On the nonlinear correlation of ML performance between data subpopulations

Anonymous Authors[1]

Abstract

Understanding the performance of machine learning models across diverse data distributions is critically important for reliable applications. Recent works empirically find that there is a strong linear relationship between in-distribution (ID) and out-of-distribution (OOD) performance, but we show that this is not necessarily true if there are subpopulation shifts. In this paper, we empirically show that out-of-distribution performance often has nonlinear correlation with in-distribution performance under subpopulation shifts. To understand this phenomenon, we decompose the model's performance into performance on each subpopulation. We show that there is a *"moon shape" correlation* (parabolic uptrend curve) between the test performance on the *majority subpopulation* and the *minority subpopulation*. This nonlinear correlations hold across model architectures, training durations and hyperparameters, and the imbalance between subpopulations. Moreover, we show that the nonlinearity increases in the presence of spurious correlations in the training data. We provide complementary theoretical and experimental analyses for this interesting phenomenon of nonlinear performance correlation across subpopulations. Finally, we discuss the implications of our findings for ML reliability and fairness.

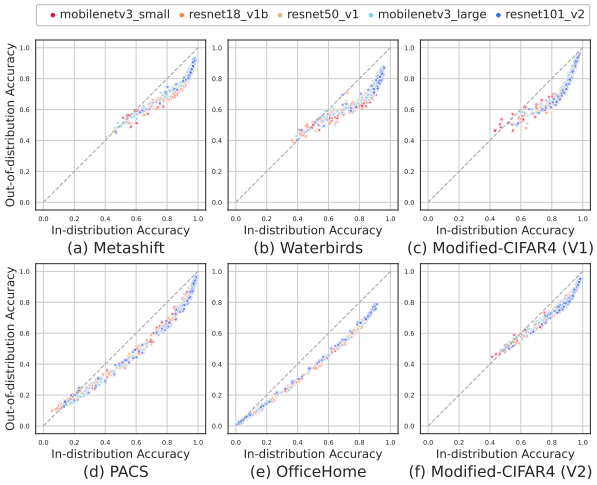


Figure 1. **Out-of-distribution accuracies vs. in-distribution accuracies under *subpopulation shifts*.** Each panel represents a dataset. On each dataset, we trained 500 different models independently, with different model architectures and hyperparameters.

1. Introduction

Subpopulation shift is a major challenge in machine learning (ML) in real-world applications: the data seen in test time often have different distribution across subgroups (e.g. different types of users or patients) compared to the training data (Daneshjou et al., 2021). Recent empirical works find that find that there is a strong linear relationship between in-distribution (ID) and out-of-distribution (OOD) performance on dataset reconstruction shifts (ImageNet-V2 (Recht et al., 2019), CIFAR-10.1 (Recht et al., 2018), CIFAR-10.2 (Lu et al., 2020)). Similar linear trends are also observed in transfer learning (Kornblith et al., 2019), cross-benchmark evaluation (Miller et al., 2021), and sub-type shifts (Santurkar et al., 2021).

In contrast, we empirically show that out-of-distribution performance has a *nonlinear* correlation with in-distribution performance under subpopulation shifts. To understand this phenomenon, we decompose the model's performance into performance on each subpopulation. We show that there is a *"moon shape" correlation* (parabolic uptrend curve) between the test performance on the *majority subpopulation* and the *minority subpopulation*. We empirically show that this phenomenon holds across a wide spectrum of model architectures, training settings, and datasets. Interestingly, the performance correlations become more nonlinear when there is a stronger *spurious correlation* in the training data. We provide rigorous theoretical explanations of how the spurious correlation and the subpopulation accuracy gap are related in a simple binary classification setting.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

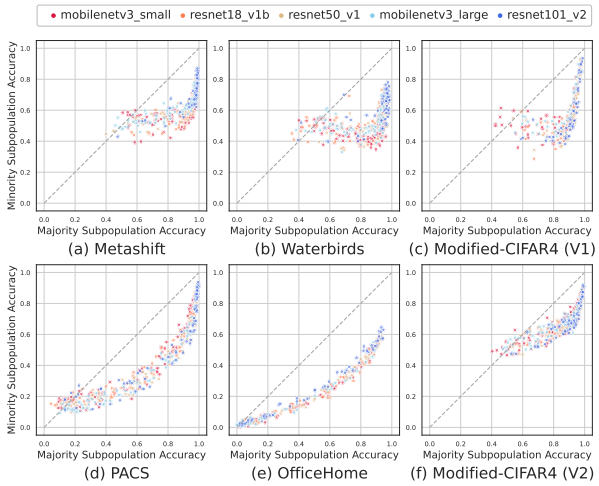


Figure 2. **Majority subpopulation accuracies vs. minority subpopulation accuracies.** There is a striking nonlinear correlation (moon-shape) between the *majority subpopulation* performance and the *minority subpopulation* performance. **Datasets with spurious correlations (top) show *more nonlinear* correlations than datasets without spurious correlations (bottom).**

2. Experimental setup

Preliminaries: ML with diverse subpopulations We consider the overall data distribution with $\mathcal{D} = \{1, \dots, D\}$ diverse subpopulations. Each subpopulation $d \in \mathcal{D}$ corresponds to a fixed data distribution P_d. In each of our main experiment, we compare the performance on two data distributions. **(1) in-distribution (ID)**, or the training distribution, $P^{tr} = \sum_{d \in \mathcal{D}} r_d^{tr} P_d$, where $\{r_d^{tr}\}$ denotes the mixture probabilities in the training set. **(2) out-of-distribution (OOD)** is also a mixture of the D subpopulations, $P^{ts} = \sum_{d \in \mathcal{D}} r_d^{ts} P_d$, where $\{r_d^{ts}\}$ is the mixture probabilities in the test set, but with a different proportion of subpopulations, i.e., $\{r_d^{ts}\} \neq \{r_d^{tr}\}$.

