# Misclassifications of Contact Lens Iris PAD Algorithms: Is it Gender Bias or Environmental Conditions?

Akshay Agarwal[1], Nalini Ratha[2], Afzel Noore[3], Richa Singh[4], and Mayank Vatsa[4]
[1]IISER Bhopal, India, [2]University at Buffalo, USA, [3]Texas A&M University Kingsville, USA,
[4]IIT Jodhpur, India
akagarwal@iiserb.ac.in, nratha@buffalo.edu, afzel.noore@tamuk.edu, {richa, mvatsa}@iitj.ac.in

## Abstract

*One of the critical steps in biometrics pipeline is detection of presentation attacks, a physical adversary. Several presentation (adversary) attack detection (PAD) algorithms, including iris PAD, have been proposed and have shown superlative performance. However, a recent study, on a small-scale database, has highlighted that iris PAD may have gender biases. In this research, we present a rigorous study on gender bias in iris presentation attack detection algorithms using a large-scale and gender-balanced database. The paper provides several interesting observations which can help in building future presentation attack detection algorithms with aim of fair treatment of each demography. In addition, we also present a robust iris presentation attack detection algorithm by combining gender-covariate based classifiers. The proposed robust classifier not only reduces the difference in accuracy between different genders but also improves the overall performance of the PAD system.*

## 1. Introduction

Biometric identification has received significant attention in recent times because of its high accuracy and non-intrusive nature. In the constrained environment iris is one of the most accurate modalities for person identification [10, 14, 28, 37, 54]. While the iris recognition algorithms show spectacular performance [36], they can easily be circumvented via presentation attack instruments (PAI)[1] [13, 29] such as 3D contact lenses and 2D printed photo [18, 52]. These presentation attack instruments are reasons of serious concern towards the secure deployment of iris recognition systems [7, 31]. In the literature, several iris presentation attack detection algorithms are developed and they have shown success in defending different PAIs [2, 3, 17, 22, 53]. However, similar to the recent literature which highlights that biometric recognition algorithms show demographic biases [20, 39, 47, 48], iris presentation attack detection algorithm are found vulnerable to gender bias. In recent work by Fang et al. [19], it has been shown that gender bias affects iris presentation attack detection (IPAD). While the accurate detection of presentation attack is essential [4, 5], the 'bias free' attribute of the algorithm is also a necessity. The bias can be described in terms of differential outcome resulting in the classification error concerning different demographic groups such as male and female. For instance, the classification error of presentation attack detection in the female group can be lower as compared to the male group. Figure 1 shows the importance of the proposed study and the impact of unfair presentation attack detection algorithm exploited by an attacker for illegal access. The proposed study has two-fold contributions: firstly, gender covariate and contact lens presentation attack instrument are used to study the fairness of PAD algorithms. Secondly, we present a iris PAD algorithm for addressing the challenges identified by the study. In brief, the contributions of this research are:

- a detailed gender bias study is performed using large-scale ocular database captured in both controlled indoor and unconstrained outdoor environmental setting;
- both pre-trained and trained from scratch CNN architectures are selected for the study and the robustness of different classifiers such as support vector machine and random decision forest are studied;
- based on the vulnerability evaluation of the classifiers, a robust iris PAD is presented.

## 2. Related Studies

In this section, we provide a brief overview of existing bias studies in biometrics in general as well as related to iris presentation attack detection algorithms.

[1] Presentation attacks are physical adversaries that intend to affect the performance by either circumventing or eluding the identity (www.iso.org/standard/53227.html).

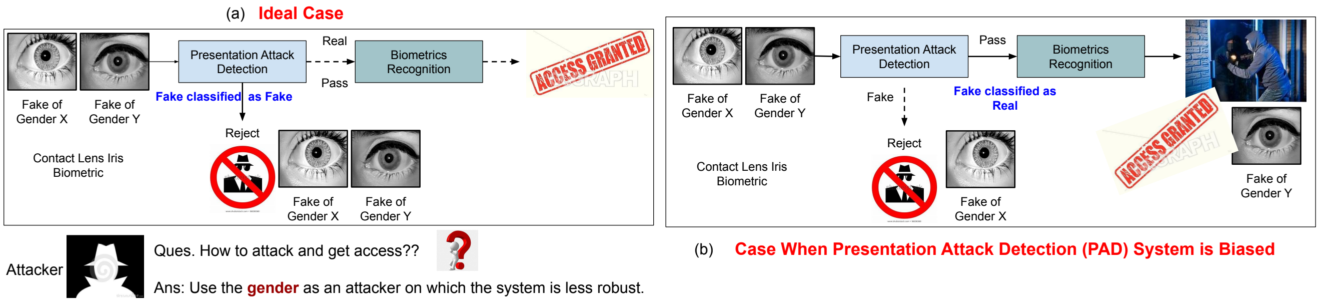


Figure 1. Illustrating the impact of bias on the recognition system. (Left) In an ideal case the fake image of any gender ($X$ or $Y$) must be classified as fake (spoof) and it will not be passed (broken dashed line) to the system for recognition. (Right) However, in the unfair/biased case, an attacked image of one gender (say $Y$) might be classified as real and the person may get illegal access to the recognition system. Dashed arrow ($-- >$) represents the broken connection, for example, in the left part, if the image is classified as fake it must not pass to the recognition part. Similarly, in right part, if the fake is classified as real it will not be rejected through PAD.

**Recognition Bias:** Recent studies have raised a serious concern towards the bias and fairness of machine learning algorithms [1, 8, 25, 42, 49]. Biometric recognition algorithms are not untouched from the bias and fairness issue [40]. *Bias* in any machine learning algorithm can be broadly described using the following terms: (i) data bias and (ii) model/classifier bias. The deep learning algorithms generally require a large amount of data and the training data bias towards one demographic can lead to a biased trained model. Therefore, the need for balanced databases has increased in recent times. Robinson et al. [43] have prepared in-the-wild database with attributes such as gender and race balanced images. Apart from the creation of databases, the one quick solution for such drawback is the use of data resampling [16, 35, 38] either through augmentation or removal of access samples of dominant demographics. The drawback of such an algorithm can be the loss of information when under-sampling is performed or the increase of noisy data under over-sampling. Another quick solution to reduce the bias in an unbalanced dataset is to apply the cost-sensitive learning by assigning a higher weight to a less representative group [9, 12]. To reduce the model bias, observing the importance of filters reflecting unbias nature in the decision-making can be one possible solution. Nagpal et al. [41] have proposed an approach to drop the bias-sensitive filters to improve the performance. Later, Nagpal et al. [39] also proposed a diversity block to improve the performance. Majumdar et al. [34] proposed an adversarial perturbation-based algorithm by learning a unique perturbation vector to mitigate the bias effect in the training database.

