Latent Subgroups: Testing the Assumptions Behind GDRO in GEORGE

Tom Siegl, Martin Becker

2026

Abstract

GEORGE is a two-stage method for robustness to hidden subclasses: it clusters the feature space of a model trained by empirical risk minimization (ERM) within each class and then trains a new model with group distributionally robust optimization (GDRO), using the clusters as groups. Both stages rest on assumptions that are rarely tested in isolation. We test two of them. First, we check whether the quality of GEORGE's clusterings predicts how much the GDRO stage improves over ERM. Across ten replications on each of MNIST, Waterbirds, ISIC and CelebA, there is not a clear trend in terms of correlation. Cluster-score gating of the GDRO stage is therefore not supported. Second, we check whether GDRO is a safe objective once a group is intrinsically hard, i.e. when its classes overlap. On a synthetic four-Gaussian problem with a linear model, GDRO finds the same mean decision boundary as ERM, but its decision boundaries vary much more from run to run once the hard groups overlap strongly. When the hard groups carry no class signal at all, GDRO's boundaries become arbitrary. This instability is not explained by lagging group weights, by groups being split by class, or by the optimizer hyperparameters. It follows from GDRO assigning near-zero weight to the easy groups, which then no longer constrain the model.

1 Introduction

Classifiers trained by empirical risk minimization (ERM) can perform well on average while failing on unlabeled, underrepresented subclasses of a class, a problem known as *hidden stratification* [12]. If subclass labels were available, *group distributionally robust optimization* (GDRO) [14] would be a fitting remedy. Given groups $g \in \mathcal{G}$ with data distributions P_g, GDRO minimizes the worst-group expected loss

$$\min_{\theta} \max_{g \in \mathcal{G}} \left\{ R_g(\theta) := \mathbb{E}_{(x,y)\sim P_g}\left[\ell(\theta;(x,y))\right] \right\}. \tag{1}$$

In practice, (1) is optimized with the online algorithm of Sagawa et al. [14]. It keeps a weight vector q on the probability simplex over groups. For each mini-batch it first updates $q_g \leftarrow q_g \exp(\eta_q \hat{R}_g(\theta))$ for all g and renormalizes $q \leftarrow q/\sum_g q_g$, and then takes a gradient step on the weighted loss $\sum_g q_g \hat{R}_g(\theta)$, where $\hat{R}_g$ is the batch loss of group g and η_q is the group step size.

GEORGE [15] offers a substitution for subclass labels. It (i) trains an ERM model, (ii) reduces the model's penultimate-layer features with UMAP [11], (iii) clusters the reduced features separately for each class with Gaussian mixture models, choosing the number of clusters by silhouette score [13], and (iv) trains a fresh model with GDRO, treating each cluster as a group. Step (iv) is limited by the groups it receives and no more reliable than GDRO itself. This report examines both dependencies.

A preceding Bachelor's thesis [6] tried to inject domain knowledge into GEORGE by oversampling every cluster to equal size with one of several oversamplers (random, SMOTE [4], LoRAS [2], ProWRAS [3], GAMUS [16]) before the GDRO stage. On MNIST [9] and FashionMNIST [17], no oversampler beat random oversampling on true-subclass robust accuracy. How well the synthetic samples matched the original feature-space clusters varied widely between oversamplers, yet it was uncorrelated with robust accuracy. Oversampling *known* subclasses even reduced robust accuracy compared to plain GDRO. These negative results raised one of the questions that this report addresses. Does the quality of the groups GDRO receives really matter much for its outcome?

Contributions. We answer the following research questions (RQs).

1. **Cluster quality as a gate.** *Does the silhouette score of GEORGE's clusterings predict the improvement of the GDRO stage over ERM?* If it did, the GDRO stage could be run only on well-clustered data. Across ten replications of GEORGE on each of four benchmarks, we find no clear relationship (Section 3.1).

