset without shortcut features is our **Baseline**. The network trained on the shortcut-corrupted dataset with standard cross-entropy optimization is referred to as **ERM**. We compare to four augmentation-based approaches: **CutOut** (Zhong et al., 2020), **MixUp** (Zhang et al., 2017), **CutMix** (Yun et al., 2019), as well as the use of random rotation (up to $15°$) and horizontal flip augmentations **(Aug)**. We also compare to two popular group-based methods: **GroupDRO (GDRO)** (Sagawa et al., 2019) and **Just Train Twice (JTT)** (Liu et al., 2021)

For each method, we implement configurations following the authors' recommendations without any specific fine-tuning or adjustments made for our data and use identical architecture backbones for fair comparison.

3.7 Implementation Details

In all experiments, we utilize an AdamW optimizer with weight decay of 0.1 and train our models with a learning rate of $1\times e^{-4}$. All intermediate layer classification probes are trained with a learning rate of 0.1. For our 2D datasets (ISIC, CheXpert, MIMIC, and Fitzpatrick17k) images are re-sized to ImageNet resolution, 224×224, while for SimBA, we resize to $96\times96\times96$. We do not apply any rotation or flipping augmentations by default. We set the maximum number of training epochs to 1000 with early stopping after 15 epochs if there is no improvement in the validation loss. In none of our experiments does training run for the complete 1000 epochs without reaching the early stop condition. We use 5-fold cross-validation, with consistent test sets across folds. Our experimental setup utilizes Python and PyTorch, and we train on an NVIDIA RTX 2080 Ti and Tesla V100s.

For the 3D experiments on SimBA data, we employ a lightweight 3D CNN consisting of five convolutional blocks, each containing a 3D convolutional layer (kernel size 3×3×3), batch normalization, and Sigmoid activation. Classification probes are attached after each convolutional block, consisting of 3D global average pooling followed by a linear layer.

4. Results

Our experimental evaluation examines several key aspects of the proposed approach. We first investigate how shortcut learning manifests in intermediate network layers, then evaluate our knowledge distillation method against alternative approaches. We also analyze the impact of partial layer distillation, the effectiveness of compact teacher architectures, and performance on realistic structural biases in 3D medical data. In the following work, our teacher networks are trained on a subset of 20% of the full, original training data. The samples in this subset are removed from the student network's training data. We later explore the efficacy of our teacher with fewer training data.

4.1 Core method validation

We begin our experimental evaluation by establishing fundamental evidence for our approach: first demonstrating how shortcut learning manifests in neural networks, then validating our knowledge distillation method's effectiveness across diverse experimental conditions.

4.1.1 Shortcut learning manifests distinctly across network layers

Considering the complexity of many medical image analysis tasks, we expect a well-trained model using clinically relevant features to exhibit lower confidence than a model relying on easy shortcut features. Additionally, we might expect that the confidence of a model reliant on shortcut features will increase in earlier layers, aligning with the expectation that the deeper layers of the network capture more sophisticated features (Baldock et al., 2021).

To test this, we train two ResNet-18 models on CheXpert following Empirical Risk Minimization (ERM), where we simply aim to optimize cross-entropy loss. One is trained on the original dataset without any synthetic biases, while the other is trained on the same dataset augmented with synthetic shortcut features associated with the disease class. In this case, the shortcut features have a 100% prevalence rate (perfectly correlated with the disease class). After training, we fine-tune our classification probes for each model.

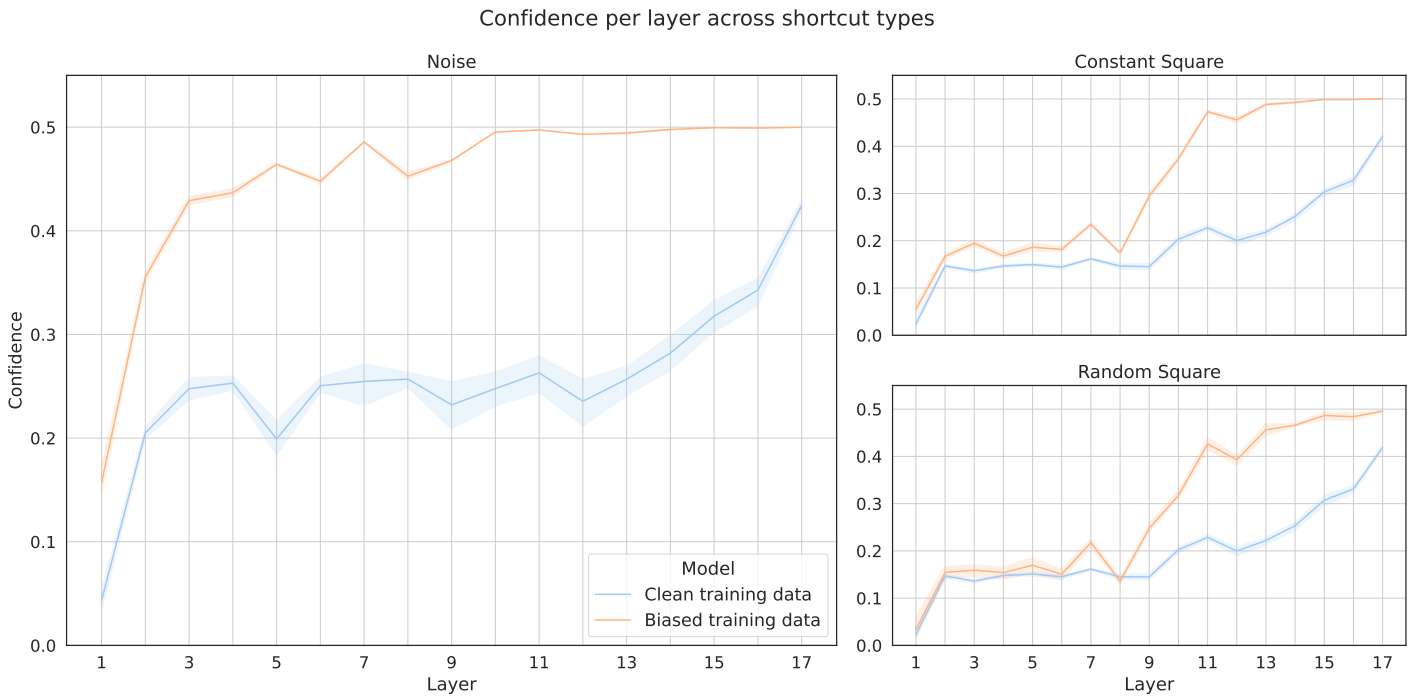

Figure 4: Intermediate-layer confidence of two ResNet-18 models trained on near-identical training sets. Confidence bands represent the standard deviation over 5-fold cross-validation. Both networks are trained on the CheXpert dataset with a learning rate of $1e^{-4}$. Intermediate layer classification probes have a learning rate of 0.1. The training data of one model has been corrupted with various synthetic shortcut features, while the training data of the other has not.

Each model is evaluated on our held-out test set, and the output of the probes are used to evaluate the influence of the bias feature on the network's predictive behavior.

Figure 4 illustrates the per-layer confidence of each model. In line with our hypothesis, the model trained on the biased data becomes overconfident compared to the baseline trained on clean data. In the case of our diffuse noise shortcut, we observe this to an extreme degree in the earliest layers, while the localized shortcuts don't result in a large degree of overconfidence until the later layers. This is likely because the diffuse shortcut is composed of low-level features that require very little disambiguation by the network, unlike the localized shortcuts. We observe similar patterns in other datasets and network architectures that we evaluate.

4.1.2 *Intermediate-layer knowledge distillation mitigates shortcut reliance*

Having established the layer-specific nature of shortcut learning, we now demonstrate that our knowledge distillation framework, utilizing a teacher trained on a small curated dataset, effectively prevents students from developing shortcut dependencies across multiple datasets and bias types. We evaluate our approach across the CheXpert and ISIC datasets and all student models are trained on data augmented with a synthetic bias feature with different degrees of correlation with the class label (prevalence). We compare the performance of our model to several augmentation-based debiasing approaches, as well as Just Train Twice (JTT) and GroupDRO.

Across all tested bias types and degrees of correlation with class labels, our student network demonstrates the most consistently low bias in its predictions, as measured by ΔTPR between bias-aligned and bias-contrasting samples (Table 2). In several cases, the TPR disparity is reduced such that it is comparable to our clean baseline. Our method remains similarly effective in reducing the bias even as its prevalence increases, while most other methods we evaluate worsen in effectiveness at higher prevalence rates. Notably, the majority of methods show significantly reduced efficacy in mitigating diffuse shortcuts across both datasets. Contrastingly, our approach is consistently effective across localized and diffuse shortcuts. We highlight a consistent drop in efficacy for the noise bias feature in the CheXpert dataset with all methods, including ours. We suspect that this is due to useful textural information being corrupted by the noise shortcut. We also typically find that our approach achieves better overall AUC compared to other methods, particularly at higher prevalence rates. As shortcut prevalence decreases from 95% to 75%, all methods show improved ΔTPR, which is expected since weaker biases will provide less misleading signal during training. However,

Table 2: ΔTPR ↓ between bias-aligned and bias-contrasting samples for a ResNet-18 trained on data with various bias prevalence rates. Results are presented as Mean±Std over 5-fold cross-validation. Models are marked as **best** and second-best. When the difference between first and second best is statistically significant ($p < 0.05$ according to a paired t-test), the best-performing model is highlighted with *.

Prev. (%)	Model	CheXpert			ISIC		
		Noise	Square (C)	Square (R)	Noise	Square (C)	Square (R)
0	Baseline	0.131±0.051	0.020±0.008	0.015±0.013	0.409±0.095	0.056±0.015	0.050±0.017
100	ERM	1.000±0.000	1.000±0.000	0.991±0.008	1.000±0.000	0.777±0.093	0.844±0.168
	MixUp	0.987±0.026	0.998±0.003	0.854±0.119	0.998±0.004	0.875±0.152	0.754±0.157
	CutOut	0.999±0.001	1.000±0.000	0.832±0.112	1.000±0.000	0.448±0.065	0.277±0.064
	CutMix	0.993±0.005	0.503±0.073	0.126±0.022	1.000±0.000	0.359±0.063	0.116±0.053
	Aug	0.957±0.092	0.979±0.013	0.984±0.006	1.000±0.000	0.161±0.052	0.515±0.133
	Ours	**0.377±0.185***	**0.079±0.017***	**0.035±0.013***	**0.068±0.055***	**0.034±0.016***	**0.074±0.042**
95	ERM	0.939±0.028	0.912±0.086	0.791±0.093	0.959±0.019	0.861±0.036	0.703±0.103
	MixUp	0.927±0.105	0.987±0.008	0.693±0.103	0.952±0.037	0.936±0.035	0.757±0.066
	CutOut	0.950±0.050	0.912±0.069	0.636±0.131	0.946±0.047	0.579±0.249	0.439±0.146
	CutMix	0.975±0.030	0.325±0.048	0.142±0.063	0.922±0.079	0.311±0.107	0.134±0.040
	Aug	0.899±0.072	0.779±0.150	0.813±0.061	0.935±0.054	0.205±0.079	0.549±0.115
	GDRO	0.986±0.010	0.978±0.015	0.702±0.091	0.967±0.014	0.800±0.035	0.477±0.029
	JTT	0.982±0.011	0.946±0.031	0.673±0.065	0.946±0.031	0.793±0.043	0.505±0.073
	Ours	**0.372±0.110***	**0.089±0.023***	**0.047±0.027**	**0.077±0.069***	**0.100±0.039**	**0.052±0.012***
85	ERM	0.745±0.145	0.735±0.113	0.423±0.063	0.274±0.037	0.376±0.094	0.294±0.083
	MixUp	0.699±0.115	0.677±0.131	0.336±0.028	0.490±0.091	0.523±0.100	0.407±0.075
	CutOut	0.810±0.072	0.656±0.214	0.374±0.104	0.511±0.227	0.471±0.158	0.262±0.115
	CutMix	0.733±0.107	0.211±0.083	0.097±0.017	0.653±0.120	0.214±0.072	0.082±0.031
	Aug	0.664±0.190	0.517±0.197	0.526±0.132	0.677±0.206	0.115±0.053	0.317±0.081
	GDRO	0.759±0.032	0.624±0.060	0.267±0.083	0.697±0.085	0.335±0.085	0.097±0.053
	JTT	0.719±0.035	0.533±0.034	0.378±0.121	0.703±0.125	0.350±0.073	0.158±0.051
	Ours	**0.348±0.151***	**0.106±0.051**	**0.059±0.044**	**0.077±0.109***	**0.057±0.037**	**0.067±0.083**
75	ERM	0.445±0.155	0.387±0.064	0.201±0.087	0.253±0.054	0.273±0.057	0.186±0.077
	MixUp	0.380±0.120	0.331±0.067	0.156±0.036	0.470±0.149	0.282±0.095	0.183±0.036
	CutOut	0.460±0.078	0.361±0.111	0.168±0.044	0.265±0.188	**0.140±0.036**	0.165±0.120
	CutMix	0.526±0.095	0.138±0.032	**0.055±0.020**	0.330±0.082	0.153±0.048	0.070±0.061
	Aug	0.446±0.131	0.408±0.158	0.345±0.044	0.324±0.142	0.149±0.063	0.256±0.066
	GDRO	0.507±0.019	0.329±0.015	0.114±0.024	0.444±0.064	0.183±0.023	**0.057±0.010**
	JTT	0.452±0.046	0.261±0.035	0.146±0.048	0.459±0.077	0.204±0.029	0.086±0.037
	Ours	**0.364±0.182**	**0.127±0.059**	0.061±0.028	**0.066±0.040***	0.145±0.088	0.063±0.024

even at lower prevalence rates, our approach maintains its advantage over other methods.