**PAD Bias:** In the presentation attack detection domain, there is limited research about bias and possible solutions, particularly in iris PAD which is one of the most critical components of the (iris) biometrics recognition pipeline. The only literature available is recently conducted study by Fang et al. [19] on iris presentation attack detection. The authors have selected three different machine learning classifiers: (i) Local Binary Patterns (LBP) a hand-crafted feature-based algorithm equipped with linear support vector machine classifier, (ii) VGG-16 as a features extractor, and linear SVM for classification, and (iii) MobileNeV3-small is trained from scratch using softmax classifier. It is found that the PAD algorithms have shown better performance on male samples as compared to the female samples. However, the attack images considered for testing are 370 for the male class and 130 for the female class. It is observed that not only the testing set of this study but the train set is also imbalanced with respect to gender distribution. Therefore, we assert that a more thorough study using a large balanced database and well defined even protocols is required to establish bias and fairness issues in PAD algorithms.

## 3. Iris Presentation Attack Detection

Popular iris presentation attack detection algorithms range from hand-crafted features with traditional classifiers to deep neural networks. Popular handcrafted image features such as Local Binary Patterns, Histogram of Oriented Gradients, Structural and Textural features along with support vector machine classifier (SVM) have been applied for the detection of presentation attack instruments [21, 23, 32]. While the hand-crafted features-based algorithms have the computational advantage, their generalizability against unseen databases and attacks is a serious concern. To overcome such issues, research works have started utilizing deep convolutional neural networks (CNN) both as feature extractors and classification. Inspired by the success of fusion from face PAD [6, 45], Yadav et al. [50] have utilized the fusion of image features and CNN to propose an effective iris presentation attack detector. Gupta et al. [22] have proposed a novel shallow CNN for iris presentation attack detection. Among multiple CNN architectures, DenseNet

[27] has received a significant attention for iris presentation attack detection both when utilized as feature extractor and trained from scratch [44, 51]. The study by Fang et al. [19] also suggest the higher accuracy of CNNs as compared to hand-crafted feature. In this research, we have also studied the role of a classifier in the classification of images belonging to a specific gender. We assert that it might be possible that a certain classifier is biased towards particular gender and hence make the entire algorithm look bias towards that gender. If it is the case, then the multiple classifiers can be combined to develop a gender unbiased PAD system.

In brief, this is the first work addressing both data and model/classifier bias in iris presentation attack detection. We have used the following CNNs and classifiers for the development of iris presentation attack detection system:

### 3.1. Feature Extractor: Transfer Learning

1. VGG-16: Similar to Fang et al. [19], we have used the pre-trained [15] VGG-16 [46] network as a feature extractor. Before extracting the features of the FC layer, a global pooling layer has been applied to obtain the features of dimension $1 \times 512$;

2. DensetNet: Similar to VGG-16 and based upon its popularity in IPAD [44, 51], the last fully connected layer values of the DenseNet are used as a feature representation of an image. A final feature vector of dimension $1 \times 1024$ is obtained from each iris image;

3. ResNet-50 [24]: With ResNet-50 architecture, a global pooling is applied after the final fully connected layer which yields the feature dimension of $1 \times 2048$.

### 3.2. Feature Extraction + Classification: CNN from Scratch

The variety of the feature learners and classifier is essential to reach any strong conclusion about fairness. Therefore, we have also utilized Wide-ResNet-16-8 [55]. In contrary to the deeper networks, this network uses the increased width with lower depth. This helps in decreasing the network parameters significantly. The network is trained from scratch using categorical cross-entropy loss and softmax classifier. The network is trained using adaptive learning rate, the batch size is set to 32 and Adam optimizer [30] has been used. We have also utilized the shallow and popular image classification network namely AlexNet [33]. The network consists of 5 convolutional layers and 3 fully connected layers. The difference between AlexNet and other networks is the use of large convolutional filters and fewer layers for feature learning and classification. The AlexNet is also trained from scratch using the similar parameter setting used for Wide-ResNet-16-8.

| **Characteristics** | **This Paper** | **Fang et al.** [19] |
|---|---|---|
| Subjects | 81 (40 male, 41 female) | 18 (14 male, 4 female) |
| Total Images | 18,706 | 3,400 |
| Real Images | 9,319 | 1,700 |
| Attack Images | 9,387 | 1,700 |
| Contact Lens | 4 | – |
| Lens Colors | Balanced (Blue, Green Violet, Brown) | Unbalanced |
| Acquisition Environment | Indoor (Controlled)<br>Outdoor (Unconstrained) | – |
| Sensors | 3 | 2 |

Table 1. Characteristics of the databases used for bias-study in iris presentation attack detection.

### 3.3. Classifiers

Based on the assertion that different classifiers might be biased towards different gender classes, in this research, we have used two classifiers along with a softmax classifier of Wide-ResNet-16-8 and AlexNet. The linear support vector machine (SVM) [11] and random decision forest (RDF) [26] classifiers are trained on the features extracted from the pre-trained CNNs mentioned in Section 3.1.

