Using generative modelling to perform diversifying data augmentation

Investigating the usage of a CycleGAN for pre-processing to decrease bias in gender classification

submitted in partial fulfillment for the degree of master of science

Dimitri Hooftman
13246038

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on 30.06.2023

	UvA Supervisor	External Supervisor
Title, Name	Sahand Mohammadi Ziabari	Joop Snijder
Affiliation	University of Amsterdam	Info Support
Email	s.s.mohammadiziabari@uva.nl	joop.snijder@infosupport.com

ABSTRACT

Deep learning algorithms have become more prevalent in real-world applications. With these developments, bias is observed in the predictions made by these algorithms. One of the reasons for this is the algorithm's capture of bias in the data set being used. This research investigates the influence of using generative adversarial networks (GANs) as a gender-to-gender data pre-processing step on the bias and accuracy measured for a VGG-16 gender classification model. A cyclic generative adversarial network (CycleGAN) is trained on the Adience data set to perform the gender-to-gender data augmentation. This architecture allows for an unpaired domain mapping and results in two generators that double the training images generating a male for every female and vice versa. The VGG-16 gender classification model uses training data to produce an accuracy that indicates its performance. In addition, the model's fairness is calculated using demographic parity and equalized odds to indicate its bias. The evaluation of the results provided by the proposed methodology in this research shows that the accuracy decreases when CycleGAN pre-processing is applied. In addition, the bias also decreases, especially when measured on an imbalanced data set. However, The decrease in bias needs to be more significant to change our evaluation of the model from unfair to fair, showing the proposed methodology to be effective but insufficient to remove bias from the data set.

KEYWORDS

Generative modelling, Generative adversarial networks, CycleGAN, Data augmentation, Bias, Gender classification

GITHUB REPOSITORY

https://github.com/HooftmanD/GAN-data-augmentation

1 INTRODUCTION

The predictive power of deep learning algorithms has led them to be widely used in real-world applications [6]. A prominent example of a widely used machine-learning technique is the supervised learning method, which can, for instance, provide a gender classifier of images by learning a mapping between an input image and a corresponding predicted class [12]. An unintended lack of diversity in the input is a problem observed in using images as data to train these algorithms [16]. For instance, this unintended lack of diversity could be introduced by a selection bias in the compilation process of a data set [24].

Using these data sets could introduce "over-fitting" in the downstream training process and lead to sub-optimal behavior in the real world where complexity and diversity are common. Examples of this sub-optimal behavior are observed for different types of training data, like images and word embeddings, induced by the learned bias existing in the training data [2, 26]. For instance, a model trained on a biased training data set shows that the activity of cooking is over 33 percent more likely to involve females than males [28].

What has been studied is using methods like over-sampling and prevailing data augmentation methods like rotations, flips, and re-scales to increase diversity in data sets used for downstream tasks [26]. In addition, GANs have been used as a data augmentation method to improve generalizability in Computerized Tomography (CT) segmentation tasks [20]. However, the impact of using cyclic GANs to increase data set diversity to reduce bias on downstream tasks needs to be explored more. This thesis project aims to quantify this approach's impact on reducing bias in the downstream task of gender classification.

In addition to reducing bias, maintaining or increasing the performance of a model trained in a downstream task is also of interest. Numerous studies have shown that including pre-processing steps using adversarial networks has increased performance when using deep neural networks downstream [19]. Using a GAN is described as one of the most promising modeling techniques for using data augmentation [27].

This thesis project hypothesizes that augmenting the training data into different domains will reduce bias and retain potential model performance on a downstream task. The reduced bias is hypothesized to be caused by the balance in training examples of different domains, for example, males and females. This hypothesis will be tested using a balanced and an unbalanced training data set to perform measurements. The retained performance is hypothesized to be caused by an increased amount of training data. A downside of the proposed data augmentation could be decreased performance on downstream tasks. To make this transparent, the performance of downstream tasks is also evaluated.

By introducing novel research into the impact of a cycling GAN on bias in a downstream task and measuring performance impact as a secondary interest, this thesis project can quantify the effect on predictive performance. As data augmentation is a pre-processing technique used to increase downstream performance, the diversification of the data set should ideally not decrease downstream performance. The difference in performance will be tested by using baseline measurements.

Obtaining this quantified knowledge benefits both research areas mentioned before. For bias reduction, it opens up a new avenue of research into complex methods like configurable data augmentation using a GAN to generate a diverse data set based on a less-diverse data set. For the pre-processing field, this research potentially verifies that using a GAN as a data augmentation method increases the performance on downstream tasks [22]. To make the mentioned knowledge objectives more concrete, the following research question is proposed:

How does pre-processing using GANs as a gender-to-gender data augmentation step influence the accuracy and bias of a VGG-16 Convolutional Neural Network performing a gender classification task?

To answer the research question, it is essential to gather measurements of the impact of the data augmentation method on the performance and bias of the downstream task of gender classification. These measurements are gathered using a well-known VGG-16 model architecture adjusted to perform binary classification [23].

On top of this, it is important to create a data augmentation method that can perform the gender-to-gender generation. Both these aspects have been captured in the following sub-research questions:

1. How can a GAN architecture perform gender-to-gender generation to diversify a data set of male and female images?

2. What is the baseline performance and bias of a VGG-16 gender classification model using established data augmentation methods on images of males and females?

3. What is the performance and bias of a VGG-16 gender classification model using a gender-to-gender data augmentation method on images of males and females?

4. What is the difference in performance and bias of a VGG-16 gender classification model using a gender-to-gender data augmentation method on an unbalanced data set of males and females?

2 RELATED WORK

The research gap described in the previous section indicates a goal of mitigating bias without reducing downstream performance. To quantitatively assess the bias reduction, the research questions mention both VGG-16 gender classification models and the image-to-image translating CycleGAN model. These areas of research are explored in this section.

2.1 Gender classification

Deep neural networks have demonstrated excellent performance in recognizing the gender of human faces [14]. Eidinger et al. use a standard linear SVM trained on the Local Binary Pattern (LBP) and Four-Patch LBP features (FPLBP) extracted from the Adience data set. They show a 77 percent accuracy when training on the near-frontal faces [5]. They, however, add that their tests leave room for future work as a drop in performance is observed when using the Adience data set as a benchmark. As the Adience data set is made available, it will be considered for this research.

