Identifying the Context Shift between Test Benchmarks and Production Data

Matthew Groh [1]

Abstract

Across a wide variety of domains, there exists a performance gap between machine learning models' accuracy on dataset benchmarks and real-world production data. Despite the careful design of static dataset benchmarks to represent the real-world, models often err when the data is out-of-distribution relative to the data the models have been trained on. We can directly measure and adjust for some aspects of distribution shift, but we cannot address sample selection bias, adversarial perturbations, and non-stationarity without knowing the data generation process. In this paper, we outline two methods for identifying changes in context that lead to distribution shifts and model prediction errors: leveraging human intuition and expert knowledge to identify first-order contexts and developing dynamic benchmarks based on desiderata for the data generation process. Furthermore, we present two case-studies to highlight the implicit assumptions underlying applied machine learning models that tend to lead to errors when attempting to generalize beyond test benchmark datasets. By paying close attention to the role of context in each prediction task, researchers can reduce context shift errors and increase generalization performance.

1. Limitations of Static Test Benchmarks

Dataset benchmarks offer a standardized method for comparing and evaluating the performance of machine learning models. While benchmarks help reveal effective statistical learning methodologies for different tasks, sole reliance on static benchmarks as measures of model performance ignores the complexity, diversity, and dynamism of real-world tasks. As a consequence, machine learning models that appear to be approaching (and sometimes surpassing) human level ability on a test benchmark will often error when shown out-of-distribution data (Torralba & Efros, 2011), which is relative to the dataset (e.g. images of people with dark skin when the dataset is images of mostly light skinned people (Buolamwini & Gebru, 2018; Groh et al., 2021), images of cows on sandy beaches and camels in green pastures when the dataset is images of mostly the opposite (Arjovsky et al., 2020), images of spontaneous facial expressions when the dataset is images of mostly posed expressions (Dupré et al., 2020)) adversarially perturbed data (e.g. a small sticker on a stop sign (Brown et al., 2017), noise or rotations in an image, and text substitution in medical notes and reimbursement codes (Finlayson et al., 2019)) or non-stationary data (e.g. the appearance of mobile phones and cars have significantly changed over the last couple decades (Recht et al., 2019)). This means that the reliance on static test benchmarks as metrics for evaluating performance (Thomas & Uminsky, 2022) inflates the accuracy of machine learning model performance relative to how the models would perform over time in a production environment leaving open questions like "Can you trust your model? Will it work in deployment?" (Lipton, 2018)

In order to demonstrate examples of the benchmark-production gap in model performance, researchers (Recht et al., 2019) built new tests sets for ImageNet (Deng et al., 2009; Russakovsky et al., 2015) and CIFAR-10 (Krizhevsky et al., 2014), two of the most commonly used dataset benchmarks for evaluating image-based object recognition. The researchers closely followed the original dataset generation processes yet found significant drops (ranging from 3% to 15% in 67 models applied to ImageNet and 34 models applied to CIFAR-10) in the state-of-the-art models' performance from the original test benchmark to the new test benchmark. In this case, the performance gap does not appear to be explained by random sampling error or hyperparameter tuning for optimizing performance on the original test set, but instead the performance gap appears to arise from changes in the distribution of the data despite the researchers best efforts to replicate the distribution of the benchmark datasets (Recht et al., 2019).

[1]MIT Media Lab, Cambridge, MA, USA. Correspondence to: Matthew Groh <groh@mit.edu>.

Presented at ICML Workshop on Principles of Distribution Shift 2022,

2. Contextualizing the Benchmark-Production Gap

In the machine learning community, the gap between models' performance on a test benchmark and performance data is often explained by *distribution shift* or *dataset shift*, which are catch-all terms to describe differences in how the data appears between the test and performance data. Sometimes, researchers focus on specific aspects of distribution shifts like *covariate shift* when the distribution of features changes but everything else remains the same, *prior probability shift* when the distribution of labels changes but everything else remains the same, *concept shift* when the distribution of labels conditional on features changes but everything else remains the same. However, distribution shift is downstream from the data generating process, which may involve *sample selection bias* "when the distributions differ as a result of an unknown sample rejection process" (Quiñonero-Candela et al., 2008), *adversarial perturbations* when the distribution of data is altered in a way that does not affect human task performance, or *nonstationarity* when the data distribution changes over time or environment. We introduce the term *context shift* to describe these three dimensions of the data generating process and center the importance of understanding how data is collected, created, and curated. While covariate shift, prior probability shift, and concept shift can be formally specified, empirically evaluated, and even sometimes mitigated (Moreno-Torres et al., 2012), the problems of context shift can only be fully addressed by learning the entire population's data distribution, the kinds of changes that are and are not perceptible to humans, and what and how things change over time and space.