## 4. Database and Experimental Protocols

For different experiments, we have used the Unconstrained Multi-sensor Iris Presentation Attack (UnMIPA) database[2] prepared by Yadav et al. [51]. In total, the database contains $18,706$ images, approximately balanced between real and attack images. The images are captured from 81 subjects, out of which 40 subjects are male and 41 subjects are female. The equal number of images belonging to real and attack classes and (almost) an equal number of subjects makes the database an ideal choice to perform the bias study. Moreover, the database is captured in both controlled indoor (*IN*) and unconstrained outdoor (*OUT*) environments to reflect the real-world conditions. The other interesting properties which make the database a good choice are different number of acquisition sensors and contact lens manufacturers. Tables 1 and 2 show the

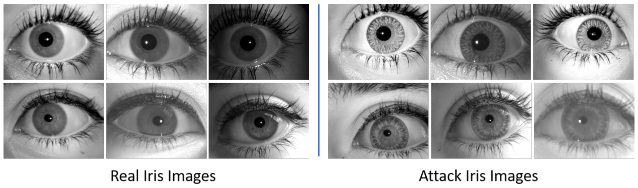


Figure 2. Samples of the database used in this research reflecting the variations required for an effective study. Fist row images are captured in the indoor controlled illumination environment and second row images are captured in the outdoor unconstrained illumination environment setting.

[2] http://www.iab-rubric.org/index.php/wvu

| Gender | Class | Fang et al. [19] | | Proposed Study (IN: Indoor) | | Proposed Study (OUT: Outdoor) | |
|---|---|---|---|---|---|---|---|
| | | Train | Test | Train | Test | Train | Test |
| Male | R. | 600 | 250 | 1138 | 1137 | 1157 | 1156 |
| | A. | 920 | 370 | 1157 | 1157 | 1161 | 1160 |
| Female | R. | 600 | 250 | 1164 | 1163 | 1202 | 1202 |
| | A. | 280 | 130 | 1190 | 1189 | 1187 | 1186 |

Table 2. Subject-disjoint number of images used for training and testing corresponding to different genders. A total of 18706 iris images are used for experimentations as compared to 3400 real and textured contact lens iris images are used in the only study by Fang et al [19]. R. And A. represent the real and attack classes, respectively. Indoor and Outdoor represents to the controlled indoor and unconstrained outdoor imaging environment, respectively.

| Train ↓ Test → | | Indoor | | Outdoor | |
|---|---|---|---|---|---|
| Env. | Gender | Male | Female | Male | Female |
| AlexNet | | | | | |
| IN | Male | **96.43** | 90.09 | **82.08** | 75.88 |
| | Female | 84.31 | **88.18** | **69.39** | 68.47 |
| OUT | Male | 71.23 | **78.66** | **69.86** | 68.76 |
| | Female | 62.69 | **64.63** | 62.52 | **66.25** |
| Wide-ResNet16-8 | | | | | |
| IN | Male | **99.48** | 96.43 | **89.90** | 86.73 |
| | Female | 97.82 | **99.70** | **91.45** | 90.28 |
| OUT | Male | 98.13 | 98.81 | **99.40** | 98.20 |
| | Female | 96.51 | 98.26 | 96.80 | **99.20** |

Table 3. Iris presentation attack detection performance of the CNNs trained from the scratch. The detector is trained and tested on the images of individual genders. IN and OUT represents the indoor and outdoor imaging environment, respectively.

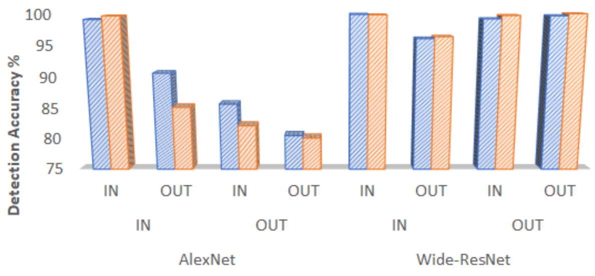
characteristics and experimental split of the database used in this research, and its comparison with the database used in the study by Fang et al. [19]. The proposed study consists of a fair split in terms of the number of images per class (real vs. attack), per gender (male vs. female), and training-testing subset. In comparison, Fang et al. have used highly unbalanced (gender-wise) train-test split. To study the gender bias in this research, the database is divided on subject disjoint training and testing condition, i.e., the images of the subjects used in the training are not used for evaluation to make the system bias-free. The characteristics such as the wide number of lenses and their colors, acquisition sensors, and imaging environment make the proposed study effective in finding the bias issue in contact lens iris PAD algorithms. We have conducted an extensive experimental study using the large-scale nature and different characteristics of the UnMIPA database.

## 5. Experimental Results and Analysis

First, we analyze the results of the IPAD algorithms where the CNN models are trained from scratch followed by the analysis on transfer learning architectures. Further, the bias of SVM and RDF classifiers is studied using the varying number of images of both genders in the training.

| CNN | Train ↓ Test → | | IN | OUT |
|---|---|---|---|---|
| AlexNet | IN | Male | 6.34 | 6.20 |
| | | Female | 3.87 | 0.92 |
| | OUT | Male | 7.43 | 1.10 |
| | | Female | 1.94 | 3.73 |
| WRN16-8 | IN | Male | 3.05 | 3.17 |
| | | Female | 1.88 | 1.17 |
| | OUT | Male | 0.68 | 1.20 |
| | | Female | 1.75 | 2.40 |

Table 4. The difference (of Table 3 values) in the iris presentation attack detection performance shows that the CNNs trained on which gender favors that gender. For instance, the Wide-ResNet trained using indoor images of the male class yields difference in the accuracy of 3.05% and has higher performance on the male class. The difference in accuracy is greater than 0% but it might be attributed to the fact that models are trained on one gender only.



| | AlexNet | | | | Wide-ResNet | | | |
|---|---|---|---|---|---|---|---|---|
| | IN | | OUT | | IN | | OUT | |
| | IN | OUT | IN | OUT | IN | OUT | IN | OUT |
| ▨ Male | 99.08 | 90.67 | 85.7 | 80.57 | 99.87 | 96.11 | 99.17 | 99.65 |
| ▨ Female | 99.62 | 85.18 | 82.19 | 80.15 | 99.79 | 96.4 | 99.7 | 99.96 |

Figure 3. Iris presentation attack detection performance of the AlexNet and Wide-ResNet-16-8 when trained from the scratch using images of both male and female.

