Identifying the Context Shift between Test Benchmarks and Production Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Benchmark datasets have traditionally served dual purposes: first, benchmarks offer a standard on which machine learning researchers can compare different methods, and second, benchmarks provide a model, albeit imperfect, of the real world. The incompleteness of test benchmarks (and the data upon which models are trained) hinder robustness in machine learning, enable shortcut learning, and leave models systematically prone to err on out-of-distribution and adversarially perturbed data. In an effort to clarify how to address the mismatch between test benchmarks and production data, we introduce context shift to describe semantically meaningful changes in the underlying data generation process. Moreover, we identify three methods for addressing context shift that would otherwise lead to model prediction errors: first, we describe how human intuition and expert knowledge can identify semantically meaningful features upon which models systematically fail, second, we detail how dynamic benchmarking – with its focus on capturing the data generation process – can promote generalizability through corroboration, and third, we highlight that clarifying a model's limitations can reduce unexpected errors.

1 Machine Learning Models are Brittle in Production

As a standard for comparison, dataset benchmarks have enabled rapid progress in computer vision and natural language processing. However, despite intentions to create and curate data that match the real-world as closely as possible, the dynamic, high-dimensional, combinatoric complexity of many real-world tasks can be difficult to fully capture in a single static benchmark. Indeed, the development and evaluation of machine learning models on benchmarks often suffer from a variety of historical, representational, measurement, aggregation, and evaluation biases which can be further exacerbated by deployment biases where the task that a benchmark is intended to measure differs from the real-world task (1; 2). Moreover, data for benchmarks are often collected at scale (3), which leaves data open to poisoning attacks (4), leakage (5), subjectivity and multiple interpretations (6) and error (7). 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 (8) data. In other words, the reliance on static test benchmarks as metrics for projecting production performance (9) inflates the accuracy of machine learning model performance and leaves open the questions, "Can you trust your model? Will it work in deployment?" (10)

The meaning of out-of-distribution data depends on a task's context. One classic example of out-of-distribution data in object detection tasks from the invariant risk minimization literature is an image of a cow on a sandy beach (and similarly a camel on a green pasture) because training data rarely contains such animal-environment pairs (11). Such a lack of uncommon animal-environment pairs leads to machine learning shortcuts where a model learns spurious correlations (e.g. cloven hoofed mammals next to sand are camels and the ones next to grass are cows). With *a priori* knowledge of likely model shortcuts, one approach for addressing this kind of model brittleness is to include

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

auxiliary labels that can serve as a causally-motivated regularization framework (12). However, for identifying previously unknown shortcuts, post hoc model explanations are often ineffective (13) (though explanations via concept traversals (14) and identifying model failures as directions in latent space via contrastive learning where images and natural language are embedded in a shared latent space show promise (15)). In contrast, human intuition can identify many out-of-distribution contexts on which shortcut learning may occur.

2 Systematic Errors Arise from Context Shift Which Drives Distribution Shift

The mismatch between two datasets (that is, a test benchmark and a sample drawn from production data) has been traditionally described as a dataset shift (16) and recently has been described more frequently as a distribution shift. Dataset shift and distribution shift both refer to the non-equivalence of the joint distributions between two datasets; formally, distribution shift describes the following equation $P_1(y, x) \neq P_2(y, x)$ where $P_n(y, x)$ is the joint distribution of labels, y, and covariates, x for a particular dataset, n (17). Based on Moreno et al 2012, the subcategories of distribution shift include 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, and concept shift (more commonly referred to as concept drift) when the distribution of labels conditional on features changes but everything else remains the same.

- Covariate shift: $P_1(x) \neq P_2(x)$ but $P_1(y|x) = P_2(y|x)$
- Prior probability shift: $P_1(y) \neq P_2(y)$ but $P_1(y|x) = P_2(y|x)$
- Concept drift: $P_1(y|x) \neq P_2(y|x)$ but $P_1(x) = P_2(x)$
- Other distribution shift: $P_1(y, x) \neq P_2(y, x)$ where none of the above three shifts applies.

Although covariate shift, prior probability shift, and concept drift can be formally specified, empirically evaluated, and in some specific cases mitigated, there has not been and will not be a perfect end-to-end solution for generalizing from a benchmark dataset that is not an independent and identically distributed (i.i.d.) sample of a data generating process. In the language of causal representation learning (18), data are often assumed to be observational (sampled from an unchanging distribution), but in the real world it is often interventional (sampled from a changing distribution), and interventions can be particularly difficult to directly learn from the data when data are raw (perceptual inputs like images and audio) as opposed to hand-crafted (semantically meaningful features) because of the combinatorial explosion of potential contexts. The core problem with the conceptual framing of distribution shift is that it is merely a symptom of changes in data-generating processes - how data are created, collected, and curated – but not part of the data-generating process itself, which needs to be directly addressed for improving model reliability and robustness.