Experimental procedure For simplicity, we consider $D = 2$ subpopulations. The training distribution has a *majority subpopulation* (e.g., $r_d^{tr} = 90\%$), and a *minority subpopulation* (e.g., $r_d^{tr} = 10\%$). As for the out-of-distribution, the majority subpopulation and minority subpopulation are equally representative (e.g., $r_d^{ts} = 50\%$). On each dataset:

1. We first train 500 different ML models $\{f_1, f_2, ...\}$ independently by varying the model architectures, training durations, and hyperparameters following the search space of commercial AutoML (AutoGluon).
2. For each trained ML model f_i, we evaluate the *ID performance*, and the *OOD performance*.

Subpopulation Shift Datasets We categorize our datasets into two categories based on the underlying reason of degraded ML performance on the minority subpopulation (Eyuboglu et al., 2022; Oakden-Rayner et al., 2020):

- **Spurious correlation.** If a target variable is correlated with another variable Z in the training distribution, the model may learn to rely on Z to make predictions. One example is that cat images can be mostly indoor and dog images mostly outdoor. We experiment with three existing datasets in the community: MetaShift (Liang & Zou, 2022), Waterbirds (Sagawa et al., 2020), and Modified-CIFAR4 V1 (Rolf et al., 2021).
- **Rare subpopulation.** Without obvious spurious correlation, ML models can still underperform on subpopulations that occur infrequently in the training set (e.g. patients with a darker skin tone), since the rare subpopulation will not significantly affect model loss during training. We adopted PACS (Li et al., 2017), Office-Home (Venkateswara et al., 2017), Modified-CIFAR4 V2 (Rolf et al., 2021) for experiments.

3. The moon shape phenomenon

3.1. Finding 1: nonlinear correlation of ML performance across data subpopulations

Out-of-distribution vs. in-distribution Figure 1 show the nonlinear correlation between the out-of-distribution performance and the in-distribution performance across multiple subpopulation shifts datasets. Moreover, datasets constructed with spurious correlations (Figure 1 top) seems to have more nonlinear correlations compared to the datasets without obvious spurious correlations (Figure 1 bottom).

Majority vs. Minority To understand this phenomenon, we decompose the model's performance into performance on each subpopulation. As shown in Figure 2, there is a *"moon shape" correlation* (parabolic uptrend curve) between the test performance on the *majority subpopulation* and the *minority subpopulation*. We show that this nonlinear correlation holds across model architectures, training durations and hyperparameters (Supp. Figure 8). It also holds for both ImageNet-pretrained or train-from-scratch (Supp. Figure 6).

Moreover, datasets with spurious correlations (Top row of Figure 1) again show more nonlinearity compared to those without obvious spurious correlations (Bottom row of Figure 1). Importantly, for datasets with spurious correlations, even with probit-transformed axes (Miller et al., 2021) as used by several prior works (Recht et al., 2019; Taori et al., 2020), the performance correlations still remain nonlinear (Supp. Figure 7). This confirms that such nonlinear correlations are indeed not captured by the previous work.

Discussion: Why the moon shape is not obvious Figure 3 demonstrates one reason why the non-linear correlation structure (i.e., the moon shape) is non-trivial. Consider a thought experiment in which we interpolate two models A, B (indicated by red circles) picked from the moon shape

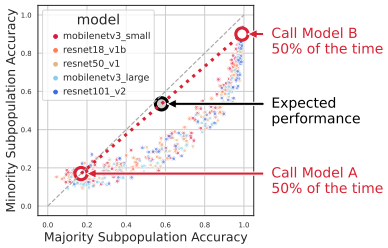


Figure 3. **Why the moon shape is not obvious. Mixture of models can fill in the moon shape.**

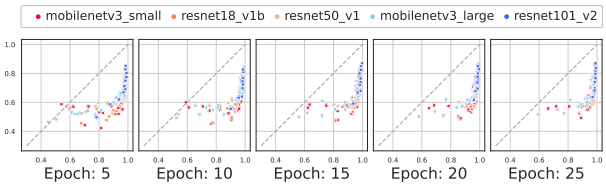


Figure 4. **The moon shape persists across different training epochs.** We stratify Figure 2 based on the number of training epochs. The x-axis indicates majority subpopulation performance. The y-axis indicates minority subpopulation performance. Most of the models have converged after 10 epochs. The moon shape is apparent in each snapshot and persists across training epochs.

curve by flipping a biased coin with probability p: If the coin lands head up, classify with model A. Otherwise classify with model B. Varying p in $[0, 1]$ gives a line between model A and model B. This thought experiment demonstrates that the interpolation line is an achievable region for the ML models, but the models deviate substantially away from this interpolation line, forming a moon shape.

Discussion: The moon shape persists within and across different training epochs In Figure 4, we stratify Figure 2 (a) by the number of training epochs. For each fixed training epoch, we still find a clear moon shape across the different models. Moreover, the similar moon shape persists across different training epochs. Results on other datasets are also similar (Supp. Figure 9). This finding motivates us to focus our analysis on comparing across *different models* rather than comparing the subpopulation performance of a single model across training epochs (which is an interesting direction complementary to our scope).

