From Creation to Detection: How Dataset Composition and Simple Augmentation Influence Deepfake Training

Lauren Matthews
Department of Computer & Information Sciences
Florida A&M University
Tallahassee, FL, USA
0009-0002-6203-2123

Dr. Idongesit Mkpong-Ruffin
Department of Computer & Information Sciences
Florida A&M University
Tallahassee, FL, USA
idongesit.ruffin@famu.edu
0000-0002-0998-3598

Dr. Deidre Evans
Department of Computer & Information Sciences
Florida A&M University
Tallahassee, FL, USA
deidre.evans@famu.edu
0009-0005-9648-2618

Dr. Chutima Boonthum-Denecke
Department of Computer Science
Hampton University
Hampton, VA, USA
Chutima.boonthum@gmail.com
0000-0003-0247-7518

Abstract—Deepfakes are sophisticated, AI-generated alterations of images and videos that pose significant threats to cybersecurity, particularly with face-swapping techniques that can deceive and spread misinformation. A major limitation in current deepfake detection strategies is that the deepfakes used to train these models are often of lower quality than those encountered in real-world scenarios [1]. This weakens model performance when tested against more sophisticated media alterations.

To bridge this gap, deepfake detection datasets must evolve to include high-quality deepfakes that better reflect real-world threats. This study examines popular datasets such as Celeb-DF and DF-1.0, which revealed that despite efforts toward attribute variability, these datasets often lack demographic and facial diversity. Our study uses FaceSwap to explore how dataset composition and simple augmentation (horizontal mirroring) relate to training behavior and output quality in a small set of faceswaps. We observed that augmented datasets were associated with slightly lower faceloss—a training-stage metric that reflects how well the model reconstructs or blends facial regions—and noted attribute-dependent differences across the subjects tested. These findings are preliminary and reflect an exploratory analysis across a small set of subjects and configurations. Because faceloss functions as a training-stage proxy rather than a perceptual or detection metric, broader replication with larger and more diverse datasets would strengthen the statistical power of future evaluations. Nevertheless, the patterns observed in this work highlight promising directions for improving dataset balance and transparency and suggest practical strategies for mitigating attribute-driven skews in both deepfake creation and detection.

The study further investigates how dataset composition affects deepfake generation by evaluating two factors: (1) the impact of horizontal-mirroring augmentation, which aims to increase facial-orientation variability, and (2) how FaceSwap performs on subjects with different attributes to reveal potential skews in the deepfake generation process.

Keywords—*Deepfake, Data Bias, AI Cybersecurity*

I. INTRODUCTION

Deepfakes are high quality alterations of images and videos created by AI algorithms. We are specifically focusing on face-swapping alterations, extracting a person's face and fusing it onto another. Deepfakes can be used dangerously by creating fake scenarios of prominent people in society. For example, a deepfake video of a politician could go viral, spreading false information or assassinating someone's character. They could be used during elections and in other ways that could affect a lot of people. It is especially dangerous because of applications, like Zao, that make it easy to create deepfakes [8]. Zao is an app that allows the user to place their face on videos of celebrities [1].

To help readers interpret our findings appropriately, we note that our experiments are exploratory and limited in scope. As such, we present our results as indicative observations that highlight emerging patterns and potential strategies, and we discuss these limitations more fully in later sections.

II. CURRENT DETECTION TECHNOLOGIES

The Defense Advanced Research Projects Agency (DARPA)'s Media Forensics program has been developing methods to detect a deepfake. Their method focuses on three types of integrity: digital, physical, and semantic. 1) Digital integrity involves analyzing the pattern of pixels in images. 2) Physical integrity involves analyzing the consistency of lighting and shadows in images. 3) Semantic integrity involves comparing the environment in images to the environment from trusted sources. The program created deepfake detectors that combine the three methods into a single integrity score. Their progress is demonstrated by a prototype Web portal that has over 20 detectors analyzing media for manipulations [8].

The success of the Media Forensics Program led to the Semantics Forensics program which was launched in 2020. It focuses on determining the importance of manipulation in media. For example, manipulating a family's vacation photo is insignificant compared to manipulating a politician's speech. The goal of the program is to utilize embedded watermarking in digital content to make the authentication process easier. The downside of this method is that every authentication tool would need to see the watermark as legitimate for it to be effective, but this would take a long time to do [8].

The AI Foundation is developing a tool for consumers to detect a deepfake, called Reality Defender. It will scan the screen and use automatic detectors to alert users of altered media. Unfortunately, social media is filled with alterations from Photoshop and other editors which would flood the user with alerts making it ineffective [8]. Deepfakes need to be authenticated before they have been uploaded and spread.

AI algorithms, specifically, Deep Neural Networks (DNNs) create deepfake videos of three categories: 1) Head puppetry replaces the head and upper shoulders of a person in a video to create the appearance of them behaving in the same way. 2) Face swapping replaces the face of a person in a video but keeps the same facial expressions. 3) Lip syncing alters the lip region of a person in a video so that they appear to say something that they have not. Deepfake detection technologies have progressed a lot in the past two years. There are many effective detection methods, but they are not without problems. The current deepfake detection methods mainly target face swapping videos. There are many categories of detection methods. The first category focuses on inconsistent physical aspects. Some of these methods observe the lack of realistic eye blinking in deepfake videos because they use pictures with open eyes. Others observe the inconsistency of head poses in deepfake videos). The second category uses signal-level alterations done during the synthesis process of deepfakes. State-of-the-art DNN splicing detection methods can be used to authenticate videos in this category. The third category detection methods are data-driven and use specifically trained DNNs [1].