Context refers to the semantically meaningful upstream factors that drive the distribution shift. Machine learning research has long considered the hidden contexts behind the distribution shift (19). In this paper, we seek to clarify the overlapping and often confusing jargon of distribution shift by introducing context shift as a term to describe the semantically meaningful features that influence data-generating processes and focusing on how to identify the changes in the creation, collection, and curation of data that lead to distribution shifts. By centering the problem of robustness and generalizability of applied machine learning on context shift, we expand and orient the problem space to data-centered human intuition and expertise.

Rather than focusing on differences in two distributions with disregard for the reasons behind the difference, we suggest researchers consider three concepts that drives semantically meaningful changes in datasets: sample selection bias (e.g. the new dataset contains images of people from a demographic not represented in the old dataset), adversarial perturbations (e.g. the new dataset contains noise injections that are imperceptible to the human eye but change model performance), or non-stationarity (e.g. the new dataset contains images of smart phones post 2018 but the old dataset only contains flip phones before 2010). While we list non-stationarity separately from sample selection bias, non-stationarity can be considered a special case of sample selection bias where the sample selection bias is over features and labels which may appear different in the future.

Unlike distribution shift, which can be measured between two datasets, 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 how the population's data distribution changes over time and space. Outside of artificially constrained spaces like synthetic datasets or games, access to the entire populations data distribution (or the rules governing the distribution) across space and time is rare. Nevertheless, people generally have intuition and the ability to reason about data distributions of combinatoric contexts that they might never experience. In fact, cognitive science research shows that intuitive reasoning about statistical power analysis begins early in childhood (20).

3 Addressing Robustness with Human Intuition and Expertise

Over the last few years, researchers have been developing data-centered frameworks to offer guidance for breaking down the data generating process into relevant component parts that reveal where context shift may lead to benchmark-production performance gaps (21; 22; 23; 24; 25; 26; 27). Likewise, meta-frameworks offer guidance for ensuring data documentation frameworks are useful and actionable (28).

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 depends largely on the context and the degree to which the data are subjective, representative, and missing (29). In ethnographic interviews with machine learning engineers, researchers find that engineers often address changes in context with "elaborate rule-based guardrails to avoid incorrect outputs" (30). Recent examples of semantically meaningful dimensions that have been demonstrated as useful for evaluating robustness in applied machine learning include skin color in face recognition (31) and dermatology diagnosis (32; 33), background scenery for affect recognition (34), number of people in a video for deepfake detection (35), number of chronic illnesses for algorithmic healthcare risk prediction (36), and data artifacts like surgical markings (37) or clinically irrelevant labels (38) for medical diagnosis classification.

Knowledge elicitation is not a solved problem, but 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. Knowledge elicitation has been historically ill-defined in artificial intelligence applications (39), but recent work developing taxonomies for knowledge elicitation helps to formalize the process and increase transparency along the way (40; 41).

Another expert intuition guided 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 of benchmark datasets working towards this goal are *BREEDS: Benchmarks for Subpopulation Shift* (42) and *WILDS: A Benchmark of in-the-Wild Distribution Shifts* (43), which includes labels for relative contexts and sub-populations for the explicit examination of context shifts.

4 Addressing Robustness with Dynamic Benchmarking

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 not evaluated on a single dataset, but rather continually evaluated on datasets produced via well-specified, quality controlled data generation processes. One example of dynamic benchmarking is dynabench (44), which is designed for natural language processing tasks. For general development of dynamic benchmarks, data generation process desiderata should include specifying the following dimensions of a dynamic benchmark:

- **Prediction task**: What are the input features and output labels? For example, inputs may be images and outputs may be lists of objects or inputs may be described more specifically as images of skin lesions photographed by dermatoscopes and outputs may be classifications of benign and malignant by board-certified dermatologists in the United States. It is important to be careful that the task matches the expected goal because unexpected mismatches between tasks and goals are relatively common (45; 46).

- **Ground truth annotation arbitration**: Who has the authority to annotate the data? How do experts differ from crowdworkers or an algorithm (47)? How should the 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 subcategories? What are the quality constraints?
- **Benchmark size and shape**: What is the minimum size of a batch of data to serve as a benchmark? How should benchmarks by different groups for the same task be combined together?