3.2. Finding 2: spurious correlation makes the moon shape more nonlinear

Setup To explore the effect of different level of spurious correlations, we conduct multiple controlled experiments on Modified-CIFAR4 V1 (Rolf et al., 2021), which was created by subsetting to the bird, car, horse, and plane classes from CIFAR-10. The task is to predict whether the image subject moves primarily by air (plane/bird) or land (car/horse), which is spuriously correlated with whether the image contains an animal (bird/horse) or vehicle (car/plane). The majority subpopulation is the vehicles: "land-vehicle(automobile)", "air-vehicle(airplane)". The minority subpopulation is the animals: "land-animal(horse)", "air-animal(bird)".

We use the same test sets and only change the training data. For the training data, we fixed the size of majority subpopulation (6,000 images) and minority subpopulation (4,000 images). Meanwhile, we also ensure that the dataset is class-balanced: i.e., 5,000 images for both $Y = 0$ and $Y = 1$. Formally, since we fix (1) $P(Y = 1) = 0.5$, (2) $P(Z = 1) = 0.6$, and (3) the total size of the training set as $10,000$ images, there is effectively only one degree of freedom left, which we vary to change the level of spurious correlation. See the theory section for the exact formula. Intuitively, we increase the level of the spurious correlation by adding more "land-vehicle(automobile)", "air-animal(bird)" (indicated by the red boxes in Figure 5), and removing "land-animal(horse)", "air-vehicle(airplane)" (gray boxes).

Results and Observations The four scatter plot panels in Figure 5 are ordered by increasingly stronger spurious correlations. As can be seen, interestingly, the performance correlations become more nonlinear when there is a stronger spurious correlation in the training data. Altogether, these results indicate that *the existence of spurious correlation* plays a crucial role in shaping how the out-of-distribution performance correlates with the in-distribution performance, which is underexplored in the previous literature. Motivated by our experimental findings, we provide rigorous theoretical explanations in the next section about how the spurious correlation and the subpopulation accuracy gap are related.

4. Theoretical analysis of the accuracy gap across subpopulations

We rigorously study the effect of the spurious correlation on the accuracy gap between the majority and the minority subpopulations in a binary classification setting.

We denote an input and an output random variable by X and Y, respectively. We denote a random variable for a subpopulation by Z, where $Z = 1$ indicates the majority subpopulation, otherwise $Z = 0$. We assume that the underlying data generating mechanism is $Z \leftarrow Y \rightarrow X$, that is X and Z are conditionally independent given Y, *i.e.*, $X \perp Z \mid Y$. We assume

$$\mathbb{P}(Z = 1 \mid Y = 1) = \pi_1, \quad \mathbb{P}(Z = 1 \mid Y = 0) = \pi_0.$$

Note that $\pi_1 = \pi_0$ is a necessary and sufficient condition for $Y \perp Z$. In this respect, the level of spurious correlation can be expressed as $|\pi_1 - \pi_0|$. With these notations, the subpopulation accuracy gap for a model g is defined as the absolute difference between the two subpopulation accuracy:

$$\Big|\mathbb{E}[\mathbb{1}(Y = g(X)) \mid Z = 1] - \mathbb{E}[\mathbb{1}(Y = g(X)) \mid Z = 0]\Big|.$$

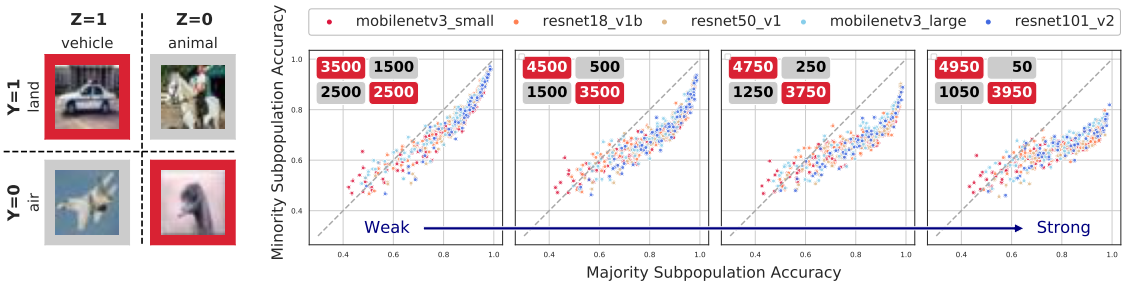


Figure 5. **Stronger spurious correlation creates more nonlinear performance correlation. Left:** We vary the level of the spurious correlation between the classification target label (air/land) and spurious feature (vehicle/animal). Y is the class label. $Z = 1$ indicates the majority subpopulation while $Z = 0$ indicates the minority subpopulation. **Right: As indicated by the blue arrow from left to right, the performance correlations become more nonlinear when there is a stronger spurious correlation in the training data.**

In the following theorem, we explicitly show that the subpopulation accuracy gap is proportional to the level of spurious correlation $|\pi_1 - \pi_0|$. The proof is deferred to the Appendix.

Theorem 1 (The higher the level of spurious correlation, the larger the accuracy gap)**.** *The subclass accuracy gap for a binary classifier g is expressed as follows.*

$$\text{Accuracy Gap} = \frac{\mathbb{P}(Y=1)\mathbb{P}(Y=0)}{\mathbb{P}(Z=1)\mathbb{P}(Z=0)}|\pi_1 - \pi_0|\,|\mathrm{TPR} - \mathrm{TNR}|\,,$$

where TPR *and* TNR *denote the true positive rate* $\mathbb{E}(g(X) = 1 \mid Y = 1)$ *and the true negative rate* $\mathbb{E}(g(X) = 0 \mid Y = 0)$*, respectively.*

Theorem 1 shows that the subpopulation accuracy gap is expressed as a function of $|\pi_1 - \pi_0|$ and $|\mathrm{TPR} - \mathrm{TNR}|$. A direct consequence is that the accuracy gap gets larger when the level of spurious correlation $|\pi_1 - \pi_0|$ increases. It is possible to keep $\mathbb{P}(Z = 1)$ and $\mathbb{P}(Y = 1)$ as constants while $|\pi_1 - \pi_0|$ changes. In particular, it occurs when π_1 and π_0 are related as $\pi_1 = (\mathbb{P}(Z = 1) - \mathbb{P}(Y = 0)\pi_0)/\mathbb{P}(Y = 1)$, which captures the setting of Figure 5.