Large datasets of deepfake videos helped develop valuable methods of detection but, the quality of deepfake videos in these datasets are of significantly lower quality than those on the Internet. These low quality deepfake videos are unconvincing and won't have the impact of better deepfake videos. Also, the methods created using specific datasets are less effective when used on others [1].

Multimedia Forensics ensures authenticity and origin of an image or video. Early methods focus on expected statistical or physical features that happen during image formation, but more recent methods focus on CNN-based solutions. CNN-based solutions use Convolutional Neural Networks (CNNs) to identify manipulated images. They are a type of deep neural network. For deep fake videos, these methods focus on specific types of manipulations. 1) Duplicated frame manipulation where each frame of a video is duplicated or replicated. This type of manipulation creates videos like looping or slow motion. 2) Varying interpolation types manipulate a video by changing the movement of objects. 3) Chroma-key compositions are also known as green screen compositions. They are popular in filmmaking and visual effects and involve the background of a video to be replaced with different images. Other methods focus specifically on manipulations of faces and differentiate the types of facial manipulations; like, Face splicing, face swapping, and deepfakes [5].

III. DETECTION SKEWS IN DEEPFAKE DETECTION

Detection skews in deepfake detection methods are a major concern. Many studies have brought attention to skews with age, gender, and ethnicity. These skews can cause deepfake detection to fail, leaving them to spread potentially harmful media. The main cause of these detection imbalances is the lack of variety in the training dataset. Detection models also make skewed assumptions that would lead to labeling real images as fake. For example, if the person is smiling or wearing a hat many detectors will detect it as fake. There are two ways to detect media manipulation: 1) Spatial features extracted from video frames and analyzed for unnatural facial features, blending, and CNN-generated/GAN-generated fingerprints. 2) Temporal features are used across video frames to detect inconsistent color, facial landmarks, and head poses. For example, LipForensics targets inconsistencies in mouth movements to indicate a video has been altered [11].

There are several recent datasets that have made efforts to have more variability and balance. Datasets Celeb-DF and DF-1.0 have equal numbers of male and female images/videos. Celeb-DF has a wide range of ages. DF-1.0 has a wide range of skin types. The neural network models use attribute-related information from the training data which may be the reason for skewed AI models. Of the many people studying AI, only a few are analyzing skewed datasets in deepfake detection. Hazirbas measured deepfake AI models for age, gender, apparent skin type, and lighting. They analyzed the top 5 winners of the Deepfake Detection Challenge and concluded all the methods were skewed for lighter skin tones because the majority failed on darker skin. Loc and Yan measured the performance of popular Deepfake detectors, MesoInception-4, Xception, and Face X-Ray on racially aware datasets with subjects of balanced race and gender. Results showed significant differences across races and large representation skews in genders. Five annotation databases are proposed to completely analyze and lessen data imbalance in deepfake detection models: A-DFD, A-FF+, A-DFDC, A-Celeb-DF, and A-DF-1.0. This work annotates 41 demographic and non-demographic attributes for five deepfake detection databases [12].

FOMM (First Order Motion Model) is a Principal Component Analysis (PCA)-based approach for motion estimation, background motion representation, and animation through disentanglement. It operates in two main stages:

coarse motion estimation and dense motion prediction. Coarse motion captures small movements between separate objects, while dense motion produces the optical flow of the image. In this formulation, S denotes the source frame, and D denotes the driving frame extracted from the same video. The model estimates motion for K distinct parts, each represented by a transformation matrix $A_k\hat{}(X \leftarrow R)$ that maps image coordinates X (either S or D) to an abstract reference frame R. Each pixel location is represented by $z \in Z$, where Z is the set of all pixel coordinates, and the key point predictor generates K heatmaps $(M_1 \ldots M_k)$ that represent per-part probability distributions across the image.

The motions for each part are defined by the transformation matrices:

$$A^k_{X \leftarrow R} \in R^{2\times 3} \quad (1)$$

The heatmaps satisfy the normalization constraint:

$$\sum_{z\in Z} M^k(Z) = 1 \quad (2)$$

And each heatmap spans the spatial dimensions of the input frame:

$$M^k \in [0,1]H \times W \quad (3)$$

Finally, the translation component for each part is estimated using the soft argmax operator [7]:

$$\mu^k = \sum_{z\in Z} M^k(Z)Z \quad (4)$$

Another deepfake detection model, DFT-MF, focuses on mouth features. It analyses lip and mouth movements to verify the video authenticity. DFT-MF was tested with datasets that contained both fake and real videos and performed well compared to other work in this area. For example, a work presented by Xin Yang, Yuezun Li, and Siwei Lyu used a Support Vector Machines (SVM) based method. It compares face landmarks between the real and fake images. They found that face landmarks' locations in the fake images were large but in the real images, they were small. They also used the head's pose to detect deepfake videos. They noticed the difference between central and whole face boundaries in fake videos are larger than real ones. They applied this finding to two datasets: UADFV and DARPA MediFor. The UADFV dataset is made up of 49 real videos and 49 deepfake videos. The DARPA MediFor contains 241 real images and 252 Deepfake images. The datasets were put into the SVM and evaluated using the Area Under the Receiver Operating Characteristic curve (AUROC). A ratio of 0.89 was attained against the UADFV dataset and 0.843 against the DARPA MediFor dataset [2].

