Benchmarking Robustness to Natural Distribution Shifts for Facial Analysis

Jessica Deuschel[1*], **Andreas Foltyn**[1*], **Leonie Anna Adams**[1], **Jan Maximilian Vieregge**[1], **Ute Schmid**[2]

[1]Smart Sensing and Electronics , Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany,
{jessica.deuschel, andreas.foltyn}@iis.fraunhofer.de
[2]Cognitive Systems, University of Bamberg, Germany, ute.schmid@uni-bamberg.de

Abstract

During the deployment of machine learning models, performance degradation can occur compared to the training and validation data. This generalization gap can appear for a variety of reasons and be particularly critical in applications where certain groups of people are disadvantaged by the outcome, e.g. facial analysis. Literature provides a vast amount of methods to either perform robust classification under distribution shifts or at least to express the uncertainty caused by the shifts. However, there is still a need for data that exhibit different natural distribution shifts considering specific subgroups to test these methods. We use a balanced dataset for facial analysis and introduce subpopulation shifts, spurious correlations, and subpopulation-specific label noise. This forms our basis to investigate to what extent known approaches for calibrating neural networks remain reliable under these specified shifts. Each of the modifications leads to performance degradation, but the combination of ensembles and temperature scaling is particularly useful to stabilize the calibration over the shifts.

1 Introduction

Machine Learning is a central tool for many tasks in the area of computer vision and facial analysis. However, most approaches are evaluated on data that are identically and independently distributed (i.i.d.). This i.i.d. assumption often cannot be guaranteed during deployment. Thus, a relatively wide range of possible shifts between training and inference can occur that have a detrimental effect on the generalization performance. This can be particularly important in areas where predictions of machine learning systems have a direct effect on humans, as it is the case with facial analysis. Subpopulations may be underrepresented in the training data compared to the overall population [9, 16], or there may be spurious correlations between subpopulations and labels during training that have no real causal relationship [1, 15]. Since this poses a significant challenge to the development of real-world machine learning and pervades all areas and applications of machine learning, efforts have been made to mitigate this problem [1, 15, 4, 12, 11]. Furthermore, it is also desirable to get well-calibrated uncertainty estimates, that are reliable across shifts such that we can at least determine when to trust the predictions of the model [13, 3]. In general, however, we need data sets that go beyond the classic train/validation splits from i.i.d data to evaluate these methods. Thus, several benchmarks have been presented to evaluate different shifts across various domains [9, 16]. For facial analysis, Sagawa et al. [15] generate spurious correlations using annotated attributes in CelebA. However, to the best of our knowledge, no structured benchmark currently exists to evaluate different types of naturally occurring shifts in the domain of facial analysis.

[*] Both authors contributed equally to this research.

Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Consequently, our first contribution is a framework to systematically create various subpopulation-specific shifts on facial data, depicted in Figure 1: underrepresented/missing subpopulations, spurious correlations, and group-specific label noise. For this, we use FairFace [7], a balanced dataset in terms of age, race and gender, as a base dataset. Furthermore, we want to address the following research question: What impact do these shifts have on the subpopulation-specific accuracy and to what extent do they affect the calibration? To answer this question, we compare a baseline model with several known approaches for building well-calibrated models. For reproducibility, we provide our code for creating the shifts and training.[1]

