This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

An Empirical Investigation into Benchmarking Model Multiplicity for Trustworthy Machine Learning: A Case Study on Image Classification

Prakhar Ganesh
Mila, Quebec AI Insitute
prakhar.ganesh@mila.quebec

Abstract

Deep learning models have proven to be highly successful. Yet, their over-parameterization gives rise to model multiplicity, a phenomenon in which multiple models achieve similar performance but exhibit distinct underlying behaviours. This multiplicity presents a significant challenge and necessitates additional specifications in model selection to prevent unexpected failures during deployment. While prior studies have examined these concerns, they focus on individual metrics in isolation, making it difficult to obtain a comprehensive view of multiplicity in trustworthy machine learning. Our work stands out by offering a one-stop empirical benchmark of multiplicity across various dimensions of model design and its impact on a diverse set of trustworthy metrics.

In this work, we establish a consistent language for studying model multiplicity by translating several trustworthy metrics into accuracy under appropriate interventions. We also develop a framework, which we call multiplicity sheets, to benchmark multiplicity in various scenarios. We demonstrate the advantages of our setup through a case study in image classification and provide actionable insights into the impact and trends of different hyperparameters on model multiplicity. Finally, we show that multiplicity persists in deep learning models even after enforcing additional specifications during model selection, highlighting the severity of over-parameterization. The concerns of under-specification thus remain, and we seek to promote a more comprehensive discussion of multiplicity in trustworthy machine learning.

1. Introduction

Deep learning has experienced a remarkable rise in recent years [29, 30, 32], with highly sophisticated and over-parameterized models leading the way [11, 31]. Consequently, these cutting-edge models find application across a diverse set of domains, including image processing [39], natural language [21], healthcare [14], finance [2], judiciary systems [42], and more, showcasing their versatility and potential impact. However, the increasing deployment of these models has sparked concerns about their trustworthiness. To confront these issues head-on, the global community has embraced a range of trustworthy machine learning practices and metrics [19, 38]. These efforts are geared towards ensuring that these models are not only accurate in their predictions, but are also fair to various groups in the dataset [3, 25], robust to distribution shifts [36], maintain the privacy of the individuals whose data was collected [13], and secure against adversarial attacks [7]. These metrics collectively strive to make deep learning deployments more reliable, fostering trust and acceptance in its widespread applications.

Alongside the discussion of trustworthy ML, the presence of multiplicity in deep learning has emerged as a significant concern yet a welcome opportunity [5, 10]. Model multiplicity is the existence of multiple high-performing models that achieve similar accuracy on a given task but can display diverse predictive behaviours due to varying decision boundaries and underlying learned functions. Model multiplicity is the result of an under-specified and over-parameterized training regime, and can be affected by design choices like model architectures, hyperparameters, training configurations, or even arbitrary choices like the randomness in training.

Model multiplicity in deep learning has significant implications. For instance, it has been shown that changes in the training configuration can lead to considerable variations in the biases present in a model [15, 27]. Deploying such models without considering the impact of multiplicity can result in the unintentional deployment of unfair models in real-world applications. Conversely, if we manage multiplicity with appropriate constraints, it presents an opportunity to deploy fairer models without compromising its utility. Thus, addressing the challenges of model multiplicity is a crucial step towards creating trustworthy systems.

Existing literature on investigating model multiplicity is limited to specialized settings that do not generalize. For instance, Somepalli *et al*. [35] provides an empirical quantification of similarity in the decision boundary of two models. However, the similarity of decision boundaries may not nec-

essarily provide any information about its trustworthiness. Models with significantly different decision boundaries can still provide similar accuracy, fairness, robustness, security, and privacy. Similarly, Ganesh *et al*. [15] investigates the impact of random seeds on fairness, but their discussion focuses on model predictions, and thus may not extend to other trustworthy metrics like robustness, security, or privacy. Furthermore, these investigations are not directly comparable to each other. For example, a 70% agreement between the decision boundaries of two models as defined by Somepalli *et al*. [35] has no comparative value to a 10% gap in equalized odds (a fairness metric) between the same set of models [15]. Thus, while these works provide a deeper investigation into a single metric in isolation, they fail to provide a comprehensive view of the overall trends of multiplicity.

In this paper, we address this gap by proposing a framework to measure multiplicity that can not only dive deeper into multiplicity trends for a single metric but also provide comparisons across different metrics. We start by converting various trustworthy metrics to a common scale, which we refer to as *accuracy under intervention*, facilitating the comparison of multiplicity across different metrics. We then create *multiplicity sheets* that capture the multiplicity of accuracy under intervention for each metric separately. To illustrate our framework, we present an image classification case study that compares multiplicity across model hyperparameters, random seeds, and architecture choices, repeating the setup for various trustworthy ML metrics, namely fairness, robustness, privacy, and security.

We end our discussion by presenting the results of combining various metrics together to improve the model specification and reduce multiplicity. However, despite following the recommendations of recent literature and providing additional specifications using trustworthy metrics, our study reveals that model multiplicity can still create unforeseen failure cases. This highlights the need for future research to gain a more holistic understanding of model multiplicity.

Setting Expectations and Our Contributions: Before delving into our contributions, it is essential to first clarify the scope of our work. Our goal is not to present novel findings on the multiplicity of any specific metric. In fact, we will revisit many existing results in the literature during our case study. Rather, we seek to establish a normative language to record model multiplicity that can be used to highlight multiplicity trends across different metrics, thus providing an overall picture of multiplicity in deep learning.