### 5.1. CNNs Trained form Scratch

Two CNNs are selected for this study are Wide-ResNet-16-8 and AlexNet. To the best of our knowledge, the Wide-ResNet architecture is never explored in the study of iris presentation attack detection. The results of AlexNet [33] and Wide-ResNet [55] are reported in the Table 3. Let us first discuss the biased study of the shallow AlexNet architecture. When the network is trained on the individual gender iris images and tested on seen imaging environment iris images, the network shows the bias towards the gender on which it is trained. For instance, when the AlexNet model is trained on the indoor images of the male class and tested on the indoor images, it yields an accuracy of 96.43% and 90.09% on male and female iris images, respectively. A similar accuracy difference can be observed when the network is trained on the female class only (Table 4). However, it cannot be termed as bias as the model is itself trained on the skewed database or only one gender

| Train ↓ Test → | | Indoor | | Outdoor | |
|---|---|---|---|---|---|
| Env. | Gender | Male | Female | Male | Female |
| VGG-16 | | | | | |
| IN | Male | **98.04** | 89.07 | **82.56** | 76.21 |
| | Female | 91.67 | **98.55** | 78.06 | **83.29** |
| OUT | Male | **91.85** | 89.24 | **96.11** | 86.31 |
| | Female | 86.10 | **91.28** | 86.01 | **96.23** |
| ResNet-50 | | | | | |
| IN | Male | **86.53** | 76.16 | **70.16** | 66.12 |
| | Female | 80.95 | **85.71** | **71.50** | 69.80 |
| OUT | Male | **80.68** | 77.93 | **79.97** | 70.52 |
| | Female | 74.72 | **79.08** | 73.57 | **78.06** |
| DenseNet-121 | | | | | |
| IN | Male | **98.04** | 91.88 | **87.82** | 84.51 |
| | Female | 92.80 | **98.68** | 82.60 | **85.09** |
| OUT | Male | **92.15** | 89.41 | **96.33** | 88.32 |
| | Female | 89.97 | **92.77** | 86.61 | **96.94** |

Table 5. Iris presentation attack detection performance when CNNs are used as features extractor and linear SVM as a classifier. The detector is trained and tested on individual genders.

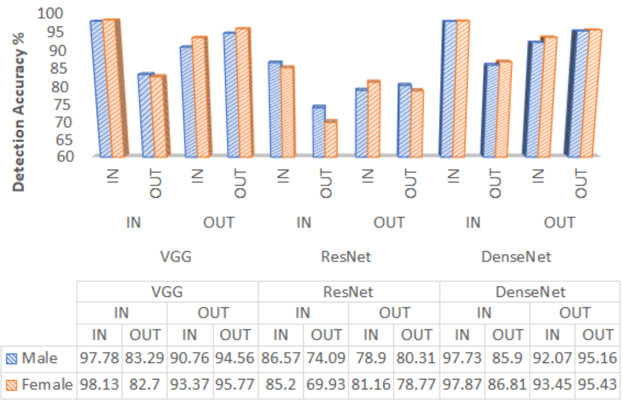
iris images. Therefore, to carefully understand the gender bias issue, in the next experiment we have used both male and female class images for training. As shown in Figure 3, AlexNet and Wide-ResNet both do not shows any significant bias. for instance, when the indoor (IN) iris images are used for Wide-ResNet16-8, the model yields an accuracy of 99.65% and 99.96% on male and female iris images, respectively. Another covariate that causes the accuracy difference even when both male and female iris images are used in training is the data shift or change the testing data characteristics. For instance, when the AlexNet is trained on the iris images of indoor environment and tested on the outdoor (OUT) iris images it shows an accuracy difference of 5.49 (90.67% and 85.18%) as compared to the difference of 0.54% (99.08% and 99.62%) when tested on indoor iris images. Therefore, we observe from these experiments that the gender bias might not be an issue, the testing data characteristics drift is.

## 5.2. Transfer Learning: CNNs as Feature Extractor

In this section, we have conducted an extensive study with multiple CNNs varying in terms of numbers of layers ranging from 16 to 121, and type of connections i.e., sequential to residual. Interestingly, the ResNet performs poorly as compared to both its shallow counterpart, i.e., VGG, and deeper counterpart, i.e., DenseNet. The prime reason might be the dimensionality of the feature vector which is 4 times and 2 times higher than the VGG and DenseNet model, respectively.

| CNN | Train Test | | IN | OUT |
|---|---|---|---|---|
| VGG16 | IN | Male | 8.97 | 6.35 |
| | | Female | 6.88 | 5.23 |
| | OUT | Male | 2.61 | 9.80 |
| | | Female | 5.18 | 10.22 |
| ResNet-50 | IN | Male | 10.37 | 4.04 |
| | | Female | 4.76 | 1.70 |
| | OUT | Male | 2.75 | 9.45 |
| | | Female | 4.36 | 4.49 |
| DenseNet | IN | Male | 6.16 | 3.31 |
| | | Female | 5.88 | 2.49 |
| | OUT | Male | 2.74 | 8.01 |
| | | Female | 2.80 | 10.03 |

Table 6. The difference (of Table 5 values) in the IPAD performance shows the performance of the CNNs trained on which gender favors that gender. For instance, the DenseNet trained using indoor images of the male class yields a difference in the accuracy of 6.16% and has higher performance on the male class when tested on indoor iris images. While the difference in accuracy is greater than 0% but it might be attributed to the fact that models are trained on one gender only.



| | VGG | | | | ResNet | | | | DenseNet | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | IN | | OUT | | IN | | OUT | | IN | | OUT | |
| | IN | OUT | IN | OUT | IN | OUT | IN | OUT | IN | OUT | IN | OUT |
| ▨ Male | 97.78 | 83.29 | 90.76 | 94.56 | 86.57 | 74.09 | 78.9 | 80.31 | 97.73 | 85.9 | 92.07 | 95.16 |
| ▨ Female | 98.13 | 82.7 | 93.37 | 95.77 | 85.2 | 69.93 | 81.16 | 78.77 | 97.87 | 86.81 | 93.45 | 95.43 |

Figure 4. Iris presentation attack detection performance of the VGG, ResNet, and DenseNet features with linear SVM classifier using images of both male and females.