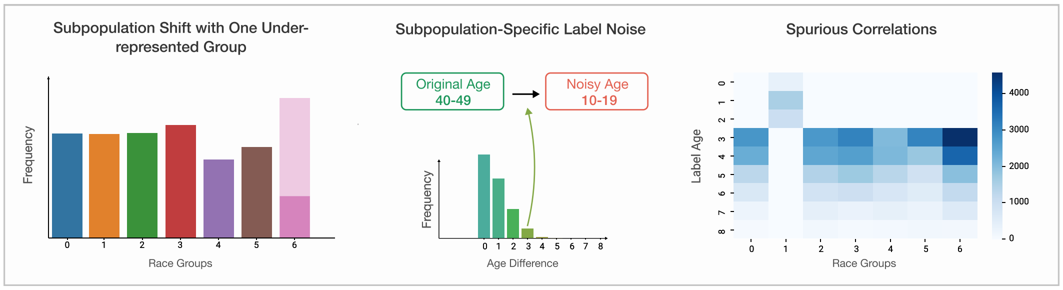


Figure 1: Overview of the introduced shifts, from left to right: subpopulation shift with one under-represented race value (here only a fraction $q^{\text{race}} = 0.25$ of race group 6 remains), group specific label noise making the subjects younger by a value z that follows a half-normal distribution with increasing standard deviation (here $\sigma = 1.5$), and spurious correlations between *young* and one race group (here equivalence-condition between *young* and race group 1).

2 Distribution shifts in facial data

In the following we describe the underlying data and modifications.

Baseline Data As a basis for our work we use the FairFace dataset [7] with 97,698 subjects and its standard train/test split. Additionally, we split off 12% of the training data as a validation set. The dataset contains information on age, gender, and race, with special attention paid to the balance of these three attributes. Following the notation of [7] the attribute *gender* is categorized into two classes ($A^{\text{gender}} = \{male, female\}$), *race* into seven ($A^{\text{race}} = \{$*Black*, *East Asian*, *Indian*, *Latino Hispanic*, *Middle Eastern*, *Southeast Asian*, *White*$\}$), and *age* into nine ($A^{\text{age}} = \{$ *0-2*, *3-9*, *10-19*, *20-29*, *30-39*, *40-49*, *50-59*, *60-69*, *more than 70*$\}$).[2] We attempt to estimate the age categories $y \in A^{\text{label}} = \{0, 1, \ldots, 8\}$ by classification, while A^{gender} and A^{race} are used to induce changes in the data distribution of the training data. The test data remains the same. We systematically apply the shifts to each attribute value so that each of them is affected once. Our goal is to examine the effects of several types of shifts without the influence of the number of training samples. Therefore, we ensure that although we reduce the number of subjects from certain groups, the total number of training samples per shift remains constant. We achieve this through a general random sampling, also on the baseline data.

Subpopulation shift with one under-represented attribute value Subpopulation shifts can occur in the real-world when certain attribute values are underrepresented in the training data compared to the data seen during deployment, e.g. when a specific ethnic group is missing during training. Let $v^i \in A^i$ be an attribute value for attributes $i \in \{race, gender\}$ and let $p_{\texttt{clean}}(v^i)$ be the probability of v^i in our clean training data. To induce the subpopulation shift, we reduce the probability of this attribute value by a factor $0 \leq q^i \leq 1$: $p_{\texttt{shift}}(v^i) = q^i \cdot p_{\texttt{clean}}(v^i)$. For the remaining attribute values $w^i \in A^i \setminus v^i$ we set the probability distribution to $p_{\texttt{shift}}(w^i) = p_{\texttt{clean}}(w^i) + \frac{p_{\texttt{clean}}(v^i) - p_{\texttt{shift}}(v^i)}{|A^i| - 1}$. This is done for all $v^i \in A^i$ and all shift intensities $q^{\text{race}} \in \{1.0, 0.75, 0.5, 0.25, 0.0\}$ and $q^{\text{gender}} \in \{1.0, 0.5, 0.0\}$. Note that the selected attribute value does not occur at the strongest shift.

[1] https://github.com/jdeuschel/DistrShiftsOnFacialData

[2] Note that we used race and gender labels as annotated in the FairFace dataset. The visual features used by the annotators are not necessarily indicative of a person's gender identity. In addition, the used labels do not reflect the identity of individuals outside the bounds of this binary categorization.