4.1.3 Generalization to clean and out-of-distribution data

A critical test of any deep neural network is whether it has learned a robust, generalizable set of decision rules. Here we evaluate our approach across three distinct evaluation scenarios: (1) a biased test set where shortcuts are present but distributed equally across classes, such that they are no longer useful predictive features; (2) a clean test set featuring none of the synthetic bias features present in our training sets; and (3) out-of-distribution (OOD) test sets to evaluate generalization. This allows us to assess both the method's ability to ignore spurious features and its capacity to learn robust, generalizable, and clinically relevant features. Our findings are highlighted in Table 3.

Comparisons on the bias-corrupted test set allow validation of how well each model learned to ignore the presence of the shortcuts at inference. Across all shortcut types on both datasets, we find that our method consistently achieves the best overall AUC and consistently matches or even outperforms the clean baseline evaluated on the shortcut-corrupted test data. We highlight that the clean baseline consistently sees a significant drop in performance when evaluated on noise-corrupted data. We hypothsize that this is likely a result of the degradation of useful texture-related information in the test set, combined with inherent bias of CNN architectures towards textural information (Geirhos et al., 2018).

Interestingly, we find that all models see significantly improved performance when tested on the clean test data. This supports previous findings that biases in the training

Table 3: AUC ↑ for ResNet-18. We compare our approach to four popular augmentation-based de-biasing techniques. Shortcuts here have a 100% correlation with the task label, so group-based methods (GroupDRO and JTT) are omitted from these comparisons. Results are presented as Mean±Std over 5-fold cross-validation. Models are marked as **best** and second-best. When the difference between first and second best is statistically significant ($p < 0.05$ according to a paired t-test), the best-performing model is marked *.

Test set	Model	CheXpert			ISIC		
		Noise	Square (C)	Square (R)	Noise	Square (C)	Square (R)
Biased	Baseline	0.709±0.024	0.755±0.013	0.752±0.015	0.749±0.024	0.809±0.019	0.808±0.017
	ERM	0.489±0.012	0.533±0.007	0.554±0.006	0.521±0.011	0.600±0.011	0.612±0.007
	MixUp	0.509±0.007	0.539±0.008	0.550±0.010	0.490±0.026	0.555±0.028	0.585±0.011
	CutOut	0.498±0.029	0.548±0.010	0.584±0.008	0.521±0.023	0.627±0.013	0.639±0.006
	CutMix	0.529±0.007	0.680±0.025	0.758±0.015	0.516±0.015	0.753±0.038	0.781±0.020
	Aug	0.483±0.009	0.585±0.017	0.550±0.013	0.518±0.006	0.731±0.017	0.631±0.010
	Ours	**0.689±0.044***	**0.747±0.008***	**0.761±0.01**	**0.775±0.023***	**0.777±0.024**	**0.805±0.016**
Clean	Baseline	0.754±0.014	0.754±0.014	0.754±0.014	0.811±0.019	0.811±0.019	0.811±0.019
	ERM	0.491±0.029	0.599±0.020	0.704±0.015	0.498±0.038	0.672±0.038	0.745±0.015
	MixUp	0.587±0.021	0.581±0.015	0.649±0.025	0.402±0.018	0.565±0.037	0.652±0.036
	CutOut	0.527±0.063	0.604±0.012	0.741±0.007	0.485±0.089	0.722±0.021	0.758±0.013
	CutMix	0.608±0.017	0.743±0.037	**0.776±0.011**	0.495±0.040	**0.791±0.037**	0.797±0.016
	Aug	0.507±0.037	0.711±0.044	0.703±0.015	0.444±0.012	0.777±0.021	0.727±0.032
	Ours	**0.741±0.010***	**0.749±0.009**	0.763±0.011	**0.767±0.028***	0.778±0.024	**0.807±0.016**
OOD	Baseline	0.737±0.014	0.737±0.014	0.737±0.014	0.677±0.024	0.677±0.024	0.677±0.024
	ERM	0.461 ± 0.042	0.548±0.015	0.688±0.027	0.645±0.031	0.557±0.012	0.556±0.052
	MixUp	0.534±0.037	0.546±0.035	0.633±0.038	0.567±0.035	0.539±0.033	0.534±0.036
	CutOut	0.424±0.066	0.571±0.049	0.730±0.013	0.635±0.016	0.585±0.008	0.583±0.020
	CutMix	0.526±0.040	0.692±0.013	0.724±0.021	0.650±0.020	0.561±0.034	0.529±0.043
	Aug	0.426±0.027	0.683±0.018	0.692±0.008	0.670±0.012	0.635±0.028	0.618±0.026
	Ours	**0.733±0.018***	**0.759±0.022***	**0.763±0.008***	**0.727±0.057**	**0.697±0.030***	**0.666±0.042**

data do not necessarily prevent models from learning underlying causal features (Stanley et al., 2025; Glocker et al., 2023); but can lead them to preferentially rely on the spuriously correlated features when they are available. Notably, across all shortcut types, our model tested on the clean dataset achieves performance that is competitive with the baseline. By comparison, most other tested methods fail to see an improvement in AUC on the clean test set when the training data was augmented with the noise shortcut. We highlight this as further evidence of the power of a teacher model fine-tuned on a small amount of task-relevant data to prevent a student from being corrupted by the spurious feature.

Finally, the OOD test sets serve to evaluate the robustness and generalizability of the decision rules learned by the network. Here, we see that our student network consistently matches the performance of the clean baseline across both datasets and all bias features, consistently outperforming all other approaches.

These findings collectively support our hypothesis that task-relevant knowledge distillation across intermediate network layers can effectively guide models toward learning more robust and clinically relevant features.

4.1.4 Effectiveness against multiple concurrent shortcuts

Our proposed approach has demonstrated promise in the mitigation of synthetic shortcuts. However, prior experiments purposefully represent a highly controlled shortcut environment. Only a single synthetic shortcut is present in the training data. Realistically, spurious features are unlikely to be constrained to a single source, particularly in large datasets. It is important, therefore, that any bias mitigation approach is able to mitigate multiple sources of bias simultaneously.

We augment our training data with two simultaneous shortcuts, one correlated with the positive class and the other with the negative class. The predictive strength of these shortcuts is varied across different training sets. As seen in Figure 5, our method remains effective in the presence of multiple bias sources in the training data, and across all tested prevalence rates, consistently outperforming all other methods. In the majority of cases, we find that our student model remains competitive with the baseline model trained on entirely clean data.

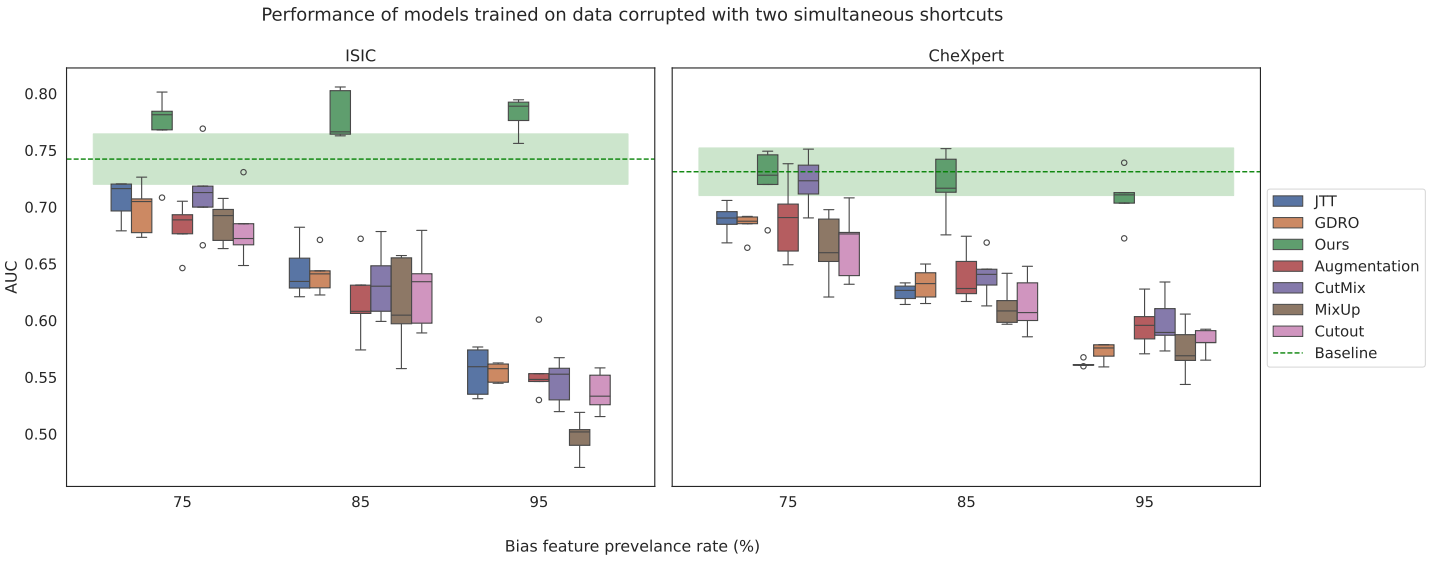

Figure 5: Performance of ResNet-18 trained on ISIC and CheXpert datasets featuring multiple simultaneous shortcuts. The green line represents a model trained on a training set before augmenting with synthetic shortcuts. We compare our student with a specialized teacher to JTT and GroupDRO.

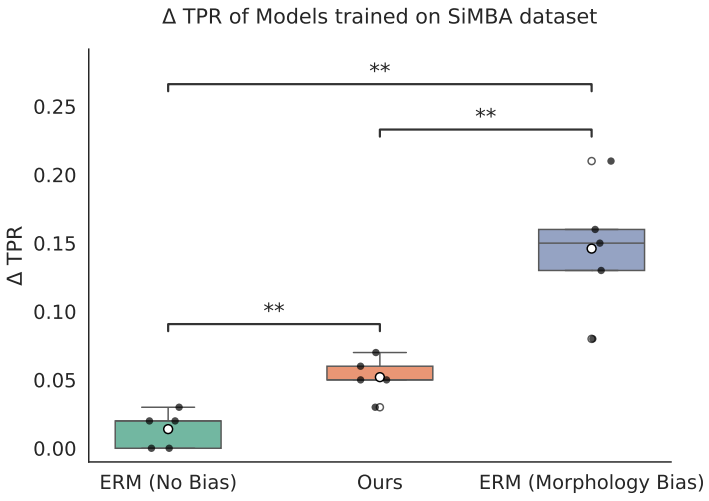

Figure 6: ΔTPR of 3D CNN models trained on SimBA data. ERM (No Bias) is trained on data without any bias features. Ours and ERM (Morphology Bias) are trained on data augmented with a synthetic morphological bias feature. All models are evaluated on test data featuring the morphology bias. ** indicates a statistically significant difference in ΔTPR according to a paired t-test with Bonferroni correction.

4.1.5 Validation on realistic 3D structural biases

While our previous experiments focused primarily on 2D image classification with synthetic shortcuts, we now extend our analysis to a more realistic scenario featuring subtle structural biases that more closely resemble real-world medical imaging artifacts. To evaluate our approach in this realistic context, we leverage the SimBA dataset — a synthetic brain MRI dataset designed specifically to study bias in 3D medical image analysis (Stanley et al., 2024). The SimBA dataset features subtle morphological deformations that correlate with disease labels at a 65% prevalence rate. These localized structural modifications represent a more nuanced and challenging form of bias compared to our previous experiments with artificial shortcuts.

Importantly, a key methodological distinction in these experiments is that all data splits in SimBA (training, validation, and test) exhibit the same bias prevalence rate. This differs from our previous synthetic shortcut experiments, where validation data contained balanced shortcut distributions. The absence of bias-balanced validation data creates a significantly more challenging scenario that closely mirrors real-world clinical settings, where validation data often shares the same biases as training data.

