values seen in Table 5, the smallest standard error in each metric is split between the two models. However, the standard error for precision is much smaller in the model with outlier exposure when compared to the model without.

Training Group	*Precision*	*Recall*	*Accuracy*	*F1*	*AUROC*
FairFace	0.59 ± 0.020	0.61 ± 0.007	0.55 ± 0.007	0.60 ± 0.009	0.57± 0.008
FairFace w/ OE	**0.62 ± 0.007**	**0.64±0.007**	**0.58±0.008**	**0.63±0.005**	**0.59±0.008**

Table 6: Averaged Results from testing on CelebA when classifying based on gender

Similarly to the LFW dataset, we set the weights for the male and female class to 1 and 1.15 respectively.

The model is also tested on the CelebA dataset to classify gender. The results shown in Table 6 are consistent with the ones seen in Table 5, where our improved model performs better in both fairness metrics and accuracy.

Due to averaging multiple trials, we also observe the standard error between both models. When comparing the values seen in Table 6, it can be seen that on average, the standard error of the precision of the model with outlier exposure is smaller than that of the model without. A lower standard error means that the results are less varied and more consistent.

Over the course of all experiments, we find that in general, our approach helps to reduce false negatives and false positives. This can be seen by observing the recall and precision in Table 5 and Table 6, in which both fairness metrics increase with our model. As such, this correlates to a lower false negative and false positive rate, which is very important as different applications suffer more from different false flagging. In contrast to the healthcare application mentioned prior, false positives can be particularly problematic in law enforcement applications, where a subject could be mislabeled as the culprit for a crime they did not commit. As such, being able to reduce both false positive and negative rates are important.

6. Conclusions. In this paper, we explore the concept of facial recognition classification. Specifically, we focus on how models struggle when confronted with data from different distributions during training and testing. Out-of-distribution data has been shown to significantly reduce the accuracy of facial recognition models. Throughout this paper, we document the results of a CNN model on different training and testing datasets. We also use KL distribution to identify outlier images from each dataset and incorporate Outlier Exposure into our model to see how it affected the model's accuracy and other metrics. We also attempt to weight the sampling of the different classes (male and female) and observe how the model's results change. Overall, we evaluate how facial recognition performs on out-of distribution data and conclude that outlier exposure can increase accuracy and other metrics of the model. We also conclude that utilizing weight sampling and outlier exposure can improve both the accuracy and the fairness of the model on out-of-distribution facial image sets. With the use of artificial intelligence and facial recognition in sectors such as law enforcement and healthcare, the lack of fairness and accurate classification of out-of-distribution data (often data & images of minority groups) has become an increasingly pressing issue. The methodologies we explore here can improve the metrics of these models, and we find that many techniques rely on having some knowledge about the dataset distributions, which requires a carefully considered

implementation of these different methods to maximize the metrics of concern.

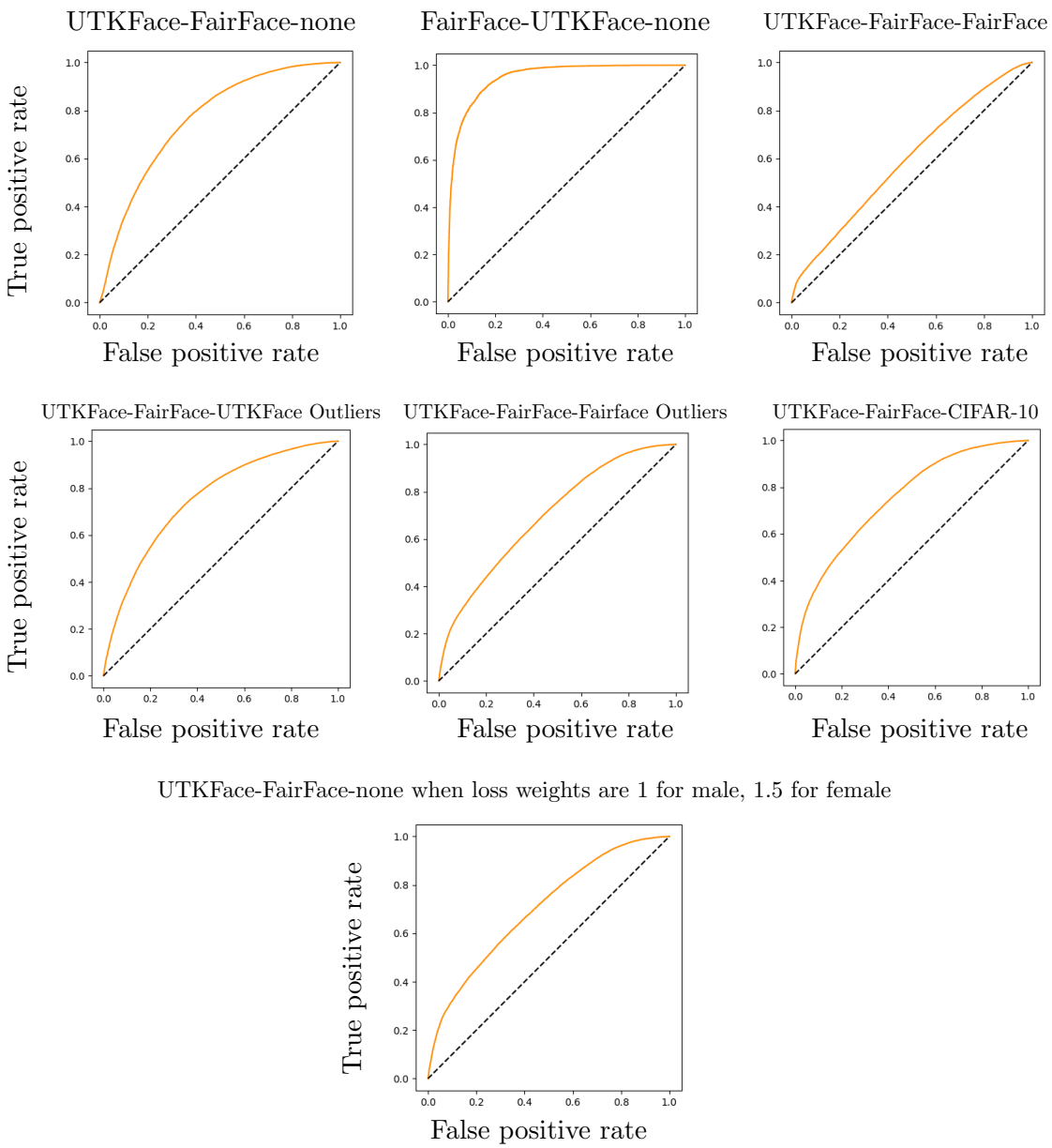
For this paper, all experiments are performed in Python using the Pytorch package, with data analysis performed using Scikit-learn package. The hardware we use is a 16 GB RAM and a 6 core, 2.6 GHz CPU. Additionally, our code will be made available upon request.

