MODEL SELECTION'S DISPARATE IMPACT IN REAL-WORLD DEEP LEARNING APPLICATIONS

Jessica Zosa Forde,* A. Feder Cooper,* Kweku Kwegyir-Aggrey, Chris De Sa, Michael Littman†

ABSTRACT

Algorithmic fairness has emphasized the role of biased data in automated decision outcomes. Recently, there has been a shift in attention to sources of bias that implicate fairness in other stages in the ML pipeline. We contend that one source of such bias, human preferences in model selection, remains under-explored in terms of its role in disparate impact across demographic groups. Using a deep learning model trained on real-world medical imaging data, we verify our claim empirically and argue that choice of metric for model comparison can significantly bias model-selection outcomes.

1 INTRODUCTION

While ML promised to remove human biases from decision making, the past several years have made it increasingly clear that automation is not a panacea with respect to fairer decision outcomes. The interaction between automation and fairness is exceedingly complex: Applying algorithmic techniques to social problems without nuance can magnify rather than correct for human bias (Abebe et al., 2020). As a result, various stakeholders, from researchers to policymakers to activists, have criticized algorithmic decision systems and advocated for approaches to mitigate the potential harms they inflict on marginalized groups.

Common to their critiques is the emphasis of data's role in unfair outcomes: Pre-existing bias in datasets results in training biased models; undersampling of marginalized populations results in worse predictive accuracy in comparison to more privileged, represented populations. In other words, regardless of the origin of bias in the data, critics have often considered this bias to be the main factor responsible for automated decision systems' reinforcement of existing inequities (Buolamwini & Gebru, 2018; de Vries et al., 2019; Denton et al., 2020). More recently, some algorithmic fairness stakeholders have begun to shift their focus to alternate sources of unfairness in ML pipelines. They contend that bias is not just in the data, and place increased scrutiny on modelling decisions. In particular, they examine the process of *problem formulation*—how data scientists construct tasks to make them amenable to ML techniques—as a source of bias in ML (Bagdasaryan et al., 2019; Cooper & Abrams, 2021; Hicks, 2019; Hooker et al., 2019; Passi & Barocas, 2019).

In this paper, we highlight another under-examined source of modelling bias—one that exists further down the ML pipeline from problem formulation. *Model selection*, the process by which researchers and data scientists choose which model to deploy after training and validation, presents another opportunity for introducing bias. During model selection for solving a particular task, the model developer compares differences in the performance of several learned models trained under various conditions, such as different optimizers or hyperparameters. This procedure is not a strictly computational; rather, the metrics used to distinguish between models are subject to human interpretation and judgement (Adebayo et al., 2018; Jacobs & Wallach, 2019). Human preferences, often geared toward the particular application domain, ultimately play an important role in choosing the model to deploy.

This role of human influence on model selection has been explored in relation to reproducibility (Henderson et al., 2018), "researcher degrees of freedom" and p-hacking (Gelman & Loken, 2014), and hyperparameter optimization (Cooper et al., 2021). **We argue that these choices also**

*Equal contribution.

†Author emails: jessica_forde@brown.edu, afc78@cornell.edu, kweku@brown.edu, cmd353@cornell.edu, michael_littman@brown.edu

have implications for fairness. We provide an intuition for model selection's possible role in disparate impact for different demographic groups (Section 3) and verify this intuition empirically on real-world medical imaging DL tasks (Section 4). To clarify the relevance and impact of this issue, we close with a call for researchers to place greater emphasis on experiments with real-world datasets and choice of model-selection comparison metrics, as commonly-used benchmark tasks and overall accuracy metrics can obscure model selection's implications for fairness (Section 5).

2 PRELIMINARIES

In ML, we typically fit a chosen model to a training dataset, iteratively improving until convergence. We evaluate the resulting model's performance for some metric, such as overall accuracy or area under the curve (AUC) on a test dataset. To further improve performance, the learning process can be repeated, changing the hyperparameters controlling the optimization process or simply using a different random seed. We then compare the performance of the different models, usually reporting or deploying the one that does the "best," where "best" is defined by the chosen performance metric (Hastie et al., 2009; Tobin, 2019). The process of repeatedly retraining until a satisfactory model is produced is what Gelman & Loken (2014) more formally call a "researcher degree of freedom" in the model-selection process. While researchers may report average performance across a small number of models, they typically report results using a specific set of hyperparameters, which can obscure a model's true performance (Bouthillier et al., 2019; Choi et al., 2019; Cooper et al., 2021; Sivaprasad et al., 2020). Moreover, when selecting a model to deploy, engineers often select a specific training run, rather than an ensemble of training runs.

