PREPRINT of:
J. Blunk, P. Bodesheim, J. Denzler: Adaptive Model Selection for Expanded Post Hoc Debiasing and Mitigating Varying Degrees of Spurious Correlations. International Conference on Computer Analysis of Images and Patterns (CAIP), pages 101–111, 2025.

Adaptive Model Selection for Expanded Post Hoc Debiasing and Mitigating Varying Degrees of Spurious Correlations

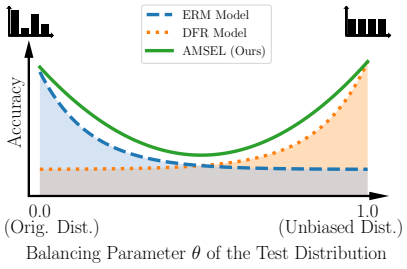
Jan Blunk[0009−0001−7037−3545], Paul Bodesheim[0000−0002−3564−6528], and Joachim Denzler[0000−0002−3193−3300]

Computer Vision Group, Friedrich Schiller University Jena, Germany
{jan.blunk,paul.bodesheim,joachim.denzler}@uni-jena.de
https://inf-cv.uni-jena.de

Abstract. Deep neural networks are prone to shortcut bias, where models rely on features that are statistically associated with the target label but lack causal relevance, leading to poor generalization under distribution shifts. To address this, debiasing methods aim to improve robustness by reducing reliance on these spurious features. Unfortunately, existing approaches typically assume unbiased test distributions, an idealized scenario that rarely holds in practice. As a result, they often underperform on the original biased distribution when compared with standard empirical risk minimization (ERM) models. We propose a novel *Adaptive Model SELection* approach for expanding post hoc debiasing called *AMSEL*, which maintains strong performance across test distributions with varying strength of spurious correlation. Using the fixed feature extractor of the biased model, AMSEL trains a family of lightweight classifier heads on simulated distributions ranging from the original biased data to a fully balanced version. At test time, it estimates the degree of spurious correlation in the test data and selects the most suitable classifier. We validate AMSEL on CelebA and ChestX-ray14, demonstrating that it matches the performance of debiased models under unbiased conditions while preserving the accuracy of the original biased model when spurious correlations are prevalent. AMSEL thus offers an adaptive solution to mitigate the impact of spurious correlations when their strength is either unknown or varies across application environments. Code and models are publicly available at https://github.com/debiasing/AMSEL.

Keywords: Post Hoc Debiasing · Shortcut Bias · Spurious Correlations · Dynamic Classifier Selection

1 Introduction



Model	Optimized For: Biased Data	Optimized For: Unbiased Data
Loss Minimization (ERM Model)	✓	✗
Classical Debiasing (DFR Model)	✗	✓
Our Approach (AMSEL)	✓	✓

Fig. 1: **Expanding Post Hoc Debiasing by Adaptive Model SELection (AMSEL):** The ERM model, trained on biased data, performs well in biased settings, whereas classical debiasing models are optimized for unbiased data and thus perform best in unbiased settings. Our method **AMSEL** bridges this gap by (i) estimating test distribution characteristics and (ii) adaptively selecting the best classifier head, achieving strong performance across the entire spectrum.

Deep neural networks often struggle to generalize when facing shifts in test distributions [10, 36]. One reason for this is *shortcut bias*, where the model relies on statistical associations present in the training data that are not causally warranted [10]. For instance, in natural image classification, models may erroneously associate background textures with certain classes rather than focusing on intrinsic object properties [3, 28, 34] while in medical imaging, they can rely on hospital-specific artifacts rather than clinically relevant features [7, 14, 37]. Consequently, when these *spurious correlations* no longer hold, the shortcut bias leads to significant performance drops.

To address this, a variety of *debiasing* methods have been proposed [2, 12, 16, 18, 21, 25, 28, 36], with some even avoiding a complete retraining. One such post hoc approach is *Deep Feature Reweighting (DFR)* [18], which observes that models extract both meaningful and spurious features when trained via standard loss minimization for empirical risk minimization (ERM) [31]. The authors of DFR thus argue that the classifier disproportionately relies on the spurious features and propose freezing the feature extractor while retraining the classifier head on unbiased data. Despite its simplicity and low computational overhead, DFR substantially improves performance on unbiased data [17, 18].

However, classical debiasing methods like DFR tend to underperform on the original biased distribution compared to the ERM model, visualized in Figure 1. This drawback stems from optimizing for an *unbiased* test distribution. This assumption is particularly strong given biased training data, as it requires test data to be free of the spurious correlations present in the training set. In real-world applications, assuming a *fixed* level of spurious correlation is further problematic,

as test data often exhibits varying levels of spurious correlation due to changes in acquisition conditions or site-specific characteristics [3, 7, 14].

We reject the assumption of a fixed test distribution and instead frame the problem of post hoc debiasing as dynamic classifier selection [1, 5, 6]. Specifically, we assume a continuum of candidate test distributions, each characterized by a different strength of spurious correlation, ranging from the original biased setting to an ideal, unbiased scenario. Building on the observations of DFR, our method trains a family of classifier heads, each optimized for a different candidate test distribution while keeping the feature extractor fixed. At test time, we select the classifier best suited to the actual conditions by comparing the predictions of the classifiers trained on the original biased data and unbiased data. Crucially, we assume that they primarily agree on simple, bias-aligned samples, allowing us to select the head optimized for the corresponding level of spurious correlation.

We validate AMSEL on CelebA [20] and ChestX-ray14 [33], a real-world medical imaging dataset, and find that it matches DFR's benefits on unbiased data while avoiding the performance degradation observed with conventional debiasing methods on biased data.

Our contributions are twofold: (i) We provide a framework for modeling varying test environments by training a family of classifier heads optimized for different levels of spurious correlation. (ii) We propose a test-time mechanism to select the estimated best classifier based on the degree of bias in the input data. Finally, we provide empirical evidence that our approach robustly adapts to varying test conditions, bridging the gap between the original level of spurious correlation and scenarios with minimal spurious correlation.

2 Related Work

Real-world datasets often exhibit *spurious correlations*, which are features that are statistically predictive in the training data but lack a causal link to the task [22, 36]. Models trained via empirical risk minimization (ERM) [31] tend to exploit these correlations, in part due to the inherent *simplicity bias* of deep networks [29, 30]. When simplicity bias aligns with spurious features, models develop a *shortcut bias* [10], also sometimes referred to as *Clever Hans phenomenon* [19]. Instead of relying on task-relevant features, they focus on these simple yet irrelevant attributes, leading to poor generalization when the spurious correlations do not persist at test time.

To mitigate shortcut bias, various debiasing strategies have been proposed, often requiring retraining of the entire model. These approaches are categorized by the availability of *subgroup labels* that provide information about spurious attributes [35]. Approaches leveraging subgroup labels involving full-model retraining include reweighting and resampling techniques [2, 16] as well as worst-case loss minimization frameworks such as GroupDRO [28]. For a comprehensive comparison of debiasing techniques, we refer the reader to [35, 36].