Frameworks like *Data Statements for Natural Language Processing* (Bender & Friedman, 2018), *The Dataset Nutrition Label* (Holland et al., 2018), *Model Cards for Model Reporting* (Mitchell et al., 2019), *Datasheets for Datasets* (Gebru et al., 2021), *Closing the AI accountability gap* (Raji et al., 2021), *The Ethical Pipeline for Healthcare Model Development* (Chen et al., 2020), and *The Clinician and Dataset Shift in Artificial Intelligence* (Finlayson et al., 2021) offer guidance for breaking down the data generating process into relevant component parts to identify potential dimensions on which context shift could lead to performance changes from a test benchmark to production environments. Likewise, meta-frameworks offer guidance for ensuring data documentation frameworks are useful and actionable (Heger et al., 2022). As a heuristic for human-centered machine learning applications, teams of conscientious, creative, and skilled model developers, data engineers, and subject matter experts may find it useful to identify a first-order, non-exhaustive list of dimensions on which context shift is likely to occur. This list of dimensions largely depends on the context and the degree to which data are subjective, representative, and missing (Mullainathan & Obermeyer, 2017). Recent examples of important contextual dimensions on machine learning tasks include skin color in face recognition (Buolamwini & Gebru, 2018) and dermatology diagnosis (Groh et al., 2021; Daneshjou et al., 2022), background scenery for affect recognition (Kosti et al., 2019), number of people in a video for deepfake detection (Groh et al., 2022a), number of chronic illnesses for algorithmic healthcare risk prediction (Obermeyer et al., 2019), data artifacts like surgical markings (Winkler et al., 2019) or clinically irrelevant labels (Oakden-Rayner et al., 2020) for medical diagnosis classification, and patients' self reports of pain for quantifying severity of knee osteoarthritis (Pierson et al., 2021). Helpful questions that may guide the identification of potential context shifts in complex, human-centered machine learning applications include (and are not limited to): who are represented in the data and as annotators of the data, when and where is the data collected, how do social, geographical, temporal, technological, aesthetic, financial incentives and other idiosyncrasies influence the creation of the data, and why the data is curated as it is.

One approach to closing the benchmark-production gap involves developing test benchmarks with adequate diversity in the data along the contextual dimensions upon which human intuition and expertise suggests model performance is most likely to vary. Recent examples benchmark datasets working towards this goal are *BREEDS: Benchmarks for Subpopulation Shift* (Santurkar et al., 2020) and *WILDS: A Benchmark of in-the-Wild Distribution Shifts* (Koh et al., 2021), which include labels for relative contexts and subpopulations for the explicit examination of context shifts.

A second approach to addressing the benchmark-production gap is to transform the practice of evaluation from static benchmarks to dynamic benchmarks where models' performance is continually evaluated on datasets produced via well-specified, quality controlled data generation processes. Dynabench (Kiela et al., 2021) is an example of dynamic benchmarking for natural language processing tasks. For developing dynamic benchmarks in general, data generation process desiderata include specifying the following:

- **Prediction task**: What are the input features and output labels? For example, inputs may be images and outputs may be lists of objects or classifications of benign and malignant.
- **Test size**: What is the minimum test size for reducing within data sampling error to an acceptably small error rate?
- **Ground truth annotation arbitration**: Who has the

authority to annotate the data? How should data be annotated? How should inter-annotator disagreement be represented? What categories should be included?

- **Data inclusion and exclusion criteria**: What are the possible data sources? How are data curated from these sources? What is the data distribution of categories and sub-categories? What are the quality constraints?

These desiderata enable the development of dynamic benchmarks that further enable quantitative evaluation of model robustness via corroborated accuracy, which is the distribution of accuracy scores across dynamic benchmarks. Rather than simply evaluating a model on a single or a few static test benchmarks, we might consider a well-corroborated model to be one that meets two criteria: first, it is reasonably available for evaluation, and second, all attempts to uncover systematic errors in well-specified contexts reveal no significant accuracy disparities. The practice of dynamic benchmarking could be particularly relevant for addressing the *AI Knowledge Gap* (Epstein et al., 2018) characterized by the disparity between the number of machine learning models and the number of studies evaluating these models' performance. Furthermore, dynamic benchmarking can be combined with benchmark-task misalignment methodologies (Tsipras et al., 2020) for evaluating how aligned (or misaligned) model predictions are with human annotations and qualitative approaches for evaluating ethical implications and societal impact (Paullada et al., 2021).