Remark 1 (Models on the similar ROC curve)**.** *Suppose that there is a trained binary classification model and its ROC curve is not a straight line, which is typically the case. We can think of different points on the curve as different models whose predicted probability outputs are only different by constant shifts. Given that a point on the ROC curve is described as* $(1 - \mathrm{TNR}, \mathrm{TPR})$*,* TPR *changes nonlinearly with respect to* TNR*. Hence, the* $|\mathrm{TPR} - \mathrm{TNR}|$ *changes nonlinearly, and so does the accuracy gap by Theorem 1. This can provide one explanation for our experimental observations that different models form the moon shape curve.*

The setting considered in this remark is admittedly simplified to provide some intuition. In practice, different models (with different architectures and hyperparameters) may not correspond to different points on one ROC curve. However, if the different models do approximately trace out an ROC curve, then the intuition here can apply.

5. Discussion

This work demonstrates that the performance of different models on the majority and minority data subpopulations can have nonlinear correlations. We show that this nonlinear correlation phenomenon is persistent across different datasets, different types of models, and both within and across training epochs. This intriguing phenomenon also leads to nonlinear correlations in models' performances in subpopulation shifts. Our findings complement and contrast previous empirical studies showing a linear correlation in model performance during other types of distribution shifts. Our experiments and theory also provide insights into how spurious correlations in the data can increase this nonlinear pattern.

Our finding has implications for model selection. As ML model building is becoming increasingly turn-key with technologies such as automated machine learning (AutoML), selecting the best model that performs well across diverse data subpopulations is increasingly a major challenge. Our results suggest that, when there is no spurious correlation, models with higher aggregate performance (which is largely skewed by the majority subpopulation performance) generally also perform better on the minority subpopulation. However, with spurious correlation, the situation becomes more nuanced: there exists a phase transition point with negative correlation before, and positive correlation after. In settings where subpopulations performance is important (e.g. fairness considerations), we recommend autoML practitioners to use similar type of scatter plots as our Figure 2 to diagnose model selection.

More generally, subpopulation shift is ubiquitous in ML applications. Our work highlights how model improvement in one subpopulation may have nonlinear effects on performance in other subpopulations. Further analysis and understanding of this nonlinear pattern is an important direction of future work.

References

Andreassen, A., Bahri, Y., Neyshabur, B., and Roelofs, R. The evolution of out-of-distribution robustness throughout fine-tuning. *CoRR*, abs/2106.15831, 2021.

AutoGluon. AutoGluon: AutoML for Text, Image, and Tabular Data. https://auto.gluon.ai/, 2022.

Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., McConnell, M. V., Percha, B., Snyder, T. M., and Dudley, J. T. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2(1):1–10, April 2019.

Blodgett, S. L., Green, L., and O'Connor, B. T. Demographic dialectal variation in social media: A case study of african-american english. In *EMNLP*, pp. 1119–1130. The Association for Computational Linguistics, 2016.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 2018.

Daneshjou, R., Vodrahalli, K., Liang, W., Novoa, R. A., Jenkins, M., Rotemberg, V. M., Ko, J. M., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Zou, J., and Chiou, A. S. Disparities in dermatology ai: Assessments using diverse clinical images. *ArXiv*, abs/2111.08006, 2021.

DeGrave, A. J., Janizek, J., and Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, May 2021.

Eyuboglu, S., Varma, M., Saab, K. K., Delbrouck, J.-B., Lee-Messer, C., Dunnmon, J., Zou, J., and Re, C. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=FPCMqjI0jXN.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. Predictably unequal? the effects of machine learning on credit markets. *Regulation of Financial Institutions eJournal*, 2017.

Grother, P. J., Grother, P. J., Phillips, P. J., and Quinn, G. W. *Report on the evaluation of 2D still-image face recognition algorithms*. Citeseer, 2011.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1934–1943. PMLR, 2018.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR (Poster)*. OpenReview.net, 2019.

Jurgens, D., Tsvetkov, Y., and Jurafsky, D. Incorporating dialectal variability for socially equitable language identification. In *ACL (2)*, pp. 51–57. Association for Computational Linguistics, 2017.

Koenecke, A., Nam, A. J. H., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117:7684 – 7689, 2020.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *CVPR*, pp. 2661–2671. Computer Vision Foundation / IEEE, 2019.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551, 2017.

Liang, W. and Zou, J. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=MTex8qKavoS.

Lu, S., Nott, B., Olson, A., Todeschini, A., Vahabi, H., Carmon, Y., and Schmidt, L. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020. http://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-101.pdf.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6905–6916. PMLR, 2020.

Miller, J., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7721–7735. PMLR, 2021.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Re, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pp. 151–159, New York, NY, USA, 2020. Association for Computing

Machinery. ISBN 9781450370462. doi: 10.1145/3368555.3384468. URL https://doi.org/10.1145/3368555.3384468.

Pfohl, S. R., Zhang, H., Xu, Y., Foryciarz, A., Ghassemi, M., and Shah, N. H. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *Scientific reports*, 12(1):1–13, 2022.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do CIFAR-10 classifiers generalize to cifar-10? *CoRR*, abs/1806.00451, 2018.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019.