In contrast, *post hoc* debiasing techniques modify a pre-trained ERM model with minimal adjustments [12, 18, 25, 26]. For instance, Deep Feature Reweight-

ing (DFR) [17,18] retrains only the classifier on a group-balanced validation set to recover robust performance on unbiased test distributions. This approach is motivated by the observation that while ERM-trained feature extractors encode both biased and unbiased information, the classifier tends to over-rely on spurious features [18,26]. Similarly, He et al. [12] directly adjust classifier weights to suppress the influence of spurious correlations.

Unlike these methods, which optimize for a fixed, unbiased test setting, our approach considers a spectrum of test distributions characterized by varying strengths of spurious correlations. This leads us to draw connections to *dynamic classifier selection (DCS)* [1, 5, 6], where the classifier is chosen based on the input data. However, unlike previous work in DCS focusing on feature space competence estimation and ensemble generation, our approach is tailored for debiasing. We constrain test distributions to a spectrum between the original biased and ideal unbiased settings, enabling a competence measure based on the estimated ratio of bias-conflicting samples.

3 Method

Preliminaries and Notation. We consider a standard supervised learning setting with input space $\mathcal{X}$ and class labels $\mathcal{Y}$. The corresponding ERM model [31] $m : \mathcal{X} \to \mathcal{Y}$ is decomposed as $m = h \circ e$, where $e : \mathcal{X} \to \mathcal{F}$ is the feature extractor and $h : \mathcal{F} \to \mathcal{Y}$ is the classifier. The training dataset $\mathcal{D}$ of size $n \in \mathbb{N}_{\geq 0}$ is given by $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} \subseteq \mathcal{X} \times \mathcal{Y}$, and is *biased* due to the spurious correlations.

Following standard practice in debiasing [2,17,18,28], we formalize this bias as an imbalance between subgroups, which the model may exploit to make predictions based on the subgroup rather than the label. Formally, we define a surjective subgroup labeling function $g : \mathcal{X} \times \mathcal{Y} \to \mathcal{G}$, where $\mathcal{G} = \{1, 2, \ldots, l\}$ indexes the subgroups defined by combinations of (class label, bias attribute). Probabilistically, the subgroup imbalance is equivalent to an imbalance in the joint distribution $P(Y, B)$ of class labels Y and bias attributes B. In line with other post hoc debiasing approaches [12,18], we assume that subgroup labels are available only on the validation dataset.

Method Overview. AMSEL addresses robust test-time generalization by constructing a family of candidate models, each optimized for a different test distribution, and by adaptively selecting the appropriate model based on the input data. Our method consists of three main steps:

(i) *Candidate Model Construction:* With a fixed feature extractor e, construct a family of classifier heads $\{h_\theta\}_{\theta \in \Theta}$ by training on subsets $\mathcal{D}_\theta$, where θ controls the degree of spurious correlation between class label Y and bias attribute B. Thus, we also refer to θ as the *balancing parameter*.

(ii) *Estimation of Test Distribution Characteristics:* Given a score function $s(\cdot)$ that measures aspects of spurious correlation from an n-tuple of inputs, learn a mapping $M : s(\mathcal{X}^n) \to \Theta$ from these scores to the corresponding balancing parameter $\theta \in \Theta$.

(iii) *Adaptive Inference:* At test time, estimate the balancing parameter $\theta_{\text{test}} = M(s(\mathcal{D}_{\text{test}}))$ from the test data and select the corresponding model $m_{\theta_{\text{test}}} = h_{\theta_{\text{test}}} \circ e$ for making predictions.

Following standard practice, we train the feature extractor e on $\mathcal{D}$ via ERM [31], and then learn the candidate classifier heads $\{h_\theta\}$ and the mapping M on disjoint validation set halves, denoted by $\mathcal{D}_{\text{candidate}}$ and $\mathcal{D}_{\text{mapping}}$, respectively.

Candidate Model Construction. Standard DFR [18] retrains the classifier on a balanced validation subset, implicitly assuming an unbiased test distribution. We instead assume a parameterized family of candidate joint distributions $\{P_\theta\}_{\theta\in\Theta}$, where θ controls the degree of spurious correlation between class label Y and bias attribute B. In our experiments, we use the following instantiation:

$$P_\theta(Y,B) = \theta\, P_{\text{bal}}(Y,B) + (1-\theta)\, P_{\text{orig}}(Y,B), \quad \theta \in [0,1], \tag{1}$$

where P_{orig} is the original (biased) joint distribution and P_{bal} is the uniform (unbiased) joint distribution. Note that due to our definition of subgroups, Equation (1) is equivalent to interpolating between subgroup distributions, thereby allowing θ to control the degree of subgroup balance.

For each θ, we sample a subset $\mathcal{D}_\theta \subseteq \mathcal{D}_{\text{candidate}}$ without replacement to approximate the distribution P_θ. For binary labels and bias attributes, assuming class balance, the strength of spurious correlation in $\mathcal{D}_\theta$ decreases monotonically with increasing θ, vanishing at $\theta = 1$ (see Appendix A2 for details).

With the fixed feature extractor e, AMSEL then trains the corresponding classifier heads h_θ via logistic regression [4] on features extracted from $\mathcal{D}_\theta$:

$$h_\theta = \arg\min_{h\in\mathcal{H}} \frac{1}{|\mathcal{D}_\theta|} \sum_{(x,y)\in\mathcal{D}_\theta} \ell\big(h(e(x)), y\big) + R(h), \tag{2}$$

where $\ell(\cdot,\cdot)$ denotes the cross-entropy loss, $\mathcal{H}$ is the set of classification functions obtainable via logistic regression, and $R(h)$ is a regularizer. Following DFR [18], we average the coefficients over multiple runs to reduce sensitivity to initialization. The final candidate model for parameter θ is given by $m_\theta = h_\theta \circ e$. This decoupled strategy enables efficient training of multiple candidate models using a single, fixed feature extractor.

Adaptive Model Selection. At test time, the test data may exhibit varying levels of spurious correlation, and the true joint distribution of class labels and bias attributes, $P_{\text{test}}(Y,B)$, is unknown. Rather than estimating the bias level directly, AMSEL identifies which candidate test distribution from the predefined set best approximates P_{test}. To this end, we define a score function s, which captures statistical properties indicative of the number of bias-conflicting samples from a given dataset of variable size, i.e., an ordered sequence of n input samples:

$$s : \bigcup_{n\in\mathbb{N}} \mathcal{X}^n \to \mathbb{R}. \tag{3}$$

In our implementation, we compare the predictions of the two most extreme candidate models: m_0 (trained on the original distribution) and m_1 (trained on an unbiased subset). Inspired by Ho et al. [13], we assume that these models agree on simple, bias-aligned samples while diverging on more challenging, bias-conflicting ones. For an input $x \in \mathcal{X}$, we define the binary consensus function:

$$\text{consensus}(x) = \begin{cases} 1, & \text{if } \arg\max(m_0(x)) \\ & \quad = \arg\max(m_1(x)) \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

For a given n-tuple of inputs $X = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$, the consensus score is defined as the percentage of top-1 prediction agreement:

$$s_{\text{consensus}}(X) = \frac{1}{n} \sum_{i=1}^{n} \text{consensus}(x_i). \tag{5}$$

AMSEL learns a mapping $M : s(\mathcal{X}^n) \to \Theta$ from this score to the balancing parameter θ on $\mathcal{D}_{\text{mapping}}$. Thus, we estimate the balancing parameter θ_{test} of the test set as $\theta_{\text{test}} = M\big(s(\mathcal{D}_{\text{test}})\big)$. Finally, for any test input $x \in \mathcal{X}$, the prediction is generated using the corresponding candidate model: $\hat{y} = m_{\theta_{\text{test}}}(x) = h_{\theta_{\text{test}}}(e(x))$. This direct model selection based on the estimated θ_{test} avoids the need to evaluate all candidate models at test time, enabling efficient adaptation to varying levels of spurious correlation.

