On the Performance of Machine Learning Fairness in Image Classification

Utsab Khakurel and Danda B. Rawat

Department of Electrical Engineering and Computer Science
Howard University, Washington DC, USA

ABSTRACT

In recent years, computer vision has made significant strides in enabling machines to perform a wide range of tasks, from image classification and segmentation to image generation and video analysis. It is a rapidly evolving field that aims to enable machines to interpret and understand visual information from the environment. One key task in computer vision is image classification, where algorithms identify and categorize objects in images based on their visual features. Image classification has a wide range of applications, from image search and recommendation systems to autonomous driving and medical diagnosis. However, recent research has highlighted the presence of bias in image classification algorithms, particularly with respect to human-sensitive attributes such as gender, race, and ethnicity. Some examples are computer programmers being predicted better in the context of men in images compared to women, and the accuracy of the algorithm being better on greyscale images compared to colored images. This discrepancy in identifying objects is developed through correlation the algorithm learns from the objects in context known as contextual bias. This bias can result in inaccurate decisions, with potential consequences in areas such as hiring, healthcare, and security. In this paper, we conduct an empirical study to investigate bias in the image classification domain based on sensitive attribute gender using deep convolutional neural networks (CNN) through transfer learning and minimize bias within the image context using data augmentation to improve overall model performance. In addition, cross-data generalization experiments are conducted to evaluate model robustness across popular open-source image datasets.

Keywords: Convolutional Neural Network, Image Classification, Transfer Learning, Equitable AI, Trustworthy AI, Fairness, Ethics

1. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) are rapidly evolving fields that have gained a lot of attention in recent years. AI refers to the development of machines that can perform tasks that typically require human intelligence, such as reasoning, learning, and problem-solving.[1,2] Machine learning, a subset of AI, involves the development of algorithms that can learn from data and make predictions or decisions based on that learning.[3,4] These technologies have been made possible by advances in computing power, data storage, and the availability of vast amounts of data.[5] Among various fields of AI/ML, Computer Vision (CV) has emerged as a highly impactful and compelling subfield with numerous applications in the modern world. Computer vision is a field of study that focuses on enabling machines to interpret and understand visual information from the environment. It involves the use of algorithms and mathematical models to analyze and interpret digital images and videos, in order to extract meaningful information and enable automated decision-making.[6] The ultimate goal of computer vision is to enable machines to see and understand the world as humans do. Computer vision has numerous practical applications in areas such as healthcare, transportation, surveillance, entertainment, manufacturing, construction, retail, agriculture, and more.

Computer vision has come a long way since its early days when the task required a significant amount of manual work from developers and human operators. With the pioneering work of LeCun in 1998, the use of deep neural networks on image recognition tasks became a turning point in computer vision. LeCun introduced

Utsab Khakurel: E-mail: utsab.khakurel@bison.howard.edu, Telephone: 1 202 660 3139
Danda B. Rawat: E-mail: danda.rawat@howard.edu, Telephone: 1 202 806 2209

the concept of Convolutional Neural Networks (CNN), which mimics the human visual cortex .[7] CNN is a deep learning algorithm that assigns importance to various objects in an image and differentiates between them. The algorithm works in four steps: image pre-processing, feature extraction, feature selection, and prediction and recognition.[8] Compared to other classification algorithms, CNN requires much lower pre-processing while having the ability to self-learn the filter/characteristics from an image. CNN consists of three essential parts: the convolutional layer, the pooling layer, and the fully-connected layer.[9] The convolutional layer extracts feature maps from the input image to identify critical object shapes and forms using a filter. The pooling layer reduces the size of the feature maps through downsampling, thereby reducing computational costs. The fully-connected layer classifies the image using activation functions with the output from the pooling layers.[10]

Computer Vision is a broad field that encompasses various tasks, including Image classification, Object detection, Object localization, and Instance segmentation.[8] Image classification involves categorizing images into different classes based on similarities and differences.[11] There are two types of image classification techniques: supervised and unsupervised. In supervised classification, CNN architectures like the one described earlier are trained using a large labeled image dataset to learn patterns and features for categorizing images into predefined classes. Thanks to advances in deep learning, image classification accuracy has improved significantly over the past decade, with models achieving up to 99% accuracy and outperforming humans in quickly processing visual information.[12] Image classification has many practical applications across fields such as healthcare, security and surveillance, sports, entertainment, and more.

Although CNNs have revolutionized various applications by providing state-of-the-art algorithms with exceptional speed and accuracy, these models are not immune to the issue of fairness. The problem arises due to the potential for these models to perpetuate and amplify bias present in the underlying dataset, leading to incorrect and unfair predictions.[13,14] Bias can be observed in speech, visual, and audio data, with visual data being particularly susceptible to bias based on object, person, and geography dimensions. Underrepresentation and stereotypical training image datasets aid in predicting images inaccurately and inflicting bias through them. While the image classification method can better predict images based on their context and co-occurrence patterns, this can also lead to learning spurious correlations between the objects and making false predictions. Contextual bias occurs when such patterns are falsely associated with certain characteristics of the image, such as geography, race, and gender.[15]

In this article, we present an empirical study that investigates gender bias in image classification using two popular convolutional neural networks, ResNet-50 and InceptionV3, implemented through transfer learning and curtails the implications of bias in the performance of the model using data augmentation. The performances of the models are evaluated through classification accuracy before and after the data augmentation. Furthermore, the study aims to analyze the generalizability of the trained image classification models and assess their adaptability to other open-source datasets through a cross-dataset generalization technique.