The DFT-MF model detects deepfake videos using CNNs. The main component in the model, image extraction, requires a lot of time and power. Using MoviePy, an open-source software in Python for editing and cutting videos, they can cut the video based on certain words where the mouth would be open and showing teeth. This allows the DFT-MF model to exclude irrelevant images and saves time, unlike other methods. Before the image frames are analyzed, they must be preprocessed to filter out image frames that do not contain faces. The Dlib classifier detects face landmarks and is used for this process. Because this model focuses on the area surrounding the mouth, the mouth is cropped from the face. Then the Dlib library is used to do this. It estimates the location of 68 (x, y) coordinates to graph facial structures. Dlib also returns the points creating the shape of the face. DFT-MF uses the points that outline the shape of the mouth to crop it from the face. The dataset has 12,500 real image frames and 12,500 fake image frames [2].

The DFT-MF model uses three variables to determine if videos are real or fake: words per sentence, speech rate, and frame rate. According to a study by the American Press Institute (API), readers understand 100% of information when sentences average 8 words or less. S, the DFT-MF model uses 5 words per sentence as a clear sentence indicator. The speech rate in conversations and official speeches is between 120 words per minute (wpm) and 150 wpm. Therefore, the model considers 120 wpm as the standard for official speeches. Lastly, the frame rate in frames per second (fps) runs at 30 fps in the DFT-MF model. An altered sentence in a deepfake video would be at least 2 seconds long, which is at least 50 frames. Because of this, the DFT-MF model classifies fake videos by having more than 50 fps [2].

Another study on data imbalance in deepfake detection models was done by the University of Southern California which focused on racial and gender differences. When evaluating facial datasets balanced by race and gender, they found large differences in the predictive accuracy across races. There was up to a 10.7% difference in error rate across racial groups. A very popular dataset, FaceForensics++, was found to be mainly composed of female Caucasian subjects. During their investigation on racial distribution of deepfakes they found that the methods used to create deepfakes as positive training signals can produce irregular faces when a person's face is swapped onto someone of a different race or gender which causes the detectors to learn false correlations. The study revealed potential challenges in fairness and emphasized the need for extensive evaluation and assessments of datasets to address data imbalance [10].

IV. OVERCOMING DATA IMBALANCE IN DEEPFAKE DETECTION MODELS

A major factor that hinders the capabilities of deepfake detection models is data imbalance. Skewed predictions are caused by the lack of variety and size in the datasets, which leads to the inability to detect accurately. Varied and balanced datasets are crucial to mitigate skewed predictions [10]. In the context of deepfakes, we are specifically focusing on face-swapping alterations, extracting a person's face and fusing it onto another. Imbalances in the training data can result in distortions that compromise authenticity and credibility of the deepfake [4]. Detection models like FOMM and DFT-MF use

motion estimation and mouth features to detect manipulation. A possible solution to address skewed predictions would be to use supervised learning models like Convolutional Neural Networks (CNNs), to categorize datasets based on their demographic attributes such as race, gender, and skin tone. CNNs classification algorithms can analyze and quantify the variety within datasets, shedding light on potential imbalances. This would strongly aid in producing comprehensive datasets which produce fair and sophisticated deepfake detection models.

V. EXPERIMENT: EFFECTS OF DATASET DIVERSITY ON DEEPFAKE QUALITY

To further investigate the impact of demographic diversity on deepfake quality, we conducted an experiment by creating four deepfakes. The first involved African American females, Lizzo and SZA, the second involved Caucasian females, Rebel Wilson and Megan Fox, the third involved African American males, Kevin Hart and Sterling K. Brown, and the fourth involved Caucasian males, Hugh Jackman and Ryan Reynolds. Each faceswap used videos of similar lengths, extracting over 1,400 images per person. We applied dataset augmentation, including horizontally mirrored images, to enhance variability. The rationale behind this augmentation was to expose the model to variations in facial orientation, potentially improving its ability to handle left-facing and right-facing faces [12]. The full dataset composition for each subject pair, including original and augmented image counts, is summarized in Table I. The original face images and their corresponding deepfake outputs for all four subject pairs are shown in Figure 1. From Top Left: Rows (A–D): Figure 1A: Hugh–Ryan; Figure 1B: Sterling–Kevin; Figure 1C: Megan–Rebel; Figure 1D: Lizzo–SZA; the third column representing their corresponding faceswap.