4 Experiments

We evaluate AMSEL on both a widely used debiasing benchmark (CelebA [20]) and a real-world medical imaging dataset (ChestX-ray14 [33]) and compare it against standard ERM [31] as well as existing post hoc debiasing methods [12,18].

4.1 Implementation Details

Datasets and ERM Models. We evaluate our approach on the following two datasets. The *CelebA* dataset [20] is a popular debiasing benchmark, which comprises celebrity face images. Following [28], we focus on the binary target attribute *blond hair*, which is spuriously correlated with gender. In contrast, *ChestX-ray14* [33] is a real-world dataset containing chest X-rays of patients with various diseases. We consider the binary classification task of detecting *pneumothorax* (diagnosed vs. not diagnosed), where the presence of chest drains (a treatment artifact) acts as a spurious feature [23, 27]. We use automatically generated chest drain labels from [22]. For both datasets, we first train an ERM model [31], which serves as the baseline and provides the fixed feature extractor e for subsequent debiasing. For CelebA, we follow the training procedure of [17] using a ResNet50 [11] pre-trained on ImageNet [8]. For ChestX-ray14, we adopt the approach described in [22] with a DenseNet121 [15] architecture.

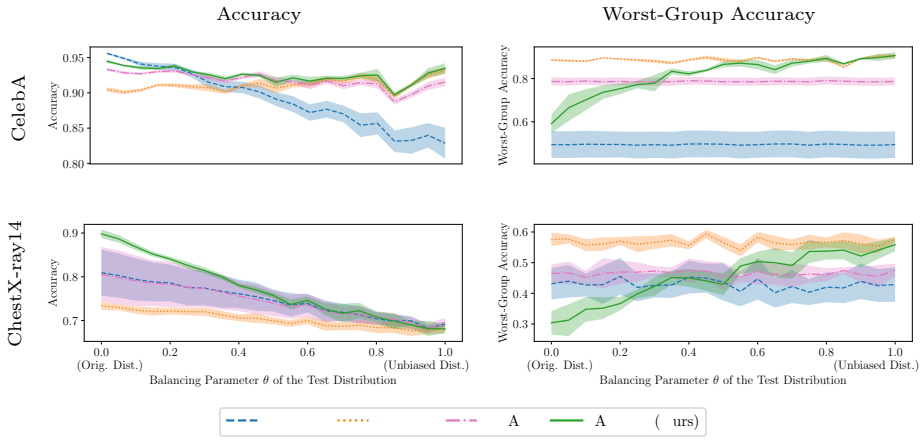
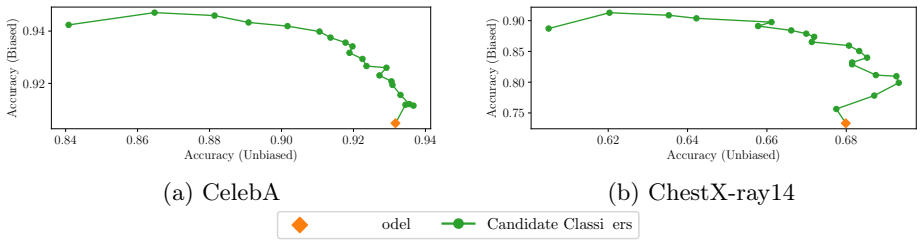


Fig. 2: **Performance Trade-Off Across Test Distributions:** We evaluate accuracy (↑, *left*) across test sets with varying spurious correlation (cf. Equation (1)), with ERM [31] performing best on biased data, while conventional post hoc debiasing methods (DFR [18] and EvA [12]) outperform the ERM model on balanced data. In contrast, AMSEL maintains competitive accuracy across all distributions by adaptively selecting the estimated best classifier. The adaptive nature of AMSEL is further evident in its worst-group accuracy [28] (↑, *right*), a proxy for model bias, demonstrating an increased model bias for biased distributions. Shaded regions around each line represent standard deviations.

Conventional Debiasing and AMSEL. In practice, we implement the mapping $M : s(\mathcal{X}^n) \to \Theta$ via linear regression and clip its predictions to $[0, 1]$. The parameter space is discretized as $\Theta = \{0.0, 0.05, 0.1, \ldots, 1.0\}$ to reduce the number of classifier heads. Following Kirichenko et al. [18], both the standard DFR models and our candidate classifier heads are trained via logistic regression using the `liblinear` solver [9] with an ℓ_1 penalty (see Appendix A1). To reduce sensitivity to initialization, each classifier head is obtained by averaging coefficients over 20 independent runs. The inverse regularization strength c is selected by training on $\mathcal{D}_{\text{candidate}}$ and tuning based on the worst-group accuracy [28] on $\mathcal{D}_{\text{mapping}}$, considering $c \in \{a \cdot 10^b \mid a \in \{1, 3, 7\}, b \in \{-3, -2, -1, 0, 1, 2, 3, 4\}\}$. For EvA [12], we follow the original procedure and retrain the classifier head using SGD (see Appendix A3 for additional details). Sample-wise reweighting is applied to all classifiers to address class imbalance. All reported results include standard deviations computed over five independent runs.

4.2 Results

Trade-Offs Between ERM and Debiasing Methods. To motivate the need for adaptive selection, Figure 2 compares the ERM model [31] with the debiased models obtained via DFR [18] and EvA [12], reporting both overall accuracy



(a) CelebA (b) ChestX-ray14

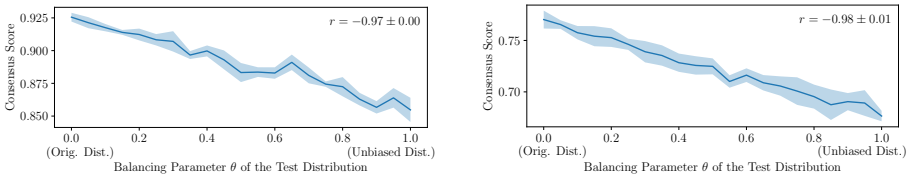
Fig. 3: **Comparison of Candidate Classifiers:** AMSEL trains classifier heads on a range of candidate test distributions ranging from the original, biased distribution to the unbiased distribution, which corresponds to standard DFR. When comparing their accuracy (↑) on the original (biased) and the unbiased distribution, we observe that there is no single uniformly dominant candidate classifier.

and worst-group accuracy [28] across test sets with varying spurious correlation strength (see Equation (1)). A clear trade-off emerges: the ERM model excels on biased data by exploiting shortcuts, whereas debiased models, by ignoring these shortcuts, outperform it on balanced sets where such correlations no longer hold.The debiased models' insensitivity to spurious correlations is reflected in their higher worst-group accuracy, indicating more uniform performance across subgroups. Since the test set's level of spurious correlation is typically unknown, the optimal model choice is ambiguous. AMSEL addresses this by adaptively selecting the classifier head based on the estimated test distribution.