More specifically, our contributions are:

- We propose a standardized framework to measure and study model multiplicity in deep learning.
  - We introduce a new class of metrics called accuracy under intervention. We showcase techniques to convert any metric into accuracy using appropriate interventions, thus providing a common scale of comparison.
  - We suggest using multiplicity sheets, a comprehensive yet compact method to record and study model multiplicity for any target metric.
- We present a case study of model multiplicity in image classification, by providing an empirical benchmark to highlight the advantages of our framework.
  - We take an all-encompassing view of model multiplicity and its impact on trustworthy ML by comparing multiplicity across fairness, robustness, privacy, and security.
  - We study the influence of various axes of model variations on multiplicity, including model architecture, training randomness, and hyperparameter choices.
- We combine several trustworthy metric specifications to challenge over-parameterization and assess its impact on multiplicity. Despite this, we see persistent multiplicity on trustworthy issues not seen during model selection, underscoring the need for better safeguards against multiplicity when deploying models in the real world.

2. Measuring Multiplicity

We will start by discussing our framework to study multiplicity. We introduce the concept of *accuracy under intervention* to translate any metric into an accuracy metric, followed by our proposal to use *multiplicity sheets* to record and compare the said accuracy under intervention.

2.1. Accuracy Under Intervention

We want to establish a standardized way to measure model multiplicity that would allow easy comparison across different scenarios. However, measuring multiplicity for various trustworthy objectives relies on vastly different metrics, making it a complex task. For instance, comparing accuracy multiplicity (difference in accuracy) with security multiplicity (difference in minimum adversarial distance to flip the label), is not straightforward. These two metrics are not comparable since they are based on different factors, one on performance and the other on distance.

We need a method that can translate these metrics to a common scale for a fair comparison. To achieve this, we propose converting each metric into accuracy through appropriate interventions. Simply put, we want to measure model accuracy under a well-designed intervention that represents a proxy for our original trustworthy metric. For instance, when testing the security of a model against adversarial attacks, instead of measuring the minimum adversarial distance to flip the label, we can measure the model accuracy under a fixed adversarial distance budget. Once a metric is translated to accuracy under such an intervention, we can now compare

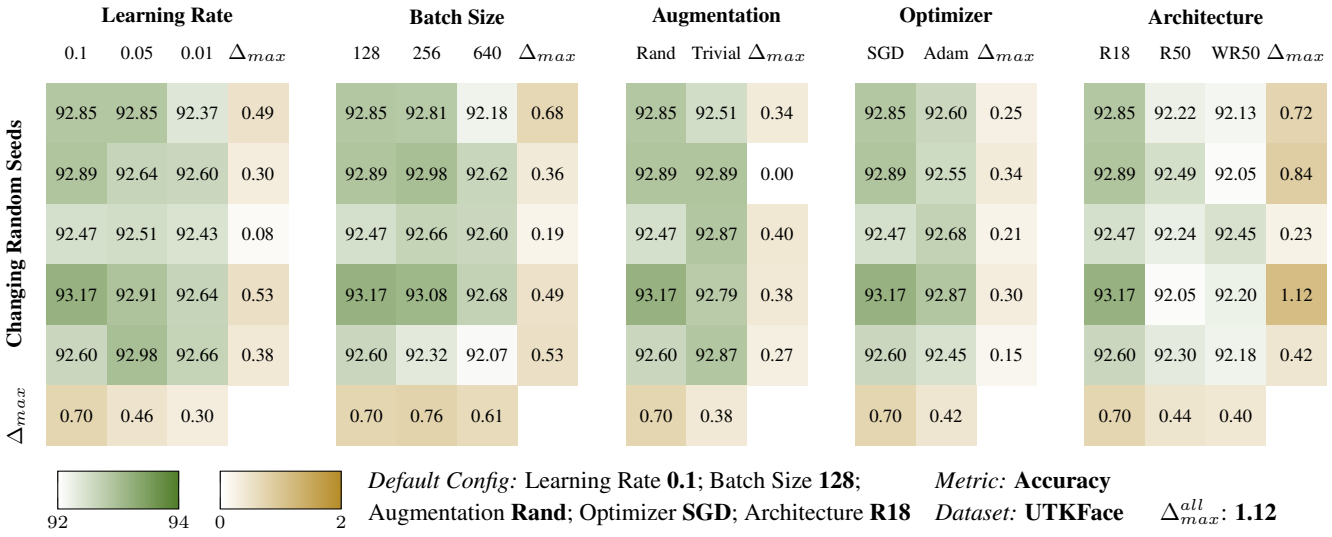

Figure 1. *Multiplicity sheet* for **Accuracy** on **UTKFace** dataset. R18/50: ResNet-18/50; WR50: WideResNet-50x2. This multiplicity sheet records accuracy scores across different hyperparameter choices and random seeds, representing the first level of readability without any loss of information. It then aggregates multiplicity by measuring Δ_{max} across random seeds for each hyperparameter and across hyperparameter choices for each random seed. This is the second level of readability, vital for extracting multiplicity trends. For instance, by studying the Δ_{max} values, we see the equal importance of both random seeds and hyperparameter choices on accuracy multiplicity. Finally, we aggregate the overall multiplicity Δ^{all}_{max}, i.e., the third level of readability, condensing accuracy multiplicity for UTKFace into a single value.

them directly to each other and get a comprehensive understanding of the multiplicity. More details on the specific interventions for each metric are present in Section 3.