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* (48) characterized by the disparity between the large number of machine learning models and the small number of studies evaluating these models' performance. Furthermore, dynamic benchmarking can be combined with benchmark task misalignment methodologies (49; 50) to assess how aligned (or misaligned) model predictions are with human annotations and considering diverse examples that bring transparency to the ethical implications and societal impact to model development (51).

The transition from static benchmarks on a particular instance (or set of instances) to dynamic benchmarks on data generation processes defined by explicit desiderate may be useful for addressing the fundamental issue of construct validity that arises in singular, static benchmarks (25).

5 Addressing Robustness by Clarifying a Model's Limitations

A third approach to reducing the benchmark-production gap is to appropriately specify the contexts in which a model is expected to work via a limitations section (52).

To clarify domain-specific limitations driving the benchmark-production gap, we consider implicit assumptions that lead to a context shift in three real-world computer vision tasks: facial expression recognition, deepfake detection, and medical diagnosis.

6 Model Organism Domains for Addressing Context Shift in Applied Machine Learning

6.1 Facial Expression Recognition

In the field of affective computing, facial expression recognition (FER) is a task to classify human facial expressions with affective labels (53; 54), which can be a useful component in designing human-AI interactions with computational empathy (55; 56; 57). Model-based FER is similar to how humans recognize the emotions of others (called empathic accuracy in affective science (58) and emotion reasoning in developmental psychology (59)) 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, what happened the day before, and more.

Consider an example from relatively recent research (60) where a standard neural network architecture, AlexNet (61), is trained on a large number of images of spontaneous and posed facial expressions to classify images into seven categories (anger, disgust, fear, happiness, sadness, surprise, and neutral) and achieves accuracy scores ranging from 48.6% in SFEW (62) to 56.0% in MMI (63) to 56.1% in DISFA (64) to 61.1% in FER2013 (65) to 77.4% in FERA (66; 67) to 92.2% in CK + (68) to 94.8% in MultiPie (69). While this model's accuracy is significantly better than random guessing, which would be 14.2%, it varies dramatically depending on the chosen benchmark dataset. How should we

interpret a performance gain of 21.9 percentage points on one dataset and an average performance gain of 3.5 percentage points 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? 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 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 expression features that map to human annotations of a handful of emotion categories to classify images at significantly better than chance rates. But, it is not reported nor clear how changes in lighting, head pose, occlusion, skin tone, ethnicity, age, gender, and background scenery influence both the model's performance or human annotations. It is also underexplored how well FER models would perform if humans of diverse cultures annotated these images. Likewise, it is unclear how the model would perform on more fine-grained emotion categories (70) 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.

6.2 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 they have not said (71). These types of manipulation can be qualitatively characterized as face swapping where two people's faces are swapped, head puppetry where facial landmarks are 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 (72).

The largest deepfake detection benchmark dataset to date is the Deepfake Detection Competition Dataset (DFDC) (73; 74), 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: These are all 10 [second] videos depicting unknown actors making uncontroversial statements in nondescript locations" (35). This deepfake test benchmark is designed to evaluate algorithmic performance in 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 that involves both searching for artifacts that reveal that 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 (75)), then it is important to evaluate models on videos with these kinds of dimensions, such as those from the Presidential Deepfakes Dataset (76; 77) and the Protecting World Leaders against Deepfakes Dataset (78). 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.

6.3 Medical Diagnosis

As a third case study of context shift, we consider medical diagnosis in store-and-forward telederm-atology settings where clinical data are collected at one site and sent electronically for evaluation at another site. Recent research on machine learning applied to skin disease classification has demonstrated the human expert-level performance of models in a number of specific tasks (79; 80). However, it is unclear how these models will perform on people with dark skin because the first paper does not describe the distribution of ethnicity or skin tone in the evaluation benchmark (79) and the evaluation benchmark in the second paper contains only 2.7% of people with the second darkest of the six Fitzpatrick Skin Types (FSTs) and 1 person with the darkest of the FSTs (80). Given the accuracy disparities that appeared across skin types in facial recognition, expert intuition suggests that systematic errors are likely to also appear in skin disease classifiers.