Table 1: Training configuration of the GEORGE replications. rw: reweighted so that all groups count equally; worst-cl.: worst cluster.

		MNIST	Waterbirds	ISIC	CelebA
shared	model	LeNet-4	ResNet-50	ResNet-50	ResNet-50
	batch size	128	128	16	128
ERM	optimizer	Adam	SGD	SGD	SGD
	momentum	–	0.9	0.9	0.9
	learning rate	$2 \cdot 10^{-3}$	10^{-3}	10^{-3}	10^{-4}
	weight decay	10^{-5}	10^{-4}	10^{-3}	10^{-4}
	LR $\times 0.1$ at epochs	50, 75	–	15	–
	epochs	100	300	20	50
	model selection (val.)	acc. (rw)	acc. (rw)	AUROC	acc.
GDRO	optimizer	Adam	SGD	SGD	SGD
	momentum	–	0.9	0.9	0.9
	learning rate	$2 \cdot 10^{-3}$	10^{-5}	10^{-3}	10^{-5}
	weight decay	10^{-5}	1	10^{-3}	0.1
	LR $\times 0.1$ at epochs	50, 75	–	15	–
	epochs	100	300	20	50
	model selection (val.)	worst-cl. acc. (rw)	worst-cl. acc. (rw)	worst-cl. AUROC	worst-cl. acc.
	GDRO loss used	no	yes	yes	yes
	group step size η_q	0.01	0.01	0.01	0.01
	group size adjustment	0	2	0	3
	AUROC variant	no	no	yes	no
	uniform group sampling	no	yes	no	yes
clustering	UMAP dimensions	2	1	2	1
	normalize features	yes	no	yes	no
	loss component in features	no	yes	no	yes
	GMM max. k (silhouette)	10	10	10	10
	overclustering factor	5	–	5	–
	superclasses not clustered	–	–	1	–

2. **GDRO under unequal group hardness.** *Does GDRO increase errors relative to ERM when one group is intrinsically hard?* On controlled synthetic data, GDRO matches ERM's *mean* solution, but its solutions vary much more across training runs. It collapses entirely once the hard groups carry no signal (Section 3.2).

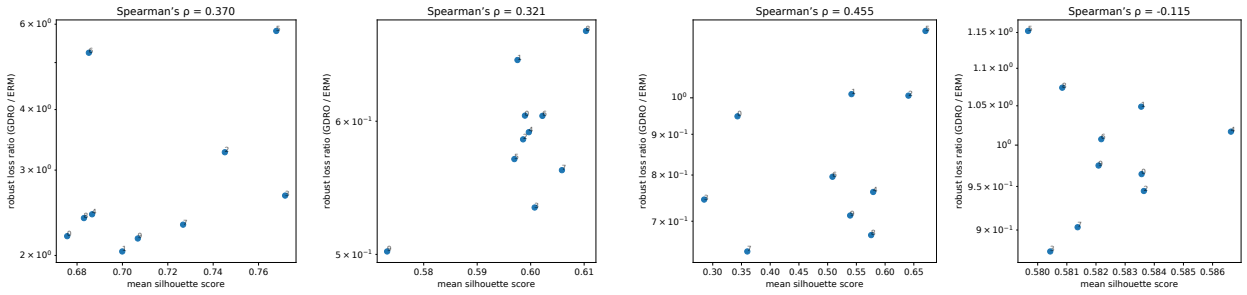
3. **Causes of the GDRO instability.** *Is the instability caused by (a) group weights lagging behind the model, (b) splitting a hard region into one group per class, or (c) unsuitable hyperparameters?* We rule out all three (Section 3.3) and attribute the instability to the easy groups dropping out of the effective objective (Section 4).