We train a lightweight 3D CNN with linear classification probes attached after each convolutional layer. Our teacher model is trained on a 20% subset of the unbiased training data, while the student model is trained on the biased dataset. All volumes were resampled to $96\times96\times96$ voxels.

Figure 6 presents the performance comparison between three models: our student model guided by a teacher fine-tuned on some task-relevant data (Ours), a model trained on the full unbiased dataset (ERM (No Bias)), and a model trained on the biased dataset using standard Empirical Risk Minimization (ERM (Morphology Bias)). The results demonstrate clear performance differences among these approaches.

Statistical analysis using repeated measures ANOVA confirms significant differences between the models. Subsequent pairwise comparisons using paired t-tests with Bon-

ferroni correction reveal statistically significant differences in ΔTPR between the ERM model trained on the morphologically biased dataset and both alternative models. Notably, while a statistically significant difference in ΔTPR remains between our model and ERM (No Bias), the disparity difference is significantly reduced compared to ERM (Morphology Bias).

Our findings demonstrate that our approach can effectively mitigate bias even without the benefit of a balanced validation set to guide the training process. This is significant for real-world medical imaging applications, where obtaining bias-balanced validation data is often infeasible. These results further validate the applicability of our method to complex 3D medical imaging tasks featuring realistic bias patterns, suggesting broader potential for clinical applications.

4.2 Method design and optimization

Having validated our core approach, we now explore key design choices that optimize its effectiveness and practical applicability.

4.2.1 Partial layer distillation preserves student performance

While our initial implementation applied knowledge distillation across all batch normalization layers of our ResNet-18 students, this comprehensive approach might over-constrain the students' learning process. We investigate whether more selective application of distillation can maintain or even enhance performance. We systematically evaluate distillation applied at varying numbers of intermediate layers in a ResNet-18 network, from all 17 intermediate layers to 0 intermediate layers (final classification head only).

For partial-layer configurations, we employ a random sampling approach where we independently select n layers from both the student and teacher networks during each training epoch. Importantly, these selections are made independently, meaning the specific layers chosen may differ between networks. We pair the selected layers sequentially based on their relative depth to establish meaningful knowledge transfer despite potentially different architectural positions.

Table 4 reveals key insights about the value of our intermediate-layer distillation. We note that applying distillation at fewer intermediate layers $(5-9)$ leads to comparable performance to distillation applied at all intermediate layers (17), both in terms of AUC and ΔTPR. In some cases on the ISIC dataset, we see improvements in the AUC when the loss is applied at fewer layers. We hypothesize that in these cases, the reduced regularization of the KD loss facilitates an improved ability of the network to learn task-relevant features without sacrificing the useful guidance away from spurious features. Importantly, when distillation is applied solely at the final classification head and not in the intermediate layers $(n = 0)$, performance declines significantly and bias increases significantly across all experiments. This dramatic deterioration highlights the critical role of intermediate-layer guidance in mitigating shortcut learning.

4.2.2 Low-capacity teachers effectively guide larger student networks

Training the teacher using a small, curated subset of data can pose challenges when applied to significantly larger models. In this study, we examine whether a low-capacity model can effectively serve as a teacher for a higher-capacity student. Specifically, we distill knowledge from an AlexNet teacher to a ResNet-18 student, and from a ResNet-18 teacher to a DenseNet-121 student. To apply knowledge distillation from a low-capacity teacher, we follow a similar protocol to Section 4.2.1. We randomly sample n layers from the student network each epoch, where n is equal to the number of classification probes in the teacher model. The n layers of the student network are paired sequentially with the classification probes of the teacher.

We train our student on our datasets augmented with synthetic biases and present these results in Table 5. We compare our student to an identical network trained following a standard Cross Entropy optimization protocol (ERM). Even with a small teacher network, knowledge distillation from the intermediate layers proves capable of effectively mitigating the influence of shortcuts present in the student training data.

In real-world applications, it is more likely that larger models, such as DenseNet-121—often considered state-of-the-art—are employed instead of smaller networks such as a ResNet-18. Training a much larger teacher network on a very limited clean subset increases the likelihood that the teacher will overfit to its training data, negatively impacting its ability to guide the student network toward robust and generalizable features. The efficacy of compact teacher networks is, therefore, significant for the practical implementation of our approach.

4.2.3 Task-specific teacher fine-tuning outperforms alternative approaches

We propose that a teacher network fine-tuned on a small subset of task-relevant data can provide sufficient insight to deter a student network from learning bias features. Here, we validate this choice. We consider two alternative approaches to our proposed fine-tuned teacher to evaluate the importance of task-specific knowledge in the teacher:

1. **ImageNet pre-trained teacher:** We use a teacher network pre-trained on the ImageNet dataset without any

Table 4: Performance of a ResNet-18 student network tested on the shortcut-corrupted test sets with our distillation loss. Distillation loss is applied at different numbers of intermediate layers between 0 and 17. When loss is applied at 0 intermediate layers, we only apply KD between the student and teacher's final outputs. Results are presented as Mean±Std over 5-fold cross-validation. Models are marked **best** and second-best.

	# layers	CheXpert			ISIC		
		Noise	Square (C)	Square (R)	Noise	Square (C)	Square (R)
AUC ↑	17	**0.694±0.034**	0.742±0.009	0.762±0.008	0.754±0.019	0.761±0.035	0.754±0.023
	13	0.688±0.034	0.746±0.007	0.762±0.005	**0.780±0.018**	0.767±0.012	0.780±0.011
	9	0.687±0.034	0.747±0.008	0.756±0.014	0.768±0.013	0.762±0.033	0.783±0.015
	5	0.689±0.044	**0.747±0.008**	**0.762±0.010**	0.775±0.023	**0.777±0.024**	**0.807±0.016**
	0	0.606±0.008	0.640±0.015	0.684±0.014	0.632±0.019	0.667±0.013	0.713±0.017
ΔTPR ↓	17	**0.272±0.07**	0.083±0.024	**0.028±0.020**	0.091±0.037	0.041±0.038	**0.028±0.02**
	13	0.378±0.188	0.107±0.018	0.049±0.030	0.074±0.034	**0.019±0.021**	0.049±0.030
	9	0.301±0.115	0.083±0.053	0.046±0.023	0.077±0.027	0.054±0.033	0.046±0.023
	5	0.377±0.185	**0.079±0.017**	0.032±0.020	**0.068±0.055**	0.038±0.032	0.032±0.020
	0	0.831±0.091	0.662±0.126	0.424±0.074	0.813±0.139	0.617±0.145	0.424±0.074

Table 5: AUC↑ of a ResNet-18 and DenseNet-121 trained and evaluated on shortcut-corrupted data. We compare student models trained following our knowledge distillation protocol using a low-capacity teacher network (Ours) to models trained following standard cross-entropy optimization (ERM). The best-performing model is in **bold**.

		CheXpert		
		Noise	Square (C)	Square (R)
ResNet-18	Ours	**0.68±0.01**	**0.74±0.02**	**0.75±0.02**
	ERM	0.49±0.01	0.53±0.01	0.55±0.01
DenseNet-121	Ours	**0.63±0.02**	**0.69±0.01**	**0.70±0.02**
	ERM	0.50±0.01	0.52±0.01	0.53±0.01
		ISIC		
		Noise	Square (C)	Square (R)
ResNet-18	Ours	**0.76±0.02**	**0.77±0.03**	**0.77±0.03**
	ERM	0.52±0.01	0.60±0.01	0.61±0.01
DenseNet-121	Ours	**0.72±0.01**	**0.72±0.03**	**0.74±0.03**
	ERM	0.51±0.01	0.63±0.01	0.62±0.01

task-specific fine-tuning. This teacher possesses general visual recognition capabilities from training on diverse natural images but lacks task-specific or domain-specific medical imaging knowledge. Knowledge distillation is performed identically as with our fine-tuned teacher, with KL divergence minimization between corresponding intermediate layers of the student and the pre-trained teacher. This comparison helps us understand whether general visual features from a diverse dataset are sufficient for guiding the student away from shortcuts or if task-specific knowledge is essential.

2. **Confidence Regularization:** The teacher model is removed entirely in favor of a form of self-regularization. Rather than distilling knowledge from a teacher, we encourage the student network to maintain low confidence in its intermediate layer predictions by minimizing the KL divergence between each layer's predictions and a uniform class probability distribution. This forces the model to avoid becoming overconfident in any particular features too early in the network, potentially discouraging reliance on simple shortcut features. By comparing against this approach, we can determine whether the specific guidance from a teacher model provides advantages beyond simply preventing early layer overconfidence.

Both alternatives represent reasonable approaches towards mitigating shortcut learning: the ImageNet teacher by transferring robust general visual representations, and confidence regularization by directly discouraging overconfidence in features at any particular layer. Our fine-tuned specialist teacher consistently outperforms both alternatives across most datasets and shortcut types, as shown in Table 6. This performance is achieved without sacrificing fairness. This highlights that a teacher with task-specific knowledge is better equipped to guide the student away from simplistic shortcut features and toward more robust, task-relevant features. This is further supported by our findings in Figure 7, which shows that our student trained with a fine-tuned teacher network achieves consistently higher AUC in the intermediate layers, and that the AUC of our student reaches higher levels earlier in the model.

4.3 Practical considerations for teacher model training

Finally, we address critical practical questions about teacher model requirements that determine the real-world viability of our approach.

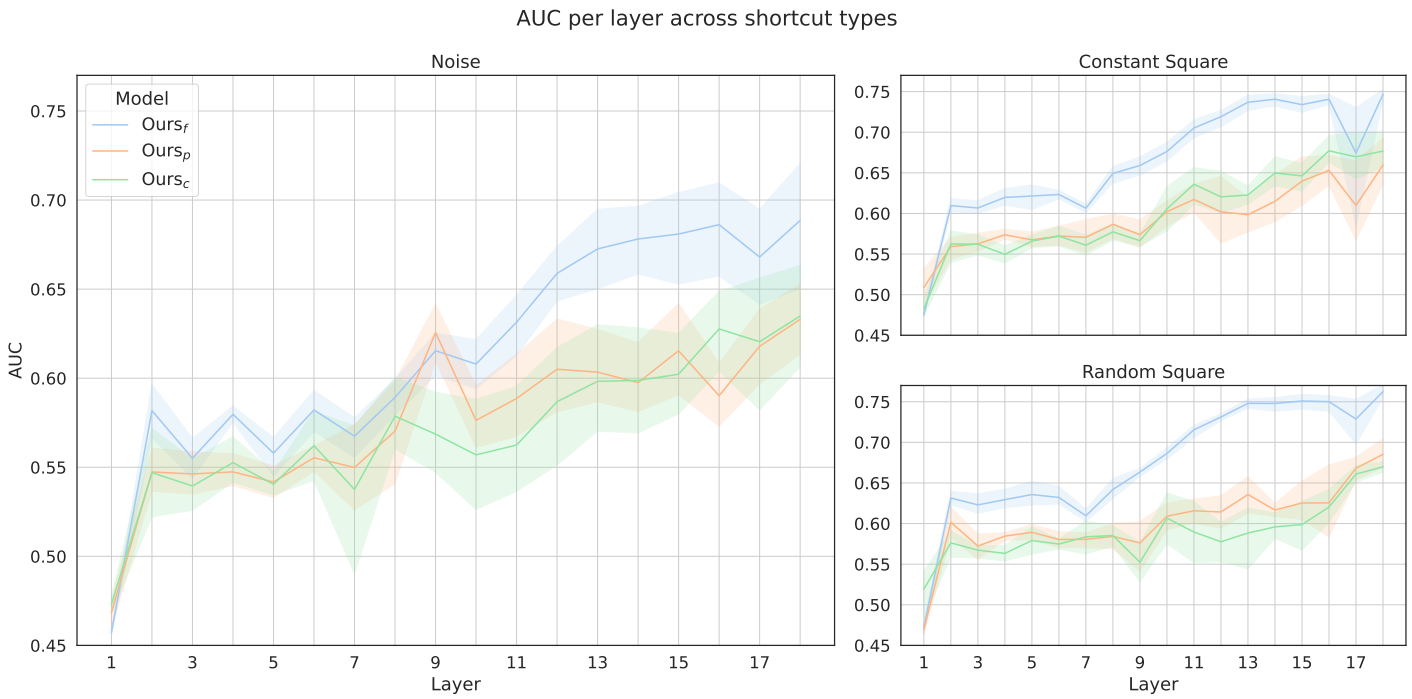

Figure 7: Per-layer AUC of ResNet-18 students trained on CheXpert data featuring various synthetic shortcuts. Ours$_f$ is our fine-tuned teacher model, Ours$_p$ uses an ImageNet-pretrained ResNet-18 as a teacher, and Ours$_c$ applies pure confidence regularization in the intermediate layers.