The remainder of the paper is organized as follows. Section II presents the overview of bias in image classification and the relevant work on the topic. Section III discusses the research approach including dataset, transfer learning, hyperparameter tuning, data augmentation, and cross-dataset generalization. Section IV presents the findings of the experiment and analyzes the result. Finally, conclusions are presented in Section V.

2. BIAS IN IMAGE CLASSIFICATION

2.1 Overview

Artificial Intelligence is becoming increasingly ubiquitous in our daily lives and is increasingly taking over human decision-making processes. This success has opened up innumerable possibilities in the world of innovation for the future. However, the key to high-performing algorithms is the availability of large datasets for training these models. Deep learning methods are heavily reliant on vast amounts of data to learn from. However, this data is a reflection of us as humans, and thus, the biases we hold are often reflected in these datasets. Consequently, deep neural networks tend to not just learn but also amplify bias in their results. In image classification, bias refers to the propensity of the algorithms to be influenced by various factors, such as race, gender, geography, and other sociocultural factors that may be present in the data. Such biases can lead to inaccurate or unjust predictions, which can have real-world consequences, such as perpetuating discriminatory practices or harming

marginalized groups.[16] However, regulating such systems is challenging since it is difficult to agree on a universal definition of fairness. Therefore, the need to identify underlying biases in visual datasets and develop methods to address them is crucial.[17]

The presence of biases in visual datasets used for computer vision has become increasingly apparent. These biases are easily recognizable due to the visual nature of the datasets. Torralba et al.[18] have identified four types of biases in visual datasets, which include selection bias, capture bias, category/label bias, and negative set bias. Selection bias occurs when the dataset is biased towards a certain pattern or representation of the real world. For instance, the Caltech images dataset is biased toward side views while ImageNet tends to favor racing cars.[18] Capture bias, on the other hand, results from the camera operator's preference in lighting, camera orientation, positioning, angles, and other factors that may influence how the images are captured. Category/label bias occurs when different labels are assigned to similar objects in the images. Negative set bias refers to what the dataset considers as 'the rest of the world'. If the negative set is unbalanced or unrepresentative, the accuracy of the results might be affected. In addition, image classifiers tend to leverage co-occurrences between objects and their context to improve performance. Strongly relying on context can hurt the accuracy and generalizability of the model when the co-occurrence patterns are absent.[15] Although it enables the algorithm to predict objects more accurately, it introduces false positives and false negatives in the prediction in the absence of a related object.

Instances of bias in the visual world are becoming increasingly common and well-documented. For example, in 2015, there were reports of gender bias in Google search results for 'CEO' images, with only white men being shown in top results. In addition, Google showed high-paying executive job ads to male groups 1,852 times compared to only 318 times to female groups.[16] In another case, several facial recognition and gender classification systems from recognizable companies such as IBM and Microsoft were found to be biased as it revealed that darker-skinned females were the most misclassified group, with error rates as high as 34.7%, while lighter skin males had an error rate of only 0.8%.[19] Similarly, a study found gender and age biases present in state-of-the-art pedestrian detection algorithms.[20]

Over the years, researchers have proposed various methods to mitigate bias in image classification models. These include methods for dataset selection, preprocessing, and augmentation, as well as algorithms that explicitly correct for bias during training. Despite these efforts, bias in image classification remains a pressing issue in computer vision research, and there is a need for ongoing research and the development of new strategies to address this problem.

2.2 Related Works

Schaaf et al.[21] proposed a method to explain the biased decisions made by deep neural networks using the feature attribution technique. By visualizing the significance of each pixel in the input image for the final prediction, visualization techniques like attribution maps can help in understanding image classification models like CNNs. The paper assesses several feature attribution map techniques to determine their effectiveness in identifying bias. In a separate study, Li et al.[22] developed a technique to uncover unknown biased attributes of an image classifier, where the generative model's latent space contains an optimizable hyperplane as an unknown biased attribute, followed by human interpretation.

Singh et al.[15] devised a unique approach to address the contextual bias that exists in image classifiers. Their study proposed a method to improve object prediction in images where the context is missing by separating the object from the context. The aim was to enhance the accuracy of predictions when the object is not in its proper context without compromising the performance when the context is available. REVISE[13] is a tool built by Wang et al. that identifies bias in an image based on object, person, and geography domain. In an experiment conducted by Model et al.,[23] the classifier was tested on a small area of a large image that lacked any interpretable information. The results showed that all classification accuracy was higher than random chance, despite the images having no significant visual information. This suggests that there is a consistent bias in images that influences the results. Torralba et al.[18] are credited with producing groundbreaking work that aimed to address the problem of dataset bias. Their research led to a significant shift in the community's focus toward reevaluating their algorithms to ensure they are fair and just in the real world. Despite the rapid evolution and improvement of computer vision performance, the focus has primarily been on developing the best-performing

classifiers, while the underlying motives for creating such models have been overlooked. Through techniques such as 'Name that Dataset!' and cross-dataset generalization, this work exposed the intrinsic bias prevalent in the large databases used to train vision models.