2 Experimental Setup

2.1 GEORGE replications

We re-run GEORGE with the configurations of the original paper [15] on MNIST [9] (superclasses: digits < 5 vs. ≥ 5), Waterbirds [14] (superclasses: land-bird vs. water-bird), ISIC [5] (superclasses: "malignant" vs. "benign") and CelebA [10] (superclasses: "blond" vs. "not blond"). Each dataset gets ten replications with seeds $0, \ldots, 9$ and otherwise identical configurations, listed in Table 1. MNIST uses a LeNet-4; the other datasets use an ImageNet-pretrained ResNet-50 [7].

For each replication we record (a) the clustering quality, defined as the silhouette score of the final clustering averaged over the two superclasses, and (b) the relative outcome of the GDRO stage, defined as the ratio of the validation robust loss of the GDRO model to that of the ERM model. The robust loss is the worst-cluster loss, i.e. it uses GEORGE's inferred groups, which are the groups the GDRO stage actually optimizes. A ratio below 1 means GDRO improved on ERM. We compute Spearman's rank correlation ρ between (a) and (b) across the ten replications of each dataset. The gating hypothesis predicts $\rho < 0$: better clusters should lead to a lower relative robust loss.



(a) MNIST (b) Waterbirds (c) ISIC (d) CelebA

Figure 1: Clustering quality vs. GDRO outcome across ten GEORGE replications per dataset. Horizontal axis: silhouette score of the clustering on activations of the ERM-trained model, averaged over both superclasses. Vertical axis (log scale): validation cluster-robust loss of the GDRO model divided by that of the ERM model (lower is better for GDRO). Point labels give the replication index.

2.2 Synthetic groups of unequal hardness

To isolate GDRO from the clustering stage, we use a two-dimensional mixture of four equally likely isotropic Gaussians with covariance $0.05\,I$. Each Gaussian is one group (Figure 2). The two *easy* groups have means $(0.5, \pm 0.5)$. The two *hard* groups have means $(-0.5, \pm(0.5 - \delta))$. The upper Gaussians form class 0 and the lower ones class 1, so the Bayes-optimal boundary is $x_2 = 0$ for every δ. The shift δ controls the hardness of the hard groups. At $\delta = 0$ all groups are equally easy. We report $\delta = 0.4$, where the two hard Gaussians still differ in their means but overlap strongly, and $\delta = 0.5$, where they coincide and are indistinguishable except for their label. Note that this design splits each region into one group per class, as GEORGE does by clustering within classes.

The model is a linear layer (2 features $\to$ 2 logits) trained with cross-entropy, SGD (learning rate 10^{-3}, momentum 0.9), batch size 128, for 100 epochs, on 10 000 samples. GDRO uses the reference implementation of Sagawa et al. [14] with $\eta_q = 0.01$ and no group size adjustment. Every configuration is trained in 100 replications, each with a fresh data sample and initialization (seed $42 + i$). Our measure of reliability is the spread of the resulting models, which we assess qualitatively. For every configuration we plot the decision boundaries of all 100 replications together with their mean boundary and a shaded band of ± 2 standard deviations of the boundary's x_2 position at each x_1. A wider band means less reliable training. Three variations probe the causes of the instability at $\delta = 0.4$:

(a) *Weight dynamics:* We log the GDRO group weights q per epoch for five replications.

(b) *Merged groups:* We merge the two hard Gaussians into one group and the two easy Gaussians into another, so that groups no longer coincide with classes.

(c) *Hyperparameter grid:* We vary the learning rate $\in \{10^{-4}, 10^{-3}, 10^{-2}\}$, batch size $\in \{128, 512\}$ and, for GDRO, $\eta_q \in \{10^{-3}, 10^{-2}, 10^{-1}\}$. We compare the ERM run with the narrowest band to the narrowest GDRO run with otherwise identical settings, and the narrowest GDRO run overall to the ERM run with identical settings.

