Bias Analysis for Deepfake Detection: A Case Study of the Impact of Facial Attributes in Deepfake Detection

Anonymous IJCB 2025 submission

Abstract

Bias analysis in the detection of deepfakes is bound to become a critical topic in the coming years. Although many detection models have been developed and several datasets have been released to reliably identify deepfake content, one crucial aspect has been largely overlooked: these models and training datasets can be biased, leading to failures in detection for certain demographic groups and raising significant social, legal, and ethical issues. In this work, we introduce an evaluation framework to contribute to the analysis of bias of deepfake detectors with respect to several facial attributes. This framework exploits synthetic data generation, with evenly distributed attribute labels, for mitigating any skew in the data that could otherwise influence the outcomes of bias analysis. We build on the proposed framework to provide an extensive case study of the bias level of five state-of-the-art detectors in synthetic datasets with 25 controlled facial attributes. While the results confirm that, in general, deepfake detectors are biased towards the presence/absence of specific facial attributes, our study also sheds light on the origins of the observed bias through the analysis of the correlations with the balancing of facial attributes in the training sets of the detectors, and the analysis of detectors activation maps in image pairs with controlled attribute modifications. The framework is available at `github.com/available-after-acceptance`.

1. Introduction

Advancements in generative artificial intelligence, particularly through Generative Adversarial Networks (GANs) [33], have greatly improved the ability to synthesize and manipulate human faces. Techniques such as face swapping, facial reconstruction, and attribute editing now generate highly realistic synthetic content, commonly referred to as deepfakes. While these innovations offer valuable applications in media, they also pose significant challenges, particularly concerning security and misinformation. Deepfake detection systems have been proposed to address these challenges.

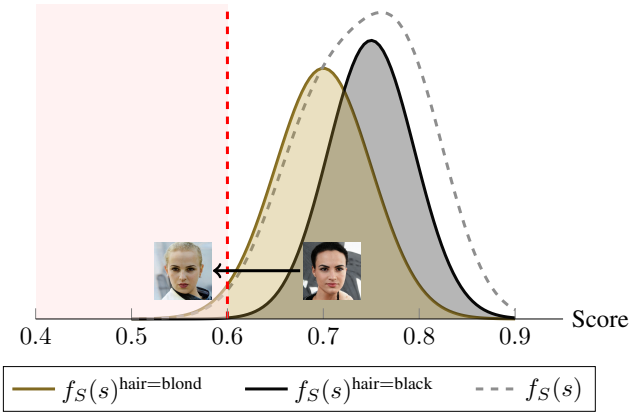

Figure 1. **Bias in deepfake detection**. The plot shows the probability density function $f_S(s)$ of deepfake detector scores for synthetic face images, conditioned on hair color attributes (blond vs. black). The distributions differ significantly by attribute, with many samples of blond-haired faces falling below the detection threshold (red dashed line). This results in a higher likelihood of synthetic blond-haired faces being misclassified as real, highlighting a critical bias issue in deepfake detection systems.

Motivations. Despite advancements in detection performance, the fairness and reliability of these systems across diverse demographic groups remain largely unexplored. This contrasts with the extensive studies on bias in facial recognition [5], and given the similarities between traditional face analysis systems and deepfake detectors, deepfake detectors are likely to show similar inconsistencies in accuracy across demographics, undermining public trust (Fig. 1). Also, the scarcity of balanced datasets with adequate demographic diversity and detailed attribute annotations remains a key issue. While the SDFD dataset [2] addresses this gap, its limited size of 1,000 images restricts its applicability.

Research questions. To address the identified gap, this work is guided by the following research questions: **RQ1.** How do the True Positive Rates (TPR) of deepfake detectors vary when analyzing images with specific facial attributes compared to those without? **RQ2.** What are the primary sources of bias in deepfake detection systems? Do these

biases originate from the synthetic data generator used for evaluation, the design and training of the detectors, or the composition of their training datasets? **RQ3.** To what extent do the biases identified in deepfake detection systems reflect those observed in pristine (non-manipulated) data?

Methodology. To systematically explore these questions, we propose a structured strategy comprising synthetic data generation, bias quantification, statistical validation, and bias source analysis. First, we develop a reproducible methodology for generating synthetic data conditioned on specific facial attributes. A novel metric is introduced to quantify bias by comparing TPRs between groups that differ in only one facial attribute, and statistical validation is carried out using paired t-tests as a bias detection approach. Bias source analysis is performed by visualizing image regions influencing classifier decisions and through the assessment of the correlation between observed biases and the distribution of facial attributes in training datasets.

In summary, our contributions are as follows:

- Identification of attribute-specific biases. To the best of our knowledge, this is the first work to investigate the biases of deepfake detection systems not only in relation to individual facial attributes but also arising from their interactions with other attribute categories.
- Framework for bias analysis. We introduce a comprehensive framework for analyzing bias in deepfake detection systems, integrating synthetic data generation, bias quantification metrics, statistical validation, and correlation analysis to provide a robust and systematic evaluation.
- Synthetic dataset. We present a balanced, labeled synthetic dataset designed specifically for certain facial attribute.
- Analysis of bias sources. Through systematic evaluation, we identify the primary sources of bias in deepfake detection systems, including synthetic data generators, training dataset composition, and model design choices.