Rolf, E., Worledge, T. T., Recht, B., and Jordan, M. I. Representation matters: Assessing the importance of subgroup allocations in training data. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9040–9051. PMLR, 2021.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

Santurkar, S., Tsipras, D., and Madry, A. BREEDS: benchmarks for subpopulation shift. In *ICLR*. OpenReview.net, 2021.

Sapiezynski, P., Kassarnig, V., and Wilson, C. Academic performance prediction in a gender-imbalanced environment. In *FATREC Workshop on Responsible Recommendation*, 2017.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020.

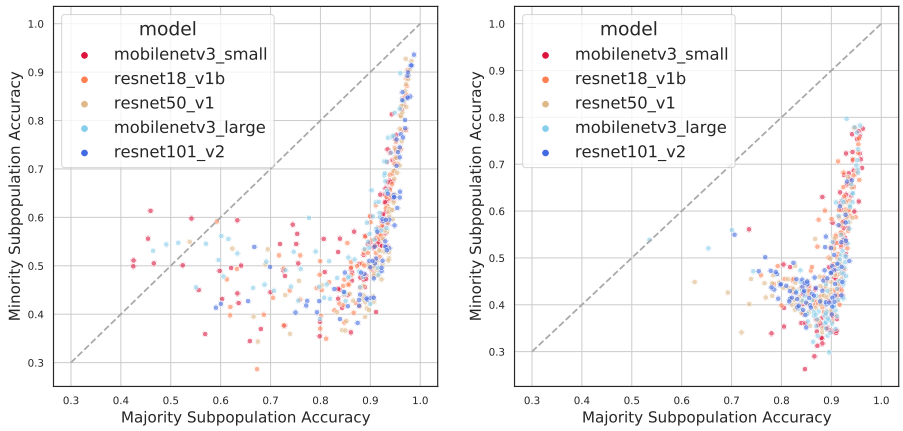
Tatman, R. Gender and dialect bias in youtube's automatic captions. In *EthNLP@EACL*, 2017.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

Yadav, C. and Bottou, L. Cold case: The lost MNIST digits. In *NeurIPS*, pp. 13443–13452, 2019.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

A. Extended Descriptions of The Moon Shape Phenomenon



(a) Modified-CIFAR4 (V2) ImageNet-pretrained

(b) Modified-CIFAR4 (V2) Trained from scratch

Figure 6. **The moon shape exists for both models pretrained on ImageNet and models trained from scratch.** Since models in other figures are all fine-tuned starting from ImageNet pre-trained checkpoints, we add an experiment of training from scratch. The moon shape is even more obvious for models trained from scratch. This shows that the moon shape is not an artifact of ImageNet pre-trained checkpoints, but a much broader phenomenon. Similarly, we also verify that this nonlinear correlation holds across model architectures, training durations and hyperparameters (Supp. Figure 8), and the imbalance between subpopulations (Figure 5).

A.1. Extended descriptions of the nonlinear correlation under probit-transformation

In Figure 2, we decompose the model's performance into performance on each subpopulation. We found that there is a *"moon shape" correlation* (parabolic uptrend curve) between the test performance on the *majority subpopulation* and the *minority subpopulation*. This nonlinear correlations hold across model architectures, training durations and hyperparameters (Supp. Figure 8), and the imbalance between subpopulations (Figure 5). We have also found that datasets with spurious correlations (Top row of Figure 2) show more nonlinearity compared to those without obvious spurious correlations (Bottom row of Figure 2).

Recent empirical works find that there is a strong *linear* relationship between in-distribution (ID) and out-of-distribution (OOD) performance under several specific types of data shifts. Most works report the linear relationship with the original performance axis-scale, while several works report that with *probit-transformed* (a non-linear transformation) accuracy (Miller et al., 2021; Recht et al., 2019; Taori et al., 2020), the linear correlation of performance would be more precise. Here the probit transform is the inverse of the cumulative density function (CDF) of the standard Gaussian distribution, i.e., $l_{\text{transformed}} = \Phi^{-1}(l)$.

Therefore, to better compare with the existing literature, we also plot with *probit*-transformed performance as shown in Supp. Figure 7. Importantly, for datasets with spurious correlations, even with probit-transformed axes as used by several prior works (Miller et al., 2021; Recht et al., 2019; Taori et al., 2020), the performance correlations still remain nonlinear (Supp. Figure 7). The fact that ours remain nonlinear with probit-transformed axes (Supp. Figure 7) is an especially exciting finding, since it confirms that the nonlinear correlations we found here under subpopulation shifts are indeed not captured by the previous work. Our findings indicate that the linear correlation trend reported by recent work is not necessarily true if there are subpopulation shifts with spurious correlation. As the presence of spurious correlation creates such an interesting phenomenon, this also motivates our theoretical analysis on the relationship between the level of spurious correlation and the moon shape.

A.2. Extended descriptions on the moon shape within and across different training epochs

In Figure 4, we stratify the dots in Figure 2 (a) (i.e. the trained models) by the number of training epochs. For each fixed training epoch, we still find a clear moon shape across the different models. Moreover, the similar moon shape persists across different training epochs. In other words, *the moon shape persists within and across different training epochs*.

Supp. Figure 9 verify that similar results are also shown when we stratify the scatter plot dots by the number of training

epochs on other datasets. This finding motivates us to focus our analysis on comparing across *different models* rather than comparing the subpopulation performance of a single model across training epochs (which is an interesting direction complementary to our scope).