Candidate Classifier Heads. Figure 3 depicts the performance trade-offs among the candidate classifier heads $\{h_\theta\}_{\theta\in[0,1]}$ (see Equation (2)), where $\theta = 1.0$ corresponds to the standard DFR setting [18]. The candidate models exhibit an almost smooth transition in performance, shifting their focus from biased to unbiased data. As no single candidate model is uniformly optimal, adaptive model selection is essential to navigate this performance trade-off.

Estimation of Test Distribution Characteristics. The effectiveness of AMSEL depends on accurately estimating the balancing parameter θ to select the estimated best classifier for a given test distribution. Figure 4 shows that our consensus-based score function $s_{\text{consensus}}$ is almost linearly related to the balancing parameter θ, as confirmed by the Pearson correlation coefficients of -0.97 ± 0.00 for CelebA [20] and -0.98 ± 0.01 for ChestX-ray14 [33]. This strong linear dependency confirms $s_{\text{consensus}}$ as a suitable score function and justifies our design choice of implementing the mapping M for estimating the test distribution's bias characteristics via linear regression.

Adaptive Model Selection (AMSEL). Finally, we combine the balancing parameter estimation with the candidate classifier heads to evaluate the full



(a) CelebA (b) ChestX-ray14

Fig. 4: **Informative Value of the Score Function:** Our consensus-based score function $s_{\text{consensus}}$ exhibits a near-linear relationship with the balancing parameter θ (Pearson correlation coefficients denoted as r), which controls the strength of spurious correlation, making it suitable for test distribution identification. Shaded regions around each line represent standard deviations.

AMSEL approach. For a test set $\mathcal{D}_{\text{test}}$, we estimate $\theta_{\text{test}} = M\big(s(\mathcal{D}_{\text{test}})\big)$ and select the corresponding model $m_{\theta_{\text{test}}} = h_{\theta_{\text{test}}} \circ e$ for prediction. As shown in Figure 2, for CelebA [20], AMSEL nearly matches the performance of the ERM model [31] on test distributions with strong spurious correlation (benefiting from shortcut bias exploitation) while also capturing the improvements of debiasing methods on balanced test distributions where spurious correlations no longer apply. For ChestX-ray14 [33], the adaptive shift from a biased to an unbiased model is most apparent when considering worst-group accuracy [28], indicated by an improved worst-group performance for higher balancing parameters. To capture overall performance across the entire spectrum of considered joint distributions, we compute the area under the accuracy curve (AuC; see Appendix A4). On CelebA [20] AMSEL achieves an AuC of 0.925 ± 0.002, outperforming DFR [18] (0.911 ± 0.003), EvA [12] (0.917 ± 0.003), and ERM [31] (0.891 ± 0.010). On ChestX-ray14 [33], similar trends are observed with AMSEL (0.771 ± 0.007) outperforming the other approaches by a margin of 0.025. These results confirm that AMSEL is highly beneficial when the strength of the spurious correlation of the test set is unknown. When it is known, AMSEL also retains high performance.

5 Conclusion

We propose a lightweight yet effective approach to spurious feature mitigation that expands debiasing with *adaptive model selection (AMSEL)*. Unlike classical debiasing methods, AMSEL preserves the high accuracy of the original ERM model in bias-dominated regimes while matching debiased models under balanced conditions by adaptively selecting its classifier from a family of candidate heads trained on simulated test distributions. This added flexibility requires minimal extra overhead, as the additional forward passes (two for estimating the test distribution characteristics) only involve the lightweight classifier head. Our experiments on both CelebA [20] and ChestX-ray14 [33] demonstrate that AMSEL outperforms ERM [31] and other debiasing approaches [12, 18] when

evaluated across subsets with varying levels of spurious correlation. For example, on ChestX-ray14, AMSEL achieves an area under the accuracy curve (AuC) of 0.771 ± 0.007, outperforming the strongest baseline by a margin of 0.025. Furthermore, we believe that the core idea of selecting from a set of candidate models, each optimized for a specific strength of spurious correlation, can extend to other debiasing strategies such as reweighting [16], resampling [16], or GroupDRO [28]. Overall, AMSEL provides an efficient solution for robust performance in environments where the degree of spurious correlation may vary significantly.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Almeida, P.R., Oliveira, L.S., Britto, A.S., Sabourin, R.: Adapting dynamic classifier selection for concept drift. Expert Syst. Appl. (2018) 3, 4
2. Asaad, I., Shadaydeh, M., Denzler, J.: Gradient Extrapolation for Debiased Representation Learning 2, 3, 4
3. Beery, S., van Horn, G., Perona, P.: Recognition in Terra Incognita. In: ECCV (2018) 2, 3
4. Bishop, C.M.: Pattern Recognit. and Machine Learning (2006) 5
5. Britto, A.S., Sabourin, R., Oliveira, L.E.: Dynamic selection of classifiers—A comprehensive review. Pattern Recognit. (2014) 3, 4
6. Cruz, R.M., Sabourin, R., Cavalcanti, G.D.: Dynamic classifier selection: Recent advances and perspectives. Inf. Fusion (2018) 3, 4
7. Dehkharghanian, T., Bidgoli, A.A., Riasatian, A., Mazaheri, P., Campbell, C.J.V., Pantanowitz, L., Tizhoosh, H.R., Rahnamayan, S.: Biased data, biased AI: deep networks predict the acquisition site of TCGA images. Diagnostic pathology (2023) 2, 3
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009) 6
9. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A Library for Large Linear Classification. JMLR (2008) 7, 1
10. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nat. Mach. Intell. (2020) 2, 3
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2016) 6
12. He, Q., Xu, K., Yao, A.: EvA: Erasing Spurious Correlations with Activations. In: ICLR (2025) 2, 3, 4, 6, 7, 9, 1, 8
13. Ho, T.K., Hull, J.J., Srihari, S.N.: Decision combination in multiple classifier systems. PAMI (1994) 6
14. Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., Cipriani, N., Grossman, R.L., Pearson, A.T.: The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat. Commun. (2021) 2, 3
15. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: CVPR (2017) 6