The dataset and framework are publicly available to support reproducibility and foster further research.

2. Related Work

Deepfake Detection Models. Deepfake detection has become a key area of focus with the rise of synthetic content creation [3]. Detection methods are typically categorized into spatial, temporal, and frequency-based approaches [27]. Spatial-based methods [6, 1, 18, 31] work by detecting visual inconsistencies and pixel-level distortions caused during the creation of deepfakes. Temporal-based methods, primarily applied in videos, detect inconsistencies in motion and coherence across frames. Frequency-based approaches analyze content in the frequency domain, identifying hidden artifacts not visible in the spatial domain. SCnet [8] and Capsule Networks [26] are prominent examples that capture frequency-based features, leveraging device fingerprints and compression artifacts. More recently, transformer-based models such as CORE [19], DFDT [13] and UIA-ViT [37] combine spatial and frequency features for improved deepfake detection.

Dataset Name	Real/Fake	Unique Combinations	Number of Attributes
FF++	Real/Fake	22529	42
CelebDF	Real/Fake	11740	30
DFDC	Real/Fake	8938	43
DFD	Real/Fake	9944	27
Ours	Fake	46656	29

Table 1. Comparison of number of unique combinations.

Deepfake Detection Datasets. Existing deepfake detection datasets, such as FF++ [24], CelebDF [15], often exhibit significant shortcomings, including imbalance, limited demographic diversity, incomplete coverage of deepfake generation techniques, and insufficient attribute annotations [11] (Table 1). In this work, to address the urgent need for controlled attributes and a more systematic evaluation of model performance across different groups and conditions, we propose generating datasets that, once specific attributes are defined, systematically include all possible combinations and interactions. This approach ensures a fully balanced dataset, allowing for a more comprehensive and unbiased evaluation of detection models across diverse scenarios.

Exploring Bias in the Detection of Deepfakes. Bias analysis in deepfake detection remains relatively unexplored. Most previous work has focused on evaluating detection algorithms based solely on the presence or absence of a single attribute, usually demographic factors such as gender or race [11]. However, the biases that can emerge from machine learning-based solutions are often more complex and involve interactions among multiple attributes that, when considered together, reveal deeper and more subtle biases. This hypothesis has been largely confirmed by studies in the literature in various other fields [16]. Some previous works evaluated bias in specific attributes such as race and gender [32, 10, 17], and skin tone [29, 35, 20, 30, 21]. Trinh and Liu [32] evaluated multiple deepfake detectors [1, 4, 14] on datasets balanced for race and gender, discovering performance differences across races [25]. Pu et al. [22] used a subset of the Face2Face dataset from FF++ concluding about the existence of bias in gender. Hazirbas et al. [10] assessed five deepfake detection models across different facial attributes and found that all methods favored lighter skin tones and underperformed on darker skin tones [29, 35, 20, 30, 21]. The study in [17] identified considerable bias in both datasets and detection models

and attempted to address gender bias by creating a balanced dataset using deepfake detection models. While this approach led to some improvement, it required extensive and time-consuming data annotation, and did not address other sensitive attributes beyond gender. Initial studies suggested that biases related to non-demographic attributes might originate from class imbalances in datasets. To address this, some evaluation methods implemented control measures by fixing a specific attribute and analyzing model performance on an equal number of random samples that excluded it, assuming the absence of the attribute was entirely independent of other factors [34]. However, this assumption oversimplifies the problem and can lead to biased conclusions about model performance and associated biases, as it fails to account for complex interactions between attributes and the influence of other contributing factors. To overcome these limitations, our framework introduces a novel bias risk evaluation metric that enables the simultaneous assessment of model performance across multiple attributes, or selected subsets, by analyzing them in combination. This approach provides a more comprehensive evaluation of detection model biases, effectively capturing the interactions between attributes and offering deeper insights into the factors influencing performance.

3. Methodology

3.1. Data Generation

To conduct a thorough analysis of the impact that different facial attributes have on the performance of deepfake detectors, we propose generating a synthetic dataset that systematically includes all possible combinations of attribute labels. This approach not only enables a comprehensive evaluation of detection accuracy but also allows for precise control over individual attributes, facilitating the isolation of their specific effects on the model performance. Let $G = \{g_1, g_2, \ldots, g_n\}$ be a set of attribute groups, where $|L_i|$ represents the number of possible attributes for the group $g_i \in G$. We generate a set of k synthetic images for each possible label combination across the n groups. The total number of unique combinations for G is thus given by $|\Omega(G)| = \prod_{g_i \in G} |L_i|$, ensuring full coverage of the attribute space. Figure 2 illustrates some examples of the synthetic images generated.