TABLE I. Dataset makeup for original & augmented datasets

Dataset	Deepfake	Face A (# of Images)	Face B (# of Images)	Dataset Size
Original	Lizzo & SZA	Lizzo (1,009)	SZA (1,442)	2,451 images
	Rebel & Megan	Rebel Wilson (1,009)	Megan Fox (1,442)	2,451 images
	Hugh & Ryan	Hugh Jackman (1,009)	Ryan Reynolds (1,446)	2,445 images
	Sterling & Kevin	Sterling K Brown (1,009)	Kevin Hart (1,446)	2,445 images
Augmented	Lizzo & SZA	Lizzo (2,018)	SZA (2,884)	4,902 images
	Rebel & Megan	Rebel Wilson (2,018)	Megan Fox (2,884)	4,902 images
	Hugh & Ryan	Hugh Jackman (2,018)	Ryan Reynolds (2,892)	4,910 images
	Sterling & Kevin	Sterling K Brown (2,018)	Kevin Hart (2,892)	4,910 images

Fig. 1. People used in faceswap datasets and their deepfake outputs

A. Model Architecture

Faceswap's training algorithm involves several components and processes. Among the crucial components of the neural network architecture are the convolutional layers, which apply learnable filters to the input images through convolutions. Convolution is a mathematical operation that involves each filter scanning across the input image, performing a dot product between the filter weights and the corresponding pixel color values in the region of the input image called the receptive field. Using filters, the convolutional layers can extract various features such as edges, textures, and patterns from the input image. Activation functions like Rectified Linear Unit (RELU) introduce non-linear transformations to the extracted features, enabling the neural network to learn complex relationships within the data [9].

B. Observations

Across 500,000-iteration runs on the four face-swap pairs we tested, we observed lower faceloss in the Hugh Jackman–Ryan Reynolds swaps relative to the Lizzo–SZA swaps under otherwise similar settings, and slightly lower faceloss when using horizontally mirrored augmentation. For example, in our runs, the Jackman–Reynolds deepfakes reached ~0.0075 faceloss on augmented datasets compared to ~0.0086 for Lizzo–SZA, suggesting smoother blending under those conditions. Average faceloss values at 100K and 500K

iterations for all four subject pairs are provided in Table II, highlighting performance differences across demographic attributes. Because faceloss is a training-stage proxy reflective of reconstruction/blending fit rather than a direct perceptual or detection metric, and because our subject set is small, these observations should be interpreted as preliminary and non-generalizable beyond our specific experimental conditions.

TABLE II. Experiment Results

Subject Pair	Dataset Iterations	Avg. Faceloss A Avg. Faceloss B	Avg. Faceloss A Avg. Faceloss B	Dataset Iterations	Subject Pair
Sterling / Kevin	Original 100,000	0.0215	0.0149	Original 100,000	Hugh / Ryan
		0.0199	0.0128		
	Original 500,000	0.0127	0.0088	Original 500,000	
		0.0120	0.0074		
Lizzo / SZA	Original 100,000	0.0194	0.0146	Original 100,000	Rebel / Megan
		0.0153	0.0158		
	Original 500,000	0.0111	0.0085	Original 500,000	
		0.0086	0.0093		
Lizzo / SZA	Original 100,000	0.0194	0.0190	Augmented 100,000	Lizzo / SZA
		0.0153	0.0147		
	Original 500,000	0.0111	0.0110	Augmented 500,000	
		0.0086	0.0083		
Hugh / Ryan	Original 100,000	0.0152	0.0151	Augmented 100,000	Hugh / Ryan
		0.0128	0.0128		
	Original 500,000	0.0088	0.0089	Augmented 500,000	
		0.0074	0.0075		

VI. EXPERIMENT METHODOLOGY

In the next phase of this experiment, we wanted to test how imbalanced datasets affect a deepfake detection model. The experiment evaluates three types of attributes in a dataset (race, gender, age) and considers overall dataset size. We collected deepfake and other facial datasets totaling approximately 145,000 real and fake faces. We created different dataset arrangements to examine how specific attribute imbalances and dataset size influence detection performance. The first tier consists of small datasets with 500 images, which serve as a baseline for model behavior. These datasets are: (1) All old (balanced race and gender), (2) All young (balanced race and gender), (3) All male (balanced race and age), (4) All female (balanced race and age), (5) All Caucasian (balanced gender and age), and (6) All African American (balanced gender and age).

At this stage, our analysis is descriptive and exploratory; we do not report hypothesis tests or confidence intervals, and we treat observed differences as signals for future, statistically powered experiments.

The second tier includes medium-sized datasets with 5,000 images that combine two to three attributes. These are: (1) Caucasian/Male (balanced ages), (2) Caucasian/Female (balanced ages), (3) African-American/Male (balanced ages), (4) African-American/Female (balanced ages), (5) Caucasian/Young (balanced genders), (6) Caucasian/Old (balanced genders), (7) African-American/Young (balanced genders), (8) African-American/Old (balanced genders), (9) Young/Female (balanced races), (10) Young/Male (balanced races), (11) Old/Female (balanced races), and (12) Old/Male (balanced races). The final tier consists of large datasets with 10,000 images and is designed to evaluate how varying attribute ratios affect model performance. These datasets include: (1) 100% male, (2) 80% male / 20% female, (3) 60% male / 40% female, (4) 50% male / 50% female, (5) 100% female, (6) 100% old, (7) 80% old / 20% young, (8) 60% old / 40% young, (9) 50% old / 50% young, (10) 100% young, (11) 100% Caucasian, (12) 80% Caucasian / 20% African-American, (13) 60% Caucasian / 40% African-American, (14) 50% Caucasian / 50% African-American, and (15) 100% African-American.

If Tier 1 or Tier 2 reveals that certain attributes have minimal effects, we may skip some ratios for those attributes in later tests to focus on the ones that demonstrate meaningful changes. We hypothesize that the model will perform better on the larger datasets and that adjusting attribute ratios will reveal a range at which the model can be trained more robustly. Ultimately, we hope that this experiment will inform dataset creation guidelines for deepfake detection and other machine-learning models so that dataset bias and skewed predictions can be minimized.