While a researcher or engineer may choose between models with similar performance, many other proprieties of models change between training runs. In classification contexts, aside from overall accuracy, *sub-population performance* serves as an important consideration for model selection and can vary even when overall accuracy falls in a small range. Consider a model trained on CIFAR-10 with 10 classes and equal training data in each class. We can consider each class as a sub-population of the overall dataset. For an overall accuracy of 94%, there are numerous possibilities for how well a model performs on each of the 10 classes. Each class could exhibit 94% accuracy or, in a pathological case, 9 of the classes could have 100% accuracy with the 10th classifying correctly at a paltry 40% rate. In other words, not all 94% accuracy models are interchangeable, even for the same model architecture.

3 THE IMPACT OF MODEL SELECTION: INTUITION

While this pathological case does not necessarily occur in practice (Appendix, Figure 2), smaller, more realistic variability in sub-population performance can still have a significant impact. Consider large-scale, high-impact applications, such as medical imaging. Seemingly minute differences in sub-population performance can multiply out to potentially huge numbers of people with missed diagnoses. Such differences can impact fairness if diagnostic misses disproportionately affect some demographic groups.

In this vein, we explore a modified example from Srivastava et al. (2019), summarized in Table 1. The authors surveyed test subjects concerning fair model selection using artificial results for 3 models trained using different algorithms, each yielding different overall accuracy and female/male accuracy rates for skin cancer detection. They asked the subjects to pick the "fairest" model, and they overwhelmingly selected the model trained by A_1. The authors reason that the subjects chose this model because, even though it exhibits the largest disparity been males and females, it has the overall highest accuracy; they posit that, in high-impact domains like medicine, it is preferable to select more accurate models, even those with a potentially high cost to fairness.

Importantly, this experiment does not take into account underlying sub-population rates of skin cancer, and why overall accuracy might be a misleading metric to consider. To underscore why this oversight is important, consider instead that Table 1 concerns detecting breast cancer instead of skin cancer. When explicitly considering accuracy in terms of sub-population, which algorithm would be fairest? Since overall accuracy only measures label alignment, we do not know the breakdown of false positives and false negatives. In medical domains, false negatives (missing a cancer diagnosis) generally incur a much higher cost than false positives (mistakenly diagnosing cancer when it is in

Table 1: Algorithms resulting in different model accuracy–fairness trade-offs for skin cancer risk prediction as presented in Srivastava et al. (2019). The authors asked test subjects to choose between the presented models and found that subjects preferred A_1 (which has highest overall accuracy).

Algorithm	Overall Accuracy	Female Accuracy	Male Accuracy
A_1	94%	89%	99%
A_2	91%	90%	92%
A_3	86%	86%	86%

fact not present). We could therefore argue A_2 now has a stronger case: Males get breast cancer, but much less frequently than females (Giordano et al., 2002); depending on these different rates of breast cancer, the extra 1% in accuracy for females using A_2 might make it a better choice than A_1 if it corresponds to an overall lower number of false negative ("missed") cancer diagnoses. We could make a similar argument about how human choices in model selection might change in the context of skin cancer prediction for individuals of different races (Adamson & Smith, 2018).[1]

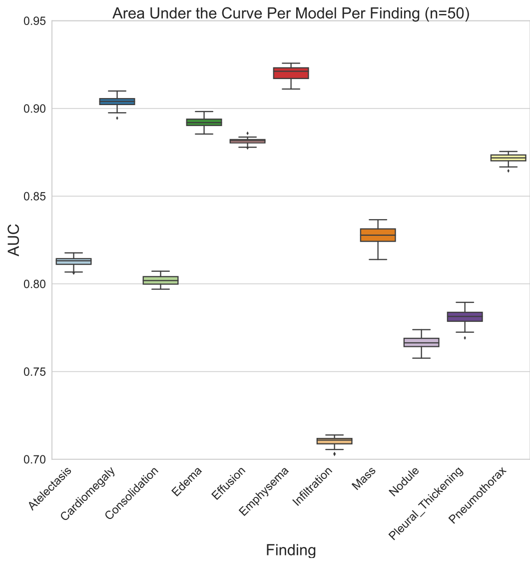
The overarching point is that contextual information is highly important for model selection, particularly with regard to which metrics we choose to inform the selection decision. Overall accuracy conceals the decomposition into false positives and false negatives. Including these metrics could bias us to make different choices concerning which model is "best" in a particular application domain (especially in medical domains, in which the relative costs of false negatives and positives is markedly different). Moreover, sub-population performance variability, where the sub-populations are split on protected attributes, can be a crucial part of that context, which in turn has implications for fairness. In medicine, this variability corresponds to the case where underlying disease rates vary by sub-population. **We therefore advocate examining sub-population performance variability as an essential component of performing fair model selection.**