B. Extended Related Work

Linear correlations between ID and OOD performances Existing research mostly reports *linear* correlations between ID and OOD performances. The linear correlations were first reported in recent dataset reconstruction settings including ImageNet-V2 (Recht et al., 2019), CIFAR-10.1 (Recht et al., 2018), CIFAR-10.2 (Lu et al., 2020), where new test sets of popular benchmarks are collected closely following the original dataset creation process. As there are subtle differences in the dataset creation pipeline, the test performance on the new test set is often lower, but appears to be linearly correlated with the performance on the original test set (Lu et al., 2020; Miller et al., 2020; Recht et al., 2018; 2019; Yadav & Bottou, 2019). Later researchers also found the linear trends in the context of cross-benchmark evaluation (Taori et al., 2020; Miller et al., 2021), and transfer learning (Kornblith et al., 2019; Andreassen et al., 2021), where a model's ImageNet test accuracy linearly correlates with the transfer learning accuracy. Similar linear trends are also observed in sub-type shifts (Hendrycks & Dietterich, 2019; Santurkar et al., 2021) (e.g., the training data for the "dog" class are all from a specific breed while the test data come from another breed). Different from these studies, we (1) focus on subpopulation shifts, where we also present the first systematic study on the performance correlation between data subpopulations, and (2) find *nonlinear* correlations of ML performance across data subpopulations, which is not captured in previous work. Importantly, we show that for datasets with spurious correlations, even with probit-transformed axes as used by several prior work (Miller et al., 2021; Recht et al., 2019; Taori et al., 2020), the performance correlations still remain nonlinear. This confirms that the nonlinear ("moon shape") correlation phenomenon is indeed not captured by previous work.

ML with diverse subpopulations A major challenge in ML is that a model can have very disparate performances even when it's applied to different subpopulations of its training and evaluation data. Models with low average error can still fail on particular groups of data points (Hashimoto et al., 2018; Buolamwini & Gebru, 2018; Blodgett et al., 2016). For example, predictive models for clinical outcomes that are accurate on average in a patient population are reported to underperform drastically for some subpopulations, potentially introducing or reinforcing inequities in care access and quality (Pfohl et al., 2022). Similar performance disparity, have also been observed in radiograph classification (Badgeley et al., 2019; Zech et al., 2018; DeGrave et al., 2021), face recognition (Grother et al., 2011; Buolamwini & Gebru, 2018), speech recognition (Koenecke et al., 2020; Blodgett et al., 2016; Jurgens et al., 2017), academic recommender systems (Sapiezynski et al., 2017)), and automatic video captioning (Tatman, 2017), among others. Worse, as model accuracy affects user retention, the minority group might shrink and thus even amplifies the performance disparity over time (Hashimoto et al., 2018; Fuster et al., 2017). These case studies highlight the importance of understanding the ML performance disparity across subpopulations.

C. Proofs

In this section, we provide a proof of Theorem 1.

Proof of Theorem 1. For any $z \in \{0, 1\}$, we have

$$\begin{aligned}
\mathbb{E}[\mathbb{1}(Y = g(X)) \mid Z = z] &= \sum_{y=0}^{1} \mathbb{E}[\mathbb{1}(y = g(X)) \mid Z = z, Y = y]\mathbb{P}(Y = y \mid Z = z) \\
&= \sum_{y=0}^{1} \mathbb{E}[\mathbb{1}(y = g(X)) \mid Y = y]\mathbb{P}(Y = y \mid Z = z) \\
&= \mathrm{TPR} \times \mathbb{P}(Y = 1 \mid Z = z) + \mathrm{TNR} \times \mathbb{P}(Y = 0 \mid Z = z).
\end{aligned}$$

Here, the second equality is due to $X \perp Z \mid Y$. Therefore, the accuracy gap between the two subpopulations is expressed as follows.

$$\begin{aligned}
\text{Accuracy Gap} &= \left|\mathbb{E}[\mathbb{1}(Y = g(X)) \mid Z = 1] - \mathbb{E}[\mathbb{1}(Y = g(X)) \mid Z = 0]\right| \\
&= \Big|\mathrm{TPR} \times \big(\mathbb{P}(Y = 1 \mid Z = 1) - \mathbb{P}(Y = 1 \mid Z = 0)\big) \\
&\quad + \mathrm{TNR} \times \big(\mathbb{P}(Y = 0 \mid Z = 1) - \mathbb{P}(Y = 0 \mid Z = 0)\big)\Big| \\
&= \left|\mathbb{P}(Y = 1 \mid Z = 1) - \mathbb{P}(Y = 1 \mid Z = 0)\right| \times \left|\mathrm{TPR} - \mathrm{TNR}\right|.
\end{aligned}$$

By the Bayes' theorem

$$\mathbb{P}(Y = 1 \mid Z = 1) = \frac{\pi_1 \mathbb{P}(Y = 1)}{\mathbb{P}(Z = 1)}, \quad \mathbb{P}(Y = 1 \mid Z = 0) = \frac{(1 - \pi_1)\mathbb{P}(Y = 1)}{\mathbb{P}(Z = 0)},$$

we have

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid Z = 1) - \mathbb{P}(Y = 1 \mid Z = 0) &= \frac{\pi_1 \mathbb{P}(Y = 1)}{\mathbb{P}(Z = 1)} - \frac{(1 - \pi_1)\mathbb{P}(Y = 1)}{\mathbb{P}(Z = 0)} \\
&= \frac{\pi_1 \mathbb{P}(Y = 1)\mathbb{P}(Z = 0) - (1 - \pi_1)\mathbb{P}(Y = 1)\mathbb{P}(Z = 1)}{\mathbb{P}(Z = 1)\mathbb{P}(Z = 0)} \\
&= \frac{\big\{\mathbb{P}(Z = 0) - (1 - \pi_1)\big\}\,\mathbb{P}(Y = 1)}{\mathbb{P}(Z = 1)\mathbb{P}(Z = 0)}.
\end{aligned}$$