The first dataset that was trained from Tier 1 was all-Caucasian dataset with balanced ages and genders. The attribute distribution for this dataset is illustrated in Figure 2. As shown in Figure 3, we noticed that overfitting happened very quickly at around 3 epochs; at 3 epochs the training accuracy was 84% but by the end of training at 20 epochs it declined to 79%. The model's classification performance on this dataset is presented in Figure 4, which displays the resulting confusion matrix. Consistent with our study's exploratory scope, this overfitting behavior reflects our specific training configuration and dataset and should not be generalized without broader replication and statistical evaluation.

Fig. 2. All Caucasian dataset breakdown

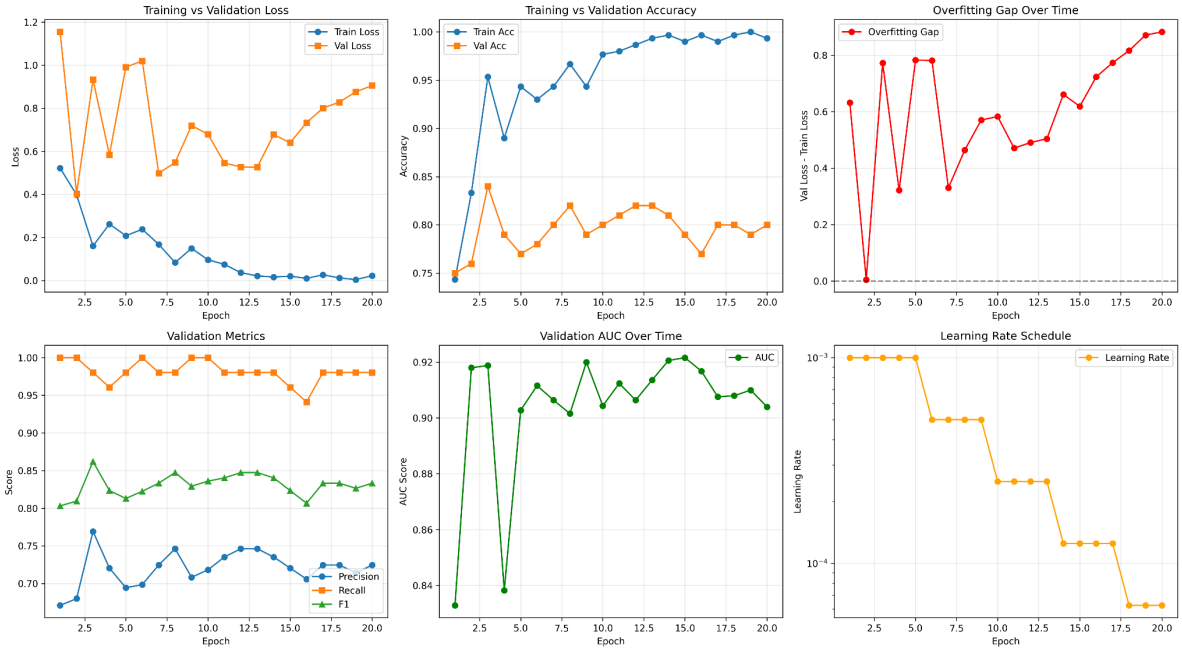

Fig. 3. All Caucasian training graphs

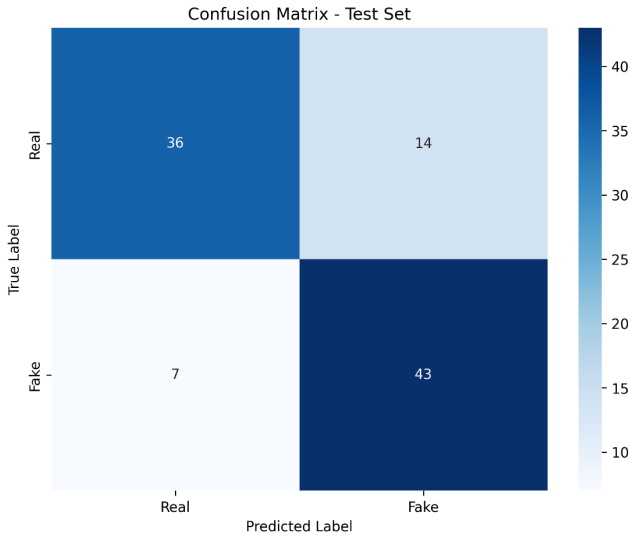

Fig. 4. All Caucasian prediction results

VII. CONCLUSION

In our experiments, augmenting training sets with horizontally mirrored images was associated with lower faceloss and qualitatively smoother blends in several runs. We also observed attribute-dependent differences among the small set of subjects tested. These patterns suggest that data composition can influence training behavior and output characteristics in this setting; however, we caution against generalizing beyond our specific configuration and subjects

Limitations. Findings are based on a limited number of subjects and runs, use faceloss as a proxy rather than a perceptual or detection metric, do not include hypothesis testing, and reflect a single toolchain (FaceSwap) and hyperparameters. Broader replication with balanced, transparently annotated datasets and formal statistical analyses is required to assess significance and generalizability.

We also note that persistent differences in faceloss across subject pairs underscores the need to address potential data imbalances. While augmentation was associated with slight improvements across our runs, the overrepresentation of certain demographics in training data likely contributes to uneven performance. Future work will expand augmentation strategies and prioritize annotated, demographically balanced datasets to better understand and mitigate skewed model behavior—toward more equitable and accurate deepfake detection systems [6].