The iris presentation attack detection results of VGG, ResNet, and DenseNet are shown in Table 5. Similar to the networks trained from scratch and as expected, the networks have shown biasness towards the gender on which they are trained. For instance, when the VGG network is trained on the indoor male iris images, it shows more than 8.93% higher accuracy on male iris images as compared to the female iris images (Table 6). Except for one situation, a similar observation of high accuracy when the training and testing genders are the same is witnessed. Therefore, to again properly understand the gender bias issue, we have now used both the gender iris images for training. The results of this experiments are reported in Figure 4. When the features of the VGG model are used for evaluation, the difference in the accuracy is 1.21% (94.56% and 95.77%) when trained and tested on outdoor iris images. Interestingly, the difference increases to 2.61% (90.76% and 93.37%) when the testing data shift occurs, i.e., when trained on outdoor iris images but tested on indoor iris images.

| Train ↓ Test → | | Indoor | | Outdoor | |
|---|---|---|---|---|---|
| Env. | Gender | Male | Female | Male | Female |
| IN | Male | **96.03** | 85.37 | **82.51** | 78.64 |
| | Female | 86.96 | **95.75** | 81.91 | **82.16** |
| | Both | **95.86** | 95.41 | **83.38** | 81.41 |
| OUT | Male | **88.49** | 84.77 | **95.16** | 85.59 |
| | Female | 84.39 | **86.60** | 85.28 | **94.97** |
| | Both | **88.32** | 87.80 | **94.64** | 94.43 |

Table 7. Iris presentation attack detection performance of the DenseNet-121 as features extractor and *RDF* for classification. The detector is trained on the images of individual genders and in combination.

| Train ↓ Test → | | Indoor | | Outdoor | |
|---|---|---|---|---|---|
| Env. | Gender | Male | Female | Male | Female |
| Equal Male and Female (EMF) | | | | | |
| IN | SVM | 97.73 | **97.87** | 85.90 | **86.81** |
| | RDF | **95.86** | 95.41 | **83.38** | 81.41 |
| OUT | SVM | 92.07 | **93.45** | 95.16 | **95.43** |
| | RDF | **88.32** | 87.80 | **94.64** | 94.43 |
| Female Dominant Scenario (FDS) | | | | | |
| IN | SVM | 96.16 | **98.12** | 86.27 | **86.64** |
| | RDF | **93.07** | 95.70 | **83.55** | 82.70 |
| OUT | SVM | 92.46 | **93.45** | **93.57** | 93.35 |
| | RDF | **87.49** | 87.33 | 92.18 | **95.10** |
| Male Dominant Scenario (MDS) | | | | | |
| IN | SVM | **97.34** | 96.98 | **86.32** | 86.22 |
| | RDF | **96.08** | 93.45 | **83.63** | 81.66 |
| OUT | SVM | 91.76 | **92.77** | **95.16** | 93.72 |
| | RDF | **88.31** | 87.37 | **94.52** | 92.63 |

Table 8. Iris presentation attack detection performance of the DenseNet-121 as features extractor. The detector is trained both genders when equal and unequal images of both genders are used to study the impact of imbalance of gender. The importance of balance of images of genders is study using linear SVM and RDF classification.

### 5.3. Key Findings

The previous study by Fang et al. [19] shows that gender bias is an issue in iris presentation attack detection. However, as mentioned earlier, the database imbalance and a limited number of classifiers are the prime factors to reach such a conclusion. In the proposed research, we have conducted an experimental evaluation using a database that is significantly larger than the database used in [19] and contains balanced gender demographic iris images. The experimental evaluation has been performed using multiple deep networks and classifiers such as softmax and SVM. Through the extensive experiments, it is observed that gender bias might not be an issue; however, the training phenomena are. In other words, when the networks are trained on a particular gender they showed higher performance on the same testing gender. When both the gender iris images are used for training, the difference in the accuracy between gender is low which shows the lack of gender bias in the algorithms. Another covariate that raises concern is the acquisition environment. The possible reason for accuracy difference can be attributed to the data characteristics variations across different environments, as shown in Figure 2.

### 5.4. SVM vs. RDF Classifiers

In literature of iris presentation attack detection (IPAD) including bias study usually SVM classifier is considered [13, 19]. To further increase the impact of our study, we considered another popular machine learning classifier namely random decision forest (RDF) [26]. RDF is an ensemble-based method where multiple weak decision trees are learned and combined to reach a strong classifier. To analyze the bias impact on the RDF classifier, the DenseNet features are utilized which performs best compared to other networks. The results are reported in Table 7. The RDF classifier shows clear biases or favoritism towards the gender on which it is trained both in seen and unseen environment train-test setting. However, interestingly, in the presence of both genders, the classifier favors the male gender over the female and shows higher detection performance. Although in the majority of the cases, the difference in the performance between gender is marginal (ranges from 0.21% to 1.97%) when both gender images are used for training which again raises a question of whether gender bias is a concern when the dataset is balanced. Surprisingly, the training and testing environment have not shown a significant impact on the bias analysis of the RDF classifier. In terms of attack detection performance, the linear support vector machine classifier yields better accuracy in comparison to the RDF classifier. For example, in an unseen environment training-testing conditions, the performance of the SVM classifier is at least 4% higher than the RDF detector.

### 5.5. Importance of a Balance Data Study

The preliminary study conducted by Fang et al. [19] concluded that the female gender yields higher detection error as compared to the male gender. We have shown through the analysis (Tables 1 and 2), that the database is highly biased towards the 'male' class and hence the performance can be expected as claimed. Therefore, to understand that phenomenon in detail, we have conducted the following experiments using DenseNet along with two classifiers SVM and RDF. In this study, we have generated the training set with varying proportions of gender images: (i) when balanced iris images of both genders are used, (ii) male dominating scenario which contains approximately double the number of male iris images as compared to female, and (iii) female dominating scenario consists of approximately double the number of female iris images as compared to males. Here, the testing iris images in each condition remain the same to make a fair comparison and are described in Table

2. The balanced gender case represents the use of training images as described in Table 2 as it is. While in the imbalance case, the 50% images of an unaggressive gender are dropped from its training set, while training images of another gender remain the same.