16. Idrissi, B.Y., Arjovsky, M., Pezeshki, M., Lopez-Paz, D.: Simple data balancing achieves competitive worst-group-accuracy. In: Proceedings of the First Conference on Causal Learning and Reasoning. PMLR (2022) 2, 3, 10
17. Izmailov, P., Kirichenko, P., Gruver, N., Wilson, A.G.: On Feature Learning in the Presence of Spurious Correlations. In: NeurIPS (2022) 2, 4, 6
18. Kirichenko, P., Izmailov, P., Wilson, A.G.: Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. In: ICLR (2023) 2, 3, 4, 5, 6, 7, 8, 9, 1
19. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking Clever Hans predictors and assessing what machines really learn. Nat. Commun. (2019) 3
20. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep Learning Face Attributes in the Wild. In: ICCV (2015) 3, 6, 8, 9, 1, 2
21. Lövdal, S., Biehl, M.: Iterated Relevance Matrix Analysis (IRMA) for the identification of class-discriminative subspaces. Neurocomputing (2024) 2
22. Murali, N., Puli, A., Yu, K., Ranganath, R., Batmanghelich, K.: Beyond Distribution Shift: Spurious Features Through the Lens of Training Dynamics. TMLR (2023) 3, 6
23. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. CHIL (2020) 6
24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine Learning in Python. JMLR (2011) 1, 8
25. Penzel, N., Stein, G., Denzler, J.: Reducing Bias in Pre-Trained Models by Tuning While Penalizing Change. In: VISIGRAPP (2024) 2, 3
26. Rosenfeld, E., Ravikumar, P.K., Risteski, A.: Domain-Adjusted Regression or: ERM May Already Learn Features Sufficient for Out-of-Distribution Generalization. In: NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications (2022) 3, 4
27. Rueckel, J., Trappmann, L., Schachtner, B., Wesp, P., Hoppe, B.F., Fink, N., Ricke, J., Dinkel, J., Ingrisch, M., Sabel, B.O.: Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs. Invest. Radiol. (2020) 6
28. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In: ICLR (2020) 2, 3, 4, 6, 7, 8, 9, 10
29. Shah, H., Tamuly, K., Raghunathan, A., Jain, P., Netrapalli, P.: The Pitfalls of Simplicity Bias in Neural Networks. In: NeurIPS (2020) 3
30. Teney, D., Abbasnejad, E., Lucey, S., van den Hengel, A.: Evading the Simplicity Bias: Training a Diverse Set of Models Discovers Solutions With Superior OOD Generalization. In: CVPR (2022) 3
31. Vapnik, V.N.: An overview of statistical learning theory. IEEE transactions on neural networks (1999) 2, 3, 4, 5, 6, 7, 9
32. Villani, C.: Optimal Transport, Old and New (2009) 8
33. Wang, X., Peng, Y., Le Lu, Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: CVPR (2017) 3, 6, 8, 9, 1
34. Xiao, K.Y., Engstrom, L., Ilyas, A., Madry, A.: Noise or Signal: The Role of Image Backgrounds in Object Recognition. In: ICLR (2021) 2

35. Yang, Y., Zhang, H., Katabi, D., Ghassemi, M.: Change is Hard: A Closer Look at Subpopulation Shift. In: ICML (2023) 3
36. Ye, W., Zheng, G., Cao, X., Ma, Y., Zhang, A.: Spurious Correlations in Machine Learning: A Survey 2, 3
37. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. (2018) 2

Appendix

In this appendix, we provide additional details on the implementation and experimental procedures used in our work. We describe the logistic regression classifiers employed in both AMSEL and standard DFR [18], the simulation of test distributions, the implementation of the Erasing with Activations (EvA) method [12], and present further experimental results. Sections are organized as follows: Section A1 details the logistic regression classifiers and the associated hyperparameter settings; Section A2 describes formal properties of the candidate test distributions and a procedure for simulating them; Section A3 outlines the EvA method [12]; and Section A4 provides additional experimental results.

A1 Logistic Regression Classifiers

Both AMSEL and standard DFR replace the classifier head with a logistic regression classifier whose coefficients are averaged over multiple runs. In the case of AMSEL, we even train a set of logistic regression classifiers, with the optimal one selected at inference time. In both approaches, the hyperparameters that must be determined include:

- number of repetitions,
- solver for logistic regression,
- type of regularization,
- strength of regularization, and
- strategy for class balancing.

Number of Repetitions, Solver, and Type of Regularization for Logistic Regression. We follow the logistic regression setup detailed by standard DFR [18]. For the logistic regression, we employ the `LIBLINEAR` solver [9] with an ℓ_1 penalty, using the implementation from `scikit-learn` [24]. To reduce sensitivity to initialization, the classifier's coefficients are averaged over 20 independent runs.

Inverse Regularization Strength. Following the protocol in [18], we determine the inverse regularization strength by training DFR models on one half of the validation dataset ($\mathcal{D}_{\text{candidate}}$) and evaluating them on the other half ($\mathcal{D}_{\text{mapping}}$). The selection criterion is based on the worst-group accuracy. In our experiments, the inverse regularization strength c is chosen from

$$c \in \{a \cdot 10^b \mid a \in \{1, 3, 7\}, b \in \{-3, -2, -1, 0, 1, 2, 3, 4\}\}. \tag{6}$$

For both CelebA [20] and ChestX-ray14 [33], we select $c = 0.7$ (for detailed results, see Table 1 and Table 2).

Table 1: **Inverse Regularization Strength for CelebA:** To select the inverse regularization strength parameter c of the logistic regression classifiers for CelebA [20], we train DFR models [18] with varying values of c. The models are trained on one half of the validation dataset and evaluated on the other. Based on the best worst-group accuracy, we set $c = 0.7$ in subsequent experiments.

Inverse Regularization Strength c	Accuracy (%, ↑)	Worst-Group Accuracy (%, ↑)
$7 \cdot 10^4$	81.44 ± 0.26	77.93 ± 0.58
$3 \cdot 10^4$	81.46 ± 0.27	77.96 ± 0.58
$1 \cdot 10^4$	81.61 ± 0.25	78.15 ± 0.62
$7 \cdot 10^3$	81.69 ± 0.26	78.22 ± 0.65
$3 \cdot 10^3$	82.11 ± 0.34	78.69 ± 0.81
$1 \cdot 10^3$	82.93 ± 0.27	79.48 ± 0.78
$7 \cdot 10^2$	83.41 ± 0.28	79.89 ± 0.85
$3 \cdot 10^2$	85.22 ± 0.34	81.61 ± 0.79
$1 \cdot 10^2$	86.30 ± 0.45	82.70 ± 0.89
$7 \cdot 10^1$	86.26 ± 0.42	82.55 ± 0.82
$3 \cdot 10^1$	86.98 ± 0.29	83.28 ± 0.70
$1 \cdot 10^1$	88.07 ± 0.25	84.44 ± 0.62
$7 \cdot 10^0$	88.43 ± 0.29	84.85 ± 0.66
$3 \cdot 10^0$	89.07 ± 0.28	85.62 ± 0.59
$1 \cdot 10^0$	89.56 ± 0.29	86.11 ± 0.67
$7 \cdot 10^{-1}$	**89.69 ± 0.29**	**86.18 ± 0.58**
$3 \cdot 10^{-1}$	89.64 ± 0.28	85.86 ± 0.79
$1 \cdot 10^{-1}$	88.30 ± 0.39	83.41 ± 1.00
$7 \cdot 10^{-2}$	87.27 ± 0.43	81.86 ± 1.07
$3 \cdot 10^{-2}$	80.67 ± 0.97	73.18 ± 1.52
$1 \cdot 10^{-2}$	74.65 ± 1.72	63.81 ± 2.77
$7 \cdot 10^{-3}$	62.74 ± 23.94	49.46 ± 24.86
$3 \cdot 10^{-3}$	14.87 ± 0.00	0.00 ± 0.00
$1 \cdot 10^{-3}$	14.87 ± 0.00	0.00 ± 0.00

Class Balancing. In standard DFR [18], class balancing is implicitly achieved when subgroups are defined as combinations of (label, bias attribute). In contrast, AMSEL trains logistic classifier heads on datasets with varying subgroup balances (i.e., including $\theta \neq 1.0$). Thus, we enforce class balancing by applying class weights inversely proportional to the class frequencies.