In fact, empirical research corroborates this intuition (32), and the Diverse Dermatology Images (DDI) dataset (33) reveals that state-of-the-art skin disease classification models make systematically more errors on dark skin than on light skin. The DDI represents a more comprehensive benchmark than previous datasets, and as a result, the DDI exposed errors that should guide and motivate the future development of machine learning models towards more robustness. However, the DDI is not perfectly comprehensive; the dataset is de-identified for privacy reasons and lacks free text clinical notes and other information that physicians would acquire via an in-person examination (33). Given that many skin diseases appear similarly and expert diagnoses are based on clinical history and non-visual features, expert intuition would expect, once again, that systematic errors lurk in the state-of-the-art machine learning models for store-and-forward skin disease classification.

7 Towards Robustness in Applied Machine Learning

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 nonstationarity. 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 quickly restore and date ancient texts (81), content moderation teams can more accurately distinguish between real and fake videos (35), and general practitioners can more accurately diagnose skin conditions from images (82) (although AI advice can also mislead experts; see (83; 84; 85; 86; 87)). In fact, initial evidence suggests that human intuition is fairly accurate in predicting model misclassifications on common object detection tasks (88). 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 contain those missing contexts, and third, adjust the model accordingly. 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. Similarly, 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

[1] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and access in algorithms, mechanisms, and optimization*,

2021, pp. 1–9.

[2] R. Srinivasan and A. Chander, "Biases in ai systems," *Communications of the ACM*, vol. 64, no. 8, pp. 44–49, 2021.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[4] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," *arXiv:1712.05526 [cs]*, Dec. 2017, arXiv: 1712.05526. [Online]. Available: http://arxiv.org/abs/1712.05526

[5] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in ml-based science," *arXiv preprint arXiv:2207.07048*, 2022.

[6] M. L. Gordon, M. S. Lam, J. S. Park, K. Patel, J. T. Hancock, T. Hashimoto, and M. S. Bernstein, "Jury Learning: Integrating Dissenting Voices into Machine Learning Models," *arXiv:2202.02950 [cs]*, Feb. 2022, arXiv: 2202.02950. [Online]. Available: http://arxiv.org/abs/2202.02950

[7] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks," *arXiv:2103.14749 [cs, stat]*, Apr. 2021, arXiv: 2103.14749. [Online]. Available: http://arxiv.org/abs/2103.14749

[8] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*. IEEE, 2011, pp. 1521–1528.

[9] R. L. Thomas and D. Uminsky, "Reliance on metrics is a fundamental challenge for ai," *Patterns*, vol. 3, no. 5, p. 100476, 2022.

[10] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, Sep. 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3233231

[11] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant Risk Minimization," Mar. 2020, number: arXiv:1907.02893 arXiv:1907.02893 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1907.02893

[12] M. Makar, B. Packer, D. Moldovan, D. Blalock, Y. Halpern, and A. D'Amour, "Causally motivated shortcut removal using auxiliary labels," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 739–766.

[13] J. Adebayo, M. Muelly, H. Abelson, and B. Kim, "Post hoc explanations may be ineffective for detecting unknown spurious correlation," in *International Conference on Learning Representations*, 2021.

[14] A. Ghandeharioun, B. Kim, C.-L. Li, B. Jou, B. Eoff, and R. W. Picard, "Dissect: Disentangled simultaneous explanations via concept traversals," *arXiv preprint arXiv:2105.15164*, 2021.

[15] S. Jain, H. Lawrence, A. Moitra, and A. Madry, "Distilling model failures as directions in latent space," *arXiv preprint arXiv:2206.14754*, 2022.

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

[17] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[18] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[19] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine learning*, vol. 23, no. 1, pp. 69–101, 1996.

[20] M. C. Pelz, K. R. Allen, J. B. Tenenbaum, and L. E. Schulz, "Foundations of intuitive power analyses in children and adults," *Nature Human Behaviour*, pp. 1–12, 2022.

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

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

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

[24] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

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

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

[27] S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, J. Zittrain, I. S. Kohane, and S. Saria, "The Clinician and Dataset Shift in Artificial Intelligence," *New England Journal of Medicine*, vol. 385, no. 3, pp. 283–286, Jul. 2021. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMc2104626

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

[29] S. Mullainathan and Z. Obermeyer, "Does Machine Learning Automate Moral Hazard and Error?" vol. 107, no. 5, p. 5, 2017.

[30] S. Shankar, R. Garcia, J. M. Hellerstein, and A. G. Parameswaran, "Operationalizing machine learning: An interview study," 2022. [Online]. Available: https://arxiv.org/abs/2209.09125

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

[32] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri, "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)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 1820–1828. [Online]. Available: https://ieeexplore.ieee.org/document/9522867/

[33] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert *et al.*, "Disparities in dermatology ai performance on a diverse, curated clinical image set," *Science advances*, vol. 8, no. 31, p. eabq6147, 2022.