2.2. Multiplicity Sheets

After converting every metric to accuracy under intervention, we propose a method for recording these values to facilitate easy comparison and visualization. We want a method that can provide both summaries for a quicker scan and detailed results for a more in-depth analysis. To achieve this, we create multiplicity sheets, a straightforward and highly intuitive approach to documenting multiplicity.

A multiplicity sheet is a collection of tables, where each table compares two axes of multiplicity. The information in our multiplicity sheets has three levels of readability. The first level shows the raw metric scores, in this case, accuracy under intervention, ensuring no loss of information. The second level aggregates multiplicity across each axis in every table by taking the maximum difference in scores denoted by Δ_{max}. This allows easy visualization of various trends and the influence of different hyperparameters on model multiplicity. Finally, the third level further aggregates the complete multiplicity sheet by taking the maximum difference across all raw metric scores to get a single value representing the overall multiplicity of the given metric, denoted as Δ^{all}_{max}. Given that we use accuracy under intervention for all metrics, Δ^{all}_{max} serves as a useful measure to compare multiplicity across different metrics, i.e., different multiplicity sheets.

An example of a multiplicity sheet can be seen in Fig. 1, where we record the accuracy multiplicity on the UTKFace dataset under various training configurations (more details in Section 3). Throughout our paper, we will designate one axis of multiplicity in each table to be the random seeds. This is to filter chance trends when comparing different hyperparameters by balancing them against multiple runs with changing random seeds. It should be noted that multiplicity sheets can be created for any metric, not just accuracy under intervention. However, using accuracy under intervention allows us to compare the multiplicity trends across different sheets, which wouldn't be possible with just any metric. We will now move to our case study to highlight the benefits of our framework, while also providing an empirical benchmark of multiplicity in image classification that can be directly useful for researchers and practitioners.

3. Image Classification on UTKFace

To demonstrate the utility of our framework, we will perform a case study of the model multiplicity in image classification on the UTKFace dataset. We first outline our experimental setup, followed by a comprehensive discussion of multiplicity in fairness, robustness, privacy, and security. To provide diversity in experiments, we also perform a separate case study on the CIFAR10 dataset in Appendix A.

3.1. Experiment Setup

Dataset We will be studying the UTKFace dataset, containing facial images that have been labelled according to

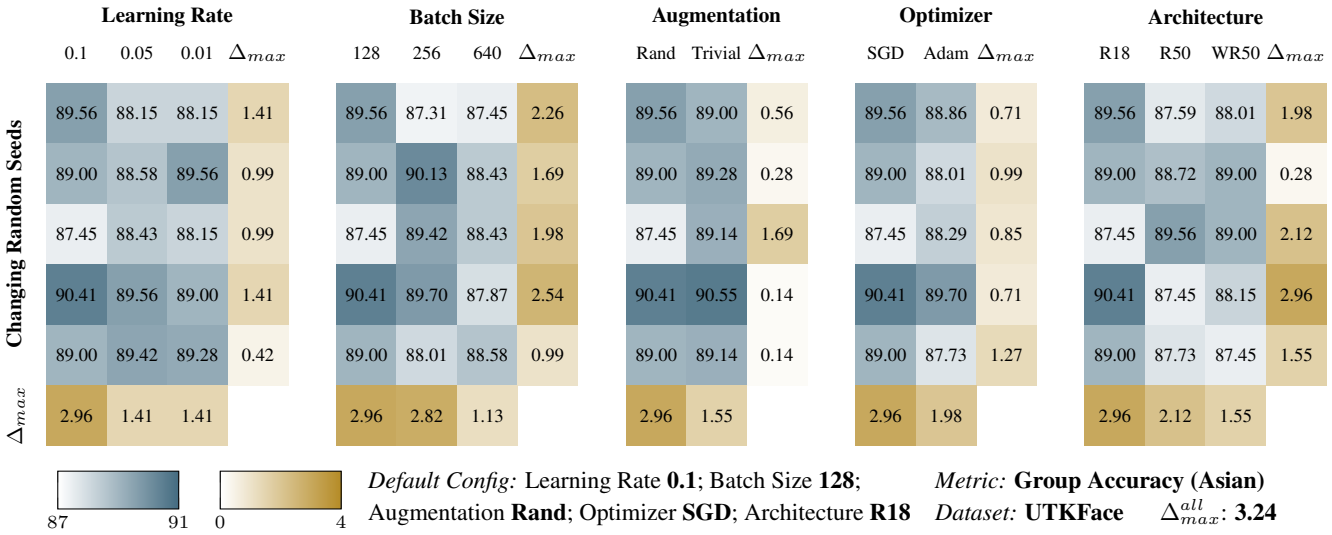

Figure 2. *Multiplicity sheet* for **Group Accuracy (Asian)** on **UTKFace** dataset. R18/50: ResNet-18/50; WR50: WideResNet-50x2. Among various hyperparameter choices, the batch size and architecture stand out in their influence on fairness multiplicity. However, it is the variance across changing random seeds that overshadows all other sources of multiplicity, making it the most important factor for fairness during model selection. The overall fairness multiplicity (Δ^{all}_{max}) is also almost 3 times higher than the accuracy multiplicity seen earlier.