A2 Candidate Test Distributions

In this section, we provide additional details on the construction and theoretical properties of the candidate test distributions used in our evaluation. In particular, we analyze the statistical behavior of the parameterized family of joint distributions defined by the interpolation parameter θ, which controls the degree of spurious correlation between the class label and the bias attribute. We also

Table 2: **Inverse Regularization Strength for ChestX-ray14:** To select the inverse regularization strength parameter c of the logistic regression classifiers for ChestX-ray14 [33], we train DFR models [18] with varying values of c. The models are trained on one half of the validation dataset and evaluated on the other. Based on the best worst-group accuracy, we set $c = 0.7$ in subsequent experiments.

Inverse Regularization Strength c	Accuracy (%, ↑)	Worst-Group Accuracy (%, ↑)
$7 \cdot 10^4$	66.77 ± 0.53	54.26 ± 3.31
$3 \cdot 10^4$	66.78 ± 0.53	54.26 ± 3.31
$1 \cdot 10^4$	66.80 ± 0.51	54.15 ± 3.12
$7 \cdot 10^3$	66.84 ± 0.50	54.36 ± 3.32
$3 \cdot 10^3$	66.96 ± 0.58	54.36 ± 3.32
$1 \cdot 10^3$	67.20 ± 0.54	54.56 ± 3.43
$7 \cdot 10^2$	67.44 ± 0.52	54.46 ± 3.20
$3 \cdot 10^2$	68.16 ± 0.65	55.08 ± 3.33
$1 \cdot 10^2$	70.28 ± 0.75	55.90 ± 2.90
$7 \cdot 10^1$	70.92 ± 0.83	55.79 ± 3.20
$3 \cdot 10^1$	71.57 ± 0.84	56.51 ± 3.01
$1 \cdot 10^1$	72.25 ± 0.94	56.72 ± 3.22
$7 \cdot 10^0$	72.45 ± 0.99	57.23 ± 3.22
$3 \cdot 10^0$	72.91 ± 1.02	57.54 ± 2.94
$1 \cdot 10^0$	73.46 ± 1.14	57.85 ± 2.62
$7 \cdot 10^{-1}$	73.74 ± 1.12	**58.05 ± 2.33**
$3 \cdot 10^{-1}$	**74.41 ± 1.33**	57.33 ± 3.03
$1 \cdot 10^{-1}$	72.74 ± 1.13	51.49 ± 1.47
$7 \cdot 10^{-2}$	68.37 ± 2.51	46.97 ± 3.29
$3 \cdot 10^{-2}$	66.26 ± 1.18	36.31 ± 6.01
$1 \cdot 10^{-2}$	4.96 ± 0.00	0.00 ± 0.00
$7 \cdot 10^{-3}$	4.96 ± 0.00	0.00 ± 0.00
$3 \cdot 10^{-3}$	4.96 ± 0.00	0.00 ± 0.00
$1 \cdot 10^{-3}$	4.96 ± 0.00	0.00 ± 0.00

describe the procedure by which we generate test distributions corresponding to a specific value of θ.

A2.1 Properties of Candidate Test Distributions

We generate our candidate test distributions based on the joint distribution $P(Y, B)$ of the ground truth label Y and the bias attribute B. In our experiments, both Y and B are binary variables, and subgroups are defined as combinations of (label, bias attribute); hence, the empirical joint distribution of $P(Y, B)$ coincides with the empirical distribution of the dataset subgroups.

Our candidate test distributions are generated via a linear interpolation between the joint distribution from the training data and a balanced (unbiased)

joint distribution (see Equation (1)):

$$P_\theta = \theta\, P_{\text{bal}} + (1-\theta)\, P_{\text{orig}}, \quad \theta \in [0,1], \tag{7}$$

where P_{orig} denotes the original (biased) joint distribution and P_{bal} is perfectly balanced.

In this section, we demonstrate that this interpolation is a sound method for generating candidate test distributions. In particular, we prove that (i) the covariance between Y and B is linear in θ, with $\text{Cov}_1(Y,B) = 0$ (i.e. Y and B are independent in the balanced distribution), and (ii) the spurious correlation $|\text{Corr}_\theta(Y,B)|$ decreases monotonically with increasing θ, vanishing when $\theta = 1$. In other words, as we transition from the original training distribution to the balanced distribution, the level of spurious correlation decreases continuously.

Note that our proofs assume that the marginal distribution of Y is held fixed. This is a valid assumption since we perform class balancing when training our classifier heads, which effectively fixes the marginal distribution of the labels $P(Y)$ across all candidate distributions.

Covariance. Let Y and B be binary random variables taking values in $\{0,1\}$. Define the following probabilities:

$$p^0_{11} = P_{\text{orig}}(Y=1, B=1), \tag{8}$$
$$p^1_{11} = P_{\text{bal}}(Y=1, B=1), \tag{9}$$
$$p^0_{1\bullet} = P_{\text{orig}}(Y=1), \tag{10}$$
$$p^1_{1\bullet} = P_{\text{bal}}(Y=1), \tag{11}$$
$$p^0_{\bullet 1} = P_{\text{orig}}(B=1), \tag{12}$$
$$p^1_{\bullet 1} = P_{\text{bal}}(B=1). \tag{13}$$

As described above, we assume that the marginal distribution of Y is fixed across $\theta \in [0,1]$ (e.g., due to class balancing):

$$p^0_{1\bullet} = p^1_{1\bullet} =: p_{1\bullet}. \tag{14}$$

Then, following Equation (1), the candidate distribution P_θ satisfies:

$$p^\theta_{11} = \theta\, p^1_{11} + (1-\theta)\, p^0_{11}, \tag{15}$$
$$p^\theta_{1\bullet} = \theta\, p^1_{1\bullet} + (1-\theta)\, p^0_{1\bullet} = p_{1\bullet}, \tag{16}$$
$$p^\theta_{\bullet 1} = \theta\, p^1_{\bullet 1} + (1-\theta)\, p^0_{\bullet 1}. \tag{17}$$

The covariance between Y and B under P_θ is given by:

$$\text{Cov}_\theta(Y,B) = p^\theta_{11} - p^\theta_{1\bullet}\, p^\theta_{\bullet 1}. \tag{18}$$