3.2. Bias Risk

The classical approach to bias assessment typically measures how model predictions differ across demographic groups, which fails to consider the influence of other relevant attributes that may contribute to bias. We propose an alternative measure, where bias comparisons are conducted within subgroups where both demographic and non-demographic attributes are alternately fixed, allowing only

Figure 2. **Generated synthetic data**. Samples from the synthetic datasets created by two generators (top row: StyleGAN, bottom: StableDiffusion) representing a diverse range of facial attributes.

the attribute under analysis to vary. This method provides a more comprehensive understanding of bias by accounting for the interactions between multiple attributes. Building on this concept and utilizing the synthetic data we generated, which allows for precise control and clear understanding of individual attributes, we propose a novel metric, termed bias risk ($brisk$). This metric quantifies bias by calculating the expected variation in True Positive Rates (TPR) across subgroups. These subgroups are defined by fixing the attribute under investigation and systematically varying all other attributes. Formally, let $A = \{a_1, a_2, \ldots, a_n\}$ represent a set of attributes, where a_i is the attribute under analysis, and let $A_{-i} = A \setminus \{a_i\}$ denote the set of all other attributes except a_i. Given a classification model S, let $f_S^{(a_i=x)}$ represent the probability density function of the scores produced by S for the subgroup where the attribute a_i takes the value x (with $x = 1$ indicating the presence and $x = 0$ indicating the absence of a_i). The TPR for each subgroup of A_{-i} is defined as:

$$\text{TPR}(a_i = x, A_{-i}, t) = \int_t^\infty f_S^{(a_i=x)}(s \mid A_{-i}, \text{TP})\, ds, \quad (1)$$

where t represents the threshold score at which the classifier operates and $f_S^{(a_i=x)}(s \mid A_{-i}, \text{TP})$ represents the probability density function of the classifier's scores s, conditioned on the instance being a True Positive (TP) for the fixed attribute a_i, and further conditioned on the remaining attributes A_{-i} . To quantify bias within each subgroup, we propose comparing the TPRs for groups with and without the attribute a_i:

$$\begin{aligned}\Delta_{\text{TPR}}(a_i, A_{-i}, t) = \text{TPR}(a_i = 1, A_{-i}, t) \\ - \text{TPR}(a_i = 0, A_{-i}, t).\end{aligned} \quad (2)$$

To summarize this comparison across all possible subgroups defined by A_{-i}, we calculate the average TPR difference across all combinations of A_{-i}, denoted by $\Omega(A_{-i})$:

$$\Delta_{\text{TPR}}(a_i, t) = \frac{1}{|\Omega(A_{-i})|} \sum_{A_{-i} \in \Omega(A_{-i})} \Delta_{\text{TPR}}(a_i, A_{-i}, t). \quad (3)$$

Here, $|\Omega(A_{-i})|$ is the number of possible subgroups formed by the different combinations of attributes in A_{-i}.

Attribute Group	Attribute
Attractiveness	Attractive, Not Attractive
Gender	Man, Woman
Age	Child, Young, Old
Hair Color	Black Hair, Blonde Hair, Brown Hair, Gray Hair
Hair Type	Straight Hair, Wavy Hair, Bald
Skin Tone	Black Skin, White Skin
Eye Color	Black Eyes, Blue Eyes, Green Eyes
Nose Shape	Pointy Nose, Big Nose
Face Shape	Oval Face, Round Face, Square Face
Mustache	Mustache, No Mustache
Beard	Beard, No Beard
Makeup Type	No Makeup, Makeup, Heavy Makeup

Table 2. Facial attributes considered in this study organized with respect to the attribute groups.

The function $\Delta_{\text{TPR}}(a_i, t)$ estimates bias as a function of the operational threshold t, reflecting how the TPR difference between groups changes with different decision boundaries. To provide a comprehensive measure of bias across all possible thresholds, we propose calculating the expected value of $\Delta_{\text{TPR}}(a_i, t)$ over the entire range of thresholds:

$$\text{brisk}(a_i) = \int_0^1 \Delta_{\text{TPR}}(a_i, t)\, dt. \tag{4}$$

This bias risk metric condenses the bias associated with a specific attribute into a single value, capturing variations in performance across different thresholds and subgroups, thereby providing a more robust estimate of the model fairness. Additionally, we introduce the worst-case bias risk, $\text{brisk}^\star(a_i)$, to assess the most extreme bias that might occur between groups:

$$\text{brisk}^\star(a_i) = max(\Delta_{\text{TPR}}(a_i, t)). \tag{5}$$

This alternative metric provides a useful insight for applications where even rare occurrences of extreme bias could have substantial consequences.

3.3. Framework Evaluation Tools

Chart of $brisk$ Values. A chart displaying the $brisk$ values enables direct comparison across different attributes, facilitating the identification of significant biases among various detectors. This visualization allows for a quick assessment of which attributes exhibit pronounced differences in bias, providing critical insights into model performance.

Detector Activation Map. An interpretable visualization tool designed to identify the most relevant image regions in the model's predictions, thereby providing insights into the model's decision-making process. The heatmaps generated by a saliency-based visualization method illustrate the importance of different regions in relation to the classifier's score. In the visualizations, each image is accompanied by a score that reflects the model's confidence in its prediction. These scores range from 0 to 1, with higher values indicating increased confidence in the classification. The images are organized in a grid format, with each row representing a specific set of attributes where only one attribute is modified while keeping all other attributes constant in the subsequent row. This arrangement allows for the observation of how the model's focus changes with alterations in a single attribute, offering valuable insights into the consistency of the model's predictions. This methodology enhances our understanding of how different regions of the image impact the model's decisions, facilitating the identification of potential biases and providing direction for further analysis and improvements in model design.