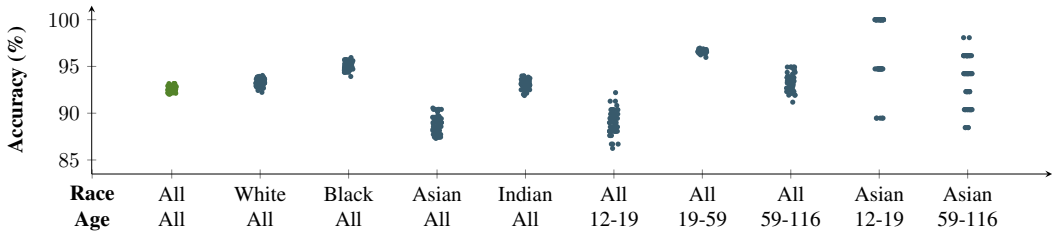

Figure 3. Distribution of fairness multiplicity (i.e., group accuracy) across different intersections of racial and age groups in the UTKFace dataset. Each distribution is a condensed representation of a multiplicity sheet, containing the group accuracy of 65 independently trained models across all axes of multiplicity described in Section A.1. Different groups have varying ranges of multiplicity, with a specially amplified variance from intersectional groups, highlighting the concerns and opportunities of multiplicity in fairness. Note: The minor perturbations along the x-axis for any group are only present for enhanced visualization and do not convey any additional signal.

their perceived gender, race, and age. We will focus on the binary classification task of perceived gender. We split the dataset into 80% training and 20% testing, and we maintain the same split throughout our paper, i.e., we do not consider potential multiplicity introduced by the train-test splits.

Training Details By default, we train our models using a learning rate of 0.1, a batch size of 128, the data augmentation RandAugment [9], the SGD optimizer, and the ResNet-18 architecture [16]. For a simpler analysis, all models are trained from scratch, i.e., without the use of pre-trained weights. We use a single random seed to control all forms of randomness in model training. We leave the decoupled analysis of various sources of randomness for future work. Finally, all models are trained with cross-entropy (CE) loss for 50 epochs, without any early stopping.

Axes of Multiplicity We will investigate the following different axes of multiplicity, (i) *Learning Rate:* $\{0.1, 0.05, 0.01\}$, (ii) *Batch Size:* $\{128, 256, 640\}$, (iii) *Data Augmentation:* RandAugment [9] and TrivialAugment [26], (iv) *Optimizer:* SGD and Adam, (v) *Model Architecture:* ResNet-18 [16], ResNet-50 [16], and WideResNet-50x2 [43]. We also compare multiplicity across changing randomness in model training.

3.2. Group Fairness

Group fairness is a measure of performance disparity between different protected groups, rooted in concerns of algorithmic bias propagated from the dataset to the model [1, 4, 8, 44]. Traditionally, group fairness is measured as the difference in performance between different groups in the dataset. For accuracy under intervention in our setup, we calculate the accuracy on the minority group (which can also

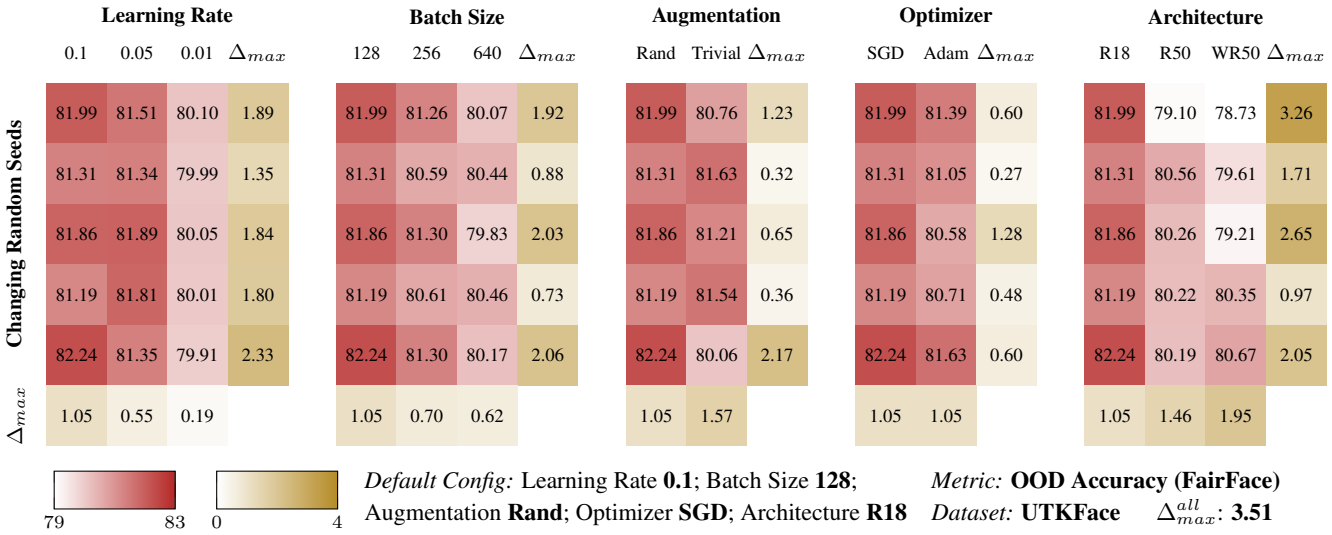

Figure 4. *Multiplicity sheet* for **OOD Accuracy (FairFace)** on **UTKFace** dataset. R18/50: ResNet-18/50; WR50: WideResNet-50x2. The learning rate and batch size both seem to noticeably influence the OOD robustness, while it's the architecture choice that dominates any other factor for robustness multiplicity. The overall multiplicity (Δ^{all}_{max}) is slightly higher than the fairness multiplicity seen previously.

be extended to other groups). More specifically, we consider racial labels for fairness and measure the accuracy of the racial minority in the UTKFace dataset, i.e., *Asians*.