4 A Deep Learning Example from Medical Imaging

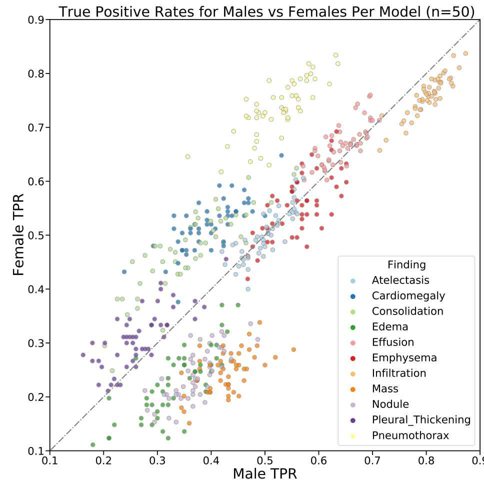
Prior work has demonstrated that medical imaging, like other computer-vision applications, exhibits disparate impact, which can lead to severe disparities in quality of care (Seyyed-Kalantari et al., 2021). We measure the baseline variability of the medical imaging classifier CheXNet, which is used to identify pathologies in chest X-rays (Rajpurkar et al., 2017). CheXNet fine-tunes DenseNet-121 (Huang et al., 2017) with fixed, pretrained ImageNet weights (Russakovsky et al., 2015) on ChestX-Ray8 (Wang et al., 2017), a standard, anonymized medical imaging dataset that identifies 14 radiological classes, known as *findings*. Findings are not mutually exclusive, and the incidence of positive test cases can be as low as 0.21% in the test set. Patient age and sex are included for each image. In this example, we focus on findings with over 50 test examples each for male and female patients. We applied a training and evaluation procedure similar to Zech et al. (2019) and Seyyed-Kalantari et al. (2021): We trained 50 CheXNet models with the same optimizer and same hyperparameters, varying the random seed across runs. Because the pre-trained weights are fixed based on ImageNet pre-training, the resulting models differ only in the order in which the images are input in minibatches to the model. **Our results suggest that, even among models with the same training procedure, we can observe significant differences in sub-population performance (Figure 1b), while exhibiting little variability in overall performance concerning findings (Figure 1a)**.

In particular, we measure overall performance by looking at the AUC for each finding; the largest range in AUC between all models is 0.0227 for the Mass finding. As each model returns a score for each instance for each finding, classification requires a threshold on this score. We select the prediction threshold for each model to maximize the F1 score on training data (Chinchor, 1992). Given each model's specific prediction threshold, we apply each threshold to the test set and measure the true positive rate (TPR) of each model. The choice of TPR over overall accuracy is due to the

[1]This example is illustrative only; we recognize that training one model is not the only solution for building an automated decision system. In cases like this one, it is perhaps desirable to train multiple, decoupled models (Dwork et al., 2018). This regime, however, has its own issues, such as how to assign individuals that span multiple demographic groups to different classifiers.



(a) Boxplot of Area Under the Curve of CheXNet trained with fixed hyperparameters 50 times, varying random seed with each retraining. Each boxplot represents 50 models' predictions for the given radiological class ("finding").



(b) True-positive-rate (TPR) disparity among 50 CheXNet models, varying random seed. Each point represents the TPR for male patients versus female patients for a finding. Points closer to the gray line have lower disparate impact between sexes.