Table 6: Performance of student models with different knowledge distillation approaches. Ours$_f$ is our fine-tuned teacher model, Ours$_p$ uses an ImageNet-pretrained ResNet-18 as a teacher, and Ours$_c$ applies pure confidence regularization in the intermediate layers. Results are presented as Mean±Std over 5-fold cross-validation. Models are marked as **best** and second-best. When the difference between first and second best is statistically significant, the best-performing model is highlighted with *.

	Model	CheXpert Noise	CheXpert Square (C)	CheXpert Square (R)	ISIC Noise	ISIC Square (C)	ISIC Square (R)
AUC ↑	Ours$_f$	**0.689±0.044**	**0.747±0.008***	**0.763±0.010***	**0.775±0.023**	**0.777±0.024***	**0.807±0.016**
	Ours$_p$	0.633±0.025	0.660±0.039	0.684±0.025	0.727±0.034	0.680±0.029	0.782±0.016
	Ours$_c$	0.635±0.037	0.677±0.029	0.673±0.007	0.753±0.019	0.731±0.023	0.696±0.071
ΔTPR ↓	Ours$_f$	0.377±0.185	**0.079±0.017**	**0.034±0.016**	0.068±0.055	**0.038±0.032**	**0.070±0.043**
	Ours$_p$	**0.285±0.082**	0.239±0.196	0.115±0.094	0.218±0.123	0.318±0.162	0.084±0.057
	Ours$_c$	0.351±0.073	0.109±0.089	0.140±0.062	**0.036±0.033**	0.070±0.067	0.197±0.172

4.3.1 *Teacher effectiveness scales with training data volume*

A critical question for practical implementation is the volume of unbiased data required to train an effective teacher model. In our previous experiments, we evaluated our approach using a teacher network trained on 20% of the full training data from each dataset. To better understand the amount of required data, we now assess the efficacy of our approach when the teacher network is trained on as little as 5% and 10% of the total training data. In each case, the teacher training data is excluded from the student's training. As a result, each student network is trained on 95%, 90%, and 80% of the full training data, depending on the amount of data used to train the teacher.

As illustrated in Figure 8, we observe a clear relationship between teacher training data volume and student performance. Bias metrics and overall model performance both improve consistently as the amount of training data used for the teacher increases. Notably, even when our teacher network is trained on as little as 5% of our original training data (56 images for ISIC), we still observe a substantial reduction in bias compared to ERM training.

The consistent performance advantage observed with minimal unbiased data has significant implications for real-world applications. In clinical settings, where it is often challenging to obtain large amounts of bias-free data, our re-

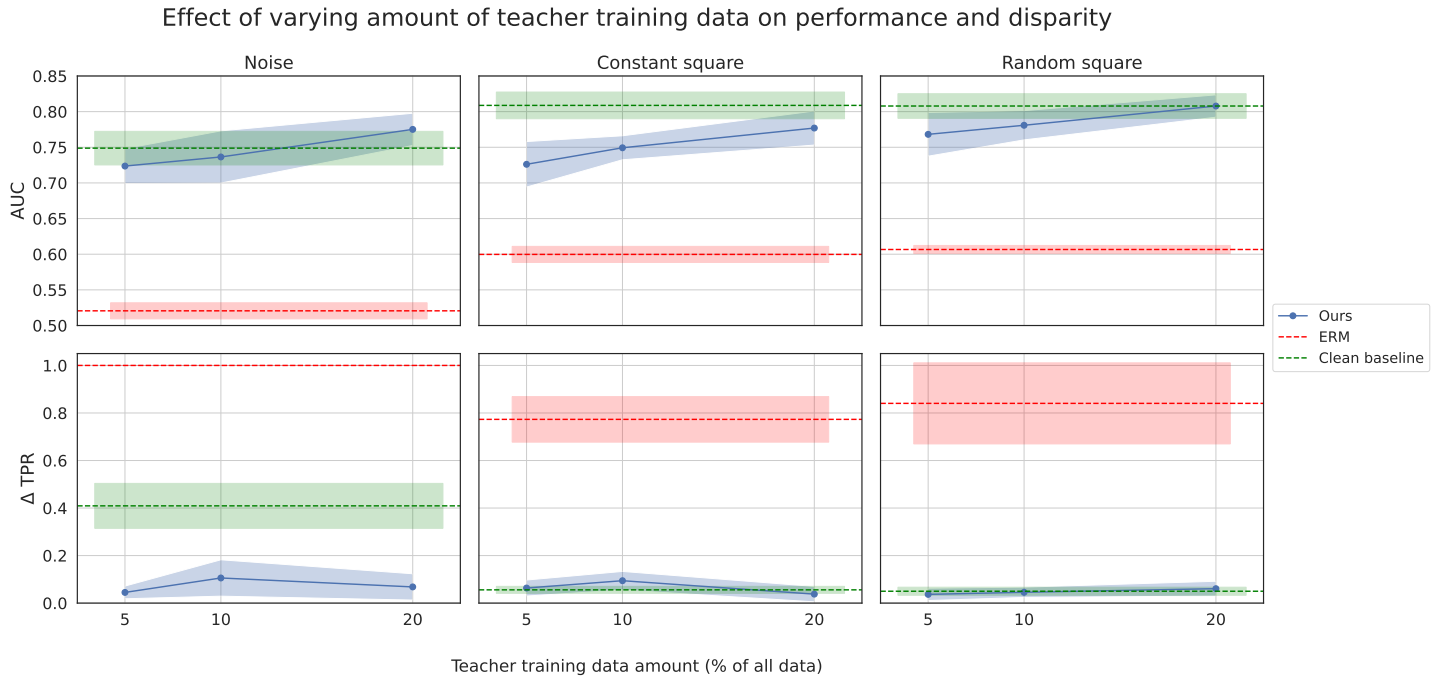

Figure 8: AUC ↑ (top) and ΔTPR ↓ (bottom) of ResNet-18 students trained on ISIC data featuring various synthetic shortcuts. We vary the proportion of the original training data used to train the teacher network, using subsets consisting of 5%, 10%, and 20% of the original training data. In each case, teacher training data is excluded from the student's training. All shortcuts have a 100% prevalence in student training data. As shortcut reliance increases, overall performance (AUC) declines and performance disparity (ΔTPR) increases.

sults indicate that even a small, carefully curated dataset can effectively guide the mitigation of shortcut learning. This finding greatly enhances the practical applicability of our approach, making it more feasible in resource-constrained environments where extensive manual annotation or bias identification could be prohibitively expensive. Although the need to curate an unbiased training set is not completely eliminated, the amount of teacher training data required may be modest enough to be achievable in many practical scenarios.

4.3.2 *Leveraging OOD data for teacher training maintains effectiveness*

In practice, obtaining curated teacher training data from the same distribution as the student's may not always be feasible. We investigate whether teacher models whose training data is OOD from the student's can still effectively guide bias mitigation. For this, we focus on the task of pneumothorax detection, training the teacher on MIMIC while the student is trained on CheXpert: both chest X-ray datasets, but from different institutions.

Figure 9 demonstrates that our approach remains effective when teacher training data is OOD relative to the student. Performance improvements scale with teacher data volume, though OOD teachers require substantially more training data than in-distribution teachers. For example, our teacher trained on 10% of the MIMIC training split is trained on approximately 400 images. By comparison, we see superior performance in a student trained with an in-distribution CheXpert teacher trained on 10% of the CheXpert train split (approximately 140 images). This increased data requirement likely reflects the underlying distribution shift between the datasets and the requirement for the teacher network to have learned robust features that transfer across institutional differences in imaging protocols and patient populations. We also observe that ResNet-18 teachers struggle on OOD test sets when they have been trained on very little data, while ResNet-34 models perform better under the same circumstances. This suggests that the increased model capacity facilitates learning more generalizable features, particularly on smaller training sets.

These findings enhance practical applicability by demonstrating that OOD training data can be used to train the teacher model where it is not possible to curate bias-free in-distribution data. However, when using OOD data to train the teacher it is important to consider to larger data requirements required to achieve comparable performance.

4.3.3 *Robustness to shortcut features in teacher training data*

A fundamental assumption of our work up until this point is the availability of perfectly clean training data for our

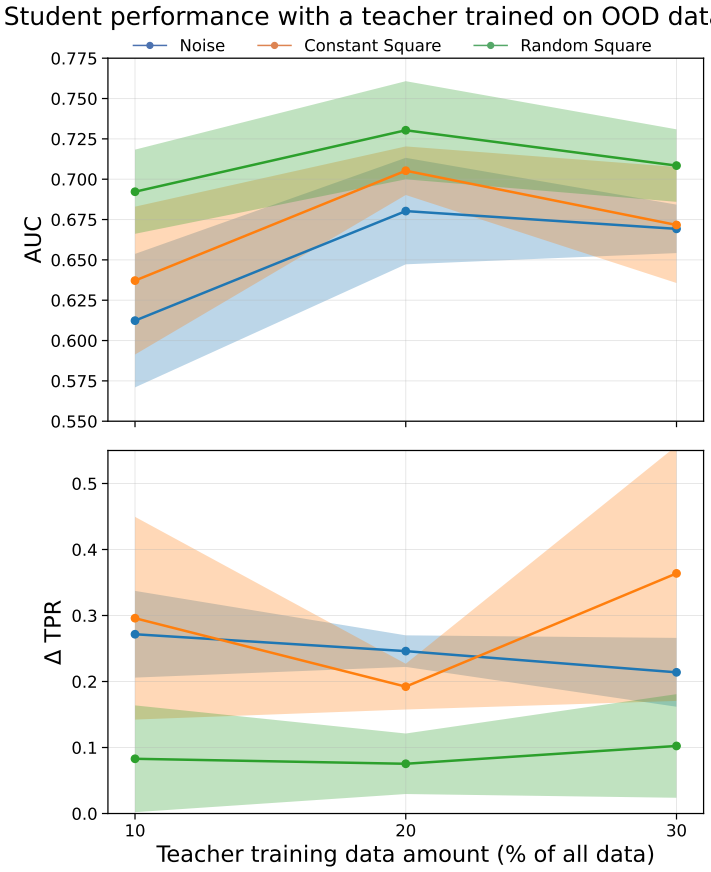

Figure 9: ΔTPR and AUC of a ResNet-34 student trained on CheXpert with various synthetic shortcuts. Teacher model is trained on MIMIC at various subset sizes (between 10% - 30%). The student is trained on the full CheXpert training split with a shortcut prevalence of 95%.

teacher, free of all shortcuts present in the student's training data. However, this assumption may be unrealistic in practice, where subtle biases, such as demographic features or complex acquisition artifacts, can interact in unexpected ways that make the identification and removal of all shortcut features extremely challenging or impossible. To address this limitation, we investigate the robustness of our approach when the teacher's training data contains residual shortcut features at low prevalence rates.

We evaluate scenarios where shortcut features appear in 5%, 10%, and 15% of positive-class samples (and in no negative-class samples) in the teacher's training data, while maintaining much higher prevalence in the student's training data. This simulates realistic conditions where shortcut learning mitigation efforts may not be able to guarantee, even with smaller, manually curated training sets, that the teacher's training data is entirely bias-free. We focus our analysis on the CheXpert dataset, training ResNet-18 models using the same protocol as in Section 4.1.2.

The visual complexity of the shortcut has a material impact on the prevalence at which we begin to observe disparities in performance. Figure 10 illustrates both overall performance (AUC) and disparity (ΔTPR) of a student model as the prevalence of shortcut features in the teacher's training data increases. For complex shortcuts like the random square pattern, even with 15% prevalence in teacher data, both AUC and ΔTPR of the student remain comparable with the clean baseline. In contrast, simpler shortcuts (noise and constant square) show greater sensitivity to teacher data contamination, with noticeable degradation even at a prevalence of 5%. Such findings illustrate that very simple shortcut features significantly influence model learning even at very low prevalence in the training data.

Our findings align with the concept of "availability" introduced by Hermann et al. (2023), who demonstrate that deep learning model's preferentially utilize the most *available* features of their training data (i.e., those which are most easily identifiable), even if they are less *predictive* than more challenging features. The greater availability of our low-level, simpler shortcut features (noise and constant square) compared to the random square shortcut or any disease feature leads the network to rely more heavily on these features, even if they are present in as little as 5% of positive-class samples.

While the curation of bias-free teacher training data remains ideal, where the identification and removal of all possible shortcuts may be impossible or prohibitively time-consuming and costly, teacher dataset curation should focus on identifying and removing the most easily identifiable shortcut features (e.g., treatment devices, hospital logos, obvious markings). Prioritizing these most available features provides the greatest benefit for teacher effectiveness.