Label noise In addition to normal subpopulation shifts, there may also be biases in the labels for individual groups. For example, certain races are estimated to be younger than they really are. This can have isolated effects for these specific groups or for all others during the deployment. To investigate this, we introduce attribute-specific label noise in the training data: for each race group separately, we reduce the age class by subtracting samples from an approximated half-normal distribution with mean 0 and increasing standard deviation $\sigma \in \{0.5, 1.0, \ldots, 4.0\}$ from the ground truth labels. Thus for a label $y \in A^{\text{label}}$ the new label is $\tilde{y} = \max(0, \lceil y - z \rceil)$, where $z \sim \mathcal{N}_{[0,\infty)}(0, \sigma^2)$.

Spurious correlations Our last shift consists of spurious correlations between *race* and the label *age*. We combine the lowest three age groups into the attribute value *young*, thus we create a new attribute set $A^{\text{generation}} = \{\textit{young}, \textit{not young}\}$. We systematically correlate the attribute *young* with each race group in the training data. Two variants for this correlation are considered: sufficiency (e.g. *"young* $\Rightarrow$ *East Asian"*) and equivalence (e.g. *"young* $\Leftrightarrow$ *East Asian"*, depicted in Figure 1). We create these variants by omitting subjects with certain *race-age* combinations: For example, for the spurious correlation *"young* $\Rightarrow$ *East Asian"* we omit all young subjects that are not East Asian. In the case of *"young* $\Leftrightarrow$ *East Asian"* we additionally remove all non-young East Asians. Thereby for each $v^{\text{race}} \in A^{\text{race}}$ separately we ensure that $p_{\texttt{shift}}(v^{\text{race}} | v^{\text{generation}} = \textit{young}) = 1$ holds in the training data for the case sufficiency. For the case of equivalence we additionally have $p_{\texttt{shift}}(v^{\text{generation}} = \textit{young} | v^{\text{race}}) = 1$.

For this shift we created a baseline with a reduced number of young subjects over all races to match the overall number of young subjects to isolate the effect of spurious correlations and exclude that of data set size (for *young*).

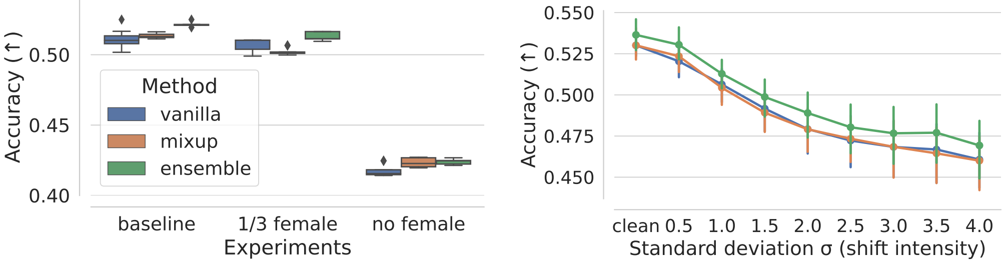
3 Experimental setup

In our experiments we investigate the impact of the distribution shifts specified in section 2 on the classification performance and calibration. For this, we use a ResNet34 [6, 14] pretrained on ImageNet as a baseline model (denoted vanilla). For comparison, deep ensembles [10] of three independently trained networks are used, as they are generally considered a simple but effective way to improve the calibration [13]. Also the influence of mixup [18] is investigated, which can improve the calibration [17] and performs a data-level change compared to the ensembles. Furthermore, we combine each of them with temperature scaling [5], a post-hoc recalibration method that divides the logit outputs by a scaling parameter before calculating the softmax. We use a clean and balanced validation set for learning the scaling parameter. Each of the three approaches offers comparably good performance in their respective categories. Therefore, in this paper we focus only on those.