Hassner et al. report a 79.3 percent accuracy when using the same features but adjusting the Adience data set by a Frontalization process. This process detects facial features and rotates them to create a frontal face [9]. This accuracy is improved upon by Levi and Hassner using a deep-convolutional neural network for gender classification [15]. They used a network architecture comprising three convolutional layers and two fully-connected layers with a small number of neurons. They used all rotations of the original images in the Adience data set to show the performance of the network architecture instead of the improved performance by pre-processing. This approach has been shown to have an 86.8 percent accuracy.

Dehgan et al. improve on this by using a larger amount of data to train on though it is unclear how this data set was aggregated [3]. In addition, it is unclear how the images' labels were provided. They state that a team of human annotators was used through a semi-supervised procedure. On top of this, the data does not seem to be provided, which is why this data set is not considered further for this research. For pre-processing the images, they use techniques that perform horizontal flips and random crop augmentation. They apply a specific but undisclosed deep network architecture for gender classification on the Adience benchmark and obtain an accuracy of 91 percent.

Lapuschkin et al. considered the influence of model initialization with weights per-trained on a real-world data set. Their results were reported using a VGG-16 model, pre-trained on the well-known ImageNet and IMDB-WIKI data sets. In addition, it was fine-tuned on the Adience data set using both face alignment techniques simultaneously [14]. They state that three major factors contribute to performance improvements on the gender classification task. (1) Changes in architecture. (2) Prior knowledge via pre-training. (3) Optional data set preparation via alignment pre-processing. The VGG-16 model consists of 13 convolutional layers of small kernel size followed by two fully connected layers. Using the VGG-16 model, an accuracy of 92.6 percent is attained using a weight initialization based on ImageNet.

Improved results on the gender classification task have also been published. These, however, use the improved ResNet model architecture to increase accuracy [13]. As the difference in performance and bias is to be measured based on pre-processing techniques, the model architecture is kept constant in both measurements, and this is why performance and bias using models that differ from VGG-16 on the Adience data set are not considered further for this research.

2.2 CycleGAN

Image-to-image translation using a cycle-consistent adversarial network was introduced by Zhu et al. [29]. This method makes it possible to perform unpaired image translation without paired training data, as obtaining this data can be difficult and expensive [29]. They apply this method to various applications, including collection style transfer, object transfiguration, season transfer, and photo enhancement [29].

The cycle-consistent adversarial network is an architecture built from two GANs. These networks are optimized using a loss function that first includes an adversarial loss allowing it to learn the domain mapping. Second, it includes a cycle consistency and identity loss that ensure that the source and generated results are related. This additional loss optimizes the generators to produce a translation instead of a random output in the target domain.

Almahairi et al. build on this idea by introducing an augmented CycleGAN, which can perform many-to-many mapping [1]. They capture variations in the generated domain by learning stochastic mapping by inferring information about the source which is not captured in the generated result. Qualitative results show the effectiveness of the many-to-many mapping approach in generating multiple females for a given male and vice versa. This research shows the viability of successfully generating male-to-female and female-to-male images [1].

Using a CycleGAN architecture as a data augmentation method for pre-processing images has already been shown adequately for the task of CT segmentation [20]. They use the GANs to render a non-contrast version of training images based on the original contrast CT image. They observed that segmentation performance significantly improved when additional synthetic images were used for training. Hammami et al. use a CycleGAN as an unsupervised

method that generates images of different modalities in a similar domain. These images are used to train a downstream model that can perform multi-organ detection, which has been shown to improve the intended task significantly [8].

2.3 Bias reduction

Mehrabi et al. describe that, like people, algorithms are vulnerable to biases that render their decisions "unfair" [16]. *Fairness* is defined as the "absence of prejudice or favoritism towards an individual or group based on their inherent or acquired characteristics [16]. An example is given about facial recognition software in digital cameras, which over-predict Asians as blinking. These biased predictions are said to stem from the hidden or neglected biases in data or algorithms [16].

Research has been carried out into the reduction of bias in data sets. Wu et al. describe that recent studies found substantial disparities in the accuracy rate of classifying gender of dark-skin females [26]. In their research, Wu et al. describe the usage of preprocessing to balance the skin-type composition of a data set. Using the *ImageDataGenerator*, familiar data augmentation techniques like horizontal flips and re-scaling can be performed. By using the *ImageDataGenerator*, they increased the percentage of dark-skin males from 1.3 percent to 15.21 percent and dark skin females from 2.5 percent to 16.03 percent.

3 METHODOLOGY

The methodology provided in this section is followed to answer the research questions stated in section 1. Figure 1 shows a broad overview of the applied methodology. Both the CycleGAN pipeline and the gender classification pipeline use the Adience benchmark data set as their input. It is used to both train and evaluate the CycleGAN and VGG-16 CNN architecture. Both of these pipelines are designed as well-known machine learning systems.

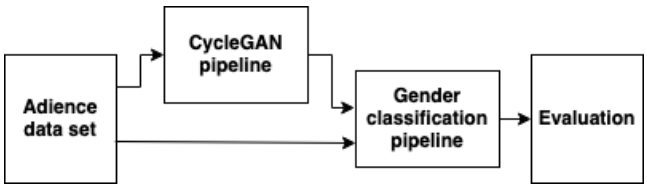


Figure 1: Overview of methodology.

To further detail these parts of the methodology, this section will first describe the data set being used. After this, the CycleGAN pipeline will be explained, answering sub-research question one. Next, the gender classification pipeline will be shown to incorporate this data set and the CycleGAN pipeline results. Finally, the evaluation will be elaborated so it is clear what is measured to answer sub-research questions two, three, and four.

3.1 Adience data set

The data set used in this classification pipeline is the Adience data set published in 2014. This data set contains photos of faces with binary gender labels and has been used in similar classification pipelines [5, 14]. The data set contains faces of different angles, with different light settings, and of different sharpness. Examples of this can be seen in figure 2a and figure 2b.

A fundamental design principle of this data set is that it is as accurate as possible to challenging real-world conditions. As such, it presents all the variations in appearance, noise, pose, lighting, and more that can be expected of images taken without careful preparation or posing [5]. The images are collected using a face detector described by Viola and Jones [25] based on images collected from Flickr albums. All images were manually labeled for gender using both the image themselves and any available contextual information [5].

(a) Male **(b)** Female

Figure 2: Sample images from Adience data set provided by [5].

From the Adience data set, 19.370 images have been used. These images come from 2284 unique individuals. The coarse and landmark images provided by the data set have been used; this brings the total up to 38.740. Of these, 16.240 images have a male gender label. The other 22.500 have a female gender label. For further usage in downstream tasks, the data set is split into a training/validation and test set based on the task for which the model is trained.