There has been an increasing interest in identifying and mitigating bias in visual machine learning (ML) research, however, there is still a need for seminal work in this area. Common mitigation strategies to minimize bias include undersampling or oversampling, data augmentation, and using large and representative datasets. Our study aims to empirically investigate bias in image classification based on gender, using two popular Convolutional Neural Networks ResNet-50 and InceptionV3, through transfer learning. The research aims to identify intrinsic and contextual bias in images based on gender and to utilize data augmentation techniques to minimize the effect of implicit bias on the results. Additionally, the paper intends to test the generalizability of the trained model through cross-dataset generalization tests. The objectives of the research are clearly outlined below.

- Identify the contextual bias prevalent in the image dataset based on sensitive attribute gender.
- Utilize data augmentation to alleviate the effects of bias in the classification performance.
- Conduct a comparison of the model's performance pre and post-data augmentation.
- Evaluate the generalizability of the models through cross-dataset generalization on open-source image datasets, including random test set, PASCAL VOC, CelebA, and MSCOCO.

3. METHODOLOGY

3.1 Dataset

The Men/Women classification dataset[24] used in our research is a manually curated and cleaned collection of images featuring 3,354 individuals who are categorized into the Men and Women category. While several benchmark datasets are available for computer vision research, very few of them include attributes such as gender, race, or age. Assigning such labels to images raises privacy concerns,[25] which is why manual collection and labeling of such images are necessary for research purposes.

Our research dataset comprises a total of 3309 images, of which 1409 are of men and 1900 are of women. The images are labeled based on gender, with a value of 1 assigned to men and 0 assigned to women. The training set for our experiment has an equal distribution of both gender labels to prevent any bias resulting from imbalanced learning samples. It includes 1,158 images for each gender label. The validation and test sets have unbalanced target labels, representing real-world data. The validation set has 496 samples, including 364 men and 132 women labels. The test set includes 378 men and 119 women labels with a total of 497 images. All the images were resized to a uniform size of 224 x 224.

3.2 Transfer Learning

Transfer learning is a technique that seeks to enhance the performance of a model on a target domain by leveraging the knowledge learned from a related but different source domain. The idea is to transfer the learned representations from the source domain to the target domain, allowing the model to adapt more efficiently and accurately to the new task.[26] However, there needs to be a connection between the learning tasks for transfer learning to be beneficial. For instance, a model trained on animal images would not perform well when applied to a dataset labeled for humans.

In the computer vision field, pre-trained models are often utilized for transfer learning, as they have already been trained on large benchmark datasets to solve problems similar to the current task. Since training such models from scratch can be computationally expensive, using published models is a common practice. Transfer learning enables building accurate models with minimal labeled training data, saving time and effort.

In our study, we adopted transfer learning by fine-tuning popular open-source Convolutional Neural Networks pre-trained on ImageNet, a large-scale hierarchical image database.[27] ImageNet includes thousands of training, validation, and test images organized into WordNet hierarchy, where each node contains relevant images based on the assigned class. We repurposed the pre-trained models by removing their original classifiers and adding a new

classifier to suit the experiment's goals. The added fully-connected layer contains a sigmoid function for binary classification, and multiple dropout layers were incorporated to regularize the model and reduce overfitting. The convolution layer weights of the pre-trained models were not modified. During training, the trainable convolution layers were fed with the training samples to enable the model to learn about gender differences in the images.

3.2.1 ResNet-50

ResNet, or Residual Network, is CNN introduced in 2016 by Kaiming et al.[28] to tackle the vanishing gradient problem in deep neural networks. Skip connections were added to allow the gradient to take an additional path, skipping some layers and feeding the output of one layer to the other layer, except only the next one. These connections are called residual blocks, and they are stacked to form the ResNet architecture.

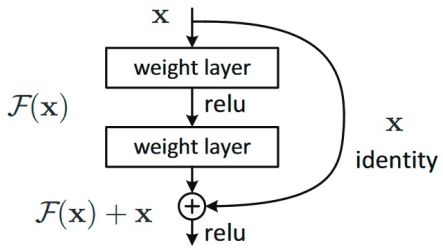

Figure 1. Residual block.

Figure 1 demonstrates the functioning of ResNet where the input from the previous layer and the layer before it is added and then fed to the current layer. ResNet architecture follows two primary design principles: first, the number of filters in each layer is constant and depends on the output feature map's size; second, the number of filters are doubled when the feature map size is halved to maintain the layer's time complexity. For our study, we utilized the ResNet-50 architecture which is pre-trained on the ImageNet database.