The results of the DenseNet features extractor with SVM and RDF classifiers are given in Table 8. Under the male dominating scenario, the classifiers on DenseNet features show a higher bias towards the male class. Out of 8 conditions, in seven conditions male accuracy slightly dominates female accuracy. Surprisingly, in the female dominating situation, mixed results are observed. In the exciting analysis, it is seen from Table 8 (through color boxes) that the SVM classifier favors the female class; however, the RDF classifier favors the male class under the balanced gender case study. This is observed not only when the seen environmental images are used for training and testing but also in the unseen image acquisition setting. We assert that other external factors such as the acquisition environment may have a bigger role in performance variations.

## 6. Proposed Robust Iris PAD Algorithm

In this research, we present a robust and fair presentation attack detection algorithm by selecting the classifier which favors individual gender even the network is trained using both genders. From the set of machine classifiers such as SVM, RDF, Logistic Regression, and Neural Network, the classifiers which show bias towards one gender are selected. In other words, one classifier favoring the male class is selected; whereas another selected classifier aims to favor the female class. The classifiers are trained using the features of the DenseNet model due to its effectiveness as compared to other evaluated deep CNNs including VGG and ResNet. The flowchart of the proposed algorithm is shown in Figure 5. On the training set containing both gender images, a DenseNet feature representation is calculated for each image. Later, through the evaluation of the pooled classifiers, linear SVM and RDF classifiers are found biased towards male and female gender-attribute, respectively. The selected classifiers are completely in-line with the findings as reported in Table 8. We postulate that the fusion of information of these biased classifiers which have strong confidence in their respective genders can rescind the impact of gender attributes even in the presence of an unseen acquisition environment. Hence, in turn, this annulment can provide us with a fair presentation attack detection classifier. The prediction probabilities of the selected classifiers are concatenated together to form a score vector of dimension $n \times 2$, where $n$ is the number of training images and 2 represents the score of both classifiers on a single image. Once the fused score vector is obtained, a PAD classifier is trained on that for the final robust iris presentation attack detector. At the time of testing, an image is passed through the first level (L1) classifiers (SVM and RDF), which produces the initial decision probabilities which passed to the second level (L2) PAD classifier for the final prediction of the image.

| Train | Test | Classifier | Male | Female | Diff. |
|---|---|---|---|---|---|
| IN | IN | SVM-L1 | 97.73 | 97.87 | 0.14 |
| | | RDF-L1 | 95.86 | 95.42 | 0.44 |
| | | SVM-L2 | **97.82** | **98.09** | **0.27** |
| | OUT | SVM-L1 | 85.90 | 86.81 | 0.91 |
| | | RDF-L1 | 83.38 | 81.41 | 1.97 |
| | | SVM-L2 | **88.69** | **88.27** | **0.42** |
| OUT | IN | SVM-L1 | 92.07 | 93.45 | 1.38 |
| | | RDF-L1 | 88.32 | 87.80 | 0.52 |
| | | SVM-L2 | **93.98** | **94.47** | **0.49** |
| | OUT | SVM-L1 | 95.16 | 95.43 | 0.37 |
| | | RDF-L1 | 94.64 | 94.43 | 0.21 |
| | | SVM-L2 | **95.37** | **95.52** | **0.15** |

Table 9. Iris presentation attack detection accuracy (%) of the proposed de-biased algorithm. L1 and L2 represents the level of the classification stage as shown in Figure 5. The robustness of the proposed algorithm (SVM-L2) can be seen from the significant improvement in accuracy under unseen environment training-testing conditions. Diff. shows the difference in the accuracy between both the genders which shows the proposed L2 classifier gives equal importance to both the genders and significantly bias free than L1 classifiers.

Table 9 shows the detection accuracy on the best performing CNN architecture, i.e., DenseNet features when utilized with first level SVM and RDF classifiers and second level (proposed) SVM classifier. It is observed that when the seen environment images are used for testing, the proposed approach shows lower improvement due to less difference in the accuracy; however, in unseen environment testing settings, the proposed multi-level fusion algorithms yield significant improvement and demonstrated the gender impact nullification. For example, when the indoor (IN) images are used in the training and outdoor (OUT) images are used in the testing, the best classifier at level 1 yields 85.90% and 86.81% detection accuracy on male and female classes, respectively. When the proposed multi-level classifier is used, the accuracy improves to 88.69% and 88.27% for the male and female classes, respectively. Interestingly, the accuracy difference between the male and female at level 2 classification is close to zero, in comparison to the first level gender favoring classifiers. It shows the proposed multi-level fusion algorithm yields better accuracy, nullifies the impact of gender attribute in PAD and improves overall performance as well, generalized in the cross-environment, and is **'bias-free'**. Figure 6 shows the equal error rate (EER) performance of the gender favor level 1 classifier and de-biased level 2 proposed classifier. Similar to the detection accuracy, in most of the cases, the EER shows improvement when the level 2 fusion of the gender-specific classifiers is