The coarse and landmark images were used as Lapuschkin et al. state that all models benefit the most from combining the coarse-aligned and the landmark-aligned data sets for training [14]. To demonstrate the rotation of the coarse-aligned images, figure 2a shows a slight adjustment with black corners for the information unavailable in the original image. Finally, the data in the Adience data set is divided into five folds. These folds have been created to evenly distribute the individuals into subsets to prevent over-fitting on a fold, as multiple images of the same unique individual are in the Adience data set.

3.2 CycleGAN pipeline

The GAN gender-to-gender data augmentation method uses the CycleGAN architecture described in [29] trained on the Adience data set. This architecture uses a GAN to learn two mappings. The first mapping is a generator function G that takes an image in domain X and generates an image that is indistinguishable from domain Y. This is done by optimizing the generator and the discriminator D_y, which learns to label images in domain Y as real or fake based on the generated images from generator G and images from domain Y.

The image resulting from equation G: X -> Y is then used as input for the second mapping, which is a generator function F that takes the image in domain Y and generates an image that is

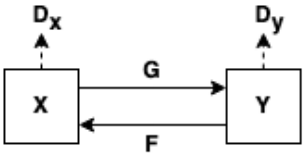


Figure 3: CycleGAN architecture as shown by [29]. This figure illustrates the cyclic nature between generator X and Y, which translates between domain X and Y.

indistinguishable from domain X. This is again through optimizing a generator F and discriminator D_x, essentially learning the inverse of G shown as equation F: Y -> X. The result of this CycleGAN architecture thus results in two generators and two discriminators.

For this specific application, generator G can generate a female image based on the image of a male. Furthermore, generator F can generate a male image based on the image of a female. The discriminator D_y determines whether the female image generated by generator G is real or fake. Moreover, the discriminator D_x determines whether the male image generated by generator F is real or fake. Both these generators and discriminators are trained simultaneously, as described in common GAN architectures [7].

3.2.1 Generator. The generator model used for this CycleGAN architecture is similar to a residual neural network. The down-sampling before the Residual Blocks is done through a layer of 2D Convolution with a vertical and horizontal stride of two. An Instance Normalization layer and a ReLu Activation layer follow this. After the residual blocks, the up-sampling is done through a layer of 2D Transposed Convolution with a vertical and horizontal stride of two. This convolutional layer is followed again by an Instance Normalization and ReLu Activation layer.

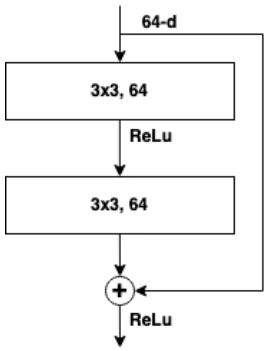


Figure 4: Overview of ResNet Block as proposed by [10]. This figure illustrates the two convolutions and the skip connection, which make a residual block used in the CycleGAN generator model architecture.

The Residual Blocks use a familiar layer configuration similar to figure 4 starting with Reflection Padding, followed by a 2D Convolution, Instance Normalization, and a ReLu Activation layer. These layers are repeated twice, but instead of a second Activation layer, the input to the Residual Block is added to the result provided by the Residual Block. The complete overview of the generator is shown in table 1.

Layer (type)	Output Shape	Parameter #
Input Layer	[224, 224, 3]	0
2D reflection padding	[230, 230, 3]	0
2D Convolution	[224, 224, 64]	9.408
Instance normalization	[224, 224, 64]	128
ReLu activation	[224, 224, 64]	0
ResNet Block (9x)	[56, 56, 256]	12.400.640
2D transpose convolution	[112, 112, 12]	294.912
Instance normalization	[112, 112, 12]	256
ReLu activation	[112, 112, 12]	0
2D transpose convolution	[224, 224, 64]	73.728
Instance normalization	[224, 224, 64]	128
ReLu activation	[224, 224, 64]	0
2D reflection padding	[230, 230, 64]	0
2D Convolution	[224, 224, 3]	9.411
Tanh Activation	[224, 224, 3]	0

Table 1: Layers of generator in CycleGAN architecture used for generating images. Architecture adopted from [29], using the example provided by Keras at [18]. The input image size for all components of the CycleGAN architecture was reduced to decrease the memory usage required to train it.

3.2.2 Discriminator. The discriminator model used for this CycleGAN architecture uses a layer configuration shown in table 2. The down-sampling is again done through a layer of 2D Convolution with a vertical and horizontal stride of two. An Instance Normalization layer and a Leaky ReLu Activation layer follow this. The last down-sample does not use a stride of two but a stride of one. A 2D Convolution layer finally follows this.

Layer (type)	Output Shape	Parameter #
Input Layer	[224, 224, 3]	0
2D Convolution	[112, 112, 64]	3.316
Leaky ReLu activation	[112, 112, 64]	0
2D Convolution	[56, 56, 128]	131.072
Instance normalization	[56, 56, 128]	256
Leaky ReLu activation	[56, 56, 128]	0
2D Convolution	[28, 28, 256]	524.288
Instance normalization	[28, 28, 256]	512
Leaky ReLu activation	[28, 28, 256]	0
2D Convolution	[28, 28, 512]	2.097.162
Instance normalization	[28, 28, 512]	1024
Leaky ReLu activation	[28, 28, 512]	0
2D convolution	[28, 28, 1]	8.193

Table 2: Layers of discriminator in CycleGAN architecture used for evaluating images. Architecture adopted from [29], using the example provided by Keras at [18].

Both the architecture of the generator and the architecture of the discriminator could potentially be improved upon by investigating optimizations to the layers described in table 1 and table 2. This optimization would, however, introduce the need to compare generated results past the current qualitative approach. The objective of this research is not to find an improved CycleGAN architecture. The CycleGAN is used to generate domain translations, so the impact of using these in the training data set can be examined. For this reason, an improved CycleGAN architecture is not considered further.