Since $\mathbb{P}(Z = 0) = 1 - (\pi_1 \mathbb{P}(Y = 1) + \pi_0 \mathbb{P}(Y = 0)) = 1 - \pi_1 + (\pi_1 - \pi_0)\mathbb{P}(Y = 0)$, we have

$$\text{Accuracy Gap} = \frac{\mathbb{P}(Y = 1)\mathbb{P}(Y = 0)}{\mathbb{P}(Z = 1)\mathbb{P}(Z = 0)} |\pi_1 - \pi_0| \times \left|\mathrm{TPR} - \mathrm{TNR}\right|.$$

It concludes a proof. □

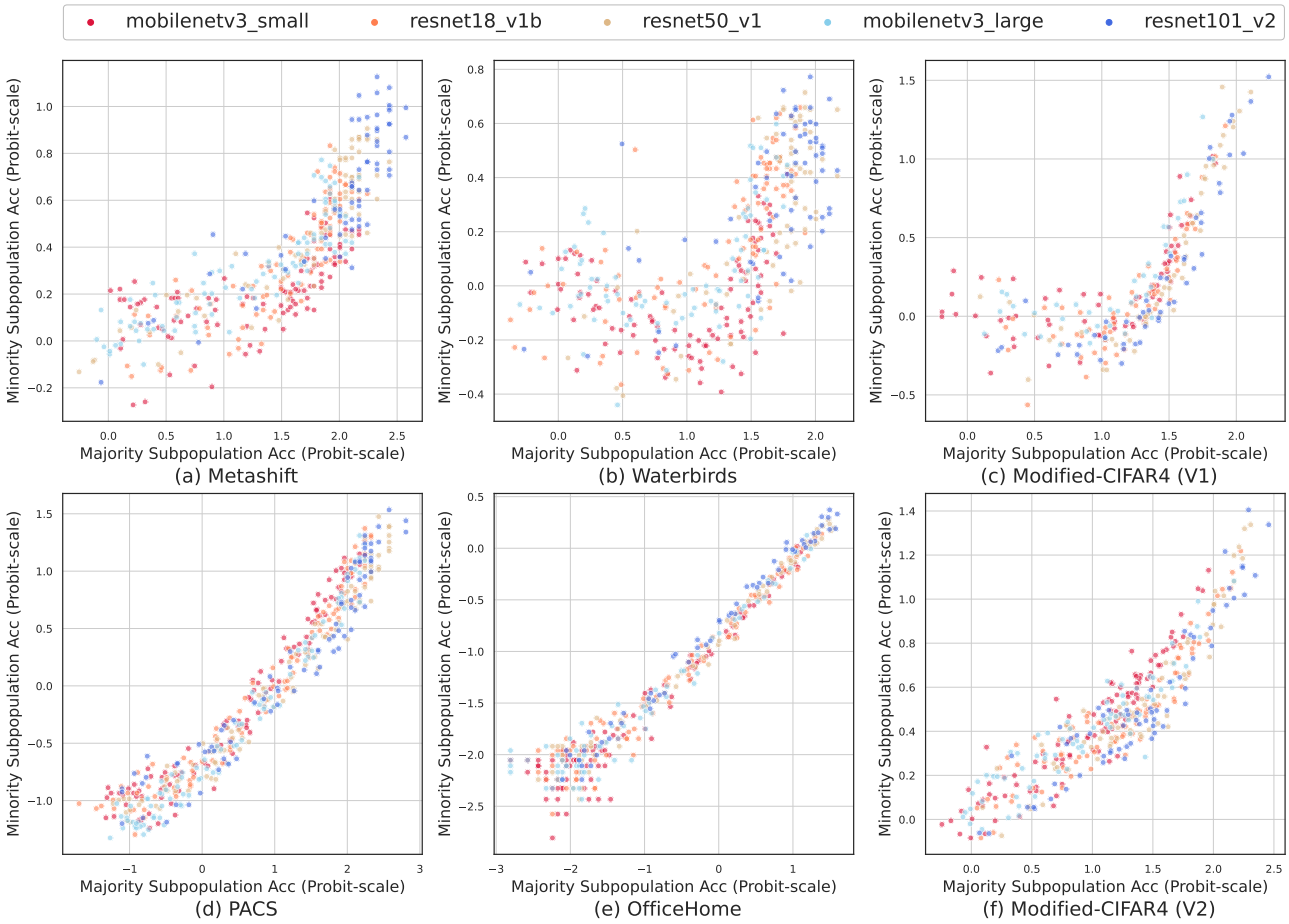


Figure 7. **Majority subpopulation accuracies vs. minority subpopulation accuracies (Figure 2) plotted in probit-transformed axes. (a-c) Results on three datasets with explicit spurious correlation by dataset construction.** The correlation between the majority subpopulation performance and the minority subpopulation performance is still *non-linear* even after the probit-transform of axes. This finding significantly broadens the literature, where previous work mostly reports a precise linear trend between the out-of-distribution performance and in-distribution performance in probit scale. Based on prior research, one might expect the majority subpopulation performance and the minority subpopulation performance to be linearly correlated. In contrast, we show that the majority subpopulation performance and the minority subpopulation performance can be correlated in a non-linear way even in probit scale. **(d-f) Results on three datasets without explicit spurious correlation.** Without spurious correlation, the majority subpopulation performance and the minority subpopulation performance appears to be in a linear trend, which is aligned with the previous literature. All together, these results indicate that *the existence of spurious correlation* plays an crucial role in shaping how the out-of-distribution performance correlates with the in-distribution performance, which is largely ignored by the previous literature.