In Fig. 2, we present the multiplicity sheet for Group Accuracy (Asian). We use the sheet to highlight the importance of random seeds in fairness and also contrast it to the multiplicity sheet for Accuracy in Fig. 1. As can be seen clearly from the Δ_{max} values of changing random seeds compared to different hyperparameter choices, random seeds have the most significant impact on fairness multiplicity. Moreover, we can also observe that the overall fairness multiplicity ($\Delta^{all}_{max} = 3.24$) is three times higher than the accuracy multiplicity ($\Delta^{all}_{max} = 1.12$). These trends of fairness variance and the impact of random seed have been previously noted in literature [15, 34], however here we show the ease with which they can be spotted in our multiplicity sheets.

We repeat the experiment for various groups and plot the distribution of fairness multiplicity across all axes of multiplicity, in Fig. 3, with groups formed at the intersection of two different protected attributes, i.e., race and age. Our findings reveal that the severity of fairness multiplicity is even higher for intersectional groups. To put this into perspective, consider selecting a model from the distribution of models in Fig. 3. While the choice may only affect the overall accuracy in the range of 92.05% to 93.17%, it can significantly alter the accuracy for older Asian individuals, ranging from 88.46% to 98.08%. To sum up, our analysis clearly shows the alarmingly high fairness variance, and the need to address this multiplicity such that deep learning models treat diverse groups fairly during deployment.

3.3. Out-of-Distribution Robustness

Out-of-distribution (OOD) robustness refers to the ability of a machine learning model to perform well on data points that are different from those it was trained on. Models that lack OOD robustness might make unreliable or incorrect predictions when faced with new, unfamiliar data, potentially leading to undesirable outcomes after deployment. Traditionally, OOD robustness is measured as the model's performance on an OOD dataset. Since it is already an accuracy metric, we do not perform any additional intervention for robustness. More specifically, we simply use the model's accuracy on the FairFace dataset [18], a facial image dataset with a different distribution than UTKFace, as the measure of OOD robustness multiplicity.

In Fig. 4, we present the multiplicity sheet for Accuracy on the FairFace dataset. We see the impact of learning rate, batch size, and architecture on robustness emerge from the multiplicity sheets, an unsurprising result based on existing work on the benefits of smaller batch size, larger learning rate, and higher risks of overfitting in bigger models [17, 24]. The range of overall robustness multiplicity ($\Delta^{all}_{max} = 3.51$) is of the same range as fairness multiplicity, i.e. three times higher than accuracy multiplicity. Thus, addressing multiplicity in OOD robustness is essential to making sure the model doesn't fail even under minor distribution shifts.

3.4. Differential Privacy

Deep learning models tend to memorize data points from their training dataset, compromising the privacy of the individuals in the dataset. For instance, an adversary with access to only the outputs of the model is capable of extracting

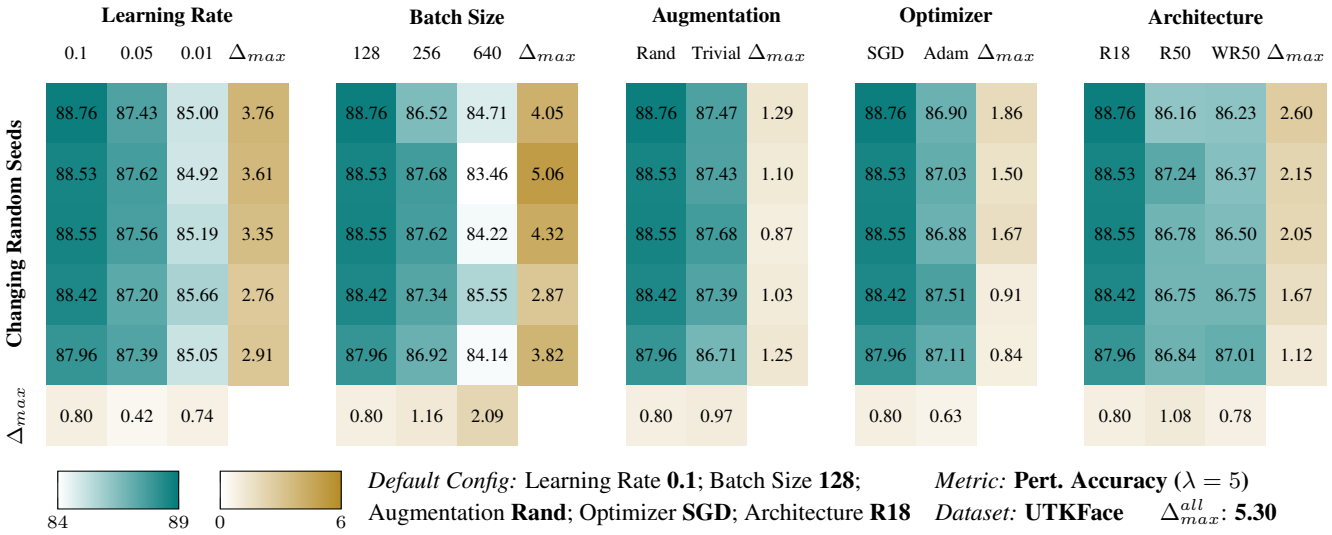

Figure 5. *Multiplicity sheet* for **Perturbation Accuracy** ($\lambda = 5$) on **UTKFace** dataset. R18/50: ResNet-18/50; WR50: WideResNet-50x2. The random seed has very little influence on the perturbation accuracy, while the default hyperparameter choices are noticably dominant over other hyperparameters, with the biggest drop caused by using a large batch size. The overall multiplicity (Δ^{all}_{max}) is also quite high, five times larger than accuracy multiplicity, but clearly dependent on the choice of the rate parameter λ.