Paired t-Test for Bias Detection. While the $brisk$ metric quantifies the expected difference in TPR when a specific attribute is present, it does not provide a definitive criterion for when a detector should be considered biased towards that attribute. To rigorously determine the presence of bias, we introduce a statistical hypothesis testing approach, specifically employing a paired t-test. Our synthetic dataset is structured to include k samples for each possible combination of attributes. This balanced design ensures that comparisons between groups are not confounded by unequal sample sizes or attribute distributions. By calculating the average difference in TPR over all thresholds using our metric, we obtain an overall measure of disparity that is straightforward to interpret. The primary objective of our statistical analysis is to test whether this mean difference in TPR between the two groups is statistically significant, which would indicate potential bias. This approach not only quantifies the mean performance difference between groups but also statistically validates whether the observed differences are significant. Formally, we determine the t-test statistic for determining the presence of bias in the attribute a_i by testing whether $\int_0^1 \Delta_{\text{TPR}}(a_i, A_{-i}, t)\, dt$ has zero mean.

Correlation Analysis. This part of our methodology involves two types of correlation analyses. First, it regards the assessment of the relationship between the bias values of a single detector or across multiple detectors and the proportions of samples for each attribute within the training dataset. This provides insights how the distribution of attributes in the dataset may influence the biases observed in the detectors. Second, it evaluates the inter-detector bias correlation, which conveys information about the source of bias, particularly if the bias may be caused by the architectural choices of deepfake detectors.

4. Experiments

4.1. Case Study Setup: Impact of Facial Attributes

This section details the specific setup for a case study, conducted to demonstrate the application of the framework we have proposed. The study is designed to explore how different facial attributes affect the detection capabilities of

	Xcep.		UIA		CNN LSTM		NPR		Caps. Net	
	S	D	S	D	S	D	S	D	S	D
Attractive	•	•	•		•		•	•	•	
Man	•	•	•		•	•			•	
Child			•					•		
Young	•	•			•					
Old	•	•	•		•			•		
Black hair	•	•	•		•	•			•	
Blonde hair	•	•	•			•		•	•	
Brown hair	•	•			•	•		•		
Gray hair	•	•	•						•	
Straight hair	•		•		•	•		•	•	
Wavy hair	•	•	•		•	•	•	•	•	
Bald	•	•	•		•	•	•	•	•	
White skin	•	•	•		•	•	•	•	•	
Black eyes	•	•	•						•	
Blue eyes	•	•	•					•	•	
Green eyes		•	•					•		
Big nose	•	•						•	•	
Oval face	•		•						•	
Round face	•		•				•	•		
Square face			•					•	•	
Mustach	•		•			•		•	•	
Beard	•	•	•		•	•	•	•		•
No makeup	•	•	•		•	•	•	•	•	
Makeup	•				•	•			•	
Heavy makeup	•	•	•		•	•		•	•	

Table 3. Bias detection results for each attribute across detectors evaluated in synthetic data from StyleGAN (S) and Diffusion generators (D) with (•) indicating significant bias (p-value < 0.01).

	Xception	**UIA**	**CNN-LSTM**	**CapsuleNet**
StyleGAN	0.03	-0.16	0.63	-0.09
Diffusion	-0.93	-0.81	0.64	0.34

Table 4. Correlation between the bias of detectors trained on FF++ and the proportion of samples of each attribute in this set.

deepfake detection models. It outlines the construction of the synthetic dataset, the selection of deepfake generators, the deployment of deepfake detectors, and the metrics used to assess bias.

Dataset. For this study, we utilize two synthetic datasets specifically constructed to enable a comprehensive analysis of biases in deepfake detection models. In particular, we construct two synthetic datasets by organizing 25 facial attributes into 12 distinct groups, as detailed in Table 2. For each dataset, all possible combinations of attributes — one selected per group — are systematically generated to ensure comprehensive coverage. These combinations condition a synthetic face generator, with the process repeated $k = 4$ times to produce multiple samples for each combination. This exhaustive approach allows for a controlled and thorough analysis of attribute-driven biases, capturing the full spectrum of attribute interactions.

Deepfake Generators. Two state-of-the-art generators are considered for the creation of the synthetic datasets: a) StyleGAN [7] pre-trained on the Flickr-Faces-HQ (FFHQ) [12] dataset; and b) Stable Diffusion v1.5 [23] pre-trained on the LAION-5B dataset [28]. These methods were selected for representing the major families of image generation strategies (GAN-based and Diffusion Models).

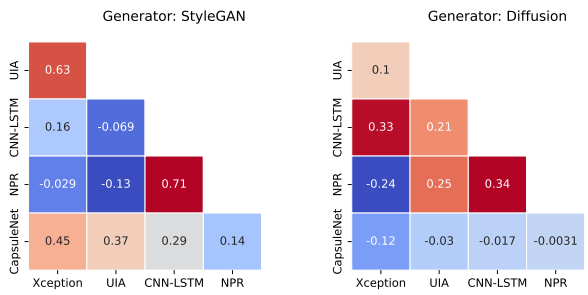

Figure 3. Correlation between the bias values of the different deepfake detectors.

Deepfake Detectors. In our experiments, we used five state-of-the-art deepfake detection models, namely XceptionNet, CapsuleNet-V2 [18], LSTM+ResNext model [9], UIA-VIT [37], and NPR [31]. All models were trained on the FF++ dataset [24], except for NPR that was trained on the GenImage dataset [36]. UIA a Vision Transformer-based model, while the others are CNN-based models. To assure that the detectors are not biased towards real or fake class, we assessed their accuracy in a balanced set (see supplementary material).