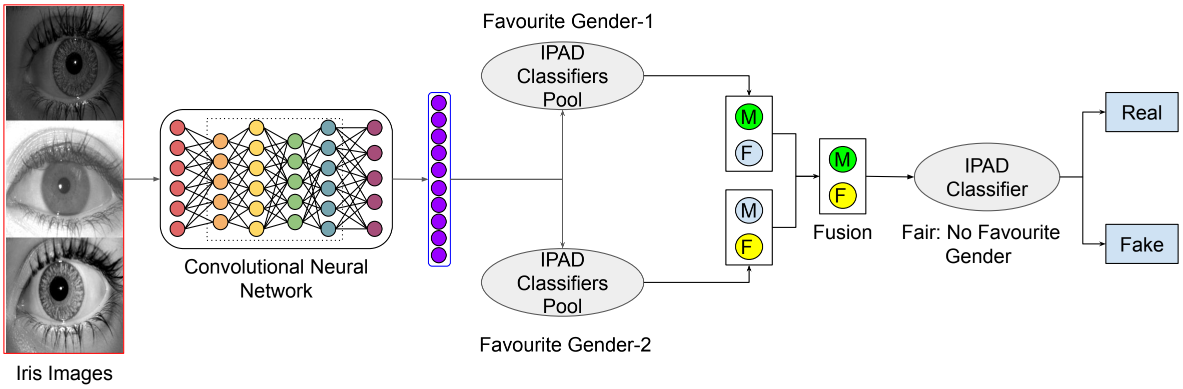
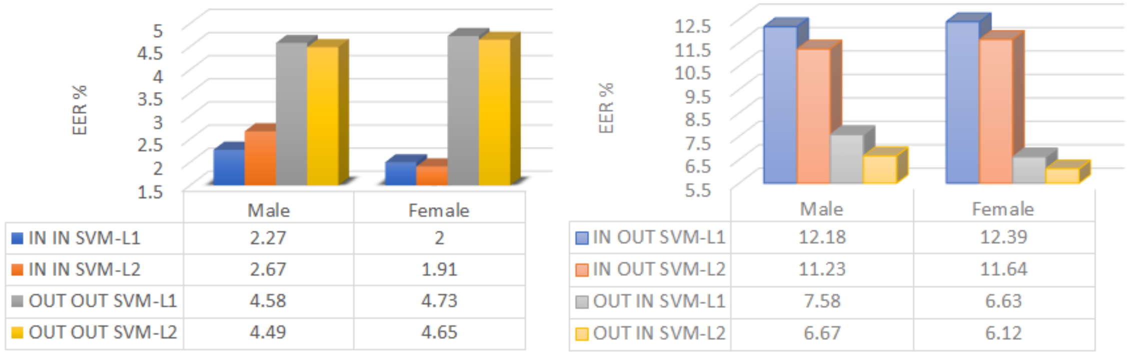


Figure 5. Robust iris presentation attack detection algorithm based on the fusion of classifiers biased (inclined) towards particular gender covariate. Among the pool of classifiers, the classifiers favoring individual gender covariate is selected and their prediction probabilities are fused together to train a classifier which has no gender bias covariate.



| | Male | Female |
|---|---|---|
| 🟦 IN IN SVM-L1 | 2.27 | 2 |
| 🟧 IN IN SVM-L2 | 2.67 | 1.91 |
| ■ OUT OUT SVM-L1 | 4.58 | 4.73 |
| 🟨 OUT OUT SVM-L2 | 4.49 | 4.65 |

| | Male | Female |
|---|---|---|
| 🟦 IN OUT SVM-L1 | 12.18 | 12.39 |
| 🟧 IN OUT SVM-L2 | 11.23 | 11.64 |
| ■ OUT IN SVM-L1 | 7.58 | 6.63 |
| 🟨 OUT IN SVM-L2 | 6.67 | 6.12 |

Figure 6. EER (%) of the proposed de-biased/fair iris presentation attack detection (IPAD) algorithm. Left and right figure shows the IPAD performance in seen and unseen environment training-testing, respectively. The level 2 SVM classifier shows the significant improvement especially when the testing environment is unseen from the training environment. This establish the robustness of the proposed solution.

performed.

## 7. Conclusion and Impact

This paper presents a detailed gender-bias study on multiple iris presentation attack detection algorithms. It is observed through the proposed study that in the case of balanced gender training, there might not be a gender-bias issue in the iris presentation attack detection algorithms. However, other factors such as image acquisition environments play a significant role in the existence of accuracy differences between genders. The proposed research also presents a robust iris presentation attack detection algorithm to bridge the gap between the difference between the accuracy of genders. An important point we would like to highlight is the need for multiple databases captured in both controlled and unconstrained environments. It is seen from this study that cross-environment training-testing creates an extra challenge in the detection and the lack of multiple databases significantly limits the progress in the field.

Impact of bias/fairness in machine learning models is a major concern across areas including biometrics and computer vision. While this research focuses on iris PAD, bias/fairness issues of other defense algorithms protecting critical security mechanism is a serious threat. Therefore, we assert that being the first large-scale study has a lot of potentials to advance bias-free future research.

## 8. Acknowledgements

M. Vatsa is partially supported through the Swarnajayanti Fellowship.

## References

[1] Amazon ditched ai recruiting tool that favored men for technical jobs, 2018. https://www.theguardian.com/technology/2018/oct/10/\amazon-hiring-ai-gender-bias-\recruiting-engine.

[2] Akshay Agarwal, Afzel Noore, Mayank Vatsa, and Richa Singh. Enhanced iris presentation attack detection via contraction-expansion cnn. *Pattern Recognition Letters*, 159:61–69, 2022.

[3] Akshay Agarwal, Afzel Noore, Mayank Vatsa, and Richa Singh. Generalized contact lens iris presentation attack detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pages 1–1, 2022.

[4] Akshay Agarwal, Akarsha Sehwag, Richa Singh, and Mayank Vatsa. Deceiving face presentation attack detection via image transforms. In *IEEE BigMM*, pages 373–382, 2019.

[5] Akshay Agarwal, Akarsha Sehwag, Mayank Vatsa, and Richa Singh. Deceiving the protector: Fooling face presentation attack detection algorithms. In *IEEE ICB*, pages 1–6, 2019.

[6] Akshay Agarwal, Richa Singh, and Mayank Vatsa. Face anti-spoofing using haralick features. In *IEEE BTAS*, pages 1–6, 2016.

[7] Sarah E Baker, Amanda Hentz, Kevin W Bowyer, and Patrick J Flynn. Degradation of iris recognition performance due to non-cosmetic prescription contact lenses. *Computer Vision and Image Understanding*, 114(9):1030–1044, 2010.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, pages 4349–4357, 2016.

[9] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019.

[10] Yifeng Chen, Cheng Wu, and Yiming Wang. Whether normalized or not? towards more robust iris recognition using dynamic programming. *I&VC*, 107:104112, 2021.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[12] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE/CVF CVPR*, pages 9268–9277, 2019.

[13] Adam Czajka and Kevin W Bowyer. Presentation attack detection for iris recognition: An assessment of the state-of-the-art. *ACM CSUR*, 51(4):1–35, 2018.