3.2.2 InceptionV3

InceptionV3 is a CNN architecture that addresses overfitting in deep neural networks by using multiple filters of different sizes on the same level. This widens the layers and allows the network to learn various features at different scales, making it more robust to input variations.[29] InceptionV3 is also computationally efficient due to techniques such as factorized convolutions, regularization, dimension reduction, and parallelized computations. InceptionV3 has been pre-trained on the ImageNet database, containing over a million labeled images from 1000 classes, enabling the network to learn general features useful for various computer vision tasks, including gender classification for our research.

3.3 Hyperparameter Tuning

Hyperparameter tuning is an important step in training a Convolutional Neural Network. It involves finding the optimal values for hyperparameters that can impact the model's performance, such as the learning rate, number of epochs, optimizer, dropout rate, and more.[30]

Learning rate controls weight updates at each iteration, with a higher rate leading to faster convergence while a lower rate yielding a more precise solution but slower convergence. Therefore, finding the optimal learning rate is crucial to achieving the best model performance. The number of epochs determines how often the model sees training data, and over/underfitting models may need to adjust this. The optimizer controls the update rule for the model's parameters. The dropout rate is a regularization technique that randomly drops out some neurons during training to prevent overfitting, but high rates may reduce accuracy.

We conducted a series of experiments in our study, adjusting different hyperparameters such as learning rate, epochs, number of dropout layers, and dropout rates. A higher number of epochs led to the overfitting of the model. We found that the ideal number of epochs was either 10 or 15, with 15 providing better results in terms of training and validation accuracy and loss. When it came to dropout layers, we observed that a model without

any dropout layer or with only one dropout layer resulted in overfitting, while three dropout layers prevented this issue by causing the validation loss to stabilize and match the training loss. We tried dropout rates of 0.2 and 0.5 and found that the latter led to an increase in validation loss over epochs. Therefore, we chose a dropout rate of 0.2 as optimal for our training. Finally, we discovered that a learning rate of 0.0001 was ideal for our CNN model. The optimization algorithm used in the model is Adam, with binary cross-entropy employed as the loss function. Mean classification accuracy is adopted as the performance evaluation metric for the model.

Hyperparameter tuning is a crucial step in optimizing the performance of a neural network, and it involves iteratively adjusting the values of hyperparameters, to find the combination that results in the best performance on a validation set.

3.4 Data Augmentation

Data augmentation is a technique used in Convolutional Neural Networks for training to artificially increase the size of the dataset by creating new examples through various transformations of existing data. The goal of data augmentation is to improve the performance and generalization of the model by exposing it to more diverse variations of the training data.[31]

Selecting appropriate augmentation techniques is crucial for improving the performance of the model. In our study, we employed several techniques such as horizontal and vertical shift, horizontal flip, rotation up to 30°, and zoom in/out by 50%. During training, image augmentation is performed in real-time, resulting in the generation of 2,304 augmented images per epoch with balanced target labels. Augmentation was only applied to the training set to enhance the model's ability to distinguish gender. The validation and test sets were not augmented to evaluate the performance of the model trained with augmented images on real-world data.

3.5 Cross-dataset Generalization

Cross-dataset generalization refers to the ability of a machine learning model to perform well on datasets that are different from the dataset it was trained on. In other words, it is the ability of the model to generalize to unseen data from different sources or domains. This is an important aspect of machine learning as real-world datasets often contain variations in the data, such as lighting conditions, camera angles, and other factors, that are not present in the training dataset. A model that has good cross-dataset generalization can handle these variations and still perform well on the new data. Cross-dataset generalization is often used along with transfer learning. Torralba et. al[18] highlighted that the model trained on the benchmark visual dataset can't generalize well when tested across other datasets. To conduct the cross-data generalization test, we randomly selected an equal number of images as the original test set and labeled them as men and women for PASCAL VOC, CelebA, and MSCOCO dataset. Below are the benchmark datasets used in our work to evaluate the generalization of ResNet-50 and InceptionV3 before and after augmentation.

3.5.1 Random Edge

The model's performance and generalizability are tested using random images collected from the Internet. To ensure variety, images featuring each gender out of commonly occurring context are chosen, such as men using beauty products, men with long hair, women in mugshots, and women wearing backward hats.

3.5.2 PASCAL VOC

PASCAL VOC is a widely used dataset for image classification, detection, segmentation, and localization.[32] The PASCAL VOC challenge was organized every year from 2005-2012 enabling the computer vision community an opportunity to understand the realm better. The dataset and challenge are hosted by PASCAL2 community.

3.5.3 CelebA

The CelebA dataset comprises facial images of famous personalities, often used for facial attribute detection, and was first collected by researchers at MMLAB, The Chinese University of Hong Kong.[33]

3.5.4 MSCOCO

Common Objets in Context or COCO is an extensive dataset created by Microsoft Research for object recognition and classification task.[34]

4. EXPERIMENTS AND FINDINGS

4.0.1 ResNet-50

Fig. 2 presents examples of misclassified instances from the experiment with ResNet-50, highlighting how the classification can be context-dependent. The initial image illustrates a woman wearing a backward baseball hat wrongly identified as a man. In the following image, a woman with her hair tied in a mugshot is labeled as a man. Lastly, the third image presents a woman engaged in a sports activity incorrectly identified as a man. These misclassifications exhibit a repeating pattern in the ResNet-50 model, including women wearing formal attire or glasses being wrongly identified as men, and most topless men being categorized as women. Additionally, women participating in sports activities are frequently classified as men.