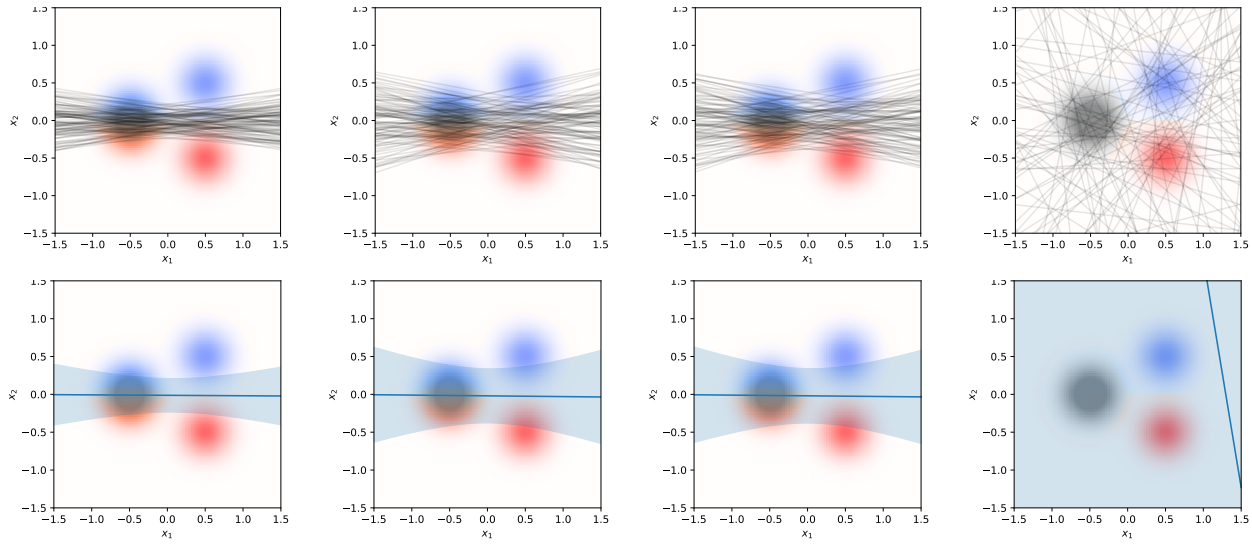
3 Results

3.1 Cluster quality does not predict the GDRO outcome

Figure 1 shows the relationship between clustering quality and the outcome of the GDRO stage. The correlation is positive on MNIST ($\rho = 0.37$), Waterbirds (0.32) and ISIC (0.46). On these three datasets, better-separated clusterings went along with a *smaller* benefit of GDRO over ERM, the opposite of what gating requires. Only CelebA shows the expected sign, and only weakly (-0.12). **Answer to RQ 1: no.** The silhouette score does not predict the robust outcome of GEORGE, and gating the GDRO stage on it is not supported.

3.2 GDRO becomes unreliable as the hardest group gets harder

Figure 2(a,b) compares ERM and GDRO at $\delta = 0.4$. The *mean* GDRO boundary coincides with the Bayes-optimal boundary $x_2 = 0$, just like ERM's. However, GDRO's band is considerably wider. Individual GDRO



(a) ERM, $\delta = 0.4$ (b) GDRO, $\delta = 0.4$ (c) GDRO, merged, $\delta = 0.4$ (d) GDRO, $\delta = 0.5$

Figure 2: Decision boundaries of 100 independently trained linear models on the synthetic data. Top: all boundaries. Bottom: mean boundary (line) and ±2 standard deviations (blue band). Background: data density (lightness) and class posterior (blue: class 0, red: class 1). The easy groups are on the right, the hard groups on the left. (c) uses one group per side instead of one group per Gaussian.

boundaries tilt much further and cut much deeper into the easy groups on the right than ERM's boundaries do. At $\delta = 0.5$, where the hard groups carry no class information, GDRO fails completely (Figure 2d). Its boundaries are scattered in all directions, the band covers the whole plot, and the mean boundary no longer separates even the easy groups. **Answer to RQ 2: yes, in distribution.** GDRO does not shift the expected solution, but the probability that a single GDRO run lands close to the optimal model shrinks as the hardest group gets harder. In the limit, GDRO ignores the learnable part of the problem altogether.