3.2.3 Loss function. The total loss of this CycleGAN architecture shown in figure 3 comprises the loss function for both the aforementioned GANs using domain X and Y as input; these networks adopt the architecture described by [11]. Only optimizing this loss function still has the potential to create realistic but unrelated images. This behavior happens because the loss of the discriminators optimizes how well the generated images fits in the other domain, but this does not say anything about how related the image is to the original. This behavior requires a third component to the loss function penalizing a difference between the input x and the output of F(G(x)) and vice versa. This loss function still allows the generators to change the tint of input images when there is no need to [29]. This is why a fourth component is added to the loss function, which is the identity mapping loss.

$$\begin{aligned}&\mathcal{L}(G, F, Dx, Dy) = \mathcal{L}_{GAN}(G, Dy, X, Y)\\&+\mathcal{L}_{GAN}(F, Dx, Y, X) + \lambda\mathcal{L}_{CYC}(G, F)\\&+\lambda\mathcal{L}_{CYC}(F, G) + \mathcal{L}_{identity}(G, F)\end{aligned} \tag{1}$$

Calculating the loss of the generator is done by calculating the mean square error between the continuous evaluation of the discriminator and a vector of the same shape containing an evaluation for which all of the generated images are classified as being real, essentially evaluating how real the discriminator perceived the generated image to be and converging towards a real prediction for every input.

Calculating the discriminator's loss is also done by first calculating the mean square error of evaluating a real image and a vector of the same shape containing an evaluation for which all real images are classified as real. Secondly, the mean square error is calculated for evaluating a fake image and a vector of the same shape containing an evaluation for which the fake images are classified as fake. This loss function evaluates how well the discriminator can identify the fake image as fake and how well the discriminator can identify the real image as real, converging towards all correct predictions. The cycle loss is calculated using the mean square error between the real image and the corresponding generated image in the same domain, converging toward related images. Finally, the identity loss is calculated using the mean absolute error between the original image and the corresponding generated image in the same domain, converging towards similar tints in the generated images.

Using this loss function, the two generators and two discriminators in the CycleGAN architecture are trained by performing a forward pass and propagating the calculated gradient back through the trainable variables. The optimization is done through an Adam optimizer, and the model is initialized using a random normal distribution based on the implementation described by [29]. 90% of the images are used as training data for male and female domains. The remaining 10% will remain available for quantitative inspection. The model is trained on the training data for 150 epochs. After training this architecture using the three-part loss function shown in equation 1 for optimization, it can be used at inference time to generate a translated image for a corresponding source image.

3.2.4 Pre-processing. Before training the CycleGAN architecture, the input images go through a pre-processing stage. The default implementation of the CycleGAN architecture provided by Zhu et al. performs both a random crop and random flip, followed by normalizing the images [29]. In the default implementation, the random crop uses a resize based on the nearest neighbor resize method. The resize method was changed to the bi-linear because the nearest neighbor resize method would overemphasize the black

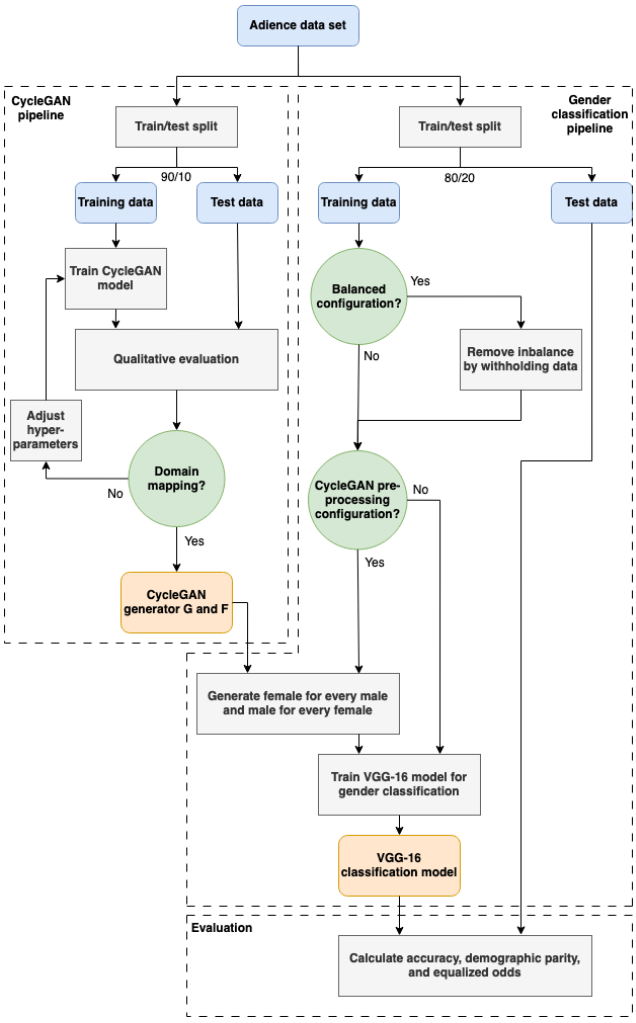


Figure 5: A detailed overview of data sets, models, decisions, and actions that are part of the methodology. Based on the configuration chosen by making the decisions shown as blue circles, results are reported based on evaluating the gender classification model's performance on the test data set.

corners in the coarse images, resulting in an increasingly dark image generated by the CycleGAN generators.

3.2.5 Learning rate. One important hyperparameter for training neural networks using stochastic gradient descent is the learning rate [7]. Figure 5 shows a qualitative evaluation step that could lead to the adjustment of hyper-parameters. This qualitative evaluation is done through a subjective investigation of the generated images after each training epoch. After 100 epochs of training, the learning rate of both discriminators was manually adjusted from 2e-4 to 5e-4. This adjustment is made because the discriminators seemed to underperform in classifying images as fake in a specific domain. The tolerant discriminator led to generators that would not consistently provide a mapping from gender to gender but instead learned to map to the same gender as this obtains the highest cycle consistency. These trained generators, however, do not lead to the intended gender mapping, which is why the discriminators' learning rate was increased while optimizing the GANs.

3.3 Gender classification pipeline

To investigate the impact of the pre-processing using the CycleGAN model, a gender classification pipeline is created for four configurations. The first serves as a benchmark with pre-processing as it was done in [14]. The second compares the benchmark with the CycleGAN pre-processing in the training stage. The last two serve as a comparison when there is an imbalance in the training data; this essentially means the same two configurations are run for a different data set.

The gender classification of facial images is done using a VGG-16 Convolutional Neural Network (CNN) architecture described by Simonyan and Zisserman [23]. This architecture is chosen because [14] provides a performance benchmark, which can be used to evaluate how well the chosen methodology performs compared to existing literature. In addition, the architecture suits the task of gender classification while having a straightforward implementation, confining the complexity of the methodology. The VGG-16 model is initialized using the ImageNet pre-trained weights. Because the model has to perform binary classification of males and females, the fully connected top layers of the VGG-16 model are removed, and the remaining weights are frozen so they will not change while training the model. To this base of the VGG-16 model, additional layers are added to provide the single result in the final layer. Table 4 in appendix A shows the complete model architecture.