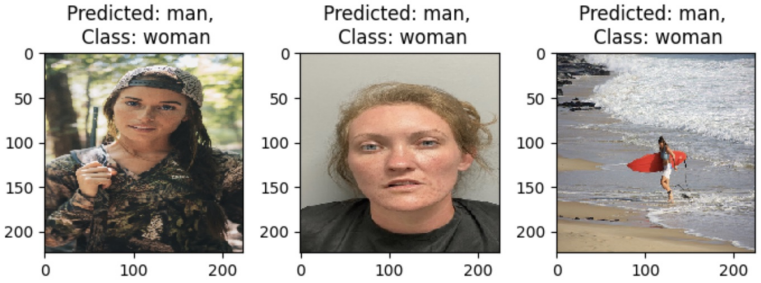

Figure 2. Instances of misclassifications in ResNet-50 model.

The training and validation accuracy pre and post-image augmentation is shown in Fig. 3. To simplify, we will refer to pre-augmentation and post-augmentation as pre-aug and post-aug, respectively, going forward. Pre-aug training and validation accuracy appear consistent and gradually improve. Although, validation accuracy dips initially, it later recovers in subsequent epochs. On the other hand, post-aug validation accuracy improves significantly compared to pre-aug, indicating better performance due to data augmentation and the inclusion of dynamic data.

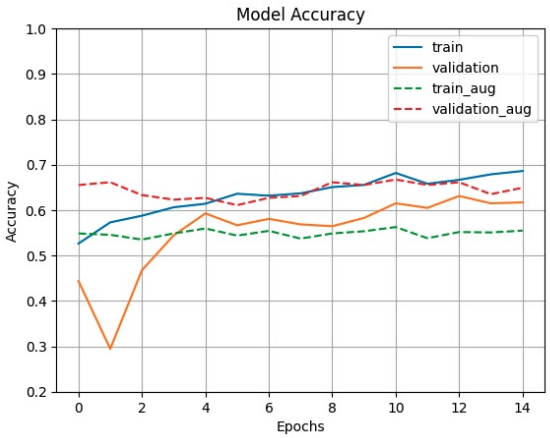

Figure 3. Training and Validation accuracy for ResNet-50 model pre and post-augmentation.

However, it is notable that post-aug training accuracy is lower than pre-aug training accuracy, which seems counterintuitive. This can be attributed to the fact that deep learning models tend to memorize patterns if they lack enough data. With the addition of more data through dynamic angles, flips, and rotations, the model learns more about the patterns and reduces overfitting. Remember, data augmentation is a regularization technique after all. Therefore, training accuracy may decrease as the model becomes more generalizable.

In addition, the post-aug training accuracy is lower than post-aug validation accuracy because the training set includes augmented images. This means that the model not only has to classify images into men and women but also figure out their orientation from the transformed images, making the task more challenging. Meanwhile, the validation set is not augmented, making classification easier, especially when trained on an augmented training set.

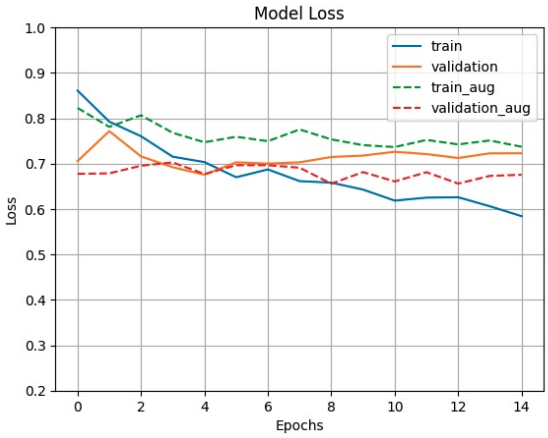

Figure 4. Training and Validation loss for ResNet-50 model pre and post-augmentation.

Figure 4 illustrates the training and validation loss before and after image augmentation. With the increase in epochs, the pre-aug training loss is decreasing and the validation loss is following suit. The post-aug validation loss is improved compared to the pre-aug validation loss. Similar to the accuracy plots, post-aug training loss is performing less compared to pre-aug training and post-aug validation loss for the reasons explained previously. Overall, the validation loss of the model has improved, providing us with a glimpse of how data augmentation can enhance the performance of deep neural networks. Fig. 5 shows instances where the ResNet-50 model corrected its classification after image augmentation.

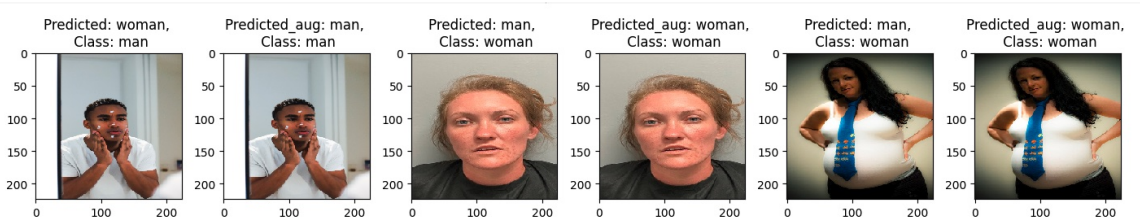

Figure 5. Classification results of ResNet-50 model pre and post-augmentation.