Substituting the expressions from above, we obtain:

$$\text{Cov}_\theta(Y,B) = \theta\, p^1_{11} + (1-\theta)\, p^0_{11} - p_{1\bullet}\Big(\theta\, p^1_{\bullet 1} + (1-\theta)\, p^0_{\bullet 1}\Big) \tag{19}$$
$$= \theta\Big(p^1_{11} - p_{1\bullet}\, p^1_{\bullet 1}\Big) + (1-\theta)\Big(p^0_{11} - p_{1\bullet}\, p^0_{\bullet 1}\Big). \tag{20}$$

Since P_{bal} is balanced, Y and B are independent under it, and hence:

$$p^1_{11} = p_{1\bullet}\, p^1_{\bullet 1}. \tag{21}$$

Thus, the term multiplying θ vanishes:

$$\theta\Big(p^1_{11} - p_{1\bullet}\, p^1_{\bullet 1}\Big) = 0. \tag{22}$$

Therefore, we obtain the final expression:

$$\mathrm{Cov}_\theta(Y, B) = (1-\theta)\Big(p^0_{11} - p_{1\bullet}\, p^0_{\bullet 1}\Big). \tag{23}$$

This shows that, assuming a fixed marginal distribution of the labels Y, the covariance is linear with respect to θ.

Note that when $\theta = 0$, we recover the covariance of the original distribution,

$$\mathrm{Cov}_0(Y, B) = p^0_{11} - p_{1\bullet}\, p^0_{\bullet 1}, \tag{24}$$

and when $\theta = 1$, we have

$$\mathrm{Cov}_1(Y, B) = 0, \tag{25}$$

which confirms that Y and B are independent in the balanced distribution.

Correlation. Under the assumption of a fixed marginal distribution of Y, we now analyze the Pearson correlation between Y and B for the candidate distribution. Recall that we have:

$$\mathrm{Cov}_\theta(Y, B) = (1-\theta)\, C, \quad \text{with } C := p^0_{11} - p_{1\bullet}\, p^0_{\bullet 1}. \tag{26}$$

The variances are given by

$$\mathrm{Var}_\theta(Y) = p_{1\bullet}(1 - p_{1\bullet}), \text{ and} \tag{27}$$

$$\mathrm{Var}_\theta(B) = A(\theta)(1 - A(\theta)), \tag{28}$$

where

$$A(\theta) := p^\theta_{\bullet 1} = \theta\, p^1_{\bullet 1} + (1-\theta)\, p^0_{\bullet 1}. \tag{29}$$

Since P_{bal} is balanced, we have $p^1_{\bullet 1} = \frac{1}{2}$, so that

$$A(\theta) = a + \theta\left(\frac{1}{2} - a\right), \quad \text{with } a := p^0_{\bullet 1}. \tag{30}$$

The Pearson correlation is then given by:

$$\mathrm{Corr}_\theta(Y, B) = \frac{\mathrm{Cov}_\theta(Y, B)}{\sqrt{\mathrm{Var}_\theta(Y)\,\mathrm{Var}_\theta(B)}} = \frac{(1-\theta)\, C}{\sqrt{p_{1\bullet}(1 - p_{1\bullet})\, A(\theta)(1 - A(\theta))}}. \tag{31}$$

Taking absolute values, we define

$$f(\theta) := |\mathrm{Corr}_\theta(Y, B)| = \frac{|C|\,(1-\theta)}{\sqrt{p_{1\bullet}(1 - p_{1\bullet})\, A(\theta)(1 - A(\theta))}}. \tag{32}$$

Since $|C|$ and $p_{1\bullet}(1-p_{1\bullet})$ are constant, it suffices to study the function

$$g(\theta) := \frac{1-\theta}{\sqrt{A(\theta)(1-A(\theta))}}. \tag{33}$$

To study the monotonicity of $g(\theta)$, we consider the numerator and denominator separately:

- **Numerator:** The term $1-\theta$ is monotonically decreasing in θ.
- **Denominator:** The function $A(\theta)$ defined in Equation (30) linearly interpolates between $a = p^0_{\bullet 1}$ and $\frac{1}{2}$ Thus, $A(\theta)$ monotonically approaches $\frac{1}{2}$ as θ increases.
  Now, consider the product $A(\theta)(1-A(\theta))$, which represents the variance of a Bernoulli variable with mean $A(\theta)$. This product is maximized when $A(\theta) = \frac{1}{2}$. Taking the derivative with respect to θ:

$$\frac{d}{d\theta}\left[A(\theta)(1-A(\theta))\right] = A'(\theta)(1-A(\theta)) - A(\theta)A'(\theta) \tag{34}$$

$$= A'(\theta)(1-2A(\theta)) \tag{35}$$

$$= \left(\frac{1}{2} - a\right)(1-2A(\theta)). \tag{36}$$

  If $a \leq \frac{1}{2}$, then for $\theta \in [0,1]$ we have $A(\theta) \in [a, \frac{1}{2}]$, ensuring both $\frac{1}{2} - a \geq 0$ and $1-2A(\theta) \geq 0$. Conversely, if $a > \frac{1}{2}$, then $A(\theta) \in (\frac{1}{2}, a]$ and thus $\frac{1}{2} - a < 0$ and $1 - 2A(\theta) < 0$. In either case, the product remains nonnegative:

$$\left(\frac{1}{2} - a\right)(1-2A(\theta)) \geq 0. \tag{37}$$

  That is, the denominator $\sqrt{A(\theta)(1-A(\theta))}$ is monotonically increasing in θ.

Since the numerator of $g(\theta)$ is monotonically decreasing and its denominator is monotonically increasing, it follows that $g(\theta)$ is monotonically decreasing in θ. Therefore, from Equation (32),

$$|\mathrm{Corr}_\theta(Y,B)| \quad \text{decreases monotonically with} \quad \theta. \tag{38}$$

In particular, when $\theta = 1$, Equation (26) implies $\mathrm{Cov}_1(Y,B) = 0$, so that

$$\mathrm{Corr}_1(Y,B) = 0. \tag{39}$$

This completes the proof that the absolute correlation $|\mathrm{Corr}_\theta(Y,B)|$ decreases monotonically with θ. Consequently, as θ increases, the spurious correlation between class labels and bias attributes decreases continuously until it vanishes for the balanced distribution.

A2.2 Simulation of Test Distributions

We simulate candidate test distributions conforming to P_θ for $\theta \in [0,1]$ (see Equation (1)) for both training candidate classifiers and evaluation. For a given dataset $\mathcal{D}$, we achieve this by subsampling subsets $\mathcal{D}_\theta \subset \mathcal{D}$ whose joint distribution approximates P_θ.

Given that we define subgroups as combinations of (class label, bias attribute), interpolating between joint distributions is equivalent to interpolating between subgroups. Thus, we can rephrase Equation (1) as:

$$P_\theta^G = \theta\, P_{\text{bal}}^G + (1-\theta)\, P_{\text{orig}}^G, \quad \theta \in [0,1], \tag{40}$$

where $P_{\text{orig}}^G \in [0,1]^{|\mathcal{G}|}$ denotes the original (biased) subgroup distribution and $P_{\text{bal}}^G \in [0,1]^{|\mathcal{G}|}$ represents the balanced (uniform) distribution across the $|\mathcal{G}|$ subgroups.