3.3 The instability is not caused by weight dynamics, group splits or hyperparameters

(a) Weight dynamics. If the group weights reacted too slowly, the model could alternate between fitting the class-0 and the class-1 half of the hard region, with q swinging after it. Figure 3 shows no such oscillation. After a short warmup period, the weights of the two hard groups converge to 0.5 each, and those of the easy groups vanish. The weights are stable from about epoch 30 of 100 in all five replications.

(b) Merged groups. Merging the hard Gaussians into one group and the easy Gaussians into another leaves the result unchanged. The boundaries, mean and band in Figure 2(c) are visually indistinguishable from the split setting in Figure 2(b). Splitting a region into per-class groups is therefore not the cause.

(c) Hyperparameters. Figure 4 compares the most favorable configurations of each method. GDRO's band remains clearly wider than ERM's with the same settings, both for the best ERM configuration (a,b) and for the best GDRO configuration (c,d). Both of these GDRO runs use the smallest step size $\eta_q = 10^{-3}$, where the weights move the slowest, hindering GDRO from deviating from ERM. **Answer to RQ 3: no** to all three candidate causes.

4 Discussion

Why GDRO becomes unstable. The motivation for RQ 2 was a property of the minimax objective (1). If the hardest group has irreducible risk R^*, then every model with $R_g(\theta) \leq R^*$ for all groups g is a minimizer. The objective is therefore indifferent to how much error the model makes on easy groups, as long as it stays below R^*. The online algorithm does not optimize (1) directly but the weighted sum $\sum_g q_g R_g$ with adaptively updated weights. Figure 3 shows that these weights can concentrate entirely on the hard groups. As a result, the easy groups contribute virtually nothing to the gradient, and the model is trained on the hard region alone. There the class signal is weak: for $\delta = 0.4$ the two hard Gaussians differ by only 0.2 in their

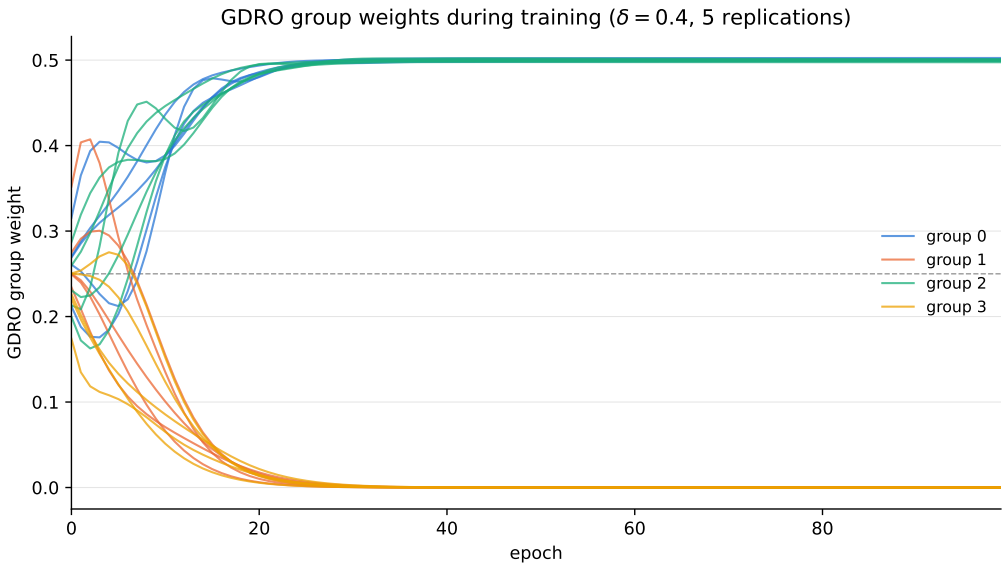
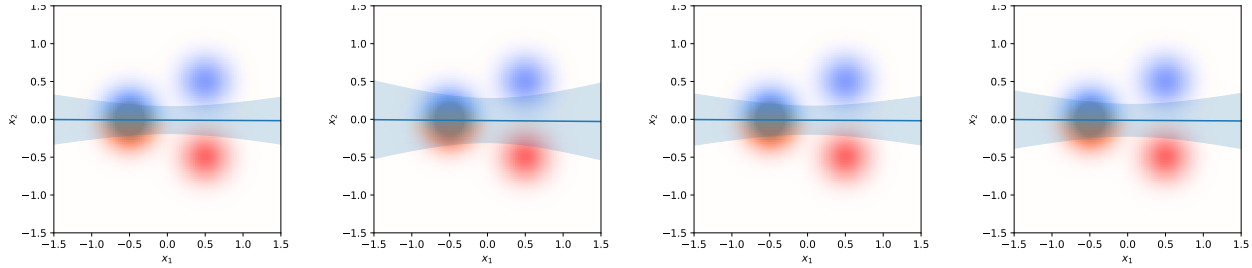