Table 1. Cross-dataset generalization. Classification Accuracy for binary class gender on ResNet-50 before and after augmentation."Self" refers to training and testing on the same dataset, and "Average" refers to averaging performance on all except self.

Classifier/Test on:	Men/Women	Random Edge	PASCAL VOC	CelebA	MSCOCO	Self	Average	% drop
ResNet-50	**72.0**	43.5	54.9	48.5	47.2	72.0	48.5	**23.5**
ResNet-50_aug	**68.4**	41.5	50.6	58.9	50.7	68.4	50.4	**18.0**
Average	70.2	42.5	52.8	53.7	49.0	70.2	49.4	20.8

Table 1 presents the outcomes of the cross-dataset generalization test, where the average classification accuracy is calculated over 10 experimental runs for each test set. It is observed that the classification accuracy declines on the Men/Women dataset after augmentation. Although the decrease in accuracy seems counterintuitive, it can be true as augmentation helps eliminate spurious relationships from the image, causing the test

accuracy to decrease. Similarly, a slight drop in test accuracy is observed on the Random Edge dataset and PASCAL VOC while the model performs better on the CelebA and MSCOCO datasets after augmentation. It is important to note that the absolute performance numbers may not hold much significance; it is the differences in performance that are more meaningful. Therefore, it is evident that the average mean classification over cross-datasets has increased after augmentation from 48.5% to 50.4%. Furthermore, the percentage drop in the difference between the mean classification on the original test set and the average mean classification on cross-datasets seems to decrease from 23.5% to 18.0% after augmentation.

4.0.2 InceptionV3

Fig. 6 presents the misclassification results of the InceptionV3 model experiments. The first image exhibits a man with long hair being identified as a woman. In the second image, a woman working on a computer is wrongly classified as a man, while in the third image, a woman jumping is labeled as a man. These misclassifications are attributed to the learned co-occurrence of the object and context by the model.

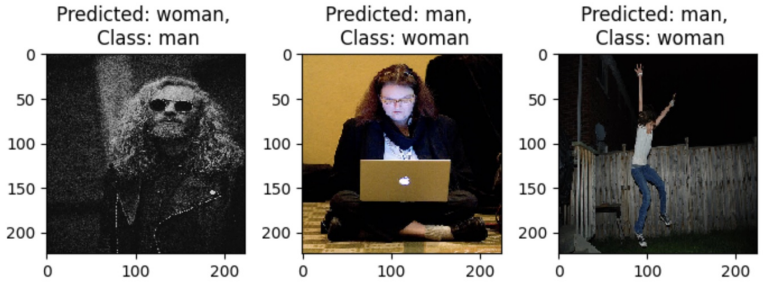

Figure 6. Instances of misclassifications in InceptionV3 model.

The training and validation accuracy of InceptionV3 before augmentation appear to be quite satisfactory across epochs, with the validation accuracy reaching a saturation point quite early and the training accuracy continuing to improve, as shown in Fig. 7. The post-aug validation accuracy also performs well and remains almost at the same level as the pre-aug validation accuracy. However, the post-aug training accuracy in the InceptionV3 model also drops down, compared to both pre-aug training accuracy and post-aug validation accuracy.

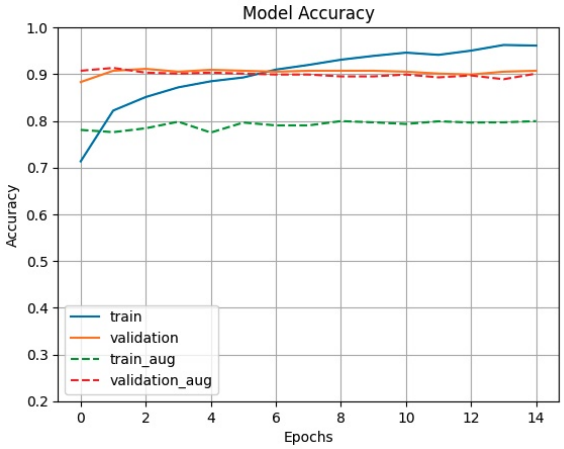

Figure 7. Training and Validation accuracy for InceptionV3 model pre and post-augmentation.