Figure 1: Training CheXNet 50 times with fixed hyperparameters while varying random seed. CheXNet uses a pre-trained initialization, so only batching is affected by random seed.

low prevalence of positive labels for the findings considered (as low as 3.5% for Edema) and for the importance of correctly classifying positive cases as discussed in Section 3. In optimizing for F1 score, we observe both substantial differences in TPRs for males and females for a given model and finding, but also substantial variability in TPR disparity between models. The biggest TPR disparity between models for a given finding is 0.2892. The largest range of TPR disparities for a given finding across models is 0.2258.

5 Discussion and Future Work

When we measure sub-population performance levels for CheXNet models with different seeds and similar overall performance, we observe not only high levels of true positive rate disparity, but also find that these disparities can significantly vary between models. We therefore contend that such variability suggests additional efforts are necessary when designing for model robustness, especially in relatively under-explored, high-impact domains. Model selection relies on a procedure (either human or automated) for comparisons and choices between algorithmic methods. While we primarily discuss sub-population performance variability within the context of algorithmic fairness, we believe that this line of work has potential for developing more accountable decision systems (Nissenbaum, 1996). Specifically, exposing model-selection decisions and the rationales that justify them enables the opportunity to hold those decisions accountable.

In future work, we intend to formalize this notion of accountability and explore if we can ameliorate the issue of sub-population performance variability that we observe in Figure 1. For the latter, we will examine how ensembling—a popular method for improving model performance and reducing prediction variability in medical imaging tasks (Halabi et al., 2019; Zech et al., 2019)—could level out differences between sub-populations. Moreover, we aim to complement this investigation by studying how AutoML methods, which divest the model-selection process from the whims of human choice, can also prevent subjective preferences from intervening in the hyperparameter optimization process (Snoek et al., 2012).

ACKNOWLEDGMENTS

We would like to thank the following individuals for feedback on this work: Harry Auster, Charles Isbell, and Michela Paganini. We would also like to thank the Artificial Intelligence Policy & Practice initiative at Cornell University and the MacArthur Foundation. This work is supported in part by the ONR PERISCOPE MURI award N00014-17-1-2699.

REFERENCES

Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. Roles for Computing in Social Change. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 252–260, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367.

Adewole S Adamson and Avery Smith. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatol.*, 154(11):1247–1248, November 2018.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 9505–9515. Curran Associates, Inc., 2018.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In H Wallach, H Larochelle, A Beygelzimer, F d'Alché Buc, E Fox, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible Research is Reproducible. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 725–734, Long Beach, California, USA, 2019. PMLR.

Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Sorelle A Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91, New York, NY, USA, 2018. PMLR.

Nancy Chinchor. MUC-4 evaluation metrics. In *Proceedings of the 4th Conference on Message Understanding*, MUC4 '92, pp. 22–29, USA, June 1992. Association for Computational Linguistics.

D Choi, C J Shallue, Z Nado, J Lee, and others. On Empirical Comparisons of Optimizers for Deep Learning, 2019.

A. Feder Cooper and Ellen Abrams. Emergent Unfairness: Normative Assumptions and Contradictions in Algorithmic Fairness-Accuracy Trade-Off Research, 2021.

A. Feder Cooper, Yucheng Lu, and Christopher De Sa. Hyperparameter Optimization Is Deceiving Us, and How to Stop It, 2021.

Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 52–59, 2019.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. In *Participatory Approaches to Machine Learning ICML Workshop*, July 2020.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 119–133, New York, NY, USA, 23–24 Feb 2018. PMLR.

Andrew Gelman and Eric Loken. The statistical crisis in science: data-dependent analysis—a "garden of forking paths"–explains why many statistically significant comparisons don't hold up. *American Scientist*, 102(6):460–466, 2014.

Sharon H Giordano, Aman U Buzdar, and Gabriel N Hortobagyi. Breast cancer in men. *Ann. Intern. Med.*, 137(8):678–687, October 2002.

Safwan S Halabi, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Artem B Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, Felipe Campos Kitamura, Hans H Thodberg, Leon Chen, George Shih, Katherine Andriole, Marc D Kohli, Bradley J Erickson, and Adam E Flanders. The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology*, 290(2):498–503, February 2019.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2 edition, 2009.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep Reinforcement Learning That Matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, April 2018.

M. Hicks. Hacking the Cis-tem. *IEEE Annals of the History of Computing*, 41(1):20–33, 2019. doi: 10.1109/MAHC.2019.2897667.