To train the VGG-16 model, a train and test split is done for the Adience data set. This split is done for all folds, so the unique individuals are distributed evenly. The training data is pre-processed by performing a random crop, a random flip, and normalization. A potential data augmenting step replaced the pre-processing stage depending on the pipeline's configuration. The generators from the CycleGAN are used to randomly provide a translated image, for which the label is changed accordingly. Multiple epochs are performed while training the model. This pre-processing approach provides the original and the translated version of the source image to the model for training.

The training data is split into folds and used for training models validated on a holdout fold. Performing training iterations for all holdout folds results in learning curves indicating at what epoch the training should stop to prevent over-fitting. The training of the VGG-16 CNN is done using the objective function of increased performance on the gender classification task using well-known optimization methods described in [7]. After determining the optimal amount of epochs obtained through k-fold cross-validation, the final VGG-16 model is trained on all training folds and used to evaluate the test data.

3.4 Evaluation

The test data is classified after training the VGG-16 model for different pre-processing configurations using the training and validation data. These classifications are evaluated to compare the performance and the bias after using the different pre-processing techniques. To measure the performance difference, the model performance is measured by its accuracy calculated as the fraction of correct decisions. In addition, the F-score is calculated to give an indication of the precision and recall performance.

The predictions on the test set are also used to evaluate the bias of the models resulting from different pre-processing configurations. Evaluating the bias is done by calculating the demographic parity and the equal odds. The demographic parity measures the balance of the positive (and negative) predictions by evaluating the predictive equality and equality of opportunity between groups [4]. The equal odds measures the balance of classification errors like false positive and false negatives rates between groups [4].

To summarize the methodology provided in this section, figure 5 shows a graphical overview of the data sets, models, decisions, and actions taken to provide the results used to answer the research questions provided in section 1.

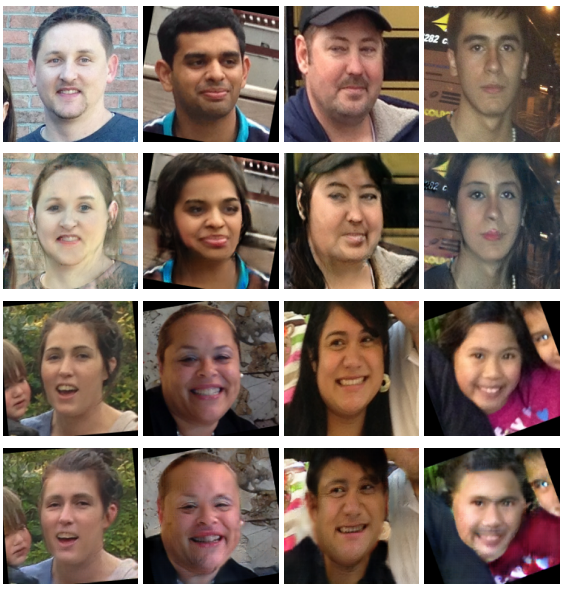


Figure 6: Source image and corresponding generated image provided by the CycleGAN generators. The first row shows source males, and the second shows the corresponding generated female images. The third row shows source females, and the fourth shows the corresponding generated male images.

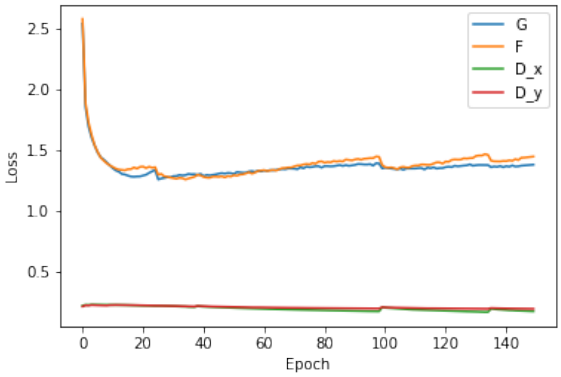


Figure 7: The loss function for generators and discriminators in CycleGAN architecture. Generators G and F show a descending loss function with a homogeneous number of errors leading to a comparable loss function. Both discriminators D_x and D_y show no initial descending loss but similar behavior for the number of errors calculated by the loss function. This could be due to the cyclical loss function where one of the two generators' massively increasing performance has a less significant impact because it does not increase the cyclical loss as much as two improving generators.

4 RESULTS

To investigate the impact of using a CycleGAN architecture as a data-augmenting pre-processing method, the sub-research questions stated in section 1 is addressed. First, the training of the CycleGAN and its resulting generators are discussed. After this, the different configurations of the gender classification pipeline are evaluated so the bias and performance can be reported.

4.1 CycleGAN

The first sub-research question enquires how a GAN can perform gender-to-gender generation. Section 3 explains how a combination of two GANs can be used with a cyclical loss function to train two generators that can perform unpaired translations of images. For this research's specific purpose, the two generators were trained for male and female images. To demonstrate the CycleGAN's effectiveness, several domain translations are shown in figure 6.

The loss values for the networks can be inspected to get a more detailed insight into the training of both the generators and discriminators. In figure 7, the development of the loss functions is shown. Here it becomes clear that the two pairs of generator and discriminator show similar internal adversarial behavior, indicating that the male-to-female and female-to-male GANs seem to learn at a similar pace. In addition, the GANs seem to avert common limitations like model collapse and non-convergence [21].

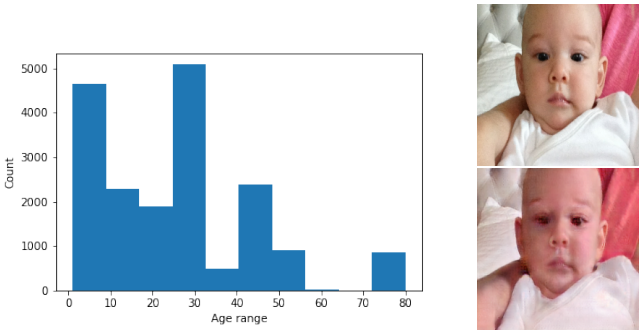
Mode collapse happens when generators learn to map different inputs to the same output. This, however, does not seem to happen for the GANs trained in this research. This is observed by inspecting generated images and subjectively evaluating that the different outputs do not show the same generated face. In addition, both generators show slight increases in their loss while the discriminators show slight decreases in loss, especially around epochs 100 and 135. The generators and discriminators, however, recover from this failure mode. The result is that none of the loss functions seems to converge towards zero, indicating that the networks keep learning on each new epoch and do not suffer from non-convergence.