In the same way, Fig.8 demonstrates that the pre-aug validation loss reaches saturation early on, while the training loss converges smoothly. The validation loss performance after augmentation appears to be comparable

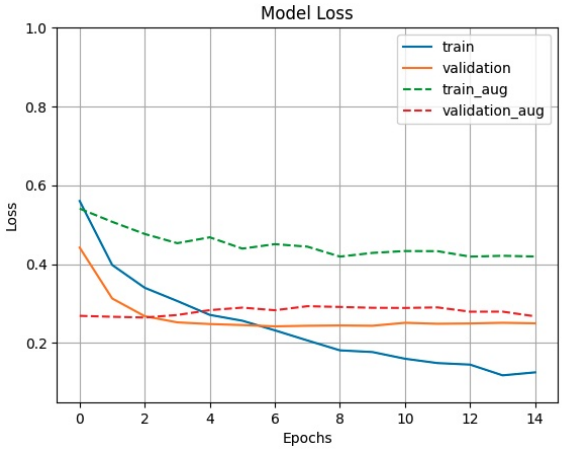

Figure 8. Training and Validation loss for InceptionV3 model pre and post-augmentation.

to the pre-aug validation loss. In addition, the performance of post-aug training loss is similar to the results obtained for ResNet-50, compared to the pre-aug training and post-aug validation loss. Fig.9 showcases some instances where the model's performance has improved after image augmentation.

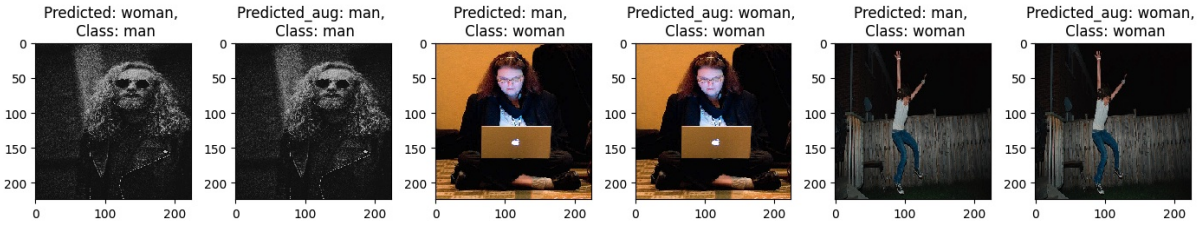

Figure 9. Classification results of InceptionV3 model pre and post-augmentation.

Table 2. Cross-dataset generalization. Classification Accuracy for binary class gender on InceptionV3 before and after augmentation."Self" refers to training and testing on the same dataset, and "Average" refers to averaging performance on all except self.

Classifier/Test on:	Men/Women	Random Edge	PASCAL VOC	CelebA	MSCOCO	Self	Average	% drop
InceptionV3	**89.3**	71.2	61.6	90.4	67.5	89.3	72.7	**16.6**
InceptionV3_aug	**87.8**	65.8	63.2	91.5	68.2	87.8	72.2	**15.6**
Average	88.6	68.5	62.4	91.0	67.9	88.6	72.5	16.1

Table 2 displays the performance of InceptionV3 on cross-datasets. The model exhibits good performance on all the test sets with a slight decline in mean classification accuracy observed on the Men/Women and Random Edge test sets, whereas improvements are seen on the PASCAL VOC, CelebA, and MSCOCO test sets. Overall, the cross-dataset generalization test suggests that the model performs on par after image augmentation, with an average mean classification over cross-datasets of 72.2%, compared to 72.7% before augmentation. It is worth mentioning that the percentage drop in the difference between the mean classification on the original test set and the average mean classification on cross-datasets decreases from 16.6% to 15.6% after augmentation.

5. CONCLUSION

In this paper, we investigated the gender classification performance of the ResNet-50 and InceptionV3 models trained in ImageNet through transfer learning on a diverse set of images, including those from the original test set and multiple external datasets - Random test set, PASCAL VOC , CelebA, and MSCOCO. Our results revealed

significant biases and misclassifications in the ResNet-50 model, particularly when it comes to images featuring women in non-stereotypical contexts. On the other hand, the InceptionV3 model demonstrated a much more balanced and accurate performance.

We also explored the impact of data augmentation on the classification performance of the models and found that it can significantly improve the generalizability and accuracy of the models, particularly in the case of ResNet-50. Overall, our study highlights the importance of diversity and inclusion in training datasets and the need for ongoing evaluation and improvement of AI models to ensure their fairness and reliability. In conclusion, our findings have significant implications for the development and deployment of AI models in various domains, including healthcare, finance, and law enforcement. As AI continues to become an increasingly integral part of our lives, it is crucial to prioritize fairness, accountability, and transparency in AI development and deployment.

Acknowledgments

This research was funded by the DoD Center of Excellence in AI and Machine Learning (CoE-AIML) at Howard University under Contract Number W911NF-20-2-0277 with the U.S. Army Research Laboratory and in part by the US NSF grant CNS/SaTC 2039583. However, any opinions, findings, conclusions, or recommendations expressed in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agencies.

REFERENCES

[1] Saghiri, A. M., Vahidipour, S. M., Jabbarpour, M. R., Sookhak, M., and Forestiero, A., "A survey of artificial intelligence challenges: Analyzing the definitions, relationships, and evolutions," *Applied Sciences* **12**(8) (2022).

[2] Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., and Lee, I., "Artificial intelligence in the 21st century," *IEEE Access* **6**, 34403–34421 (2018).

[3] El Mrabet, M. A., El Makkaoui, K., and Faize, A., "Supervised machine learning: A survey," in [*2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet)*], 1–10 (2021).