Another approach to reducing the benchmark-production gap is to appropriately specify the contexts in which a model should be expected to work. In order to clarify domain-specific limitations driving the benchmark-production gap, we consider implicit assumptions that lead to context shift in two specific but separate kinds of real-world computer vision tasks: facial expression recognition and deepfake detection.

2.1. Implicit Assumptions in Facial Expression Recognition

In the field of affective computing, facial expression recognition (FER) is a task for classifying human facial expressions with affective labels (Cohn & De la Torre, 2015; Li & Deng, 2020), which can be a useful component for designing human-AI interactions with computational empathy (Picard, 2000; Paiva et al., 2017). Model-based FER is similar to how humans recognize the emotions of others (called empathic accuracy in affective science (Ickes, 1993) and emotion reasoning in developmental psychology (Ruba & Pollak, 2020)) except that FER is based solely on facial expressions whereas affect recognition can include information about someone's gestures, language, tone, physiological measurements, and the long-tail of context, which can include factors such as the temperature outside, the social relationship between two individuals, and more.

Consider an example from relatively recent research (Mollahosseini et al., 2016) where a standard neural network architecture, AlexNet (Krizhevsky et al., 2012), is trained on a large number of images of posed and spontaneous facial expressions to classify images into seven categories (anger, disgust, fear, happiness, sadness, surprise, and neutral) and achieves accuracy scores ranging from 48.6% on SFEW (Dhall et al., 2011) to 56.0% on MMI (Pantic et al., 2005) to 56.1% on DISFA (Mavadati et al., 2013) to 61.1% on FER2013 (Goodfellow et al., 2013) to 77.4% on FERA (Valstar et al., 2011; Bänziger & Scherer, 2010) to 92.2% on CK+ (Lucey et al., 2010) to 94.8% on Multi-Pie (Gross et al., 2010). While this model's accuracy is significantly better than random guessing, which would be 14.2%, it varies dramatically depending on the benchmark dataset chosen. How should we interpret a performance gain of 21.9 percentage points on one dataset and an average performance gain of 3.5 percentage on the other 6 datasets in an alternative network architecture? How should we interpret the model's ability to achieve higher accuracy scores than non-neural network methods on three of the seven benchmark datasets? And, what does the distribution of performance tell us about how this model would perform on real-world production data? There is no clear answer to any of these three questions, yet an implicit assumption in the well-cited, peer-reviewed publication of this FER paper is the slightly improved performance on several benchmark datasets appears to mark a contribution to the field of facial expression recognition. This assumption has the potential to lead to another more pernicious and mistaken assumption: the role of contextual features for real-world performance can be ignored when assessing the state-of-the-art methodology in applied problems like FER.

Clearly, models can learn facial expressions features that map to humans annotations of 7 emotion categories to classify images at significantly better than chance rates. But, it is not reported how changes in lighting, head pose, occlusion, skin tone, ethnicity, age, gender, and background scenery influence both the model's performance and the human annotations. It is also underexplored how well FER models would do if humans from diverse cultures annotated these images. Likewise, it is unclear how the model would perform on more fine-grained emotion categories (Cowen et al., 2021) or labels based on affective dimensions like valence, arousal, and dominance. Furthermore, in many real-world settings where people may

feign smiles to appease their managers, cry to express joy, or appear neutral to hide a winning poker hand, the perspective of outside observers may be very different than the perspective of close friends or individuals themselves. We highlight these relevant contextual features to highlight the many dimensions in which context shift can occur between test benchmarks and real-world production data. While these are not an exhaustive list of contextual features, these represent intuitive, first-order contexts for conducting algorithmic audits, developing future benchmark datasets with these labeled contexts, and adapting models to handle these dimensions. While researchers build the next version of contextualized dynamic benchmarks, other researchers who are focused on developing models should at the very least include caveats in their papers about the likely contextual dimensions that may affect performance and are not explicitly examined.