For each of these methods, we use the same training protocol. We train for 30 epochs using Adam [8] with weight decay of 0.01, learning rate of 0.0001 and apply a random crop of size 224. RandAugment [2] is used for data augmentation. The weights of the epoch with the lowest validation loss are used for the evaluation. The basic hyperparameter setting was chosen based on [7] and then improved to replicate the results on the clean FairFace data. Since ensembles are independently trained networks, the same setting can be applied to them. As mixup generally needs more epochs to converge, it was verified that the loss did not change significantly after the specified epochs. In addition to the classification accuracy, we are particularly interested in the reliability of the confidence outputs, which we measure with the ECE (Expected Calibration Error) [5]. In the evaluation we include the results from 5 random seeds for all metrics and methods.

4 Results

In the following, we try to answer the research question for each of the defined shifts. First, we want to address the question of what impact the previously defined changes in the distributions have on the accuracy. We contrast the changes in the accuracy with the stability of the calibration. This allows us to answer the question of whether our model remains at least reliable even when the accuracy deteriorates. For this purpose, we focus on the relevant subpopulations that are affected by the changes. Thus, for a shift that affects attribute value v_i we consider the accuracy and ECE over all subjects with this specific attribute value. We aggregate these results over all affected attribute values for each intensity level, such that the plots below show the spread over the results of the different relevant attribute values.



(a) Subpopulation shift for *gender female*

(b) Label noise

Figure 2: Accuracy for subpopulation shift and label noise for the relevant subgroups affected by the shift.

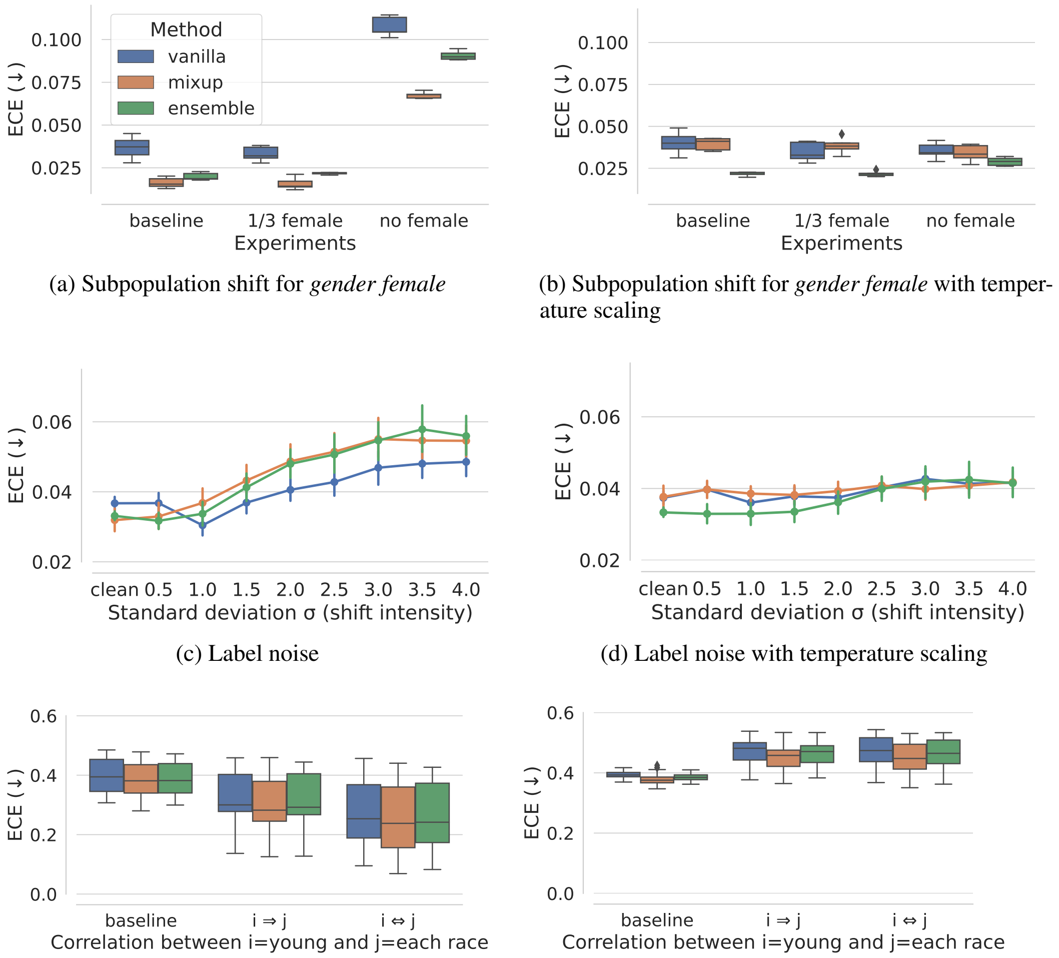
Subpopulation shift For subpopulation shifts we considered both attributes: *gender* (see Figure 2a for the accuracy and 3a and 3b for the ECE) and *race* (see Figure 5 in the Appendix). We can observe that in both cases the accuracy of all methods deteriorates with increasing shift. This effect is more pronounced for the attribute *gender* than for *race* when we compare the strongest shifts. This poses the question of whether the existing race diversity in the dataset is sufficient to learn race-invariant features, making the removal of a single race not very significant. The calibration error increases with a stronger shift. Similarly, as for the accuracy this effect is larger for the shift in *gender*. Among the compared approaches, ensembles and mixup result in an improvement of the calibration over vanilla; with mixup this effect is larger in the gender case. A recalibration causes a slight stabilization of the baseline but leads to a deterioration of mixup.