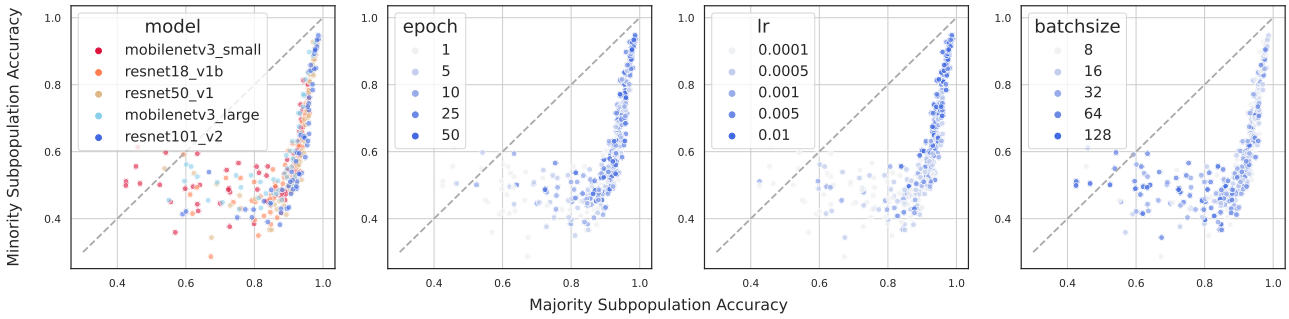


Figure 8. **The strong correlations hold across model architectures, training durations and hyper parameters.**

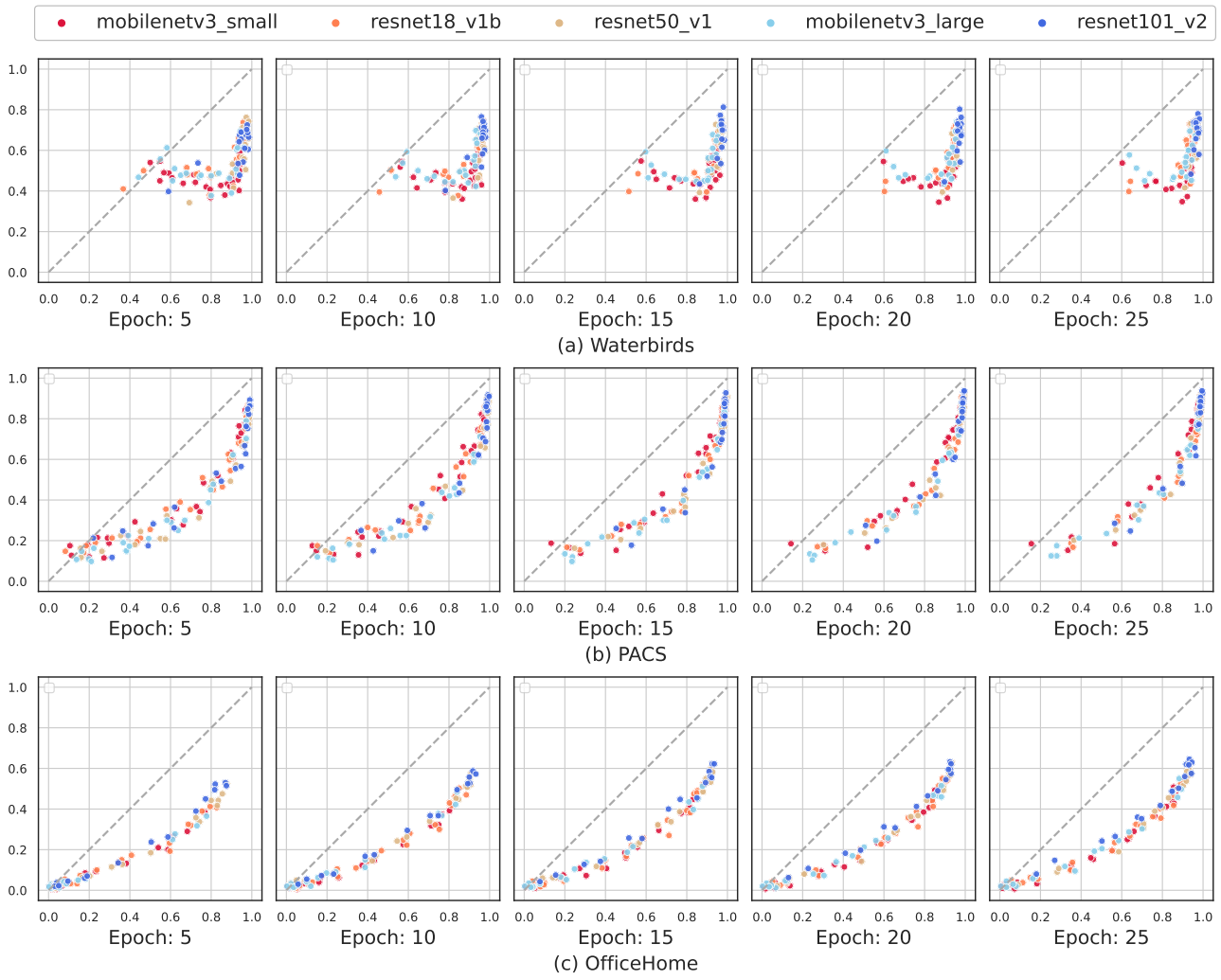


Figure 9. **The moon shape persists across different training epochs. Results on other datasets similar to Figure 4.** We stratify Figure 2 based on the number of training epochs. The x-axis indicates majority subpopulation performance. The y-axis indicates minority subpopulation performance. Most of the models have converged after 10 epochs. The moon shape is apparent in each snapshot and persists across training epochs.