This behavior could be due to the cyclical loss function discussed in section 3. Because both the GANs are optimized to have cyclical behavior for an input image, the generators are incentivized not to diverge to similar or low-quality outputs. Because of this, the discriminators also keep having relevant predictions, making both the loss functions stable. Because of the equilibrium in generator and discriminator loss, the generators are expected to keep improving, generating fake domain mappings.

An observation made by empirically inspecting the results of the CycleGAN generators is that images of infants do not translate as noticeably as images of other age categories. This behavior is shown in figure 8. This result is notable as infants are not underrepresented in the data, as shown in figure 8; the amount of labels below ten years old is almost the largest group in the Adience data set. The creators of the Adience data set report that ages between zero and two are the second largest group in the data set, behind ages between 25 and 32 [5]. Under-representation in the data is thus not an obvious explanation. An alternative explanation of this result could be that images in both the male and female domains are similar for this age range. Thus, the generators do not learn as strong a mapping between the domains because the discriminators do not judge them fake as often. Possible solutions for this limitation are discussed in section 5.

4.2 Gender classification

A baseline must be established to answer the second sub-research question stated in section 1. This is done by evaluating the balanced and imbalanced data configurations. Only the prevailing pre-processing techniques described in section 3 are used for this configuration. The training data was split into a train and validation set to determine around what epoch to stop the training process



(a) (b)

Figure 8: (a) shows an overview of the age labels provided in the Adience data set. (b) shows an example of a female source image for an infant on the top row, below the corresponding generated male image. Upon quantitative inspection, it does not seem that the generator applies gender-altering mapping.

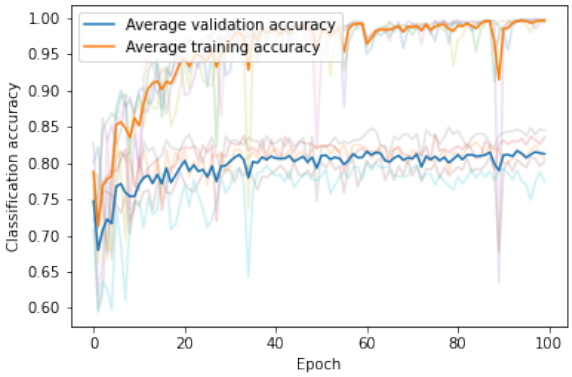


Figure 9: Accuracy of VGG-16 model at gender classification on validation set. The background of this plot shows the validation and training accuracy per fold. This allows us to observe the outlier data points. The average validation and training accuracy is shown in the foreground. The validation accuracy seems to stagnate around epoch 60 without showing signs of over-fitting in the training accuracy.

and use the model. The average validation accuracy is shown in figure 9, which increases meaningfully until around epoch 60. The average is calculated from the accuracy measured for each fold. These individual results are shown in the background of figure 9.

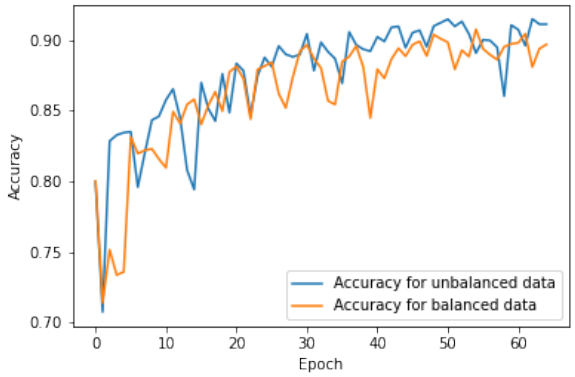
Based on the information in figure 9, the decision is made to train the final VGG-16 gender classification models for around 60 epochs. This allows it to reach the expected accuracy without potentially over-fitting the test data. The results of this training process and the accuracy as measured on the test data are shown in figure 10. As the model seems stagnated with little fluctuations around epoch 60 but does not show specific over-fitting just before or after 60 epochs, the model with the highest accuracy between epoch 55 and epoch 65 is chosen. Results for this are shown in figure 10.

To evaluate fairness, demographic parity is used to measure the probability of a particular prediction depending on sensitive group membership, as reported in table 3. The results are reported as the difference between the largest and the lowest group-level selection rate across all values of the sensitive feature. The demographic parity difference of zero means all groups have the same selection rate. In addition, an equalized odds difference of zero means that all groups have the same true positive, true negative, false positive, and false negative rate [1].

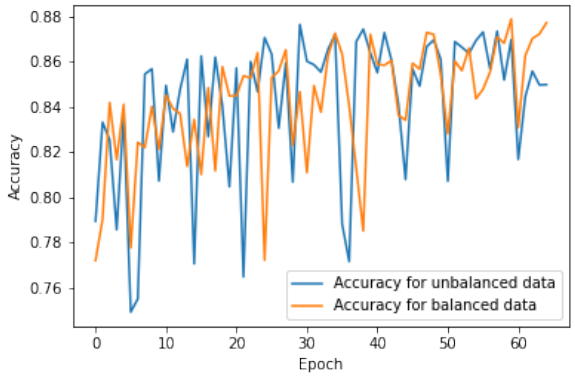
4.2.1 Common pre-processing. To evaluate the baseline for both the balanced and unbalanced data set configuration, the models are trained up until epoch 55 for the former configuration and epoch 63 for the latter configuration. As the first configuration is for a balanced data set, and the second configuration is for an unbalanced data set that is not highly unbalanced, the accuracy is measured as the fraction of correct decisions. These accuracy results are provided in table 3 for both configurations. Both results are similar to the reference paper mentioned in section 3. However, it is important to recognize that this reference paper reported results for an unbalanced data set.

The demographic parity difference and equalized odds difference are also provided in table 3. For both the balanced and unbalanced data set, values above 0.8 are observed, which is quite far from 0. This indicates that demographic parity and equalized odds have not been achieved.