Metrics. In our experiments, bias is assessed by comparing the differences in model predictions across various groups while controlling for other attributes using the metric denoted as $brisk$. To complement this analysis, we also utilize the Equal Opportunity Difference (EOD) as an approximation of $brisk$ in scenarios where all subgroups are evenly represented within the dataset. EOD serves as a practical alternative, particularly for existing fake datasets where the complete range of attribute combinations is not available, enabling bias analysis even in the presence of incomplete subgroup representations. Additionally, the methodologies and metrics outlined here utilize the visualization and analysis tools defined in Section 3.3. By integrating these tools, our approach not only quantifies but also visually represents the biases, enabling a clearer understanding of where and how these biases manifest within our models.

4.2. Results and Discussion

Bias Assessment. Our framework was adopted for the assessing the bias level in the five state-of-the-art deepfake detectors considered in this study. The results of the brisk* metric obtained for each detector along 25 facial attributes are depicted in Figure 4. The analysis of the results evidence that some methods are severely affected by bias, as several attributes exceed 5% in the absolute value of brisk*. Apart from the absolute bias, it is also worth analysing the sign of the bias, which conveys if the presence of a facial attribute increases (positive bias) or decrease (negative bias)

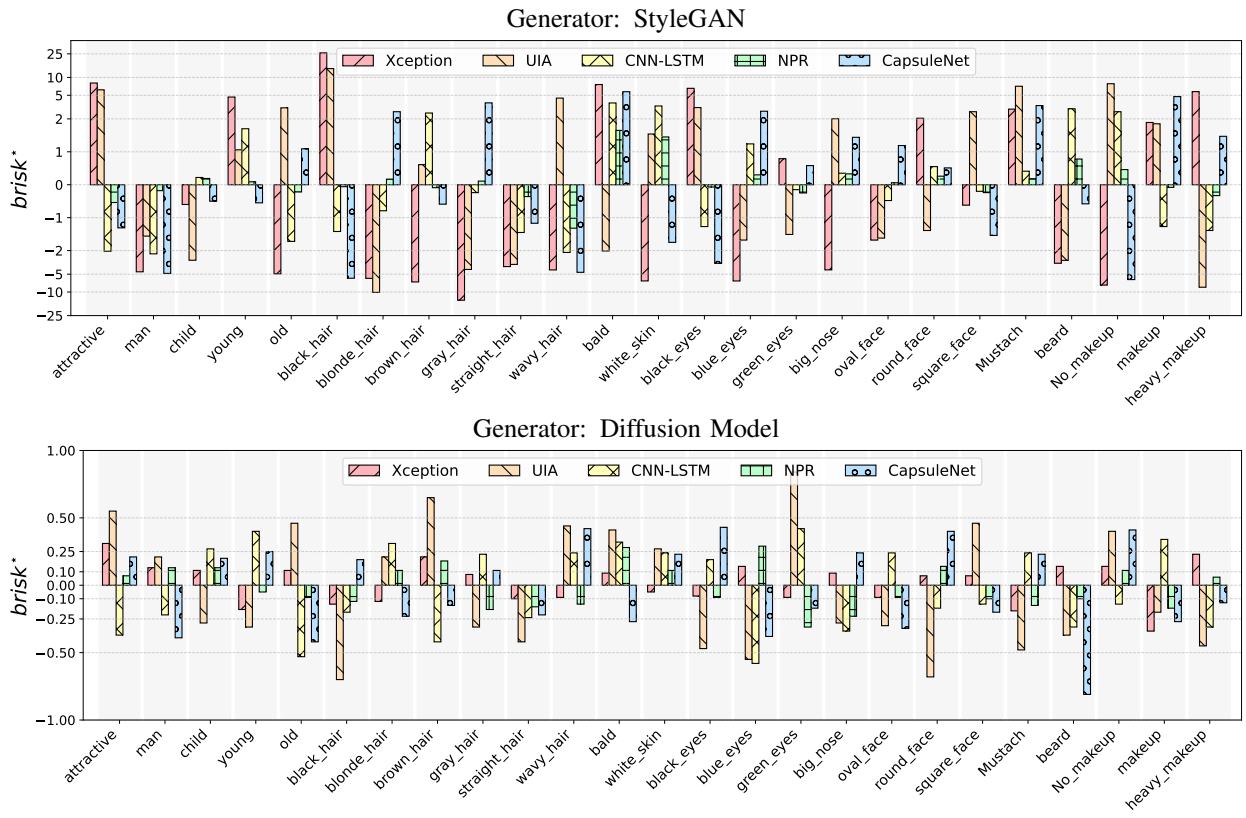

Figure 4. **Bias level of deepfake detectors along different facial attributes in synthetic data.** The bias level was determined using brisk* metric independently for each attribute. Positive values mean that the detector is biased towards the presence of the attribute, implying that samples with this attribute are more likely to be correctly classified as synthetic.

the probability of the detection method to correctly classify the image as synthetic. As an example, the attribute *man* seems to consistently decrease the TPR of the detection method compared to images where this attribute is absent, which in this case corresponds to images of women. One possible reason for this is the unbalanced representation of these two groups in the dataset of the detection methods (42% man vs 58% woman in the FF++ dataset). This topic will be further analysed when testing several hypothesis about the origin of bias. Apart from the analysis of the bias level, we report in Table 3 the bias detection results carried out using the paired t-test on the TPR difference between subgroups. The results evidence that even using a highly statistically significant level (p-value < 0.01) all detection approaches have bias in a vast amount of facial attributes, evidencing that bias in deepfake detectors is still an open problem.