sensitive information from the model [6, 33]. To address this issue, researchers often study differential privacy [12], which aims to make models trained on datasets differing at exactly one data point indistinguishable. One way to achieve this is by adding noise to the model's outputs. However, adding noise can also hurt the model's performance, thus creating a trade-off between privacy and accuracy. It is this very trade-off that we will exploit to define our accuracy under intervention, i.e., we will measure the accuracy of the model under output perturbations from an exponential distribution with a fixed *rate parameter* λ, for privacy multiplicity.

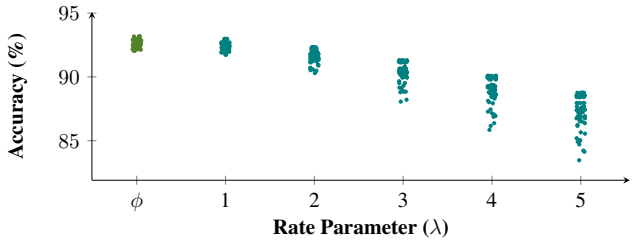

Figure 6. Distribution of privacy multiplicity (i.e., perturbation accuracy) across different values of the rate parameter λ. The higher the rate parameter, the larger the output perturbations, which in turn creates larger drops in accuracy and a larger range of multiplicity. Refer to Fig. 3 for further details on distribution visualization.

We present the multiplicity sheet for privacy by recording the Perturbation Accuracy with $\lambda = 5$ in Fig. 5. Interestingly, unlike other trustworthy metrics, the random seed has minimal impact on the privacy multiplicity. Instead, one hyperparameter choice along each axis is clearly the best when it comes to the privacy-accuracy trade-off, in line with the existing literature on practical tips for privacy [28]. Additionally, the overall privacy multiplicity range ($\Delta^{all}_{max} = 5.30$) is almost five times larger than the accuracy multiplicity. These results are clearly dependent on the rate parameter λ used to calculate accuracy under intervention and to emphasize this, we plot the distribution of privacy multiplicity for different rate parameter values in Fig. 6. It is evident that choosing the right model by accounting for privacy multiplicity is crucial to achieving better privacy-utility trade-offs during inference, and this choice becomes even more critical with a decrease in privacy budget (i.e., higher values of rate parameter λ).

3.5. Security against Adversarial Attacks

Machine learning models are vulnerable to various attacks that can manipulate the model's behaviour to suit the attacker's desires. One of the most common adversarial attacks studied in literature is the perturbation-based attack [37], which takes advantage of the brittle decision boundaries of deep learning models. In this attack, the objective of the adversary is to perturb the input image by a minimum amount that can incite adversarial outputs, while keeping the perturbation imperceptible to the human eye. Instead of measuring the minimum distance of the perturbed image to the original image (measured as L_∞), we will measure the accuracy under the intervention of a fixed distance budget represented by δ. Specifically, we use projected gradient descent (PGD) [23] to progressively move out of the local minima until we reach the given distance budget, and then measure the accuracy of the model under these perturbations.