For a target total of $N \in \mathbb{N}_{>0}$ samples in $\mathcal{D}_\theta$, the number of samples drawn from subgroup $i \in \mathbb{N}$ is:

$$n_i = N \cdot P_{\theta_i}^G, \quad i = 1, \ldots, |\mathcal{G}|. \tag{41}$$

In practice, to determine a feasible N without oversampling any subgroup, we identify the minority subgroup. Formally, let $j \in \mathbb{N}$ represent the index of the minority subgroup

$$j = \underset{i \in \{1,\ldots,|\mathcal{G}|\}}{\arg\min} \left|\{(x,y) \in \mathcal{D} \mid g((x,y)) = i\}\right| \tag{42}$$

and denote the number of available samples in this subgroup as

$$n_j = \min_{i \in \{1,\ldots,|\mathcal{G}|\}} \left|\{(x,y) \in \mathcal{D} \mid g((x,y)) = i\}\right|. \tag{43}$$

Since subsampling is performed without replacement and all samples from the minority subgroup are utilized, we must have $n_j = N \cdot P_{\theta_j}^G$, which implies

$$N = \frac{n_j}{P_{\theta_j}^G}. \tag{44}$$

Subsequently, for each subgroup i, we randomly select n_i samples without replacement from $\mathcal{D}$ to form $\mathcal{D}_\theta$. This procedure ensures that the resulting dataset $\mathcal{D}_\theta$ adheres to the desired subgroup distribution P_θ^G.

A3 Erasing with Activations (EvA)

In this section, we provide implementation details for the *Erasing with Activations (EvA)* approach [12], which we use as a baseline in our experimental evaluation. EvA is designed to mitigate the influence of spurious features in deep neural networks by first identifying such features and then retraining the classifier head with the corresponding weights erased.

In the initial phase, spurious features are identified by comparing activation distributions obtained from a biased dataset, denoted by $\mathcal{D}_{\text{biased}}$, and an unbiased dataset, denoted by $\mathcal{D}_{\text{unbiased}}$. The intuition is that if the activation distributions differ significantly, the corresponding feature is likely to be spurious, as it is sensitive to the bias. In practice, we construct these datasets from the two halves of the validation set used in our other experiments. In our experiments, we form $\mathcal{D}_{\text{biased}}$ using the candidate training set $\mathcal{D}_{\text{candidate}}$, while $\mathcal{D}_{\text{unbiased}}$ is derived from $\mathcal{D}_{\text{mapping}}$ via subsampling based on subgroup labels. In cases where an unbiased dataset is not available, He et al. [12] propose an alternative procedure; however, our experiments are conducted in the setting where $\mathcal{D}_{\text{unbiased}}$ is accessible.

For each feature dimension i, let $\phi^{(i)}$ denote the distribution of activation values $e(x)_i$ over samples $x \in \mathcal{X}$, and let $\phi^{(ik)}$ denote the distribution restricted to samples from class k. Following [12], we define the consistency measure for feature dimension i and class k as:

$$\mathbf{C}^{(ik)} \approx -\mathbf{d}\left(\phi^{(ik)}_{\text{biased}}, \phi^{(ik)}_{\text{unbiased}}\right), \tag{45}$$

where the Wasserstein distance [32] is used as the metric $\mathbf{d}$. A lower value of $\mathbf{C}^{(ik)}$ indicates a larger discrepancy between the biased and unbiased distributions, suggesting that the corresponding feature is more likely to be spurious.

To formally select spurious features, an indicator function is introduced:

$$\mathbb{I}^{(ik)}_{\mathbf{C}} = \begin{cases} 1, & \text{if } \mathbf{C}^{(ik)} \leq \epsilon, \\ 0, & \text{otherwise,} \end{cases} \tag{46}$$

where $\epsilon \in \mathbb{R}$ is a threshold hyperparameter. The set of spurious feature indices for class k is then defined as:

$$\Phi_k = \{\phi^j \in \Phi | \mathbb{I}^{(jk)} = 1\}. \tag{47}$$

In the second phase, the classifier head is retrained as a linear layer using stochastic gradient descent (SGD) with a learning rate of 0.001. The weights corresponding to the spurious features identified in Φ_k are set to zero during retraining. We implement this using the `SGDClassifier` provided by `scikit-learn` [24]. For selecting the threshold ϵ, we follow the procedure described by He et al. [12]: classifier heads are trained on $\mathcal{D}_{\text{biased}}$ with different threshold values, and the model achieving the best accuracy on $\mathcal{D}_{\text{unbiased}}$ is selected.

A4 Additional Experimental Results

To provide a single, comprehensive measure of a model's robustness to varying degrees of spurious correlation, we define the Area under the Accuracy Curve (AuC). It aggregates the performance across the considered spectrum of test distributions $\{\mathcal{D}_\theta\}_{\theta \in [0,1]}$, ranging from the original training distribution ($\theta = 0$) to an unbiased one ($\theta = 1$):

$$\text{AuC}(m) = \int_0^1 \text{acc}_m(\mathcal{D}_\theta)\, \mathrm{d}\theta, \tag{48}$$

Table 3: **Area Under the Accuracy Curve (AuC):** As an aggregated measure of model performance, we evaluate the area under the accuracy curve (↑). For both CelebA [20] and ChestX-ray14 [33], AMSEL achieves a higher AuC than both ERM [31] and the considered debiasing methods [12, 18], demonstrating a more robust performance across the considered candidate test distributions.

Model	CelebA [20]	ChestX-ray14 [33]
ERM Model [31]	0.891 ± 0.010	0.746 ± 0.028
DFR Model [18]	0.911 ± 0.003	0.702 ± 0.008
EvA Model [12]	0.917 ± 0.003	0.743 ± 0.031
AMSEL (Ours)	0.925 ± 0.002	0.771 ± 0.008

where $\text{acc}_m(\mathcal{D}_\theta)$ is the accuracy of model m on the subset $\mathcal{D}_\theta$ simulated with balancing parameter θ. In practice, we estimate this integral using the trapezoidal rule based on the considered discrete set of $\theta \in \Theta$.

The AuC metric has a direct probabilistic justification. Since the true level of spurious correlation at test time is unknown, we can model the balancing parameter θ as a random variable. Assuming a uniform prior, $\theta \sim \mathcal{U}(0, 1)$, reflects maximal uncertainty over the test condition. The AuC is then the expected accuracy under this prior:

$$\text{AuC}(m) = \mathbb{E}_{\theta \sim \mathcal{U}(0,1)} \left[\text{acc}_m(\mathcal{D}_\theta) \right]. \tag{49}$$

Consequently, the AuC provides a robust measure of generalization. It rewards models that perform consistently well across the entire range of potential strengths of spurious correlation, rather than only at the distributional extremes.

Table 3 reports the AuC for all evaluated methods on the CelebA [20] and ChestX-ray14 [33] datasets. As justified above, a higher AuC indicates superior and more robust performance across the full spectrum of spurious correlation strengths.