A. Implications for Cybersecurity Education

As deepfake technologies evolve, understanding dataset composition, annotation transparency, and model behavior has become an essential skill for students preparing for careers in cybersecurity. The exploratory findings in this study can support classroom discussions and hands-on exercises around dataset auditing, augmentation methods, and the limitations of current detection technologies—helping students develop the analytical mindset needed to evaluate AI-driven media manipulation and build more robust detection strategies.

VIII. SUPPLEMENTAL MATERIALS

Lower faceloss indicates smoother reconstruction / blending during training; faceloss is a training-stage proxy rather than a perceptual or detection metric. Δ = mean (Original) − mean (Augmented) for per-pair figures; for cross-pair figures, Δ First−Second = mean (First) − mean (Second). Observed differences across pairs may reflect attribute-linked data effects; we hypothesize that imbalances (e.g., in pre-existing landmark/face datasets) contribute to these patterns, but confirming causality requires broader, statistically powered experiments.

As displayed in Appendix Figure 5, at 100K, Lizzo: 0.01936 (Original) vs 0.01904 (Augmented), Δ = 0.00033; SZA: 0.01536 (Original) vs 0.01478 (Augmented), Δ = 0.00058. At 500K, Lizzo: 0.01110 (Original) vs 0.01102 (Augmented), Δ = 0.000079; SZA: 0.00867 (Original) vs 0.00831 (Augmented), Δ = 0.00036.

The corresponding results for the Rebel-Megan pair are displayed in Appendix Figure 6. At 100K, Rebel: 0.01461 (Original) vs 0.01502 (Augmented), Δ = −0.00041; Megan: 0.01588 (Original) vs 0.01655 (Augmented), Δ = −0.00067. At 500K, Rebel: 0.00859 (Original) vs 0.00878 (Augmented), Δ = −0.00019; Megan: 0.00934 (Original) vs 0.00969 (Augmented), Δ = −0.00035.

The faceloss outcomes for the Sterling-Kevin pair are shown in Appendix Figure 7. At 100K, Sterling: 0.02168 (Original) vs 0.02175 (Augmented), Δ = −0.00008; Kevin: 0.019999 (Original) vs 0.020912 (Augmented), Δ = −0.00091. At 500K, Sterling: 0.012696 (Original) vs 0.013003 (Augmented), Δ = −0.00031; Kevin: 0.012041 (Original) vs 0.012515 (Augmented), Δ = −0.00047.

The Hugh-Ryan faceloss trajectories are provided in Appendix Figure 8. At 100K, Hugh: 0.01502 (Original) vs 0.01515 (Augmented), Δ = −0.00013; Ryan: 0.01269 (Original) vs 0.01285 (Augmented), Δ = −0.00016. At 500K, Hugh: 0.00882 (Original) vs 0.00889 (Augmented), Δ = −0.000068; Ryan: 0.00741 (Original) vs 0.00748 (Augmented), Δ = −0.000070.

The direct comparison between the Hugh-Ryan and Sterling-Kevin pairs at the 100K iteration window is illustrated in Appendix Figure 9. Original: Hugh (0.01526) and Ryan (0.01269) < Sterling (0.02170) and Kevin (0.01999); Δ Hugh−Sterling = −0.00644, Δ Ryan−Kevin = −0.00730. Augmented: Hugh (0.01515) and Ryan (0.01285) < Sterling (0.02174) and Kevin (0.02086); Δ Hugh−Sterling = −0.00660, Δ Ryan−Kevin = −0.00801.

The 500K iteration comparison for these two pairs is shown in Appendix Figure 10. Original: Hugh (0.00882) and Ryan (0.00741) < Sterling (0.01267) and Kevin (0.01208); Δ

Hugh–Sterling = −0.00385, Δ Ryan–Kevin = −0.00468. Augmented: Hugh (0.00889) and Ryan (0.00748) < Sterling (0.012999) and Kevin (0.01254); Δ Hugh–Sterling = −0.00411, Δ Ryan–Kevin = −0.00506.

A cross-pair comparison between Lizzo-SZA and Rebel-Megan at 100K iterations appears in Appendix Figure 11. Original: Lizzo (0.01936) > Rebel (0.01460); Δ Lizzo–Rebel = 0.00476. SZA (0.01535) < Megan (0.01587); ΔSZA–Megan = −0.00052. Augmented: Lizzo (0.01904) > Rebel (0.01502); Δ = 0.00402. SZA (0.01478) < Megan (0.01656); Δ = −0.00178.

The 500K iteration comparison for these pairs is shown in Appendix Figure 12. Original: Lizzo (0.01110) > Rebel (0.00857); Δ = 0.00253. SZA (0.00867) < Megan (0.00932); Δ = −0.00065. Augmented: Lizzo (0.01102) > Rebel (0.00876); Δ = 0.00226. SZA (0.00831) < Megan (0.00964); Δ = −0.00133.

Taken together, our results show that simple choices about dataset composition and augmentation can measurably shape deepfake training behavior, with persistent cross-pair differences and selective gains under mirroring revealing where current practices help—and where they fall short. Although exploratory and limited to a small set of subjects and a training-stage proxy (faceloss), these patterns point to practical next steps: prioritize transparently annotated, demographically balanced datasets and report augmentation effects so researchers and educators can better anticipate skew and improve robustness. In short, this work offers early, evidence-based guidance that can inform both dataset design and classroom practice—moving the community toward more equitable, reliable creation and detection of AI-manipulated media.