Determining the Origin of the Bias. Apart from the bias level observed in Figure 4, it is worth noting that only a few attributes exhibit follow a consistent pattern between the value/direction of the biases observed along the different detectors. To provide additional evidence on this, we report the correlation between the bias of the different detectors in Figure 3. In general, the bias of the detectors are weakly or not correlated, suggesting that these bias do not originate on the generators, as each method presents different patterns of bias in the same synthetic data. Considering that most methods were trained in the same dataset (FF++), we inspected the possible relation between the bias of detectors trained on FF++ and the proportion of each attribute in this set. To determine the proportion of samples where each attribute is present, we relied on the annotations provided in [34], and determined the correlation with the bias values of each attribute. These results are provided in Table 4, where no strong positive correlation has been found, suggesting that the unbalanced distribution of the attributes in the training set of the detectors can not be attributed as the source of bias of the methods. Based on this conclusion, we attempt to verify whether the learning strategy can be related to the bias level of each detector. To verify this hypothesis, we inspect the activation maps of face images where only one attribute has been changed. Figure 5 depicts the heatmaps of the importance of different regions in the image to the classifier score. It is worth noting that the image of the second, fourth and sixth rows have been generated by only inverting the attribute *man*. Additional results on other attributes

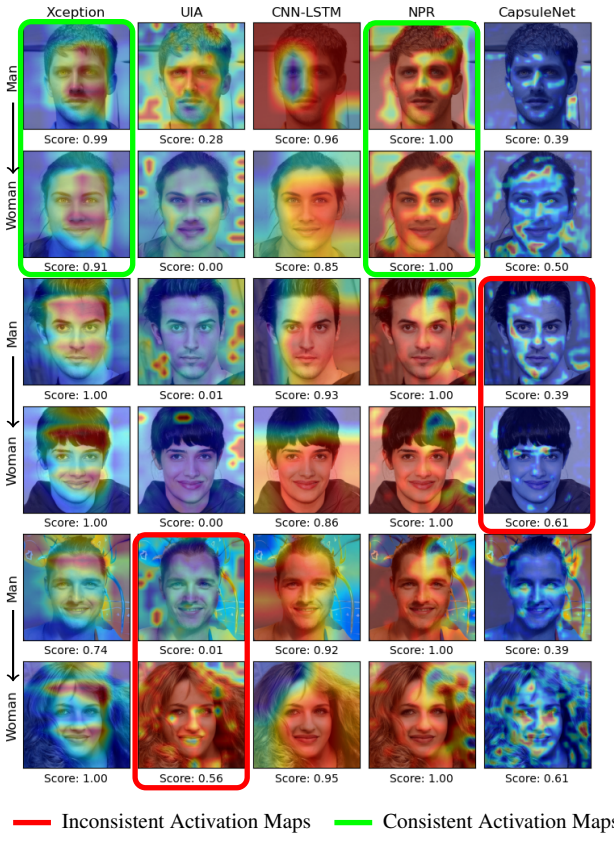

Figure 5. **Deepfake detectors activation maps.** The activation maps of deepfake detectors were inferred for pairs of images where only one attribute was changed. The comparison between the activation maps of a pair of images and the respective detection score evidences that in some methods a change in an attribute impacts the regions analysed for determining the classification score, justifying the observed bias with respect to a specific attribute.

are provided in supplementary material, showcasing the diverse influence of individual facial features on the classifier's decision-making process. This experiment allows to perceive the consistencies of the detectors to a change in a single attribute, providing insights not only about the origin of the bias but also why some methods suffer more from bias than others. As an example, NPR has the most consistent activation maps when compared with other detectors, and consequently smaller differences between the scores of the groups, and, in turn, less bias. In short, NPR seems to be agnostic to changes in the attribute *man*, while UIA and CapsuleNet present significant differences in the importance regions of the images of the two groups. The results show that architecture and learning strategies may be responsible for the observed bias and not dataset distribution. Some methods, e.g., UIA and CapsuleNet seem to be more sensitive to attribute changes, shifting their focus to background when the attribute *man* is modified. This highlights the need for bias-aware learning strategies to enforce detectors to learn invariant features and reduce bias.

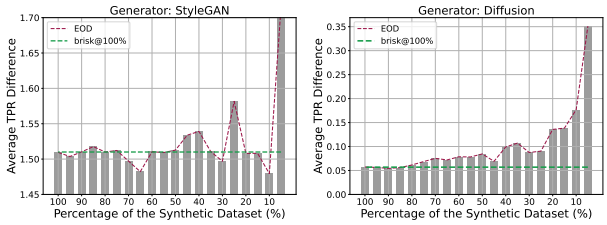

Figure 6. **Impact of the representativeness of the subgroups in the estimation of bias.** The average of the EOD metric over the 25 facial attributes was determined using different percentages of our synthtetic dataset. It can be observed that the bias estimation starts diverging from EOD and $brisk$ metrics obtained for a dataset with a balanced number of samples with respect to the different attribute combinations.