Discussion

This paper addresses the critical challenge of shortcut learning in medical image analysis, proposing a novel knowledge distillation method leveraging teacher models fine-tuned on a small amount of unbiased, task-relevant data to guide student models towards robust features of their training data and away from bias features. Our findings highlight several key insights and practical advancements:

First, we demonstrate that shortcut learning manifests as distinct patterns of overconfidence at intermediate network layers, dependent on the type of shortcut involved. Diffuse shortcuts, such as noise patterns, tend to emerge in earlier network layers, suggesting that they do not require significant disambiguation to identify. In contrast, localized shortcuts like geometric shapes manifest in later layers, indicating they require more complex feature disambiguation (Figure 4) .

This layer-specific manifestation has important implications for both shortcut detection and mitigation. The

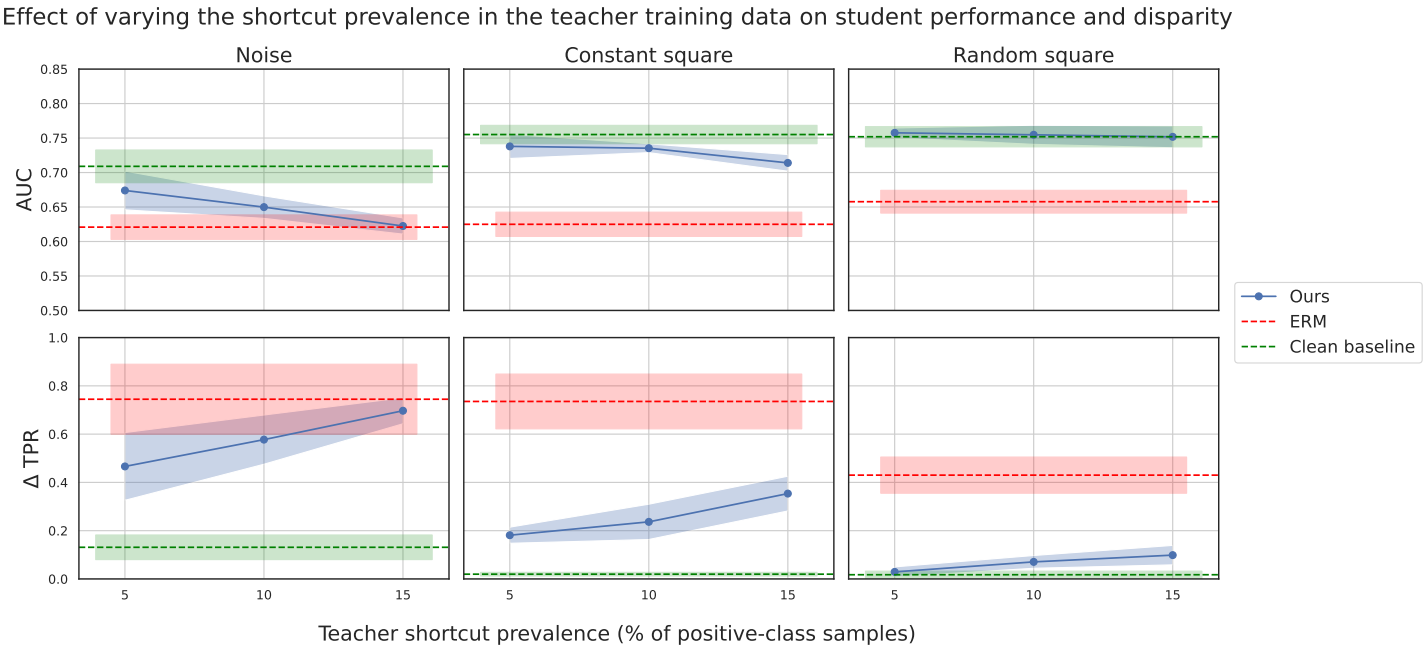

Figure 10: AUC ↑ (top) and ΔTPR ↓ (bottom) of ResNet-18 students trained on CheXpert data featuring various synthetic shortcuts. We vary the prevalence of the shortcut in the data used to train the teacher network. In each case, the teacher is trained on a subset of 20% of the full training split, and the student is trained on the remaining 80%. The shortcut feature has a prevalence of 85% in the student's training data across all experiments. As shortcut reliance increases, overall performance (AUC) declines and performance disparity (ΔTPR) increases.

early appearance of diffuse shortcuts suggests that initial network layers are particularly susceptible to learning simple, texture-related spurious correlations. This aligns with previous findings about the hierarchical nature of neural network learning, where early layers typically learn basic features while deeper layers capture more complex patterns (Baldock et al., 2021; Chen et al., 2020). The observation that different shortcuts manifest at different depths suggests that effective mitigation strategies should consider the network's entire processing pipeline rather than focusing solely on the final classification layer.

This is supported by our finding that distillation from an unbiased teacher to the intermediate layers of a student more effectively mitigates shortcut learning than distillation based solely on the final output (Table 4).

This finding offers a more nuanced understanding of how unwanted correlations manifest within the network's internal representations, and we believe that these insights are valuable beyond the specific method we propose here. For example, such an observation may serve as an effective tool to monitor the learning and performance of deep neural networks to identify when they may be relying on easy spurious features.

A key contribution of our work is demonstrating that knowledge distillation from a teacher network trained on a small curated dataset significantly outperforms traditional de-biasing approaches (Tables 2 & 3). Our method effectively prevents the student network from learning to rely on bias features present in their training data, surpassing traditional empirical risk minimization and alternative approaches such as confidence regularization or using ImageNet-pretrained teachers (Table 6). The approach consistently improves generalization and robustness, evidenced by substantial performance gains on both in-distribution and out-of-distribution test sets for the CheXpert and ISIC datasets (Table 3).

Our results demonstrate that selective intermediate-layer distillation can be as effective as comprehensive distillation across all network layers. As shown in Table 4, distilling knowledge at only 5-9 layers consistently achieved comparable or superior performance to full 17-layer distillation, both in terms of AUC and ΔTPR. This finding suggests that comprehensive distillation across all layers may be unnecessary in most cases and could even add excessive regularization to the student network's learning. While our random layer sampling approach proved effective, it represents a naive strategy that does not consider layer-specific contributions to shortcut learning. Future work should explore principled methods for identifying the layers where distillation would be most impactful. A more targeted distillation approach could further improve the effectiveness of mitigating shortcut learning.

Importantly, we demonstrate that compact architectures, such as AlexNet, can effectively guide larger, more sophisti-

cated networks (ResNet-18 and DenseNet-121), addressing practical constraints related to training high-capacity models on small, unbiased datasets (Table 5). This finding is critical for practical deployment in clinical contexts, where limited availability of unbiased data and computational constraints can limit the use of larger, resource-intensive models.

While our experiments are restricted to CNN-based architectures, transformer architectures are increasingly prevalent in medical image analysis literature. Many KD methods designed for CNNs that leverage the feature-space representations are not directly applicable to transformer networks due to the architectural differences. We suggest that since we do not leverage feature vectors directly, our method could translate to transformer architectures. Recent literature has demonstrated the efficacy of similar KD approaches in transformer architectures, suggesting that it would be possible to apply our framework to transformer architectures (Liu et al., 2024; Wang et al., 2022). However, we suggest that establishing if the distinctive intermediate-layer confidence trajectories that we see in CNN models (Figure 4) is also mirrored in transformer architectures.

The requirement for a clean, curated dataset to train the teacher model presents a potential limitation, though our approach only necessitates a small amount of training data for the teacher network. While such an approach still imposes limitations and necessitates some degree of manual data curation and knowledge of possible sources of bias, the burden of doing so for this much smaller subset is significantly reduced compared to the full training dataset.

One interesting avenue for possible future work would be the use of generative models to create clean, synthetic training data for the teacher model. Additionally, self-supervised or unsupervised techniques for student training may provide a route to remove the need for an unbiased teacher model, addressing one of the primary limitations of this work.

While our synthetic bias features provide a controlled experimental environment, a critical next step is the investigation of the effectiveness of our approach against a broader range of real-world medical image shortcuts, such as those related to patient demographics. This would further validate the practical utility of our method across diverse clinical contexts.

Our work advances both the theoretical understanding and practical mitigation of shortcut learning in medical image analysis. The demonstrated effectiveness of small specialist teachers and selective layer distillation provides a promising direction for developing robust medical AI systems that can generalize across healthcare environments. As these systems become increasingly prevalent in clinical settings, approaches like ours that can effectively prevent shortcut learning while maintaining high performance become crucial for ensuring safe and equitable healthcare delivery.

Acknowledgments

This work was supported by the UKRI EPSRC Centre for Doctoral Training in Applied Photonics [EP/S022821/1].

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

All datasets used in this study are publicly available at the following repositories:

CheXpert: `https://stanfordmlgroup.github.io/competitions/chexpert/`

ISIC: `https://challenge.isic-archive.com/data/#2017`

SimBA: `https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP3/A9SOBZ`

MIMIC: `https://www.physionet.org/content/mimic-cxr-jpg/2.1.0/`

Fitzpatrick17k: `https://github.com/mattgroh/fitzpatrick17k`

The class-balanced subsets used for training the teacher models can be reproduced following the methodology described in Section 3.

References

Kaoutar Ben Ahmed, Lawrence O Hall, Dmitry B Goldgof, and Ryan Fogarty. Achieving multisite generalization for cnn-based disease diagnosis models by mitigating shortcut learning. *IEEE Access*, 10:78726–78738, 2022.

Shuang Ao, Stefan Rueger, and Advaith Siddharthan. Confidence-aware calibration and scoring functions for curriculum learning. In *Fifteenth International Conference on Machine Vision (ICMV 2022)*, volume 12701, pages 558–567. SPIE, 2023.

Robert Baldock, Hartmut Maennel, and Behnam Neyshabur.

Deep learning through the lens of example difficulty. *Advances in Neural Information Processing Systems*, 34: 10876–10889, 2021.

Imon Banerjee, Kamanasish Bhattacharjee, John L Burns, Hari Trivedi, Saptarshi Purkayastha, Laleh Seyyed-Kalantari, Bhavik N Patel, Rakesh Shiradkar, and Judy Gichoya. "shortcuts" causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. *Journal of the American College of Radiology*, 20(9):842–851, 2023.

Pedro RAS Bassi, Andrea Cavalli, and Sergio Decherchi. Explanation is all you need in distillation: Mitigating bias and shortcut learning. *arXiv preprint arXiv:2407.09788*, 2024.

Nourhan Bayasi, Jamil Fayyad, Alceu Bissoto, Ghassan Hamarneh, and Rafeef Garbi. Biaspruner: Debiased continual learning for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 90–101. Springer, 2024.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

Christopher Boland, Owen Anderson, Keith A Goatman, John Hipwell, Sotirios A Tsaftaris, and Sonia Dahdouh. All you need is a guiding hand: Mitigating shortcut bias in deep learning models for medical imaging. In *MICCAI Workshop on Fairness of AI in Medical Imaging*, pages 67–77. Springer, 2024a.

Christopher Boland, Keith A Goatman, Sotirios A Tsaftaris, and Sonia Dahdouh. There are no shortcuts to anywhere worth going: Identifying shortcuts in deep learning models for medical image analysis. In *Medical Imaging with Deep Learning*, 2024b.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *European conference on computer vision*, pages 440–457. Springer, 2022.

Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35: 19152–19164, 2022.

Minshuo Chen, Yu Bai, Jason D Lee, Tuo Zhao, Huan Wang, Caiming Xiong, and Richard Socher. Towards understanding hierarchical learning: Benefits of neural representations. *Advances in Neural Information Processing Systems*, 33:22134–22145, 2020.

Noel C.F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *Proceedings - International Symposium on Biomedical Imaging*, volume 2018-April, pages 168–172. IEEE Computer Society, 2018.

Ramon Correa, Khushbu Pahwa, Bhavik Patel, Celine M Vachon, Judy W Gichoya, and Imon Banerjee. Efficient adversarial debiasing with concept activation vector—medical image case-studies. *Journal of biomedical informatics*, 149:104548, 2024.

Nikolay Dagaev, Brett D Roads, Xiaoliang Luo, Daniel N Barry, Kaustubh R Patil, and Bradley C Love. A too-good-to-be-true prior to reduce shortcut reliance. *Pattern Recognition Letters*, 166:164–171, 2023.

US FDA et al. Marketing submission recommendations for a predetermined change control plan for artificial intelligence. *Machine Learning (AI/ML)-Enabled Device Software Functions*, 2023.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ali Ghadiri, Maurice Pagnucco, and Yang Song. Xtranprune: explainability-aware transformer pruning for bias mitigation in dermatological disease classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 749–758. Springer, 2024.

Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. Risk of bias in chest radiography deep learning

foundation models. *Radiology: Artificial Intelligence*, 5 (6):e230060, 2023.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 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*, pages 1820–1828, 2021.

Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–26, 2022.

Md Akmal Haidar, Nithin Anchuri, Mehdi Rezagholizadeh, Abbas Ghaddar, Philippe Langlais, and Pascal Poupart. Rail-kd: Random intermediate layer mapping for knowledge distillation. *arXiv preprint arXiv:2109.10164*, 2021.

Katherine L Hermann, Hossein Mobahi, Thomas Fel, and Michael C Mozer. On the foundations of shortcut learning. *arXiv preprint arXiv:2310.16228*, 2023.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr database. *PhysioNet10*, 13026:C2JT1Q, 2024.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Patrik Kenfack, Ulrich Aïvodji, and Samira Ebrahimi Kahou. Adaptive group robust ensemble knowledge distillation. *arXiv preprint arXiv:2411.14984*, 2024.

Nicholas Konz and Maciej A Mazurowski. Reverse engineering breast mris: Predicting acquisition parameters directly from images. In *Medical Imaging with Deep Learning*, pages 829–845. PMLR, 2024.

Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9572–9581, 2019.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

Ruiping Liu, Kailun Yang, Alina Roitberg, Jiaming Zhang, Kunyu Peng, Huayao Liu, Yaonan Wang, and Rainer Stiefelhagen. Transkd: Transformer knowledge distillation for efficient semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

Nicolas M Müller, Jochen Jacobs, Jennifer Williams, and Konstantin Böttinger. Localized shortcut removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3721–3725, 2023.

Nihal Murali, Aahlad Manas Puli, Ke Yu, Rajesh Ranganath, et al. Shortcut learning through the lens of early training dynamics. 2022.

Vincent Olesen, Nina Weng, Aasa Feragen, and Eike Petersen. Slicing through bias: Explaining performance gaps in medical image analysis using slice discovery methods. In *MICCAI Workshop on Fairness of AI in Medical Imaging*, pages 3–13. Springer, 2024.

Cathy Ong Ly, Balagopal Unnikrishnan, Tony Tadic, Tirth Patel, Joe Duhamel, Sonja Kandel, Yasbanoo Moayedi, Michael Brudno, Andrew Hope, Heather Ross, et al. Shortcut learning in medical ai hinders generalization: method for estimating ai model generalization without external data. *NPJ Digital Medicine*, 7(1):124, 2024.

Nicholas Petrick, Weijie Chen, Jana G Delfino, Brandon D Gallas, Yanna Kang, Daniel Krainak, Berkman Sahiner, and Ravi K Samala. Regulatory considerations for medical imaging ai/ml devices in the united states: concepts and challenges. *Journal of Medical Imaging*, 10(5):051804–051804, 2023.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9573–9585. Curran Associates, Inc., 2020.

Raissa Souza, Anthony Winder, Emma AM Stanley, Vibujithan Vigneshwaran, Milton Camacho, Richard Camicioli, Oury Monchi, Matthias Wilms, and Nils D Forkert. Identifying biases in a multicenter mri database for parkinson's disease classification: Is the disease classifier a secret site classifier? *IEEE Journal of Biomedical and Health Informatics*, 2024.

Emma AM Stanley, Matthias Wilms, and Nils D Forkert. A flexible framework for simulating and evaluating biases in deep learning-based medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 489–499. Springer, 2023.

Emma AM Stanley, Raissa Souza, Anthony J Winder, Vedant Gulve, Kimberly Amador, Matthias Wilms, and Nils D Forkert. Towards objective and systematic evaluation of bias in artificial intelligence for medical imaging. *Journal of the American Medical Informatics Association*, 31(11):2613–2621, 2024.

Emma AM Stanley, Raissa Souza, Matthias Wilms, and Nils D Forkert. Where, why, and how is bias learned in medical image analysis models? a study of bias encoding within convolutional networks using synthetic data. *EBioMedicine*, 111, 2025.

Abdel Aziz Taha, Leonhard Hennig, and Petr Knoth. Confidence estimation of classification based on the distribution of the neural network output layer. *arXiv preprint arXiv:2210.07745*, 2022.

Huan Tian, Bo Liu, Tianqing Zhu, Wanlei Zhou, and S Yu Philip. Distilling fair representations from fair teachers. *IEEE Transactions on Big Data*, 2024.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance. *arXiv preprint arXiv:2005.00315*, 2020.

Jiahao Wang, Mingdeng Cao, Shuwei Shi, Baoyuan Wu, and Yujiu Yang. Attention probe: Vision transformer distillation in the wild. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2220–2224. IEEE, 2022.

Ryan Wang, Po-Chih Kuo, Li-Ching Chen, Kenneth Patrick Seastedt, Judy Wawira Gichoya, and Leo Anthony Celi. Drop the shortcuts: image augmentation improves fairness and decreases ai detection of race and other demographics from medical images. *EBioMedicine*, 102, 2024.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. In *International Conference on Machine Learning*, pages 37765–37786. PMLR, 2023.

Yawen Wu, Dewen Zeng, Xiaowei Xu, Yiyu Shi, and Jingtong Hu. Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 743–753. Springer, 2022.

Yuyang Xue, Junyu Yan, Raman Dutt, Fasih Haider, Jingshuai Liu, Steven McDonagh, and Sotirios A Tsaftaris. Bmft: Achieving fairness via bias-based weight masking fine-tuning. In *MICCAI Workshop on Fairness of AI in Medical Imaging*, pages 98–108. Springer, 2024.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the fairness of chest x-ray classifiers. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 204–233. PMLR, 07–08 Apr 2022.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

Appendix A. Shortcut reliance results in intermediate-layer overconfidence

This section provides comprehensive layer-wise confidence analysis extending our main findings from Section 4.1.1 to additional architectures and datasets, demonstrating the generalizability of our core observation that shortcut learning manifests distinctly across network layers across architectures and datasets.

Figures 11 and 12 present intermediate layer confidence for ResNet-18 and DenseNet-121 architectures. In our main experiments, we use validation sets where shortcuts are balanced across classes (present equally in both positive and negative samples). Under these conditions, we observe that the overconfidence signal in DenseNet-121 is noticeably weaker than in ResNet-18. We hypothesize that the substantially larger capacity of DenseNet-121, combined with early stopping on validation data where shortcuts no longer correlate with class labels, prevents the network from fully developing shortcut dependencies.

To examine how validation set composition affects these patterns, Figures 13 and 14 show the same architectures trained with validation sets that maintain the same shortcut-class correlations as the training data. Under these conditions, the overconfidence signal becomes much more pronounced even in high-capacity models like DenseNet-121, as the validation setup no longer provides feedback that discourages shortcut reliance during training.

Appendix B. Overall performance

This section provides detailed performance breakdowns, including AUC values and True Positive Rate (TPR) analysis for bias-aligned and bias-contrasting samples, supplementing the ΔTPR analysis presented in Section 4.1.2 of the main text.

Table 7 presents AUC values across all experimental conditions, showing that our method consistently achieves performance competitive with the clean baseline even when trained on heavily biased data. Notably, our approach maintains stable performance across varying bias prevalence rates, while competing methods show significant degradation at higher bias levels.

Tables 8 and 9 provide absolute TPR values for CheXpert and ISIC datasets respectively, breaking down performance for bias-aligned samples (where shortcut presence matches training correlation with the class-label) and bias-contrasting samples (where shortcuts oppose training correlation with the class-label). We demonstrate that our method achieves more balanced performance across both sample types, indicating reduced reliance on shortcut features for prediction. We note that our method demonstrates lower TPR for bias-aligned samples compared to other methods, and that the better TPR seen in other models is likely a result of shortcut reliance, causing a TPR much higher than the clean baseline. We consistently observe that our method achieves a TPR for biased-aligned samples that is closest to the clean baseline.

Appendix C. Teacher sensitivity to shortcut features

Here, we highlight the sensitivity of our teacher networks to corruption from shortcut features. Figure 15 supplements our findings in Section 4.3.3 and supports our argument that practitioners should prioritize removing the most available (easily identifiable) shortcuts from the teacher training data as these cause the most significant harm.

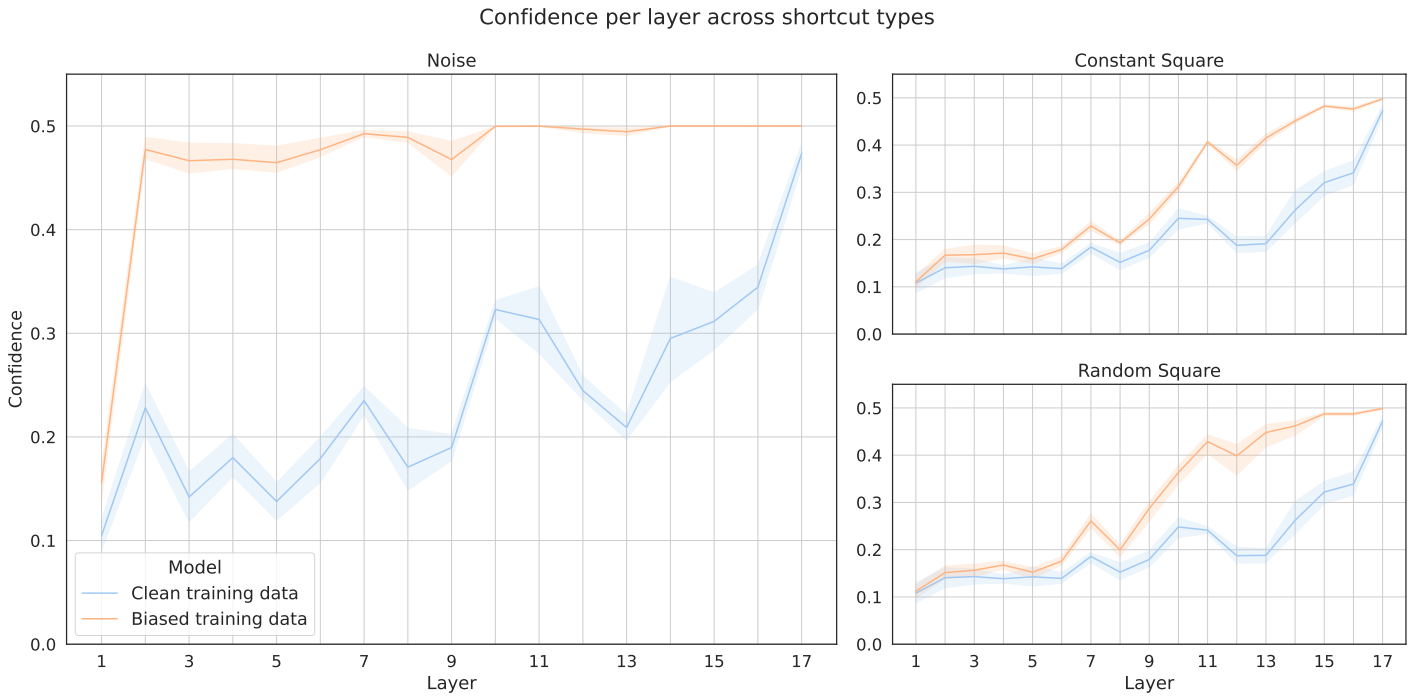

Figure 11: Per-layer confidence of ResNet-18 students trained on ISIC data featuring various synthetic shortcuts.