4.2.2 CycleGAN pre-processing. To compare the balanced and unbalanced data set configuration with CycleGAN pre-processing



(a)



(b)

Figure 10: Information about the training process of the VGG-16 gender classification model on the test data set. (a) shows the model's accuracy for the balanced and unbalanced training data set configuration. Both accuracy lines show a similar increasing curve to the results recorded on the validation set shown in figure 9. (b) shows the model's accuracy for which both the balanced and unbalanced training images were augmented using a CycleGAN. Here the accuracy in the initial phase of the training seems to be higher. However, the accuracy is not similar to the model for which no augmentation has been applied to the training data.

[1] https://fairlearn.org/

applied, the models are trained until epoch 60 for the balanced configuration and epoch 58 for the unbalanced configuration. This allows for the evaluation necessary to answer sub-research questions three and four. The accuracy is again reported as the fraction of correct decisions in table 3. Compared to the benchmark, the VGG-16 gender classification models do not seem to have improved performance when trained on data augmented with a gender-to-gender transformation. Instead, a decrease in performance is observed.

The demographic parity and equalized odds differences are again calculated and provided in table 3. A decrease is observed for the demographic parity difference, especially for the unbalanced data set configuration. For the equalized odds difference, no decrease can be observed for the balanced data set, but a decrease can be observed for the imbalanced data set. This indicates that the preprocessing using a gender-to-gender transformation does seem to positively impact the demographic parity and the equalized odds, mainly when the data is imbalanced. It is, however, important to recognize that the demographic parity difference and equalized odds difference remain far from zero. This indicates that demographic parity and equalized odds have again not been achieved.

Overall, table 3 shows that pre-processing using GANs as a gender-to-gender data augmentation step does influence the accuracy and bias of a VGG-16 Convolutional Neural Network performing a gender classification task. The accuracy decreases by around 3.5 percent for the balanced and imbalanced data set configuration. The bias, however, is measured through demographic parity and equalized odds. The demographic parity decreases by 9 percent for the imbalanced training data set configuration, and 5.7 percent for the balanced training data set configuration. The equalized odds only noticeably change for the unbalanced training data set configuration and shows a decrease of 6.1 percent. Though these results show an influence on the accuracy and bias, it does not show that bias has been removed. A result closer to zero for both the demographic parity and equalized odds is expected to evaluate the resulting model as fair.

Male	Female	GAN	Acc.	F-score	Parity	Odds
13.024	13.024	No	0.911	0.899	0.815	0.890
13.024	17.966	No	0.915	0.892	0.828	0.905
13.024	13.024	Yes	0.877	0.847	0.758	0.892
13.024	17.966	Yes	0.873	0.857	0.738	0.844

Table 3: Different configurations of the gender classification pipeline. It shows the configuration of training data used by showing the number of male and female images. When the same amount of images of both has been used, this means the data configuration was balanced. In addition, it shows the pre-processing configuration, accuracy, F-score, demographic parity, and equalized odds. The first and second rows show the baseline where no CycleGAN pre-processing and classification is performed for a balanced and imbalanced data set, respectively. The third and fourth rows show the performance of the classification when CycleGAN pre-processing is performed. The accuracy and F-score is shown to decrease after applying the CycleGAN pre-processing; however, the demographic parity decreases for both data configurations, and the equalized odds decreases for the imbalanced data configuration.

5 DISCUSSION

Comparing the benchmark provided by answering sub-research question two, to the measurements provided by answering sub-research questions three and four, it is observed that an impact can be measured but does not lead to a significantly different evaluation of the model's fairness. This indicates that bias remains measurable in the output of the model. This means that the methodology described in section 3 does not fully achieve the hypothesized impact described in section 1. This section goes deeper into the discussion of these results.

A limitation of the current approach to training the gender-to-gender mapping with the CycleGAN architecture is that images of infants do not translate. This could be due to infants' less distinguishable characteristics, making it harder for the discriminator to label a generated infant as fake. Future research into the training of a CycleGAN for generating gender-to-gender mapping could mitigate this by, for instance, having different augmentation techniques for different age ranges. Using this improved CycleGAN model to augment data could have positive implications for the downstream task of gender classification. These possible implications are illustrated by an example of the infant shown in figure 8b. During the pre-processing, the image will be translated without observable change in the gender of the infant in the image. The gender label, however, will be changed, creating a wrongly labeled training example. This could potentially negatively impact the accuracy of the gender classification model downstream and might explain the decrease in accuracy reported in table 3.

Another limitation could be the generalizability of the current methodology provided in section 3. The only input to the currently trained CycleGAN is the Adience data set. Though not all available data has been used for the training of the CycleGAN, similar data was used for training both the CycleGAN and the VGG-16 gender classification model. However, It is unclear how well the CycleGAN weights translate to a pre-processing step for a different data set, like the IMDB-WIKI data set. Future research into the generalizability and performance of the CycleGAN architecture for gender-to-gender mapping could, for instance, entail introducing multiple data sets as training data. In addition, no attention was spent on the underlying model architectures of the generators and discriminators of the CycleGAN architecture. These could potentially be optimized to both prevent over-fitting and increase performance. Measuring this performance is also not explored in this research, though it might be interesting to improve training capabilities. This, in turn, unlocks the potential to get insights into the behavior of the CycleGAN concerning potential over-fitting.

Potential future research could investigate these limitations. In addition, different approaches to generating images could be explored. A different type of model used for image generation is the variational autoencoder, which could replace the CycleGAN as the data-augmenting model. Mescheder et al., however, state that GANs generally yield visually sharper results when applied to learning a representation of natural images [17]. However, it is unclear what

the impact of losing the sharpness is, thus making this an exciting area of further investigation.

An ethical concern is that this pre-processing method should be applied knowledgeably depending on the domain and the type of predictions to make. A scenario that could occur is one where characteristics like gender can benefit the model's predictions. In this scenario, it is essential to retain this signal. Diversifying the data could have a negative impact and thus lead to negative real-world implications. An example would be detecting certain diseases that only occur for one of the two genders. Removing the distinction could allow for wrong predictions. When these wrong predictions happen more often for the group for which it is the goal to remove bias, the bias is only reinforced.

6 CONCLUSION

This research examines the impact of generative adversarial networks (GANs) as a gender-to-gender data pre-processing step on bias and accuracy in a gender classification model. A cyclic generative adversarial network (CycleGAN) is trained on the Adience dataset to augment the data by performing a gender translation. Considering the limitations discussed in section 5 and the answers to the sub-research questions provided in section 4, it allows for the answer to the research question provided in section 1. The results show decreases in demographic parity and equalized odds, indicating increases in fairness. These changes in fairness, however, are not significantly consequential to change our evaluation of the fairness of the VGG-16 gender classification model and thus do not allow this research to conclude that it removes bias in the training data. This indicates that the proposed methodology is effective but insufficient for complete bias removal.