Gen.	Detector	Attribute	*p-value*	
			Classical	Ours
D	NPR	child	0.885	0.005
S	CapsNet	heavy makeup	0.720	0.000
D	LSTM	brown hair	0.642	0.002
D	LSTM	heavy makeup	0.220	0.000
S	UIA	man	0.173	0.001
D	NPR	old	0.167	0.002
S	CapsNet	square face	0.152	0.000
S	Xception	oval face	0.139	0.007
D	NPR	mustache	0.120	0.005

Table 5. Comparison between the strategies for deriving a p-value for testing the hypothesis of the TPRs of both $a_i = 0$ $a_i = 1$ groups not having statistically significant differences.

Importance of Synthetic Data. To assess the impact of the dataset size in bias estimation, we relied on a simplification of the $brisk$ metric, the EOD metric. EOD measures the TPR difference ($\Delta_{\text{TPR}}(a_i, t)$), but without considering subgroups defined by A_{-i}. Figure 6 reports the EOD over the different attributes and detectors when using different sample sizes, as well as the $brisk$ metric obtained from the complete synthetic dataset. The results show that the bias estimation is significantly affected when sampling only a small margin of the original synthetic data. We argue that this is caused by the lack of some attribute combinations that impair the accurate estimation of bias. To provide additional evidence on the importance of bias estimation over subgroups when compared to the general strategy of comparing solely the distributions $f_S^{(a_i=1)}(s)$ and $f_S^{(a_i=1)}(s)$, we compared the results of a t-test carried out between the $a_i = 1$ and $a_i = 0$ groups, and when measuring the differences inside each subgroup sharing the same facial attributes. Table 5 reports the top-8 attributes/detectors regarding the difference between the p-values of two approaches, and the comparison between the p-value magnitudes clearly evidences that in all these cases, carrying out a statistical test on the difference between the average TPR of

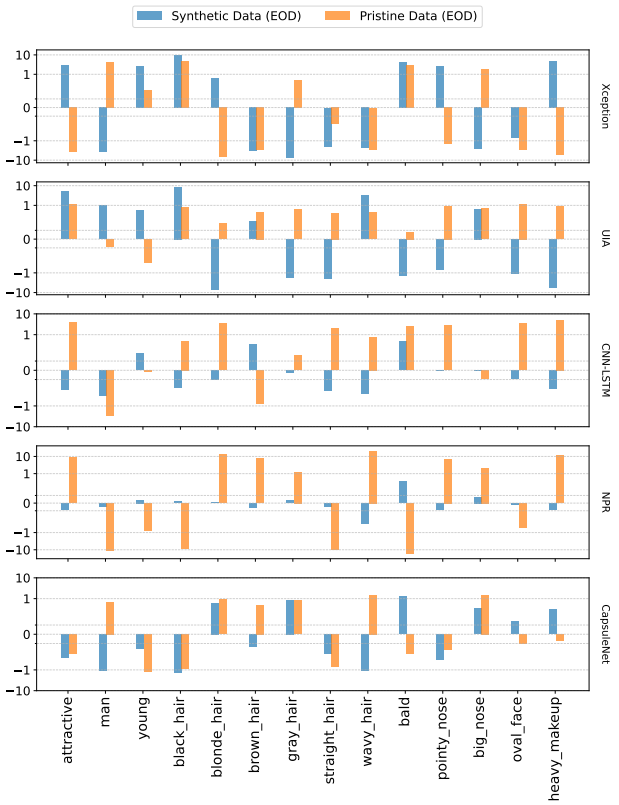

Figure 7. Bias levels of deepfake detectors in synthetic (blue) and pristine (orange) data, measured across shared facial attributes using the EOD metric.

both groups fails to identify bias. In contrast, when considering the average TPR in each subgroup, a statistical significant difference is observed between the two groups, justifying the need for adopting the proposed evaluation methodology for bias estimation.

Bias Analysis in Pristine Data. While the proposed evaluation framework thoroughly analyzes the bias of deepfake detectors in synthetic data, it does not directly evaluate bias in pristine data due to the impracticality of collecting all possible combinations of attributes. However, to explore the relationship between the biases observed in synthetic and pristine data, we relied on the EOD metric. For this comparison, we selected the CelebA dataset, as it shares several facial attributes with those considered in our study. Figure 7 visually compares the bias levels of different detectors in various facial attributes using synthetic and pristine data. The results demonstrate that biases observed in synthetic data are, in general, also present in pristine data. This discrepancy can be attributed to the varying bias patterns exhibited by different detectors on pristine datasets. Similarly to synthetic data, we observed no strong correlation between a detector's bias level and the proportion of samples with a specific attribute. We hypothesize that this behavior arises from the disparate way classifiers handle attribute information across different prediction score ranges. Specifically, the influence of a facial attribute on bias appears to depend on whether the classifier outputs low scores (pristine data) or high scores (synthetic data). This highlights the distinct dynamics of bias in real-world and synthetic contexts and underscores the value of our synthetic framework for controlled, systematic bias analysis.