6.1. Related & Future Work. While this paper is a good overview of current methods, there are alternate avenues that future researchers may want to explore to increase accuracy. One method is Geometric Sensitivity Decomposition, which works with feature norms of images after they have run through a neural network. Besides the importance weighting and outlier exposure perspectives we looked into, another region of focus is through bringing the distributions of the train and test datasets closer through the modification of the distributions. Working with this and incorporating geometric sensitivity decomposition would represent our next avenue to improve out-of-distribution images.

Along the same lines of fairness transfer, [25] develops a causal approach using conditional independence tests to characterize distribution shifts in healthcare machine learning, revealing that understanding such shifts can diagnose fairness discrepancies and suggesting potential mitigation strategies throughout the ML pipeline. Additional modifications to our CNN network could represent an area to improve the facial recognition of out-of-distribution images.

Acknowledgments. We would like to thank our mentor, Dr. Nicole Yang. This work was supported in part by the US NSF award DMS-2051019.

Appendix A. ROC Curves for Experiments.



UTKFace-FairFace-none when loss weights are 1 for male, 1.5 for female

Figure 6: ROC Curve from all experiments used, labelled in the order of "train set-test set-outlier set". Orange is the classifier performance and the dashed line is for 50% accuracy.

Appendix B. Confusion Matrix for Experiments.

Train-Test-Outlier	*Male-Male*	*Male-Female*	*Female-Male*	*Female-Female*
UTKFace-FairFace	33711	14274	12274	26484
FairFace-UTKFace	10818	1477	1573	9837
UTKFace-UTKFace	12079	226	312	11088
FairFace-FairFace	42657	3754	3328	37004
UTKFace-FairFace-UTKFace	36306	15304	9679	25454
UTKFace-FairFace-FairFace	26010	19041	19975	21717
UTKFace-FairFace-UTKFace Outliers	35693	16079	10292	24679
UTKFace-FairFace-FairFace Outliers	33636	16952	12349	23806
UTKFace-FairFace-CIFAR-10	36486	17144	9499	23614
UTKFace-FairFace (weights 1-1.5)	32003	15503	13982	25255

Table 7: Confusion Matrices from all experiments used. Columns labeled by "Predicted Label-Actual Label"

REFERENCES

[1] S. Azam, S. Montaha, K. U. Fahim, A. R. H. Rafid, M. S. H. Mukta, and M. Jonkman, *Using feature maps to unpack the cnn 'black box' theory with two medical datasets of different modality*, Intelligent Systems with Applications, 18 (2023), p. 200233.

[2] 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.

[3] J. Byrd and Z. Lipton, *What is the effect of importance weighting in deep learning?*, in International conference on machine learning, PMLR, 2019, pp. 872–881.

[4] J. Duan and C.-C. Jay Kuo, *Bridging gap between image pixels and semantics via supervision: A survey*, APSIPA Transactions on Signal and Information Processing, 11 (2022).