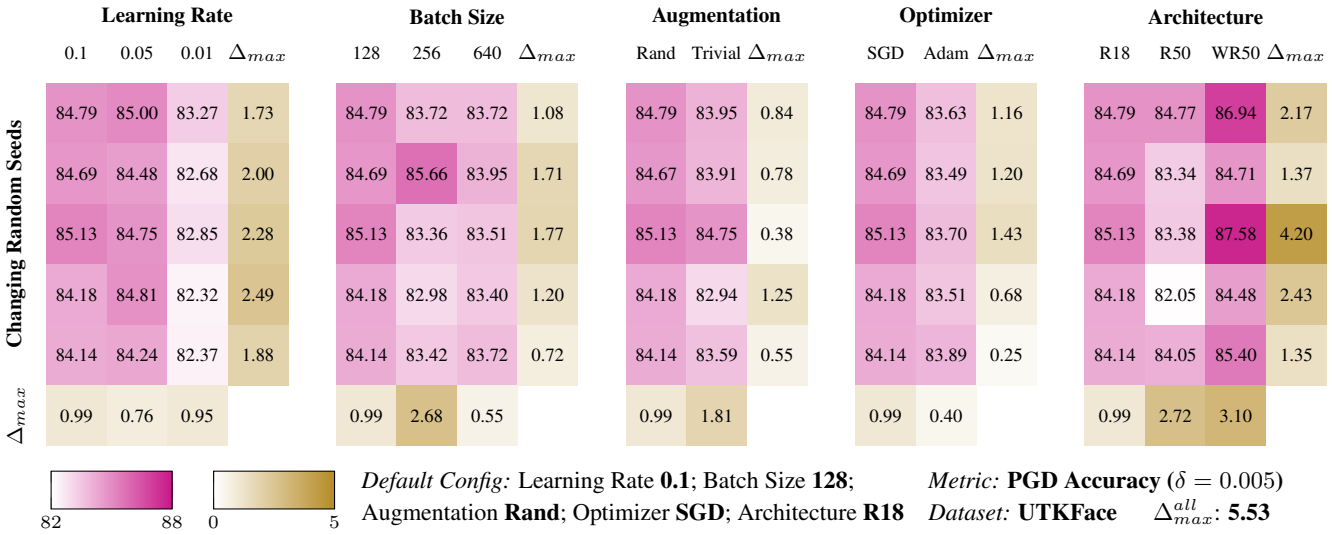

Figure 7. *Multiplicity sheet* for **PGD Accuracy** ($\delta = 0.005$) on **UTKFace** dataset. R18/50: ResNet-18/50; WR50: WideResNet-50x2. The architecture choice stands out as the most influential factor in security multiplicity. The overall multiplicity (Δ^{all}_{max}) is five times larger than the accuracy multiplicity, dependent on the choice of the adversarial distance budget δ.

We present the multiplicity sheet for PGD Accuracy with $\delta = 0.005$ in Fig. 7. The trends for security multiplicity are similar to the accuracy multiplicity trends we observed previously, i.e., no single factor dominates the multiplicity, except for architecture choice. Surprisingly, the larger model ResNet-50 had a negative impact on security multiplicity, while the even larger but wider model WideResNet-50x2 improved it, which contradicts previous findings in literature [41] and raises interesting questions for future research. Similar to privacy multiplicity, the overall multiplicity range ($\Delta^{all}_{max} = 5.53$) for security is almost five times larger than the accuracy multiplicity and depends on the adversarial distance budget δ. We plot the distribution of security multiplicity for different adversarial distance budget values in Fig. 8. Our results have shown that a model's robustness to adversarial attacks suffers from severe multiplicity, and needs to be addressed to provide robust models.

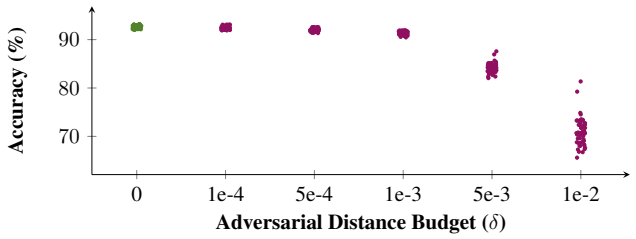

Figure 8. Distribution of security multiplicity (i.e., accuracy after PGD) across different adversarial distance budget values δ. A higher budget corresponds to a more powerful adversary, which in turn results in lower accuracy and higher security multiplicity. Refer to Fig. 3 for further details on distribution visualization.

4. Model Selection

In our case study, we found significant multiplicity in various trustworthy metrics that can hurt model deployment, if left unchecked. To address this multiplicity, the literature suggests providing appropriate specifications during model selection [5]. This involves imposing additional constraints based on some chosen metrics, in our case the trustworthy metrics. For example, one can measure the fairness scores of the model under different hyperparameters and only choose the configurations with bias scores less than some threshold. This ensures that unfair models are not selected.

These recommendations stem from the belief that implementing extra measures during the selection of a model will decrease its variability, ensuring predictable behaviour upon deployment. However, as we will demonstrate in this section, over-parameterized models can still encounter unforeseen failure cases during deployment, which are not simply solved with appropriate specifications during model selection.

Model Specifications: We first define the following criteria to simulate model selection. We choose models that rank in the top k% of every metric under varying training configurations. We assess fairness by measuring accuracy for the Asian racial group, robustness by evaluating test performance on FairFace, privacy by measuring accuracy under output perturbations with a rate parameter $\lambda = 5$, and security by measuring accuracy under PGD attacks with an adversarial distance budget of $\delta = 0.005$. Our approach ensures that we only select models that meet the high standards for all four metrics mentioned above.

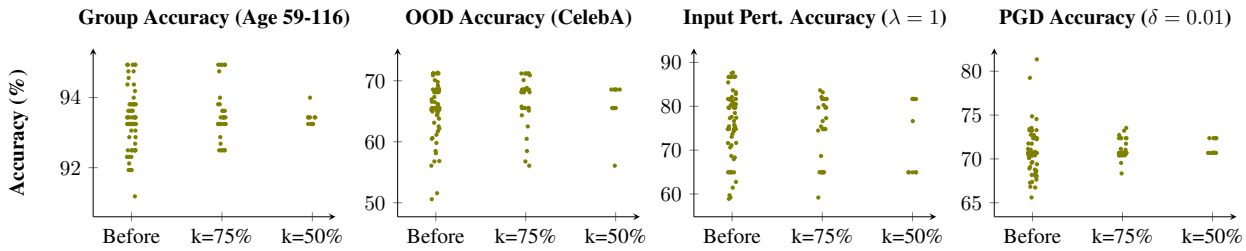

Figure 9. Distribution of multiplicity across unforeseen metrics under various degrees of model selection. The range of multiplicity across unforeseen metrics might remain unchanged even after we provide additional specifications for known trustworthy metrics, highlighting the severity of over-parameterization and the need to address multiplicity beyond just a checklist of metrics.

Unforeseen Circumstances: We now introduce a new set of metrics to account for situations that were not previously considered in our specifications. To simplify the discussion, we will just make minor adjustments to the model specifications and create these 'unforeseen circumstances'. To test fairness, we will measure group accuracy for the age group $59 - 116$ instead of the Asian racial group. To test robustness, we will evaluate the performance on the CelebA dataset [22] instead of FairFace. To test privacy, we will measure accuracy under input perturbations (with a rate parameter of $\lambda = 1$) instead of output perturbations. Finally, to test security, we will increase the distance budget from $\delta = 0.005$ to $\delta = 0.01$, thus creating a stronger adversary.