[4] Shivahare, B. D., Suman, S., Challapalli, S. S. N., Kaushik, P., Gupta, A. D., and Bibhu, V., "Survey paper: Comparative study of machine learning techniques and its recent applications," in [*2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*], **2**, 449–454 (2022).

[5] Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C.-W., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., Roepman, R., Dietmann, S., Virta, M., Kengara, F., Zhang, Z., Zhang, L., Zhao, T., Dai, J., Yang, J., Lan, L., Luo, M., Liu, Z., An, T., Zhang, B., He, X., Cong, S., Liu, X., Zhang, W., Lewis, J. P., Tiedje, J. M., Wang, Q., An, Z., Wang, F., Zhang, L., Huang, T., Lu, C., Cai, Z., Wang, F., and Zhang, J., "Artificial intelligence: A powerful paradigm for scientific research," *The Innovation* **2**(4), 100179 (2021).

[6] Bi, Y., Xue, B., Mesejo, P., Cagnoni, S., and Zhang, M., "A survey on evolutionary computation for computer vision and image analysis: Past, present, and future trends," (09 2022).

[7] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P., "Gradient-based learning applied to document recognition," *Proceedings of the IEEE* **86**(11), 2278–2324 (1998).

[8] Wu, Q., Liu, Y., Li, Q., Jin, S., and Li, F., "The application of deep learning in computer vision," in [*2017 Chinese Automation Congress (CAC)*], 6522–6527 (2017).

[9] O'Shea, K. and Nash, R., "An introduction to convolutional neural networks," *ArXiv e-prints* (11 2015).

[10] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A. Q., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L., "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data* **8** (2021).

[11] Sanghvi, K., Aralkar, A., Sanghvi, S., and Saha, I., "A survey on image classification techniques," *SSRN Electronic Journal* (01 2021).

[12] He, Z., "Deep learning in image classification: A survey report," in [*2020 2nd International Conference on Information Technology and Computer Application (ITCA)*], 174–177 (2020).

[13] Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., and Russakovsky, O., "Revise: A tool for measuring and mitigating bias in visual datasets," *Int. J. Comput. Vision* **130**, 1790–1810 (jul 2022).

[14] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A., "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," (07 2016).

[15] Singh, K. K., Mahajan, D., Grauman, K., Lee, Y. J., Feiszli, M., and Ghadiyaram, D., "Don't judge an object by its context: Learning to overcome contextual bias," in [*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 11067–11075 (2020).

[16] Yapo, A. and Weiss, J., "Ethical implications of bias in machine learning," (01 2018).

[17] Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., and Kompatsiaris, I., "A survey on bias in visual datasets," *Computer Vision and Image Understanding* **223**, 103552 (2022).

[18] Torralba, A. and Efros, A. A., "Unbiased look at dataset bias," in [*CVPR 2011*], 1521–1528 (2011).

[19] Buolamwini, J. and Gebru, T., "Gender shades: Intersectional accuracy disparities in commercial gender classification," in [*Proceedings of the 1st Conference on Fairness, Accountability and Transparency*], Friedler, S. A. and Wilson, C., eds., *Proceedings of Machine Learning Research* **81**, 77–91, PMLR (23–24 Feb 2018).

[20] Brandão, M., "Age and gender bias in pedestrian detection algorithms," (06 2019).

[21] Schaaf, N., Mitri, O., Kim, H., Windberger, A., and Huber, M., [*Towards Measuring Bias in Image Classification*], 433–445 (09 2021).

[22] Li, Z. and Xu, C., "Discover the unknown biased attribute of an image classifier," in [*2021 IEEE/CVF International Conference on Computer Vision (ICCV)*], 14950–14959 (2021).

[23] Model, I. and Shamir, L., "Comparison of data set bias in object recognition benchmarks," *IEEE Access* **3**, 1953–1962 (2015).

[24] Orsolini, J., "Men/women classification: A jpg dataset for male/female classification," (2019).

[25] Deviyani, A., "Assessing dataset bias in computer vision," (05 2022).

[26] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q., "A comprehensive survey on transfer learning," *Proceedings of the Institute of Radio Engineers* **109**, 43–76 (Jan. 2021).

[27] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE Conference on Computer Vision and Pattern Recognition*], 248–255 (2009).

[28] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 770–778 (2016).

[29] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., "Rethinking the inception architecture for computer vision," in [*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 2818–2826, IEEE Computer Society, Los Alamitos, CA, USA (jun 2016).

[30] Yeh, W. C., Lin, Y.-P., Liang, Y.-C., and Lai, C.-M., "Convolution neural network hyperparameter optimization using simplified swarm optimization," *ArXiv* **abs/2103.03995** (2021).

[31] Shorten, C. and Khoshgoftaar, T., "A survey on image data augmentation for deep learning," *Journal of Big Data* **6** (07 2019).

[32] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A., "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision* **88**, 303–338 (June 2010).

[33] Yang, S., Luo, P., Loy, C. C., and Tang, X., "From facial parts responses to face detection: A deep learning approach," 3676–3684 (12 2015).

[34] Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L., "Microsoft coco: Common objects in context," in [*European Conference on Computer Vision*], (2014).