Label noise In the case of group-specific label noise we see a clear drop of accuracy, by 7.5 percentage points (denoted %) with higher intensity (Figure 2b). However, the calibration error seems to be relatively stable (max. 1.0% increase for vanilla, as can be seen in Figure 3c). Interestingly, the calibration first improves with slight noise. An assumption would be that the otherwise overconfident networks become more uncertain, which leads to an improved calibration. We can observe that ensembles, as well as mixup, have better calibration without label noise, however, under label noise the vanilla network outperforms both of them. On the other hand, ensembles have better accuracy under noise. A recalibration (Figure 3d) reduces the gap between all methods and leads to stabilization over the noise intensities. In this case, ensembles show the lowest ECE. We also evaluated the subpopulations that did not encounter noise during training and observed that the accuracy worsens by only 1.5% for the vanilla network. Therefore, we can conclude that the label noise applied to a single group does not exhibit a similarly strong performance degradation for all other groups.

Spurious correlations We investigate the effect of spurious correlations between *young* and each value of *race*. We distinguish between the race groups that include the attribute *young* in the training data (*young-correlated*) and the remaining ones (*non-young-correlated*). The accuracy increases for the *young-correlated* groups (by 14%) and decreases for the *non-young-correlated* groups (by 7.5%) with stronger spurious correlations (see Appendix A.2), while the spread between the different *young-correlated* observations increases. In terms of calibration, we observe a significant improvement for the *young-correlated* groups and a deterioration for *non-young-correlated* groups (see Figure 3e and 3f). However, we notice that the models are poorly calibrated, thus the uncertainty estimates of the methods are not reliable. As before, temperature scaling improves the calibration. However, unlike for the other shifts, neither ensembles nor mixup can improve the calibration compared to vanilla and no clear ranking can be observed.

5 Conclusion & discussion

In this paper, we introduced several shifts for a given facial analysis dataset and investigated the impact on accuracy and calibration. In summary, each type of distributional change has an increasingly negative impact on accuracy with higher shift intensity. Also the calibration of the methods continues



(a) Subpopulation shift for *gender female*

(b) Subpopulation shift for *gender female* with temperature scaling

(c) Label noise

(d) Label noise with temperature scaling

(e) Spurious correlation *young-correlated*

(f) Spurious correlation *non-young-correlated*

Figure 3: ECE for subpopulation shift, label noise and spurious correlations for the relevant subgroups affected by the shift.

to decline as the shift increases such that the models become less reliable. Ensembles and mixup outperformed vanilla for the subpopulation shift, but for spurious correlations, there has not been a clear ranking and for label noise vanilla even performed better than the other methods. For the case that a balanced validation dataset is available, a simple post-hoc recalibration can improve the ECE significantly for all shifts. It stabilizes the calibration for all methods and leads to an advantage of ensembles over the other methods. However, getting a balanced data set poses difficulties in the wild. We did not examine this effect concerning different modifications of the validation set, thus it remains unclear what significance the specific composition of the validation set might play. We leave this investigation for future work.