[5] K. Fukushima, *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*, Biological Cybernetics, 36 (1980), pp. 193–202.

[6] S. Galbraith, J. A. Daniel, and B. Vissel, *A study of clustered data and approaches to its analysis*, The Journal of Neuroscience, 30 (2010), pp. 10601–10608.

[7] M. Hashemi, *Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation*, Journal of Big Data, 6 (2019).

[8] M. Hashemi, *Web page classification: A survey of perspectives, gaps, and future directions*, Multimedia Tools Appl., 79 (2020), p. 11921–11945.

[9] D. Hendrycks and K. Gimpel, *A baseline for detecting misclassified and out-of-distribution examples in neural networks*, in International Conference on Learning Representations, 2017.

[10] D. Hendrycks, M. Mazeika, and T. Dietterich, *Deep anomaly detection with outlier exposure*, in International Conference on Learning Representations, 2019.

[11] D. Hirahara, E. Takaya, T. Takahara, and T. Ueda, *Effects of data count and image scaling on deep learning training*, PeerJ Computer Science, 6 (2020), p. e312.

[12] J. P. Horwath, D. N. Zakharov, R. Mégret, and E. A. Stach, *Understanding important features of deep learning models for segmentation of high-resolution transmission electron microscopy images*, npj Computational Materials, 6 (2020).

[13] C. Hu, B. B. Sapkota, J. A. Thomasson, and M. V. Bagavathiannan, *Influence of image quality and light consistency on the performance of convolutional neural networks for weed mapping*, Remote Sensing, 13 (2021).

[14] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*, Tech. Rep. 07-49, University of

Massachusetts, Amherst, October 2007.
[15] D. H. Hubel and T. N. Wiesel, *Receptive fields and functional architecture of monkey striate cortex*, The Journal of Physiology, 195 (1968), pp. 215–243.
[16] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017.
[17] A. Krizhevsky, *Learning multiple layers of features from tiny images*, 2009.
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in Advances in Neural Information Processing Systems, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds., vol. 25, Curran Associates, Inc., 2012.
[19] K. Kärkkäinen and J. Joo, *Fairface: Face attribute dataset for balanced race, gender, and age*, 2019.
[20] Z. Liu, P. Luo, X. Wang, and X. Tang, *Deep learning face attributes in the wild*, in Proceedings of International Conference on Computer Vision (ICCV), December 2015.
[21] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith, *Diversity in faces*, 2019.
[22] L. Nanni, S. Brahnam, M. Paci, and S. Ghidoni, *Comparison of different convolutional neural network activation functions and methods for building ensembles for small to midsize medical data sets*, Sensors, 22 (2022), p. 6129.
[23] R. Nirthika, S. Manivannan, A. Ramanan, and R. Wang, *Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study*, Neural Computing and Applications, 34 (2022), pp. 5321–5347.
[24] T. D. Ryck, S. Lanthaler, and S. Mishra, *On the approximation of functions by tanh neural networks*, Neural Networks, 143 (2021), pp. 732–750.
[25] J. Schrouff, N. Harris, S. Koyejo, I. M. Alabdulmohsin, E. Schnider, K. Opsahl-Ong, A. Brown, S. Roy, D. Mincu, C. Chen, et al., *Diagnosing failures of fairness transfer across distribution shift in real-world medical settings*, Advances in Neural Information Processing Systems, 35 (2022), pp. 19304–19318.
[26] L. Shao, Z. Cai, L. Liu, and K. Lu, *Performance evaluation of deep feature learning for rgb-d image/video classification*, Information Sciences, 385-386 (2017), pp. 266–283.
[27] M. M. Taye, *Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions*, Computation, 11 (2023), p. 52.
[28] Q. Wei and R. L. Dunbrack, *The role of balanced training and testing data sets for binary classifiers in bioinformatics*, PLoS ONE, 8 (2013), p. e67863.
[29] K. Weiss, T. M. Khoshgoftaar, and D. Wang, *A survey of transfer learning*, Journal of Big Data, 3 (2016).
[30] F. Xing, Y. Xie, X. Shi, P. Chen, Z. Zhang, and L. Yang, *Towards pixel-to-pixel deep nucleus detection in microscopy images*, BMC Bioinformatics, 20 (2019).
[31] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, *Convolutional neural networks: an overview and application in radiology*, Insights into Imaging, 9 (2018), pp. 611–629.
[32] S. Zabala-Travers, M. Choi, W.-C. Cheng, and A. Badano, *Effect of color visualization and display hardware on the visual assessment of pseudocolor medical images*, Medical Physics, 42 (2015), pp. 2942–2954.
[33] A. Zafar, M. Aamir, N. M. Nawi, A. Arshad, S. Riaz, A. Alruban, A. K. Dutta, and S. Almotairi, *A comparison of pooling methods for convolutional neural networks*, Applied Sciences, 12 (2022), p. 8643.
[34] Z. Zhang, Y. Song, and H. Qi, *Age progression/regression by conditional adversarial autoencoder*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017.