In Fig. 9, we plot the distribution of multiplicity for all four unforeseen metrics before any model selection, and then after model selection for $k\% = 75\%$ and $k\% = 50\%$ respectively. We see a noticeable drop in unforeseen fairness and security multiplicities while maintaining decent fairness and security accuracy scores under intervention. However, we do not see this improvement in unforeseen robustness or privacy multiplicity. That is, despite the highly rigorous model selection on four different trustworthy ML metrics, the overall range of multiplicity in these two unforeseen metrics remains the same, and thus they will face the same issues during deployment. Clearly, incorporating additional specifications while selecting models can only provide limited assistance, leaving a substantial level of multiplicity that cannot be managed in the same way. Thus, addressing multiplicity with a checklist of trustworthy requirements is still likely to create models that face the same risks of failure in unforeseen circumstances, emphasizing the need for a more fundamental investigation into model multiplicity.

5. Related Work

Model multiplicity has been an active subject of research in the deep learning literature, despite not being in the spotlight. Much of the related work in multiplicity is indirect, often disguised as research on the impact of hyperparameter choices or randomness on trustworthy ML [15, 17, 24, 28, 34].

Very few works in literature have focused solely on multiplicity. Black *et al*. [5] provides a discussion on the opportunities and concerns of multiplicity within the context of machine learning. However, their work is highly qualitative and does not provide any framework to quantify and measure multiplicity. On the other hand, D'Amour *et al*. [10] offer a more quantitative perspective to underspecification in machine learning. Nevertheless, their analysis is fragmented across different case studies and does not provide a common language on multiplicity measurement that can be adapted for future works on model multiplicity.

6. Conclusion and Future Work

In this paper, we contribute to the discussion of model multiplicity, specifically in the context of image classification. By establishing a consistent and comprehensive language for multiplicity, we have created a foundation for more effective communication in the field. Our multiplicity sheets offer an intuitive and structured approach to capturing the various facets of multiplicity. Furthermore, through a detailed case study, we demonstrated the practical implementation of our framework, shedding light on the complexities that arise when dealing with model multiplicity. The insights derived from the case study not only showcased the utility of our approach but also unveiled intriguing trends within the multiplicity scores. Finally, we show empirically that the challenge of model multiplicity cannot be simply resolved by providing additional specifications or constraints.

While we emphasize a specific structure for multiplicity sheets in this paper, it is important to acknowledge that further research is required to develop more effective methods for recording multiplicity. Moreover, our recommendation to use accuracy under intervention is primarily applicable to classification tasks. Nevertheless, the challenge of model multiplicity is a major issue in deep learning that goes beyond classification alone. Consequently, it is imperative that the community engages in further discussion on the topic of model multiplicity. We must shift away from treating multiplicity as an auxiliary discussion and bring it to the forefront to address potential unforeseen failures in real-world deployment scenarios and create truly trustworthy systems.

References

[1] Mohsen Abbasi, Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Fairness in representation: quantifying stereotyping as a representational harm. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 801–809. SIAM, 2019. 4

[2] Shamima Ahmed, Muneer M Alshater, Anis El Ammari, and Helmi Hammami. Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*, 61:101646, 2022. 1

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017. 1

[4] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016. 4

[5] Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, 2022. 1, 7, 8

[6] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021. 6

[7] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021. 1

[8] Kate Crawford. The hidden biases in big data. *Harvard business review*, 1(4), 2013. 4

[9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 4

[10] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research*, 23(1):10237–10297, 2022. 1, 8

[11] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 1

[12] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006. 6

[13] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. 1

[14] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019. 1

[15] Prakhar Ganesh, Hongyan Chang, Martin Strobel, and Reza Shokri. On the impact of machine learning randomness on group fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1789–1800, 2023. 1, 2, 5, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[17] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017. 5, 8

[18] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 5

[19] Michael Kearns and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, 2019. 1

[20] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 11

[21] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 2022. 1

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 8

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 6, 11

[24] Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018. 5, 8

[25] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021. 1

[26] Samuel G Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 774–782, 2021. 4

[27] Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. Fair bayesian optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 854–863, 2021. 1

[28] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023. 6, 8, 11

[29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[31] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 1

[32] Koosha Sharifani and Mahyar Amini. Machine learning and deep learning: A review of methods and applications. *World Information Technology and Engineering Journal*, 10(07):3897–3904, 2023. 1

[33] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. 6

[34] Ioana Baldini Soares, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh, and Mikhail Yurochkin. Your fairness may vary: pretrained language model fairness in toxic text classification. In *Annual Meeting of the Association for Computational Linguistics*, 2022. 5, 8

[35] Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13699–13708, 2022. 1, 2

[36] Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In *International conference on artificial intelligence and statistics*, pages 2611–2619. PMLR, 2021. 1

[37] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014. 6

[38] Kush R Varshney. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3):26–29, 2019. 1

[39] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, et al. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018. 1

[40] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020. 11

[41] Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? *Advances in Neural Information Processing Systems*, 34:7054–7067, 2021. 7

[42] S Yassine, M Esghir, and O Ibrihich. Using artificial intelligence tools in the judicial domain and the evaluation of their impact on the prediction of judgments. *Procedia Computer Science*, 220:1021–1026, 2023. 1

[43] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016. 4

[44] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017. 4