[34] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context Based Emotion Recognition using EMOTIC Dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019, arXiv: 2003.13401. [Online]. Available: http://arxiv.org/abs/2003.13401

[35] M. Groh, Z. Epstein, C. Firestone, and R. Picard, "Deepfake detection by human crowds, machines, and machine-informed crowds," *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, p. e2110013119, Jan. 2022. [Online]. Available: http://www.pnas.org/lookup/doi/10.1073/pnas.2110013119

[36] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019. [Online]. Available: https://www.sciencemag.org/lookup/doi/10.1126/science.aax2342

[37] J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, and H. A. Haenssle, "Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition," *JAMA Dermatology*, vol. 155, no. 10, p. 1135, Oct. 2019. [Online]. Available: https://jamanetwork.com/journals/jamadermatology/fullarticle/2740808

[38] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Re, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in *Proceedings of the ACM Conference on Health, Inference, and Learning*. Toronto Ontario Canada: ACM, Apr. 2020, pp. 151–159. [Online]. Available: https://dl.acm.org/doi/10.1145/3368555.3384468

[39] D. E. Forsythe, "Engineering knowledge: The construction of knowledge in artificial intelligence," *Social studies of science*, vol. 23, no. 3, pp. 445–477, 1993.

[40] D. Kerrigan, J. Hullman, and E. Bertini, "A survey of domain knowledge elicitation in applied machine learning," *Multimodal Technologies and Interaction*, vol. 5, no. 12, p. 73, 2021.

[41] V. Chen, U. Bhatt, H. Heidari, A. Weller, and A. Talwalkar, "Perspectives on incorporating expert feedback into model updates," *arXiv preprint arXiv:2205.06905*, 2022.

[42] S. Santurkar, D. Tsipras, and A. Madry, "BREEDS: Benchmarks for Subpopulation Shift," Aug. 2020, number: arXiv:2008.04859 arXiv:2008.04859 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2008.04859

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

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

[45] S. Mullainathan and Z. Obermeyer, "On the Inequity of Predicting A While Hoping for B," *AEA Papers and Proceedings*, vol. 111, pp. 37–42, May 2021. [Online]. Available: https://pubs.aeaweb.org/doi/10.1257/pandp.20211078

[46] S. Kerr, "On the folly of rewarding a, while hoping for b," *Academy of Management journal*, vol. 18, no. 4, pp. 769–783, 1975.

[47] M. Groh, C. Harris, R. Daneshjou, O. Badri, and A. Koochek, "Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm," *arXiv preprint arXiv:2207.02942*, 2022.

[48] Z. Epstein, B. H. Payne, J. H. Shen, A. Dubey, B. Felbo, M. Groh, N. Obradovich, M. Cebrian, and I. Rahwan, "Closing the AI Knowledge Gap," *arXiv:1803.07233 [cs]*, Mar. 2018, arXiv: 1803.07233. [Online]. Available: http://arxiv.org/abs/1803.07233

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

[50] A. Ilyas, S. M. Park, L. Engstrom, G. Leclerc, and A. Madry, "Datamodels: Understanding predictions with data and data with predictions," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9525–9587.

[51] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis) contents: A survey of dataset development and use in machine learning research," *Patterns*, vol. 2, no. 11, p. 100336, 2021.

[52] J. J. Smith, S. Amershi, S. Barocas, H. Wallach, and J. Wortman Vaughan, "Real ml: Recognizing, exploring, and articulating limitations of machine learning research," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 587–597.

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

[54] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9039580/

[55] R. W. Picard, *Affective computing*, 2000.

[56] A. Paiva, I. Leite, H. Boukricha, and I. Wachsmuth, "Empathy in virtual agents and robots: A survey," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 7, no. 3, pp. 1–40, 2017.

[57] M. Groh, C. Ferguson, R. Lewis, and R. Picard, "Computational empathy counteracts the negative effects of anger on creative problem solving," *arXiv preprint arXiv:2208.07178*, 2022.

[58] W. Ickes, "Empathic accuracy," *Journal of personality*, vol. 61, no. 4, pp. 587–610, 1993.

[59] A. L. Ruba and S. D. Pollak, "The Development of Emotion Reasoning in Infancy and Early Childhood," *Annual Review of Developmental Psychology*, vol. 2, no. 1, pp. 503–531, Dec. 2020. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev-devpsych-060320-102556

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

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

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

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

[64] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

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

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

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

[68] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "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*. IEEE, 2010, pp. 94–101.