6 Acknowledgment

This work was supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy through the Center for Analytics Data Applications (ADA-Center) within the framework of "BAYERN DIGITAL II" (20-3410-2-9-8).

References

[1] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019.

[2] Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, 2020.

[3] Andreas Foltyn and Jessica Deuschel. Towards reliable multimodal stress detection under distribution shift. *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion)*, 2021.

[4] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *ArXiv*, abs/2007.01434, 2021.

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[6] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[9] P. W. Koh, Shiori Sagawa, H. Marklund, Sang Michael Xie, Marvin Zhang, A. Balsubramani, Wei hua Hu, Michihiro Yasunaga, Richard L. Phillips, Sara Beery, J. Leskovec, A. Kundaje, E. Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. *ArXiv*, abs/2012.07421, 2020.

[10] Balaji Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2017.

[11] Zachary Chase Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. *ArXiv*, abs/1802.03916, 2018.

[12] Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, 2021.

[13] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[15] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020.

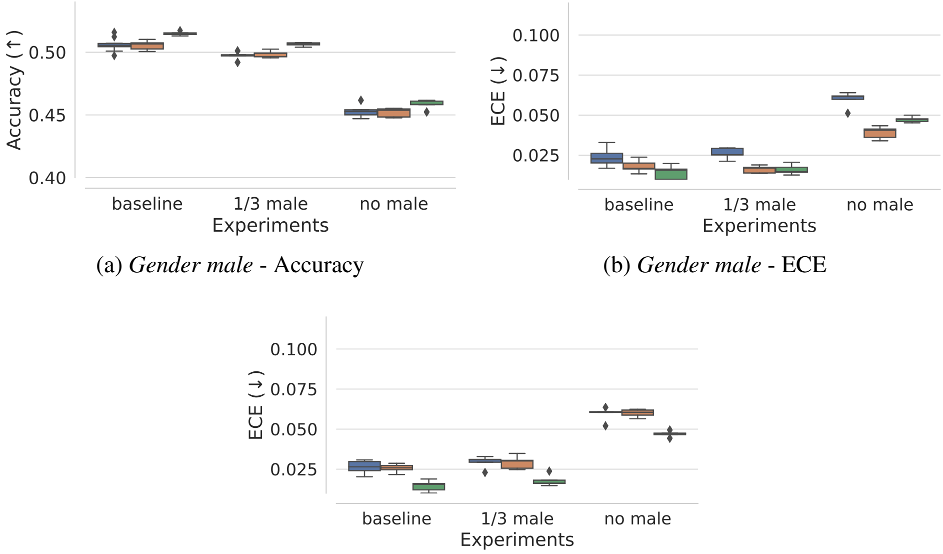
[16] Shibani Santurkar, D. Tsipras, and A. Madry. Breeds: Benchmarks for subpopulation shift. *arXiv: Computer Vision and Pattern Recognition*, 2021.

[17] Sunil Thulasidasan, Gopinath Chennupati, Jeff Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *arXiv preprint arXiv:1905.11001*, 2019.

[18] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2018.