REFERENCES

[1] He, L., Guy, J., Wang, S. 2019. New Chinese “deepfake” face app backpedals after privacy backlash | CNN business. CNN.

[2] Jafar, M. T., Ababneh, M., Zoube, M. A., & Elhassan, A. 2020. Forensics and Analysis of Deepfake Videos. 11th International Conference of Information and Communication Systems (ICICS)

[3] Lyu, S. 2020. Deepfake detection: Current challenges and next steps. IEEE International Conference on Multimedia & Expo Workshops (ICEW).

[4] Ranjan, P., Patil, S., & Kazi, F. 2020. Improved Generalizability of Deep-Fakes Detection using Transfer Learning Based CNN Framework. 2020 3rd International Conference on Information and Computer Technologies (ICICT), San Jose, CA, USA, 2020, pp. 86-90, doi: 10.1109/ICICT50521.2020.00021

[5] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. 2019. FaceForensics++ Learning to Detect Manipulated Facial Images. arXiv.1901.08971

[6] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). 10.1186/s40537-019-0197-0

[7] Siarohin, A., Woodard, O. J., Ren, J., Chai, M. 2021. Motion Representations for Articulated Animation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

[8] Strickland, E. 2023. Facebook AI launches its Deepfake Detection Challenge. IEEE Spectrum.

[9] Taye, M.M. Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. Computation 2023, 11, 52. 10.3390/computation11030052

[10] Trinh, L., & Liu, Y. 2021. An Examination of Fairness of AI Models for Deepfake Detection. International Joint Conference on Artificial Intelligence (IJCAI-21).

[11] Xu, M., Yoon, S., Fuentes, A., Park, D.S. 2022. A Comprehensive Survey of Image Augmentation Techniques for Deep Learning

[12] Xu,Y.,Terhörst, P., Raja, K., Pedersen, M. 2022. Comprehensive Analysis of AI Biases In Deepfake Detection With Massively Annotated Databases

APPENDIX

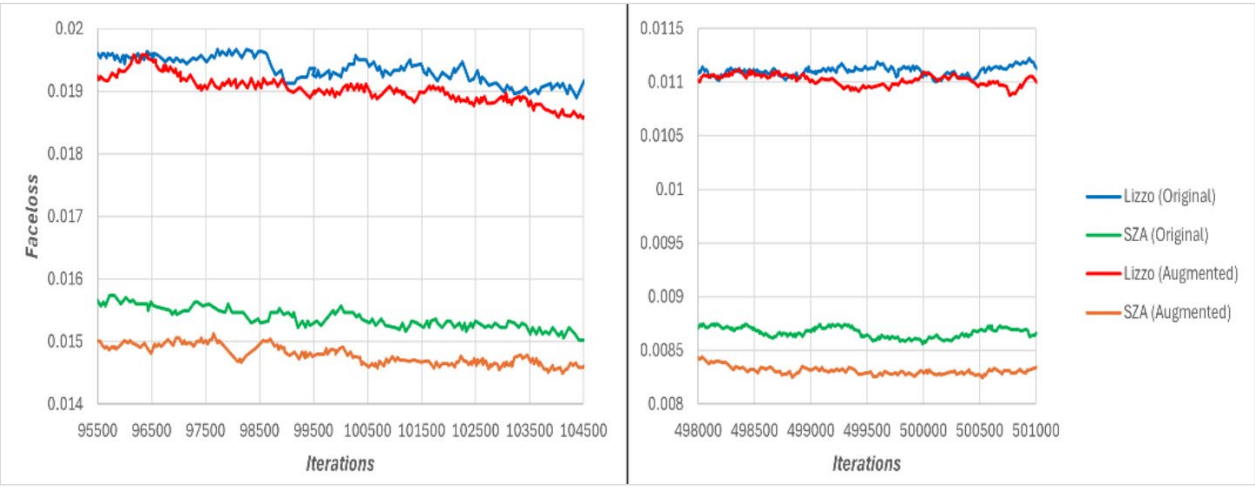

Fig. 5. Lizzo–SZA: faceloss at 100K and 500K for Original vs. Augmented.