[14] John Daugman. New methods in iris recognition. *IEEE TSMC-B*, 37(5):1167–1175, 2007.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255, 2009.

[16] Chris Drumnond. Class imbalance and cost sensitivity: Why undersampling beats oversampling. In *KDDW*, 2003.

[17] Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Cross-database and cross-attack iris presentation attack detection using micro stripes analyses. *Image and Vision Computing*, 105:104057, 2021.

[18] Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Iris presentation attack detection by attention-based and deep pixel-wise binary supervision network. In *IEEE/IAPR IJCB*, pages 1–8, 2021.

[19] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Demographic bias in presentation attack detection of iris recognition systems. In *IEEE EUSIPCO*, pages 835–839, 2021.

[20] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *ECCV*, pages 330–347, 2020.

[21] Diego Gragnaniello, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. An investigation of local descriptors for biometric spoofing detection. *IEEE TIFS*, 10(4):849–863, 2015.

[22] Mehak Gupta, Vishal Singh, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Generalized iris presentation attack detection algorithm under cross-database settings. In *ICPR*, 2020.

[23] Priyanshu Gupta, Shipra Behera, Mayank Vatsa, and Richa Singh. On iris spoofing using print attack. In *IEEE ICPR*, pages 1681–1686, 2014.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016.

[25] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, pages 771–787, 2018.

[26] Tin Kam Ho. Random decision forests. In *IEEE ICDRA*, volume 1, pages 278–282, 1995.

[27] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE CVPR*, pages 4700–4708, 2017.

[28] Anil K Jain, Patrick Flynn, and Arun A Ross. *Handbook of biometrics*. Springer Science & Business Media, 2007.

[29] Vishi Jain, Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. Robust iris presentation attack detection through stochastic filter noise. In *IEEE ICPR*, pages 1–7, 2022.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Naman Kohli, Daksha Yadav, Mayank Vatsa, and Richa Singh. Revisiting iris recognition with color cosmetic contact lenses. In *IEEE ICB*, pages 1–7, 2013.

[32] Naman Kohli, Daksha Yadav, Mayank Vatsa, Richa Singh, and Afzel Noore. Detecting medley of iris spoofing attacks using desist. In *IEEE BTAS*, pages 1–6, 2016.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25:1097–1105, 2012.

[34] Puspita Majumdar, Saheb Chhabra, Richa Singh, and Mayank Vatsa. Subgroup invariant perturbation for unbiased pre-trained model prediction. *Frontiers in Big Data*, 3:52, 2020.

[35] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Data-augmentation for reducing dataset bias in person re-identification. In *IEEE AVSS*, pages 1–6, 2015.

[36] Daniel Moreira, Mateusz Trokielewicz, Adam Czajka, Kevin Bowyer, and Patrick Flynn. Performance of humans in iris recognition: The impact of iris condition and annotation-driven verification. In *IEEE WACV*, pages 941–949, 2019.

[37] Moktari Mostofa, Salman Mohamadi, Jeremy Dawson, and Nasser M Nasrabadi. Deep gan-based cross-spectral cross-resolution iris recognition. *IEEE T-BIOM*, 3(4):443–463, 2021.

[38] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. In *IEEE/CVF ICCV*, pages 1695–1704, 2019.

[39] Shruti Nagpal, Maneet Singh, Richa Singh, and Mayank Vatsa. Diversity blocks for de-biasing classification models. In *IEEE/IAPR IJCB*, pages 1–9.

[40] Shruti Nagpal, Maneet Singh, Richa Singh, and Mayank Vatsa. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019.

[41] Shruti Nagpal, Maneet Singh, Richa Singh, and Mayank Vatsa. Attribute aware filter-drop for bias invariant classification. In *IEEE/CVF CVPRW*, pages 32–33, 2020.

[42] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *FAT**, pages 469–481, 2020.

[43] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *IEEE/CVF CVPRW*, pages 0–1, 2020.

[44] Renu Sharma and Arun Ross. D-netpad: An explainable and interpretable iris presentation attack detector. In *IEEE/IAPR IJCB*, pages 1–10, 2020.

[45] Talha Ahmad Siddiqui, Samarth Bharadwaj, Tejas I Dhamecha, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Face anti-spoofing with multifeature videolet aggregation. In *IEEE ICPR*, pages 1035–1040, 2016.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[47] Richa Singh, Akshay Agarwal, Maneet Singh, Shruti Nagpal, and Mayank Vatsa. On the robustness of face recognition algorithms against attacks and bias. In *AAAI*, volume 34, pages 13583–13589, 2020.

[48] Sanchit Sinha, Mohit Agarwal, Mayank Vatsa, Richa Singh, and Saket Anand. Exploring bias in primate face detection and recognition. In *ECCVW*, pages 0–0, 2018.

[49] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *IEEE/CVF ICCV*, pages 5310–5319, 2019.

[50] Daksha Yadav, Naman Kohli, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Afzel Noore. Fusion of handcrafted and deep learning features for large-scale multiple iris presentation attack detection. In *IEEE CVPRW*, pages 572–579, 2018.

[51] Daksha Yadav, Naman Kohli, Mayank Vatsa, Richa Singh, and Afzel Noore. Detecting textured contact lens in uncontrolled environment using densepad. In *IEEE/CVF CVPRW*, pages 0–0, 2019.

[52] Shivangi Yadav and Arun Ross. Cit-gan: Cyclic image translation generative adversarial network with application in iris presentation attack detection. In *IEEE/CVF WACV*, pages 2412–2421, 2021.

[53] David Yambay, Adam Czajka, Kevin Bowyer, Mayank Vatsa, Richa Singh, Afzel Noore, Naman Kohli, Daksha Yadav, and Stephanie Schuckers. Review of iris presentation attack detection competitions. In *Handbook of biometric anti-spoofing*, pages 169–183. Springer, 2019.

[54] Kai Yang, Zihao Xu, and Jingjing Fei. Dualsanet: Dual spatial attention network for iris recognition. In *IEEE/CVF WACV*, pages 889–897, 2021.

[55] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.