[69] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and vision computing*, vol. 28, no. 5, pp. 807–813, 2010.

[70] A. S. Cowen, D. Keltner, F. Schroff, B. Jou, H. Adam, and G. Prasad, "Sixteen facial expressions occur in similar contexts worldwide," *Nature*, vol. 589, no. 7841, pp. 251–257, Jan. 2021. [Online]. Available: http://www.nature.com/articles/s41586-020-3037-7

[71] D. Boneh, A. J. Grotto, P. McDaniel, and N. Papernot, "Preparing for the age of deepfakes and disinformation."

[72] S. Lyu, "DeepFake Detection: Current Challenges and Next Steps," *arXiv:2003.09234 [cs]*, Mar. 2020, arXiv: 2003.09234. [Online]. Available: http://arxiv.org/abs/2003.09234

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

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

[75] J. Wakefield, "Deepfake presidents used in russia-ukraine war," Mar 2022. [Online]. Available: https://www.bbc.com/news/technology-60780142

[76] A. Sankaranarayanan, M. Groh, R. Picard, and A. Lippman, "The presidential deepfakes dataset," 2021.

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

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

[79] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017. [Online]. Available: http://www.nature.com/articles/nature21056

[80] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, V. Gupta, N. Singh, V. Natarajan, R. Hofmann-Wellenhof, G. S. Corrado, L. H. Peng, D. R. Webster, D. Ai, S. J. Huang, Y. Liu, R. C. Dunn, and D. Coz, "A deep learning system for differential diagnosis of skin diseases," *Nature Medicine*, vol. 26, no. 6, pp. 900–908, Jun. 2020. [Online]. Available: http://www.nature.com/articles/s41591-020-0842-3

[81] Y. Assael, T. Sommerschield, B. Shillingford, M. Bordbar, J. Pavlopoulos, M. Chatzipanagiotou, I. Androutsopoulos, J. Prag, and N. de Freitas, "Restoring and attributing ancient texts using deep neural networks," *Nature*, vol. 603, no. 7900, pp. 280–283, Mar. 2022. [Online]. Available: https://www.nature.com/articles/s41586-022-04448-z

[82] A. Jain, D. Way, V. Gupta, Y. Gao, G. de Oliveira Marinho, J. Hartford, R. Sayres, K. Kanada, C. Eng, K. Nagpal, K. B. DeSalvo, G. S. Corrado, L. Peng, D. R. Webster, R. C. Dunn, D. Coz, S. J. Huang, Y. Liu, P. Bui, and Y. Liu, "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*, vol. 4, no. 4, p. e217249, Apr. 2021. [Online]. Available: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2779250

[83] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, J. Paoli, S. Puig, C. Rosendahl, H. P. Soyer, I. Zalaudek, and H. Kittler, "Human–computer collaboration for skin cancer recognition," *Nature Medicine*, vol. 26, no. 8, pp. 1229–1234, Aug. 2020. [Online]. Available: http://www.nature.com/articles/s41591-020-0942-0

[84] A. Abeliuk, D. M. Benjamin, F. Morstatter, and A. Galstyan, "Quantifying machine influence over human forecasters," *Scientific Reports*, vol. 10, no. 1, p. 15940, Dec. 2020. [Online]. Available: http://www.nature.com/articles/s41598-020-72690-4

[85] S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, J. F. Coughlin, J. V. Guttag, E. Colak, and M. Ghassemi, "Do as AI say: susceptibility in deployment of clinical decision-aids," *npj Digital Medicine*, vol. 4, no. 1, p. 31, Dec. 2021. [Online]. Available: http://www.nature.com/articles/s41746-021-00385-9

[86] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos, "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection," *Translational Psychiatry*, vol. 11, no. 1, p. 108, Jun. 2021. [Online]. Available: http://www.nature.com/articles/s41398-021-01224-x

[87] M. Vaccaro and J. Waldo, "The effects of mixing machine learning and human judgment," *Communications of the ACM*, vol. 62, no. 11, pp. 104–110, 2019.

[88] Z. Zhou, M. Nartker, and C. Firestone, "When will ai misclassify? human intuition for machine (mis) perception," *Journal of Vision*, vol. 20, no. 11, pp. 1325–1325, 2020.

Checklist

1. Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
2. Did you describe the limitations of your work? [Yes]
3. Did you discuss any potential negative societal impacts of your work? [Yes]
4. Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]