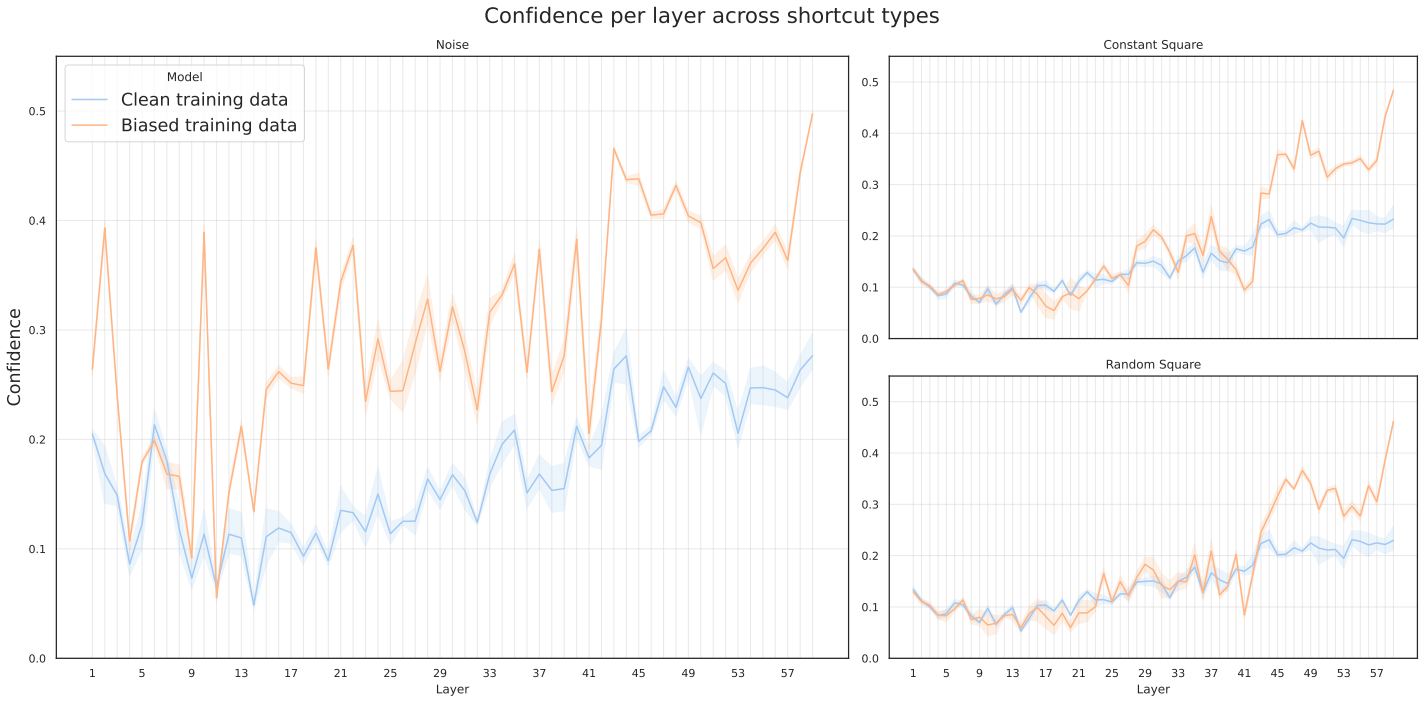

Figure 12: Per-layer confidence of DenseNet-121 students trained on CheXpert data featuring various synthetic shortcuts. Shortcut prevalence in the train split is 100%. In the validation and test split, the shortcut feature is balanced between classes.

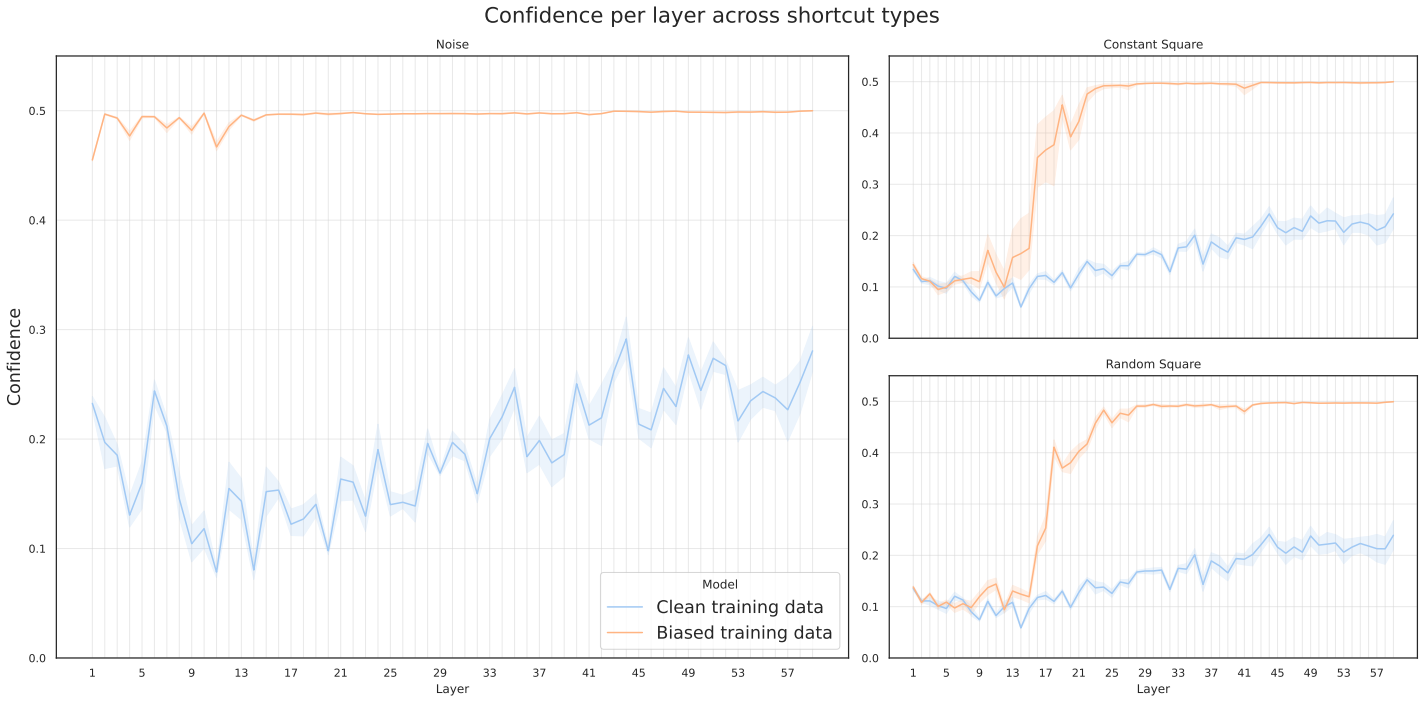

Figure 13: Per-layer confidence of DenseNet-121 students trained on CheXpert data featuring various synthetic shortcuts. Shortcut prevalence in the train and validation splits is 100%. In the test split, the shortcut feature is balanced between classes.

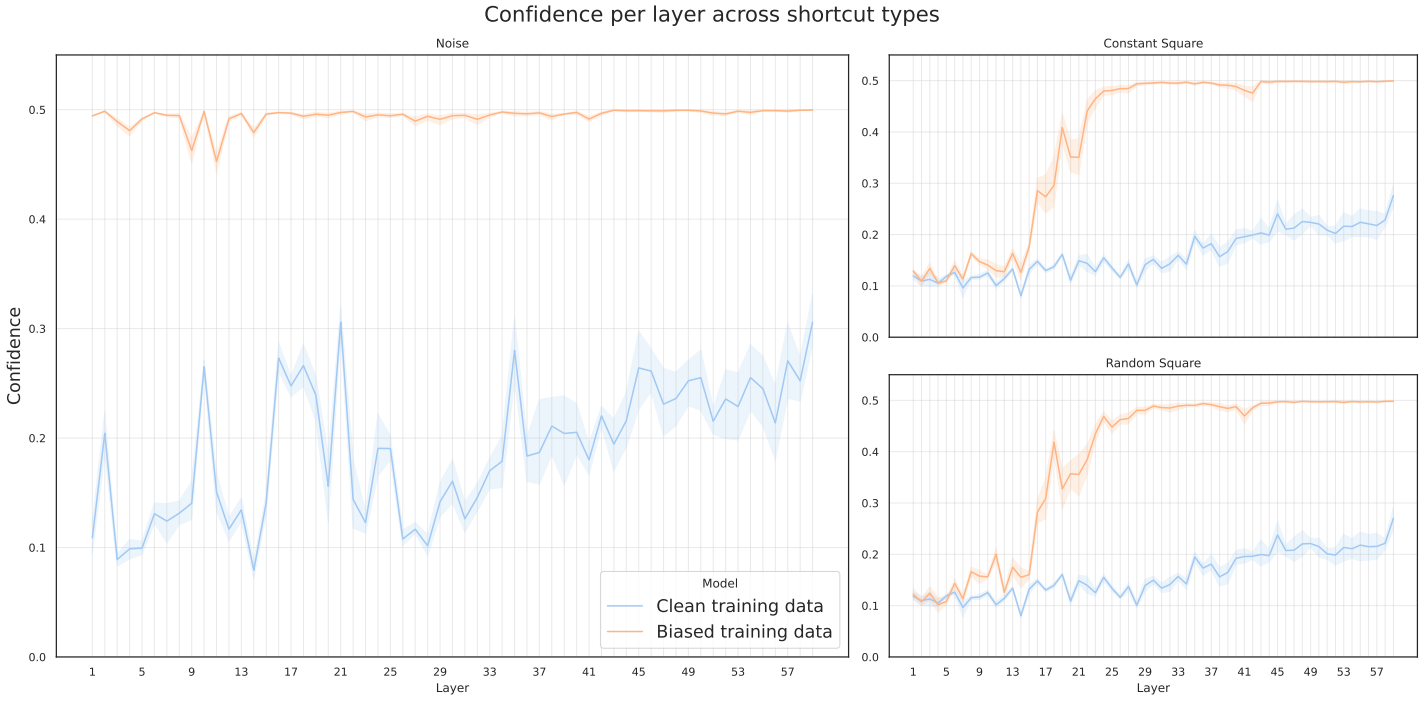

Figure 14: Per-layer confidence of DenseNet-121 students trained on ISIC data featuring various synthetic shortcuts. Shortcut prevalence in the train and validation splits is 100%. In the test split, the shortcut feature is balanced between classes.

Table 7: $AUC \uparrow$ of a ResNet-18 trained on data with various bias prevalence rates. Results are presented as Mean±Std over 5-fold cross-validation. Models are marked as **best** and second-best.

Prev. (%)	Model	CheXpert			ISIC		
		Noise	Square (C)	Square (R)	Noise	Square (C)	Square (R)
0	Baseline	0.709±0.024	0.755±0.013	0.752±0.013	0.749±0.024	0.809±0.019	0.808±0.019
100	ERM	0.489±0.012	0.533±0.007	0.554±0.005	0.521±0.010	0.600±0.011	0.612±0.007
	MixUp	0.509±0.007	0.539±0.008	0.549±0.005	0.489±0.012	0.555±0.028	0.591±0.014
	CutOut	0.498±0.029	0.548±0.010	0.583±0.011	0.521±0.023	0.627±0.013	0.639±0.014
	CutMix	0.529±0.007	0.680±0.025	0.759±0.011	0.516±0.015	0.753±0.038	0.784±0.017
	Aug	0.483±0.009	0.585±0.017	0.555±0.010	0.518±0.006	0.731±0.017	0.626±0.014
	Ours	**0.689±0.044**	**0.747±0.008**	**0.761±0.010**	**0.775±0.023**	**0.777±0.024**	**0.806±0.016**
95	ERM	0.579±0.009	0.597±0.007	0.595±0.015	0.602±0.032	0.645±0.024	0.651±0.028
	MixUp	0.559±0.008	0.578±0.008	0.603±0.024	0.542±0.043	0.609±0.021	0.603±0.024
	CutOut	0.568±0.013	0.599±0.013	0.611±0.040	0.608±0.024	0.641±0.011	0.611±0.040
	CutMix	0.567±0.020	0.739±0.024	**0.760±0.020**	0.587±0.029	0.768±0.031	0.760±0.020
	Aug	0.578±0.014	0.628±0.028	0.597±0.015	0.608±0.026	0.745±0.010	0.633±0.013
	GDRO	0.566±0.008	0.585±0.010	0.613±0.014	0.595±0.012	0.631±0.008	0.673±0.015
	JTT	0.565±0.007	0.588±0.004	0.602±0.018	0.591±0.026	0.630±0.015	0.672±0.016
	Ours	**0.693±0.029**	**0.756±0.005**	0.756±0.011	**0.778±0.020**	**0.797±0.015**	**0.803±0.009**
85	ERM	0.621±0.018	0.625±0.018	0.661±0.013	0.694±0.018	0.751±0.016	0.748±0.030
	MixUp	0.598±0.017	0.630±0.030	0.697±0.021	0.664±0.034	0.687±0.019	0.706±0.052
	CutOut	0.596±0.018	0.639±0.026	0.683±0.018	0.663±0.035	0.731±0.040	0.739±0.029
	CutMix	0.604±0.015	0.742±0.049	**0.766±0.016**	0.634±0.015	0.766±0.039	**0.796±0.019**
	Aug	0.636±0.034	0.677±0.040	0.661±0.035	0.667±0.039	0.768±0.018	0.694±0.041
	GDRO	0.616±0.012	0.649±0.015	0.725±0.015	0.655±0.027	0.732±0.022	0.780±0.018
	JTT	**0.780±0.018**	0.659±0.008	0.699±0.010	0.656±0.025	0.721±0.022	0.771±0.018
	Ours	0.708±0.037	**0.756±0.008**	0.760±0.015	**0.798±0.024**	**0.782±0.031**	0.787±0.026
75	ERM	0.681±0.023	0.686±0.038	0.718±0.015	0.711±0.028	0.763±0.013	0.784±0.013
	MixUp	0.677±0.025	0.705±0.030	0.746±0.026	0.673±0.018	0.720±0.021	0.757±0.016
	CutOut	0.675±0.027	0.707±0.026	0.744±0.011	0.715±0.034	0.726±0.020	0.757±0.026
	CutMix	0.662±0.029	0.748±0.027	0.759±0.031	0.696±0.025	0.768±0.030	0.781±0.029
	Aug	0.676±0.026	0.712±0.041	0.701±0.041	0.722±0.029	0.766±0.022	0.731±0.035
	GDRO	0.692±0.010	0.731±0.012	**0.768±0.010**	0.696±0.013	0.770±0.008	**0.796±0.008**
	JTT	0.687±0.020	0.730±0.017	0.751±0.026	0.694±0.009	0.769±0.011	0.794±0.012
	Ours	**0.709±0.039**	**0.760±0.015**	0.762±0.017	**0.792±0.023**	**0.773±0.013**	0.781±0.029

Table 8: $TPR \uparrow$ of bias-aligned and bias-contrasting samples for a ResNet-18 trained on CheXpert data with various bias prevalence rates. Results are presented as Mean±Std over 5-fold cross-validation. Models are marked as **best** and second-best. When multiple models achieve identical performance, all are highlighted.