Sara Hooker, Aaron Courville, Yann Dauphin, and Andrea Frome. Selective brain damage: Measuring the disparate impact of model pruning, November 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Abigail Z Jacobs and Hanna Wallach. Measurement and Fairness, December 2019.

A Krizhevsky. Learning multiple layers of features from tiny images, 2009.

Helen Nissenbaum. Accountability in a Computerized Society. *Science and Engineering Ethics*, 2: 25–42, 1996.

Samir Passi and Solon Barocas. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 39–48, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560. 3287567. URL `https://doi.org/10.1145/3287560.3287567`.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P Lungren, and Andrew Y Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with deep learning, November 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, December 2015.

L Seyyed-Kalantari, G Liu, M McDermott, and M Ghassemi. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *Pacific Symposium on Biocomputing*. arxiv.org, 2021.

Prabhu Teja Sivaprasad, Florian Mai, Thijs Vogels, Martin Jaggi, and François Fleuret. Optimizer Benchmarking Needs to Account for Hyperparameter Tuning. *Proceedings of the International Conference on Machine Learning*, 1, 2020.

J Snoek, H Larochelle, and R P Adams. Practical bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst.*, 2012.

Megha Srivastava, Hoda Heidari, and Andreas Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2459–2468, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330664. URL https://doi.org/10.1145/3292500.3330664.

Josh Tobin. Troubleshooting deep neural networks, January 2019.

X Wang, Y Peng, L Lu, Z Lu, M Bagheri, and R M Summers. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, July 2017.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The Marginal Value of Adaptive Gradient Methods in Machine Learning. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4148–4158. Curran Associates, Inc., 2017.

John R Zech, Jessica Zosa Forde, and Michael L Littman. Individual predictions matter: Assessing the effect of data ordering in training fine-tuned CNNs for medical imaging, December 2019.

A Appendix

We revisit the toy example of CIFAR-10 (Krizhevsky, 2009) from Section 2, where we consider the 10 classes to be different sub-populations. We ran image classification using SGD on VGG-16 following the setup of Wilson et al. (2017) (Figure 2). We did not find similar levels of sub-population performance variability to that of ChexNet. While one can observe declines in sub-population performance when decreasing the initial learning rate, these differences are by percentage points, rather than the tens of percentage points observed in Figure 1b. This outcome is unsurprising; however, given that benchmark datasets like CIFAR-10 are used for designing and tuning off-the-shelf models like VGG-16.

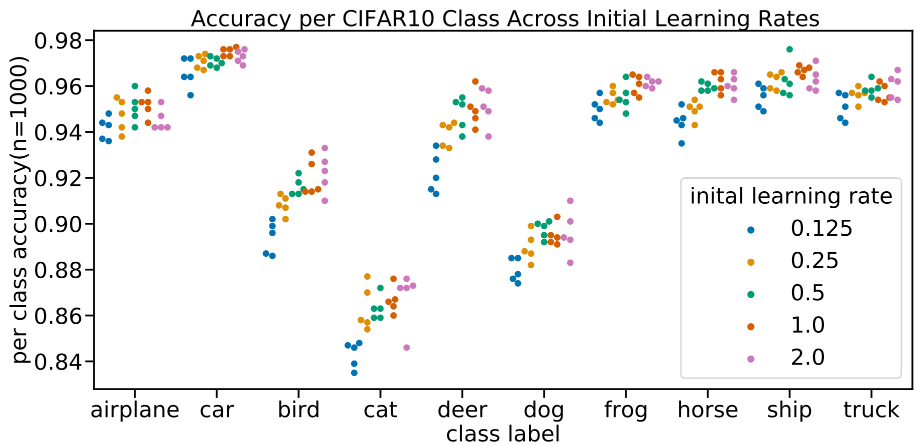


Figure 2: Per class accuracy of VGG-16 trained on CIFAR-10.

In stark contrast to toy benchmark problems like CIFAR-10, real-world datasets that capture the kinds of problems we actually want to solve with ML are not used to design and tune novel generic classification models. The same amount of collective research effort spent on hyper-optimizing models for CIFAR-10 has not gone to hyper-tuning models like ChexNet. As others have noted, benchmark tasks are not reflective of the novel, difficult problems to which want to apply ML (de Vries et al., 2019). In fact, based on our results, we hypothesize that models can be so highly tuned to benchmarks that they conceal problems like sub-population performance variability.