2.2. Implicit Assumptions in Deepfake Detection

As a second case study of context shift in real-world applications of computer vision, we consider deepfake detection. Deepfakes are videos that have been manipulated to make someone appear to do or say something that they have not said (Groh et al., 2022a). The kinds of manipulations can be qualitatively characterized as face swapping where two peoples faces are swapped, head puppetry where facial landmarks adjusted to make someone appear to be speaking, and lip-syncing where an individual's lips are moved in sync with the phonemes from an external audio track (Lyu, 2020).

The largest deepfake detection benchmark dataset to date is the Deepfake Detection Competition Dataset (DFDC) (Dolhansky et al., 2019; 2020), which consists of 128,154 videos based on performances by 960 consenting actors representing diversity across sex and ethnicity. However, Groh et al 2022 point out,"Unlike viral deepfake videos of politicians and other famous people, the videos from [this benchmark dataset] have minimal context: They are all 10 [second] videos depicting unknown actors making uncontroversial statements in nondescript locations" (Groh et al., 2022a). This deepfake test benchmark is designed to evaluate algorithmic performance at identifying videos that have (and have not) been manipulated by seven synthetic techniques.

But, the real-world deepfake detection problem is not simply identifying whether one of seven synthetic techniques has been applied to a video. Instead, the real-world problem is identifying videos that have been algorithmically altered to impersonate innocent people and deceive the viewer. This problem is more than just a computer vision problem; it is a deception detection problem which involves both searching for artifacts that reveal a manipulation has occurred and applying prior knowledge and critical reasoning to assess the likelihood that the video has been fabricated.

The DFDC does not include politicians or any scenes of news conferences or people speaking to a large audience. If we assume that harmful deepfakes will involve these kinds of contexts (like a deepfake of President Volodomyr Zelensky that appeared in March 2022 (Wakefield, 2022)), then it is important to evaluate models on videos with these kind of dimensions like ones from the Presidential Deepfakes Dataset (Sankaranarayanan et al., 2021; Groh et al., 2022b) and the Protecting World Leaders against Deepfakes Dataset (Agarwal et al., 2019). When Groh et al 2022 examined the leading state-of-the-art for detecting DFDC videos on deepfakes of Kim Jung-un and Vladimir Putin, they found the the leading model predicted a 2% and 8% likelihood these videos are deepfakes. While failure on two examples is only an anecdote, this failure speaks to an important need: diverse test benchmarks that cover the first-order dimensions where human intuition and expertise suggests context shift is most likely to occur.

3. Towards Seeking Robustness and Avoiding Brittleness

Supervised machine learning models are very good at identifying statistical regularities in a given dataset but tend to err on out-of-distribution data that may arise from sample selection bias, adversarial perturbation, or non-stationarity. On the other hand, humans can be quite good at identifying contextual examples of out-of-distribution data. By combining the strengths of machine learning models with human intuition and expertise, early career ancient historians can rapidly restore and date ancient texts (Assael et al., 2022), content moderation teams can more accurately distinguish between real and fake videos (Groh et al., 2022a), and general practitioners can more accurately diagnose skin conditions from images (Jain et al., 2021) (although AI advice can also mislead experts, see (Tschandl et al., 2020; Abeliuk et al., 2020; Gaube et al., 2021; Jacobs et al., 2021; Vaccaro & Waldo, 2019)). In fact, initial evidence suggests human intuition is fairly accurate at predicting model misclassifications on object detection tasks (Zhou et al., 2020). The integration of machine predictions with human decisions in collaborative decision making systems may be the most immediately effective way to avoid errors from context shift.

Promising future research directions for developing robust machine learning models under distribution shift involve the following iterative process: first, identify missing contexts in test benchmarks, second, collect data that contains those missing contexts, and third, adjust the model accord-

ingly. Researchers can begin to identify missing contexts by collaborating with human experts who may be able to identify first-order drivers of context shift on a task-by-task basis. Likewise, researchers can further identify missing contexts by evaluating models against data generation process desiderata rather than a single or a few datasets.

Finally, one of the most effective solutions for addressing the benchmark-production gap is for researchers to clearly communicate the contexts in which a model has been evaluated and the contexts in which the model's performance is unknown.

References

Abeliuk, A., Benjamin, D. M., Morstatter, F., and Galstyan, A. Quantifying machine influence over human forecasters. *Scientific Reports*, 10(1):15940, December 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-72690-4. URL http://www.nature.com/articles/s41598-020-72690-4.