A Additional results

A.1 Subpopulation shift

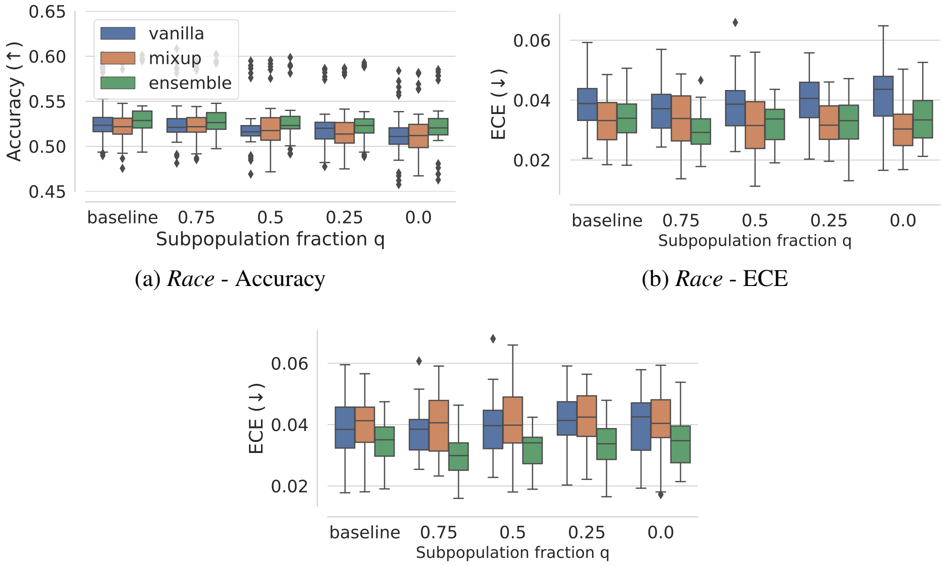


(a) *Gender male* - Accuracy

(b) *Gender male* - ECE

(c) *Gender male* - ECE with temperature scaling

Figure 4: Accuracy and ECE of subpopulation shift for *gender*.



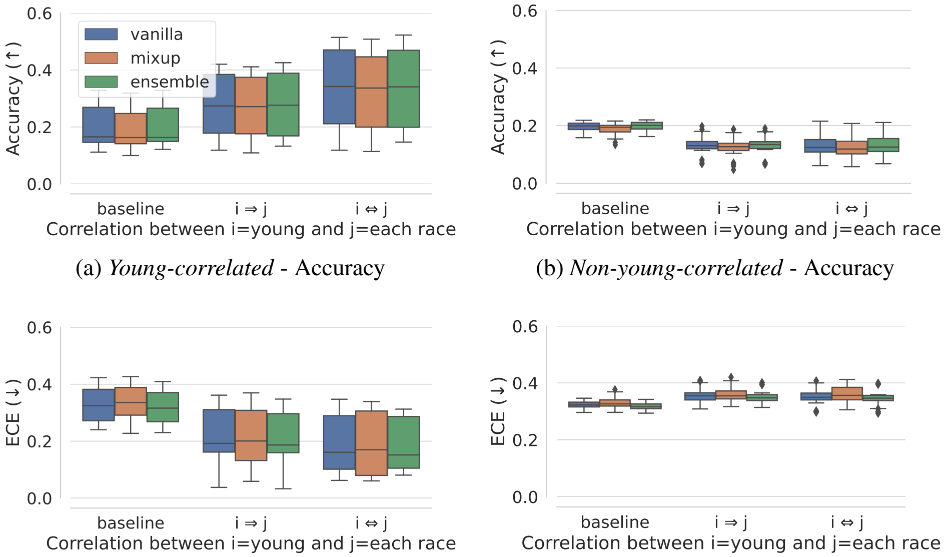
(a) *Race* - Accuracy

(b) *Race* - ECE

(c) *Race* - ECE with temperature scaling

Figure 5: Accuracy and ECE of subpopulation shift for *race*, where only the races affected by the shift respectively are considered.

A.2 Spurious correlations



(a) *Young-correlated* - Accuracy

(b) *Non-young-correlated* - Accuracy

(c) *Young-correlated* - ECE with temperature scaling

(d) *Non-young-correlated* - ECE with temperature scaling

Figure 6: Accuracy and ECE of spurious correlations between *race* and *young*.