Limitations. The proposed bias metric for deepfake detection, while useful for identifying disparities, should not be considered in isolation. For instance, a model that classifies all inputs as deepfakes would appear unbiased according to this metric, despite poor detection accuracy. Therefore, it is crucial to consider this bias measure in conjunction with standard performance metrics like accuracy for a complete understanding of the model.

5. Conclusion

This study highlights the importance of a systematic framework for uncovering and addressing biases in deepfake detection systems. Our primary contribution lies in the development of an evaluation strategy that not only identifies the facial attributes most influencing detector decisions but also provides insights into the sources of bias, whether stemming from the detector itself or the training data. While we contribute a synthetic dataset designed for bias analysis, the broader goal of our work is to establish a robust and reproducible methodology for identifying and understanding biases in these systems, particularly those resulting from interactions between facial attributes. This evaluation strategy is dependent on the existence of a dataset encompassing all attribute combinations, which can be obtained using generative methods. Importantly, only the use of a complete set of attributes ensures an accurate estimation of bias. This is evidenced in Figure 6, where the EOD metric diverges from our proposed metric as the dataset size decreases, highlighting the critical role of comprehensive attribute representation. Through a case study of five state-of-the-art detectors across 25 facial attributes, we observed significant biases in detection accuracy, with TPR differences reaching up to five percentage points in some cases. To identify the sources of these biases, we tested several hypotheses. Our results showed a weak correlation between biases across different detectors, indicating that the synthetic data generator is not the only cause of the biases. Furthermore, the lack of a positive correlation between bias levels and facial attribute distributions in the detectors' training datasets suggests that model architectures and learning strategies play a critical role in introducing biases. This conclusion was further supported by the analysis of activation maps, which highlighted variations in score stability and bias levels across detectors. These results highlight the ongoing challenges posed by biases in deepfake detection systems and underscore the need for refined methods to mitigate them.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

[2] G. Baltsou, I. Sarridis, C. Koutlis, and S. Papadopoulos. Sdfd: Building a versatile synthetic face image dataset with diverse attributes. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10, 2024.

[3] F. Becattini, C. Bisogni, V. Loia, C. Pero, and F. Hao. Head pose estimation patterns as deepfake detectors. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[4] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[5] M. Georgopoulos, Y. Panagakis, and M. Pantic. Investigating bias in deep face analysis: The kanface dataset and empirical study. *Image and Vision Computing*, 102:103954, 2020.

[6] L. Y. Gong and X. J. Li. A contemporary survey on deepfake detection: datasets, algorithms, and challenges. *Electronics*, 13(3):585, 2024.

[7] M. Guo. Face-attribute-editing-stylegan3. `https://github.com/MingtaoGuo/Face-Attribute-Editing-StyleGAN3`.

[8] Z. Guo, L. Hu, M. Xia, and G. Yang. Blind detection of glow-based facial forgery. *Multimedia Tools and Applications*, 80(5):7687–7710, 2021.

[9] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.

[10] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):324–332, 2021.

[11] Y. Ju, S. Hu, S. Jia, G. H. Chen, and S. Lyu. Improving fairness in deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4655–4665, 2024.

[12] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[13] A. Khormali and J.-S. Yuan. Dfdt: an end-to-end deepfake detection framework using vision transformer. *Applied Sciences*, 12(6):2953, 2022.

[14] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020.

[15] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020.

[16] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[17] A. V. Nadimpalli and A. Rattani. Gbdf: gender balanced deepfake dataset towards fair deepfake detection. In *International Conference on Pattern Recognition*, pages 320–337. Springer, 2022.

[18] H. H. Nguyen, J. Yamagishi, and I. Echizen. Use of a capsule network to detect fake images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[19] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12–21, 2022.

[20] NTech-Lab. Ntech-lab/deepfake-detection-challenge, 2020.

[21] L. Pan and J. Howard. Jphdotam/dfdc, 2020.

[22] M. Pu, M. Y. Kuan, N. T. Lim, C. Y. Chong, and M. K. Lim. Fairness evaluation in deepfake detection models using metamorphic testing. In *Proceedings of the 7th International Workshop on Metamorphic Testing*, pages 7–14, 2022.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[24] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

[25] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Nision*, pages 1–11, 2019.

[26] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. *Advances in Neural Information Pprocessing Systems*, 30, 2017.

[27] N. Sandotra and B. Arora. A comprehensive evaluation of feature-based ai techniques for deepfake detection. *Neural Computing and Applications*, 36(8):3859–3887, 2024.

[28] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[29] S. Seferbekov. Selimsef/dfdc_deepfake_challenge, 2020.

[30] Siyu-C. Siyu-c/robustforensics, 2020.

[31] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024.

[32] L. Trinh and Y. Liu. An examination of fairness of ai models for deepfake detection. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 567–574, 2021.

[33] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu. Gan-generated faces detection: A survey and new perspectives. *ECAI 2023*, pages 2533–2542, 2023.

[34] Y. Xu, P. Terhöst, M. Pedersen, and K. Raja. Analyzing fairness in deepfake detection with massively annotated databases. *IEEE Transactions on Technology and Society*, 2024.

[35] H. Zhou, W. Zhao, and H. Cui. Cuihaoleo/kaggle-dfdc, 2020.

[36] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, and N. Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European Conference on Computer Vision (ECCV)*, 2022.