Figure 3: GDRO group weights q over training for five replications at $\delta = 0.4$. Groups 0 and 2 are the hard (left) Gaussians of class 0 and 1, groups 1 and 3 the easy ones. The dashed line marks uniform weighting.



(a) ERM, batch size 128 (b) GDRO, batch size 128 (c) ERM, batch size 512 (d) GDRO, batch size 512

Figure 4: Selected runs of the hyperparameter grid ($\delta = 0.4$, learning rate 10^{-2}, GDRO with $\eta_q = 10^{-3}$), each shown as mean boundary (line) and ± 2 standard deviations (blue band) over 100 replications. (a) The ERM run with the narrowest band and (b) the narrowest GDRO run with the same settings. (d) The narrowest GDRO run overall and (c) the ERM run with the same settings.

means with a significant overlap due to their standard deviation, as Figure 4 shows. This signal is still enough to find the best generalizing boundary on average, but it guides the boundary less than the well-separated easy groups do for ERM. For $\delta = 0.5$ the signal vanishes and the boundary is arbitrary, since the group weights only focus on the unsolvable groups. The less label separation the hard groups carry, the more closely pronounced this effect becomes. In short, *ERM has lower variance because the easy groups regularize it, and GDRO discards exactly this regularizer.* Group weight oscillation (a) is absent in cases with elevated variance in GDRO-trained models, so it is not a necessary condition. Merging groups (b) changes nothing because the merged hard group still receives all the weight. Hyperparameters (c) cannot help because they do not bring the easy groups back into the objective, except by making GDRO more like ERM.

This agrees with the observation of Sagawa et al. [14] that GDRO needs strong regularization to outperform ERM. Our setting is complementary: their finding concerns overparameterized networks that can fit every group perfectly, whereas our model is a three-parameter-per-class linear classifier that underfits the training data. Even then, GDRO needs some force other than the worst group to determine the solution. This regime has become practically relevant with linear probing of foundation models, e.g. JEPA-style encoders [1], and with last-layer group-robust retraining [8]. We see the evaluation of our findings for linear heads on frozen high-dimensional features as a promosing future direction to translate our findings into practice.

Implications for GEORGE. GEORGE's GDRO stage concentrates on the clusters with the highest loss. If those clusters are hard for intrinsic reasons, such as class overlap or label noise, rather than because they are underrepresented, GDRO adds variance without adding robustness. A global clustering score like the silhouette score does not reveal this: it measures geometric separation in a one- or two-dimensional UMAP projection, not whether the high-loss clusters are learnable. This might also help explain why oversampling

clusters in the preceding thesis [6] did not help. Oversampling changes the group sizes, but GDRO's weights already compensate for size, and it does not change whether a group is learnable.