Prev. (%)	Model	Noise		Square (C)		Square (R)	
		Bias-Aligned	Bias-Contrasting	Bias-Aligned	Bias-Contrasting	Bias-Aligned	Bias-Contrasting
0	Baseline	0.942±0.036	0.811±0.069	0.804±0.080	0.811±0.069	0.822±0.069	0.811±0.069
100	ERM	**1.000±0.000**	0.000±0.000	**1.000±0.000**	0.000±0.000	**0.995±0.005**	0.004±0.007
	MixUp	**1.000±0.000**	0.013±0.026	0.998±0.003	0.000±0.000	0.974±0.018	0.004±0.007
	CutOut	**1.000±0.000**	0.001±0.001	**1.000±0.000**	0.000±0.000	0.982±0.011	0.004±0.009
	CutMix	**1.000±0.000**	0.007±0.005	0.929±0.045	0.427±0.114	0.894±0.014	0.765±0.015
	Aug	0.957±0.092	0.000±0.000	0.996±0.004	0.017±0.015	0.982±0.011	0.004±0.009
	Ours	0.942±0.033	**0.565±0.161**	0.862±0.005	**0.783±0.017**	0.839±0.063	**0.805±0.050**
95	ERM	0.995±0.010	0.055±0.022	0.993±0.011	0.081±0.088	0.973±0.014	0.182±0.105
	MixUp	0.997±0.006	0.070±0.107	0.992±0.007	0.005±0.010	0.948±0.015	0.253±0.100
	CutOut	0.997±0.004	0.047±0.051	0.987±0.013	0.074±0.061	0.959±0.007	0.330±0.127
	CutMix	0.999±0.001	0.025±0.029	0.908±0.022	0.583±0.047	0.845±0.029	0.709±0.070
	Aug	0.997±0.003	0.098±0.075	0.962±0.039	0.183±0.132	0.978±0.014	0.159±0.079
	JTT	**1.000±0.000**	0.018±0.011	**0.999±0.001**	0.054±0.031	0.956±0.080	0.283±0.041
	GDRO	**1.000±0.000**	0.014±0.010	0.997±0.002	0.019±0.014	**0.981±0.011**	0.279±0.096
	Ours	0.935±0.037	**0.563±0.092**	0.862±0.031	**0.773±0.019**	0.810±0.051	**0.775±0.021**
85	ERM	0.960±0.039	0.215±0.167	0.974±0.013	0.239±0.115	0.926±0.047	0.503±0.106
	MixUp	0.985±0.011	0.286±0.119	0.962±0.035	0.284±0.117	0.903±0.022	0.566±0.031
	CutOut	**0.988±0.009**	0.178±0.072	0.977±0.014	0.320±0.204	**0.938±0.029**	0.557±0.127
	CutMix	0.979±0.021	0.246±0.106	0.886±0.039	0.675±0.050	0.835±0.036	0.737±0.037
	Aug	0.957±0.039	0.292±0.208	0.954±0.007	0.437±0.197	0.927±0.030	0.397±0.144
	JTT	0.987±0.005	0.268±0.034	0.968±0.008	0.435±0.030	0.933±0.029	0.555±0.144
	GDRO	0.983±0.006	0.224±0.033	**0.977±0.011**	0.353±0.057	0.922±0.012	0.654±0.071
	Ours	0.929±0.042	**0.581±0.119**	0.861±0.060	**0.755±0.030**	0.860±0.026	**0.801±0.044**
75	ERM	0.879±0.099	0.434±0.193	0.903±0.125	0.516±0.077	**0.910±0.025**	0.708±0.078
	MixUp	0.944±0.017	**0.564±0.130**	0.888±0.066	0.557±0.088	0.849±0.044	0.692±0.032
	CutOut	0.941±0.043	0.481±0.087	0.855±0.119	0.494±0.187	0.897±0.021	0.729±0.063
	CutMix	0.926±0.045	0.399±0.085	0.852±0.072	0.714±0.091	0.819±0.028	0.763±0.012
	Aug	0.909±0.051	0.463±0.171	0.878±0.059	0.470±0.205	0.849±0.032	0.505±0.071
	JTT	0.949±0.023	0.497±0.052	**0.928±0.025**	0.667±0.036	0.871±0.028	0.725±0.060
	GDRO	**0.951±0.004**	0.444±0.015	0.904±0.023	0.575±0.022	0.866±0.022	0.752±0.019
	Ours	0.926±0.040	0.561±0.162	0.878±0.029	**0.751±0.042**	0.854±0.044	**0.793±0.056**

Table 9: $TPR \uparrow$ of bias-aligned and bias-contrasting samples for a ResNet-18 trained on ISIC data with various bias prevalence rates. Results are presented as Mean±Std over 5-fold cross-validation. Models are marked as **best** and second-best. When multiple models achieve identical performance, all are highlighted.

Prev. (%)	Model	Noise		Square (C)		Square (R)	
		Bias-Aligned	Bias-Contrasting	Bias-Aligned	Bias-Contrasting	Bias-Aligned	Bias-Contrasting
0	Baseline	0.449±0.113	0.858±0.027	0.802±0.030	0.858±0.027	0.808±0.036	0.858±0.027
100	ERM	**1.000±0.000**	0.000±0.000	**1.000±0.000**	0.223±0.093	**1.000±0.000**	0.156±0.168
	MixUp	0.998±0.004	0.000±0.000	**1.000±0.000**	0.127±0.150	0.990±0.012	0.238±0.146
	CutOut	**1.000±0.000**	0.000±0.000	**1.000±0.000**	0.552±0.065	0.986±0.016	0.712±0.066
	CutMix	**1.000±0.000**	0.000±0.000	0.905±0.032	0.546±0.083	0.856±0.021	0.729±0.050
	Aug	**1.000±0.000**	0.000±0.000	0.961±0.019	0.800±0.062	0.971±0.023	0.467±0.136
	Ours	0.792±0.066	**0.771±0.09**	0.829±0.044	**0.852±0.028**	0.829±0.042	**0.900±0.016**
95	ERM	**1.000±0.000**	0.041±0.019	0.993±0.010	0.132±0.026	0.987±0.020	0.285±0.097
	MixUp	0.973±0.047	0.021±0.019	**0.998±0.004**	0.062±0.035	0.984±0.020	0.229±0.081
	CutOut	**1.000±0.000**	0.054±0.047	0.985±0.023	0.406±0.263	0.973±0.027	0.544±0.144
	CutMix	0.998±0.004	0.076±0.079	0.909±0.017	0.598±0.098	0.880±0.024	0.757±0.050
	Aug	0.996±0.008	0.062±0.059	0.965±0.015	0.761±0.079	0.985±0.014	0.429±0.126
	GDRO	0.996±0.005	0.029±0.017	0.987±0.005	0.188±0.035	0.980±0.010	0.503±0.024
	JTT	**1.000±0.000**	0.054±0.031	0.991±0.011	0.198±0.049	**0.987±0.014**	0.482±0.070
	Ours	0.858±0.042	**0.781±0.062**	0.898±0.024	**0.798±0.035**	0.873±0.031	**0.821±0.036**
85	ERM	0.958±0.008	0.685±0.034	0.931±0.020	0.555±0.104	**0.954±0.011**	0.660±0.083
	MixUp	0.952±0.016	0.398±0.159	0.939±0.031	0.449±0.100	0.905±0.043	0.511±0.084
	CutOut	0.952±0.024	0.442±0.215	**0.954±0.022**	0.483±0.155	0.933±0.015	0.683±0.130
	CutMix	0.970±0.025	0.317±0.113	0.895±0.033	0.681±0.074	0.871±0.025	0.785±0.040
	Aug	0.950±0.029	0.274±0.185	0.883±0.056	0.768±0.066	0.913±0.051	0.594±0.123
	GDRO	**0.980±0.010**	0.283±0.083	0.945±0.034	0.609±0.052	0.903±0.039	0.806±0.017
	JTT	0.978±0.011	0.275±0.119	0.950±0.019	0.600±0.058	0.911±0.029	0.753±0.037
	Ours	0.855±0.033	**0.792±0.100**	0.863±0.047	**0.817±0.032**	0.861±0.031	**0.811±0.086**
75	ERM	0.923±0.022	0.670±0.048	0.829±0.041	0.616±0.051	0.894±0.019	0.721±0.083
	MixUp	**0.948±0.019**	0.477±0.168	0.927±0.035	0.645±0.126	**0.902±0.017**	0.732±0.040
	CutOut	0.933±0.012	0.668±0.189	**0.956±0.014**	0.816±0.047	0.896±0.041	0.724±0.136
	CutMix	0.910±0.040	0.580±0.098	0.867±0.029	0.714±0.056	0.863±0.026	0.793±0.056
	Aug	0.915±0.034	0.591±0.155	0.863±0.047	0.714±0.078	0.894±0.048	0.636±0.118
	GDRO	0.942±0.019	0.497±0.051	0.912±0.012	**0.730±0.021**	0.892±0.009	**0.834±0.018**
	JTT	0.940±0.020	0.481±0.065	0.923±0.012	0.719±0.027	0.875±0.016	0.789±0.036
	Ours	0.829±0.041	**0.840±0.060**	0.875±0.035	0.730±0.099	0.869±0.041	0.805±0.032

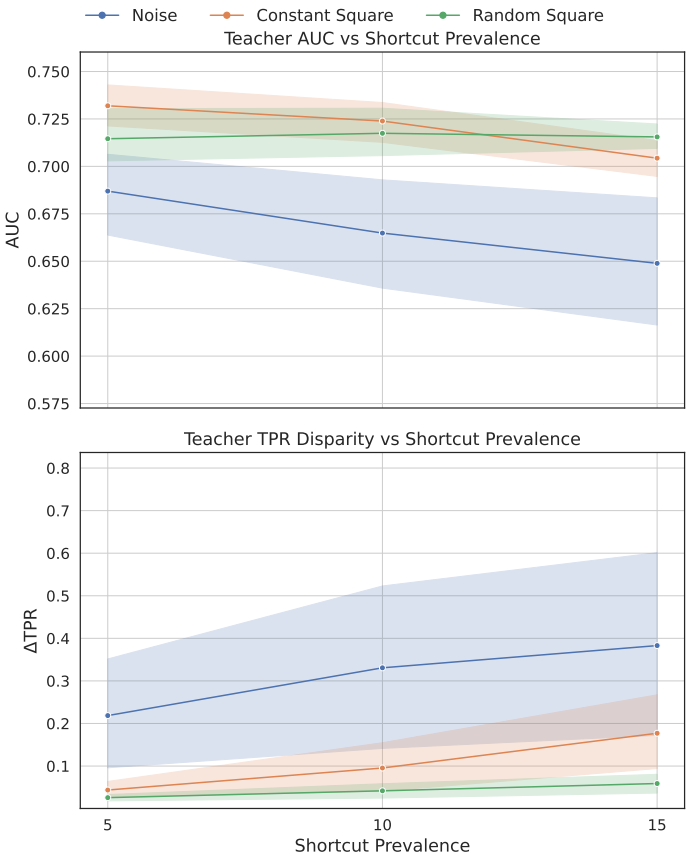

Figure 15: AUC and ΔTPR of ResNet-18 teacher networks trained on 20% subsets of the CheXpert, corrupted with shortcut features at various prevalence. All test sets feature the same shortcut feature as is present in the train split, evenly distributed between samples belonging to each class, and therefore is no longer a useful predictive feature.