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., and Li, H. Protecting world leaders against deep fakes. 2019.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant Risk Minimization, March 2020. URL http://arxiv.org/abs/1907.02893. Number: arXiv:1907.02893 arXiv:1907.02893 [cs, stat].

Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., and de Freitas, N. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900): 280–283, March 2022. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-022-04448-z. URL https://www.nature.com/articles/s41586-022-04448-z.

Bänziger, T. and Scherer, K. R. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook*, 2010:271–94, 2010.

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

Brown, T. B., Mané, D., Roy, A., Abadi, M., and Gilmer, J. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. Ethical Machine Learning in Health Care. *arXiv:2009.10576 [cs]*, October 2020. URL http://arxiv.org/abs/2009.10576. arXiv: 2009.10576.

Cohn, J. F. and De la Torre, F. Automated face analysis for affective computing. 2015.

Cowen, A. S., Keltner, D., Schroff, F., Jou, B., Adam, H., and Prasad, G. Sixteen facial expressions occur in similar contexts worldwide. *Nature*, 589 (7841):251–257, January 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-3037-7. URL http://www.nature.com/articles/s41586-020-3037-7.

Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *arXiv preprint arXiv:2203.08807*, 2022.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 2106–2112. IEEE, 2011.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., and Ferrer, C. C. The Deepfake Detection Challenge (DFDC) Preview Dataset. *arXiv:1910.08854 [cs]*, October 2019. URL http://arxiv.org/abs/1910.08854. arXiv: 1910.08854.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv:2006.07397 [cs]*, October 2020. URL http://arxiv.org/abs/2006.07397. arXiv: 2006.07397.

Dupré, D., Krumhuber, E. G., Küster, D., and McKeown, G. J. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLOS ONE*, 15(4):e0231968, April 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0231968. URL https://dx.plos.org/10.1371/journal.pone.0231968.

Epstein, Z., Payne, B. H., Shen, J. H., Dubey, A., Felbo, B., Groh, M., Obradovich, N., Cebrian, M., and Rahwan, I. Closing the AI Knowledge Gap. *arXiv:1803.07233 [cs]*, March 2018. URL

http://arxiv.org/abs/1803.07233. arXiv: 1803.07233.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., and Kohane, I. S. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., and Saria, S. The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine*, 385(3):283–286, July 2021. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMc2104626. URL http://www.nejm.org/doi/10.1056/NEJMc2104626.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., and Ghassemi, M. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1):31, December 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00385-9. URL http://www.nature.com/articles/s41746-021-00385-9.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pp. 117–124. Springer, 2013.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., and Badri, O. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1820–1828, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-899-4. doi: 10.1109/CVPRW53098.2021.00201. URL https://ieeexplore.ieee.org/document/9522867/.

Groh, M., Epstein, Z., Firestone, C., and Picard, R. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1): e2110013119, January 2022a. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2110013119. URL http://www.pnas.org/lookup/doi/10.1073/pnas.2110013119.

Groh, M., Sankaranarayanan, A., and Picard, R. Human detection of political deepfakes across transcripts, audio, and video. *arXiv preprint arXiv:2202.12883*, 2022b.

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010.

Heger, A., Marquis, E. B., Vorvoreanu, M., Wallach, H., and Vaughan, J. W. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *arXiv preprint arXiv:2206.02923*, 2022.

Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

Ickes, W. Empathic accuracy. *Journal of personality*, 61 (4):587–610, 1993.

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., and Gajos, K. Z. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry*, 11(1):108, June 2021. ISSN 2158-3188. doi: 10.1038/s41398-021-01224-x. URL http://www.nature.com/articles/s41398-021-01224-x

Jain, A., Way, D., Gupta, V., Gao, Y., de Oliveira Marinho, G., Hartford, J., Sayres, R., Kanada, K., Eng, C., Nagpal, K., DeSalvo, K. B., Corrado, G. S., Peng, L., Webster, D. R., Dunn, R. C., Coz, D., Huang, S. J., Liu, Y., Bui, P., and Liu, Y. Development and Assessment of an Artificial Intelligence–Based Tool for Skin Condition Diagnosis by Primary Care Physicians and Nurse Practitioners in Teledermatology Practices. *JAMA Network Open*, 4(4):e217249, April 2021. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2021.7249. URL https://jamanetwork.com/journals/jamanetworkopen/

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., et al. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*, 2021.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv:2012.07421 [cs]*, July 2021. URL http://arxiv.org/abs/2012.07421. arXiv: 2012.07421.

Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. Context Based Emotion Recognition using

EMOTIC Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2019.2916866. URL http://arxiv.org/abs/2003.13401. arXiv: 2003.13401.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Krizhevsky, A., Nair, V., and Hinton, G. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55(5), 2014.

Li, S. and Deng, W. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, pp. 1–1, 2020. ISSN 1949-3045, 2371-9850. doi: 10.1109/TAFFC.2020.2981446. URL https://ieeexplore.ieee.org/document/9039580.

Lipton, Z. C. The mythos of model interpretability. *Communications of the ACM*, 61(10): 36–43, September 2018. ISSN 0001-0782, 1557-7317. doi: 10.1145/3233231. URL https://dl.acm.org/doi/10.1145/3233231.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pp. 94–101. IEEE, 2010.

Lyu, S. DeepFake Detection: Current Challenges and Next Steps. *arXiv:2003.09234 [cs]*, March 2020. URL http://arxiv.org/abs/2003.09234. arXiv: 2003.09234.

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4 (2):151–160, 2013.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Mollahosseini, A., Chan, D., and Mahoor, M. H. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pp. 1–10. IEEE, 2016.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1): 521–530, 2012.

Mullainathan, S. and Obermeyer, Z. Does Machine Learning Automate Moral Hazard and Error? 107(5):5, 2017.

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*, pp. 151–159, Toronto Ontario Canada, April 2020. ACM. ISBN 978-1-4503-7046-2. doi: 10.1145/3368555.3384468. URL https://dl.acm.org/doi/10.1145/3368555.3384468.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366 (6464):447–453, October 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aax2342. URL https://www.sciencemag.org/lookup/doi/10.1126/sci

Paiva, A., Leite, I., Boukricha, H., and Wachsmuth, I. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3):1–40, 2017.

Pantic, M., Valstar, M., Rademaker, R., and Maat, L. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pp. 5–pp. IEEE, 2005.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

Picard, R. W. *Affective computing*. 2000.

Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S., and Obermeyer, Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1): 136–140, January 2021. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-020-01192-7. URL http://www.nature.com/articles/s41591-020-01192-7

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. 2008.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. AI and the Everything in the Whole Wide World Benchmark. *arXiv:2111.15366 [cs]*, November 2021. URL http://arxiv.org/abs/2111.15366. arXiv: 2111.15366.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Ruba, A. L. and Pollak, S. D. The Development of Emotion Reasoning in Infancy and Early Childhood. *Annual Review of Developmental Psychology*, 2(1): 503–531, December 2020. ISSN 2640-7922, 2640-7922. doi: 10.1146/annurev-devpsych-060320-102556. URL https://www.annualreviews.org/doi/10.1146/annurev-devpsych-060320-102556.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Sankaranarayanan, A., Groh, M., Picard, R., and Lippman, A. The presidential deepfakes dataset. 2021.

Santurkar, S., Tsipras, D., and Madry, A. BREEDS: Benchmarks for Subpopulation Shift, August 2020. URL http://arxiv.org/abs/2008.04859. Number: arXiv:2008.04859 arXiv:2008.04859 [cs, stat].

Thomas, R. L. and Uminsky, D. Reliance on metrics is a fundamental challenge for ai. *Patterns*, 3(5):100476, 2022.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., and Kittler, H. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, August 2020. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-020-0942-0. URL http://www.nature.com/articles/s41591-020-0942-0.

Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., and Madry, A. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pp. 9625–9635. PMLR, 2020.

Vaccaro, M. and Waldo, J. The effects of mixing machine learning and human judgment. *Communications of the ACM*, 62(11):104–110, 2019.

Valstar, M. F., Jiang, B., Mehu, M., Pantic, M., and Scherer, K. The first facial expression recognition and analysis challenge. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 921–926. IEEE, 2011.

Wakefield, J. Deepfake presidents used in russia-ukraine war, Mar 2022. URL https://www.bbc.com/news/technology-60780142.

Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., and Haenssle, H. A. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, 155(10):1135, October 2019. ISSN 2168-6068. doi: 10.1001/jamadermatol.2019.1735. URL https://jamanetwork.com/journals/jamadermatology/

Zhou, Z., Nartker, M., and Firestone, C. When will ai misclassify? human intuition for machine (mis) perception. *Journal of Vision*, 20(11):1325–1325, 2020.