Acknowledgements

I would like to express my gratitude to dr. Sahand Mohammadi Ziabari and Joop Snijder for their guidance throughout this research project. In addition, I would like to thank the University of Amsterdam and Info Support for providing the opportunity and the resources to perform this research project.

REFERENCES

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron C. Courville. 2018. Augmented CycleGAN: Learning Many-to-Many Mappings from Unpaired Data. *CoRR* abs/1802.10151 (2018). arXiv:1802.10151 http://arxiv.org/abs/1802.10151

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[3] Afshin Dehghan, Enrique G. Ortiz, Guang Shu, and Syed Zain Masood. 2017. DAGER: Deep Age, Gender and Emotion Recognition Using Convolutional Neural Network. *CoRR* abs/1702.04280 (2017). arXiv:1702.04280 http://arxiv.org/abs/1702.04280

[4] V. Dignum. 2020. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer International Publishing. https://books.google.nl/books?id=ghGlzQEACAAJ

[5] Eran Eidinger, Roee Enbar, and Tal Hassner. 2014. Age and Gender Estimation of Unfiltered Faces. *IEEE Trans. Inf. Forensics Secur.* 9, 12 (2014), 2170–2179. http://dblp.uni-trier.de/db/journals/tifs/tifs9.html#EidingerEH14

[6] Zhiqiang Gong, Ping Zhong, and Weidong Hu. 2018. Diversity in Machine Learning. *CoRR* abs/1807.01477. arXiv:1807.01477 http://arxiv.org/abs/1807.01477

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[8] Maryam Hammami, Denis Friboulet, and Razmig Kechichian. 2020. Cycle GAN-Based Data Augmentation For Multi-Organ Detection In CT Images Via Yolo. In *2020 IEEE International Conference on Image Processing (ICIP)*. 390–393. https://doi.org/10.1109/ICIP40778.2020.9191127

[9] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. 2014. Effective Face Frontalization in Unconstrained Images. *CoRR* abs/1411.7964 (2014). arXiv:1411.7964 http://arxiv.org/abs/1411.7964

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *CoRR* abs/1603.08155 (2016). arXiv:1603.08155 http://arxiv.org/abs/1603.08155

[12] M. I. Jordan and T. M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349, 6245 (2015), 255–260. https://doi.org/10.1126/science.aaa8415 arXiv:https://www.science.org/doi/pdf/10.1126/science.aaa8415

[13] Jun Beom Kho. 2018. Multi-Expert Gender Classification on Age Group by Integrating Deep Neural Networks. *CoRR* abs/1809.01990 (2018). arXiv:1809.01990 http://arxiv.org/abs/1809.01990

[14] Sebastian Lapuschkin, Alexander Binder, Klaus-Robert Muller, and Wojciech Samek. 2017. Understanding and Comparing Deep Neural Networks for Age and Gender Classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*.

[15] Gil Levi and Tal Hassner. 2015. Age and gender classification using convolutional neural networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2015), 34–42.

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635 [cs.LG]

[17] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. 2017. Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 2391–2400. https://proceedings.mlr.press/v70/mescheder17a.html

[18] Aakash Kumar Nain. 2020. CycleGAN. https://keras.io/examples/generative/cyclegan/. [Online; accessed 11-March-2023].

[19] Massimo Salvi, U. Rajendra Acharya, Filippo Molinari, and Kristen M. Meiburger. 2021. The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine* 128 (2021), 104129. https://doi.org/10.1016/j.compbiomed.2020.104129

[20] Veit Sandfort, Ke Yan, Perry Pickhardt, and Ronald Summers. 2019. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific Reports* 9 (11 2019). https://doi.org/10.1038/s41598-019-52737-x

[21] Divya Saxena and Jiannong Cao. 2021. Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions. *ACM Comput. Surv.* 54, 3, Article 63 (may 2021), 42 pages. https://doi.org/10.1145/3446374

[22] Connor Shorten and Taghi Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6 (07 2019). https://doi.org/10.1186/s40537-019-0197-0

[23] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[24] Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. (2011), 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

[25] P. Viola and M. Jones. 2001. Robust real-time face detection. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, Vol. 2. 747–747. https://doi.org/10.1109/ICCV.2001.937709

[26] Wenying Wu, Pavlos Protopapas, Zheng Yang, and Panagiotis Michalatos. 2020. Gender Classification and Bias Mitigation in Facial Images. *CoRR* abs/2007.06141 (2020). arXiv:2007.06141 https://arxiv.org/abs/2007.06141

[27] Suorong Yang, Weikang Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Furao Shen. 2022. Image Data Augmentation for Deep Learning: A Survey. https://doi.org/10.48550/ARXIV.2204.08610

[28] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. 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*. Association for Computational Linguistics, Copenhagen, Denmark, 2979–2989. https://doi.org/10.18653/v1/D17-1323

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *CoRR* abs/1703.10593 (2017). arXiv:1703.10593 http://arxiv.org/abs/1703.10593

Appendix A VGG-16 MODEL ARCHITECTURE

Layer (type)	Output Shape	Parameter #
Input Layer	[224, 224, 3]	0
VGG-16 base model	[7, 7, 512]	14.714.688
Dropout	[7, 7, 512]	0
2D Convolution	[7, 7, 512]	2.359.808
Batch normalization	[7, 7, 512]	2.048
Dropout	[7, 7, 512]	0
2D Convolution	[7, 7, 128]	589.952
Batch normalization	[7, 7, 128]	512
Dropout	[7, 7, 128]	0
2D convolution	[7, 7, 384]	442.752
Batch normalization	[7, 7, 384]	1.536
Dropout	[7, 7, 384]	0
2D Convolution	[7, 7, 384]	1.327.488
Batch normalization	[7, 7, 384]	1.536
Dropout	[7, 7, 384]	0
2D Convolution	[7, 7, 500]	1.728.500
Batch normalization	[7, 7, 500]	2.000
2D max pooling	[4, 4, 500]	0
Flatten	[8.000]	0
Dense	[2.048]	16.386.048
Dropout	[2.048]	0
Dense	[2.048]	4.196.352
Dropout	[2.048]	0
Dense	[2.048]	4.196.352
Dropout	[2.048]	0
Sigmoid activation	[1]	2.049

Table 4: Layers of VGG-16 model architecture adopting the VGG-16 base model from [23], using pre-trained weights.