Limitations. The correlation study only has ten replications per dataset (no significance) and only uses one error measure. Its configurations need tuning at two points: MNIST's GDRO stage actually uses ERM ("GDRO loss used" = "no") and CelebA's ERM stage does not converge at the configured 50 epochs. The synthetic study uses a single linear model on a single geometry, and judges the spread of the learned models only qualitatively, through the band of decision boundaries.

5 Conclusion

We tested two assumptions behind GDRO in GEORGE, and both turned out not to hold. First, clustering quality as measured by the silhouette score does not predict whether the GDRO stage improves on ERM, so it cannot be used to gate that stage. Second, GDRO is not safe under unequal group hardness. Its expected solution stays optimal, but its solutions vary far more from run to run than ERM's once the hard groups overlap strongly, up to complete failure when they are unlearnable. The cause is neither weight oscillation, nor class-split groups, nor hyperparameters. We suspect that it is the near-zero weight that GDRO assigns to easy groups, which removes their regularizing effect that makes ERM stable. Robust training on inferred groups might therefore improve by distinguishing groups that are hard because they are rare from groups that are hard because they are noisy.

Acknowledgments

This work was funded by the German Federal Ministry of Research, Technology and Space (BMFTR) within the junior research group THEMIS (grant numbers 01IS22077 and 16IS22077).

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, June 2023. doi: 10.1109/CVPR52729.2023.01499. URL https://ieeexplore.ieee.org/abstract/document/10205476. ISSN: 2575-7075.

[2] Saptarshi Bej, Narek Davtyan, Markus Wolfien, Mariam Nassar, and Olaf Wolkenhauer. LoRAS: an oversampling approach for imbalanced datasets. *Machine Learning*, 110(2):279–301, February 2021. ISSN 1573-0565. doi: 10.1007/s10994-020-05913-4. URL https://doi.org/10.1007/s10994-020-05913-4.

[3] Saptarshi Bej, Kristian Schulz, Prashant Srivastava, Markus Wolfien, and Olaf Wolkenhauer. A Multi-Schematic Classifier-Independent Oversampling Approach for Imbalanced Datasets. *IEEE Access*, 9: 123358–123374, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3108450. URL https://ieeexplore.ieee.org/abstract/document/9524579.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL https://www.jair.org/index.php/jair/article/view/10302.

[5] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC), March 2019. URL http://arxiv.org/abs/1902.03368. arXiv:1902.03368 [cs.CV].

[6] Konstantin Alexander Frunzek. Rebalance the hidden subclasses: Oversampling Techniques to address hidden stratification in AI Models. Bachelor's thesis, University of Rostock, May 2024.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. pages 770–778, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

[8] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. September 2022. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791. URL https://ieeexplore.ieee.org/document/726791.

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. pages 3730–3738, 2015. URL https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html.

[11] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. URL http://arxiv.org/abs/1802.03426. arXiv:1802.03426 [stat].

[12] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. 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, pages 151–159, New York, NY, USA, April 2020. Association for Computing Machinery. ISBN 978-1-4503-7046-2. doi: 10.1145/3368555.3384468. URL https://dl.acm.org/doi/10.1145/3368555.3384468.

[13] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

[14] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. September 2019. URL https://openreview.net/forum?id=ryxGuJrFvS.

[15] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 19339–19352. Curran Associates, Inc., 2020. URL https://dl.acm.org/doi/10.5555/3495724.3497346.

[16] Ayush Tripathi, Rupayan Chakraborty, and Sunil Kumar Kopparapu. A Novel Adaptive Minority Oversampling Technique for Improved Classification in Data Imbalanced Scenarios. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10650–10657, January 2021. doi: 10.1109/ICPR48806.2021.9413002. URL https://ieeexplore.ieee.org/abstract/document/9413002. ISSN: 1051-4651.

[17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms, September 2017. URL http://arxiv.org/abs/1708.07747. arXiv:1708.07747 [cs.LG].