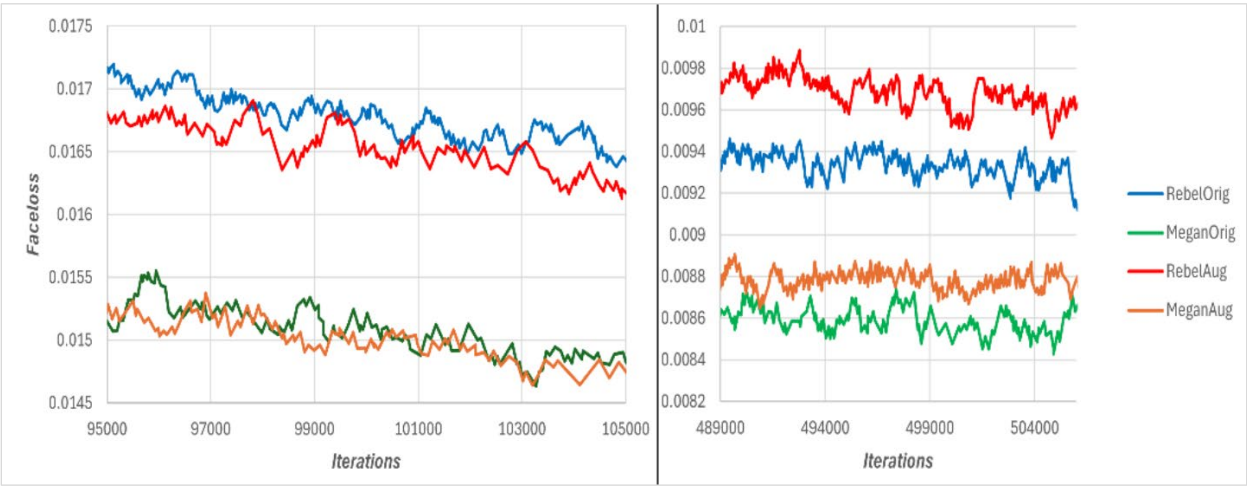

Fig. 6. Rebel Wilson-Megan Fox: Faceloss at 100K and 500K for Original vs. Augmented.

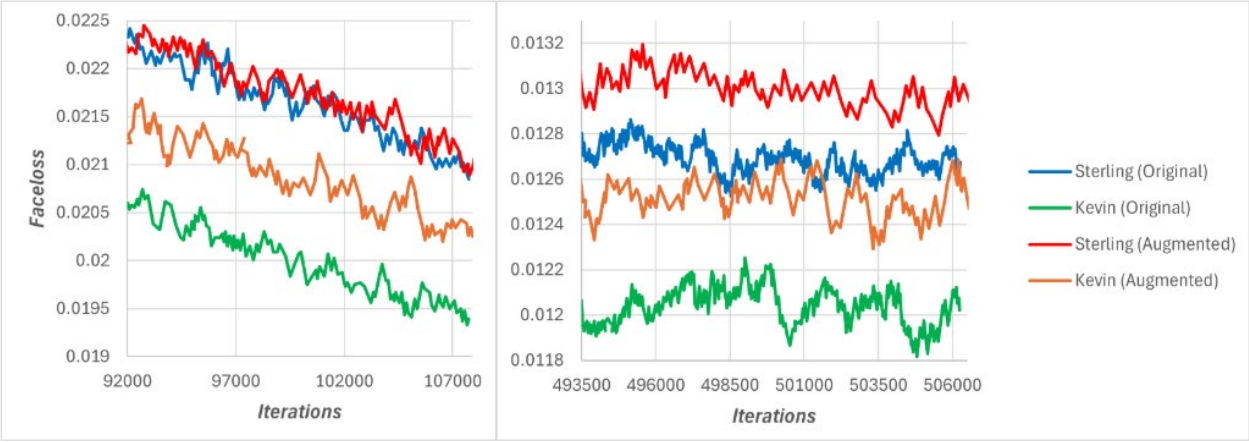

Fig. 7. Sterling K. Brown–Kevin Hart: Faceloss at 100K and 500K for Original vs. Augmented.

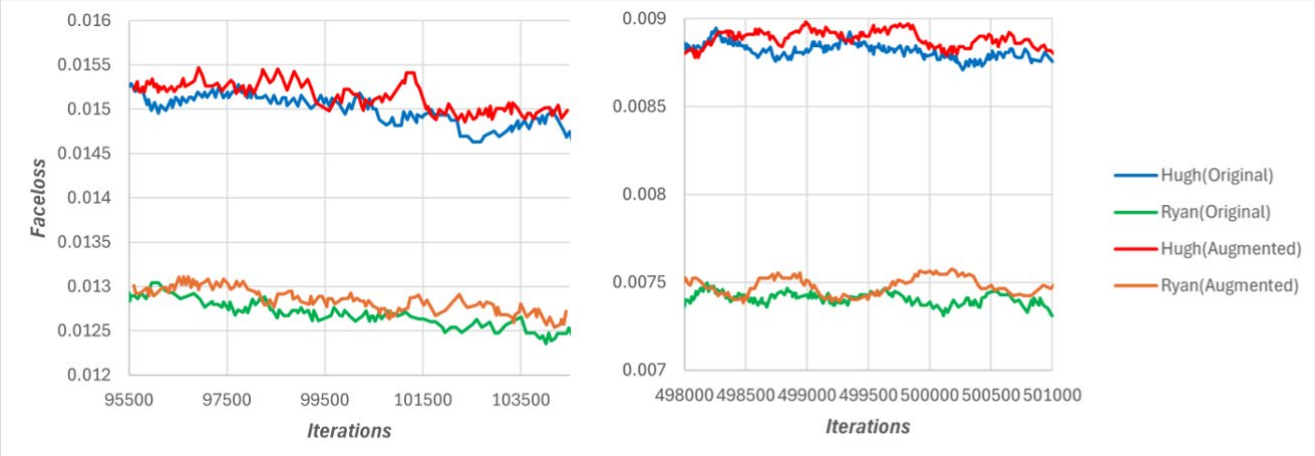

Fig. 8. Hugh Jackman–Ryan Reynolds: Faceloss trajectories at 100K and 500K iterations for Original vs. Augmented datasets. Augmentation is associated with slightly lower faceloss and smoother blending within this pair.

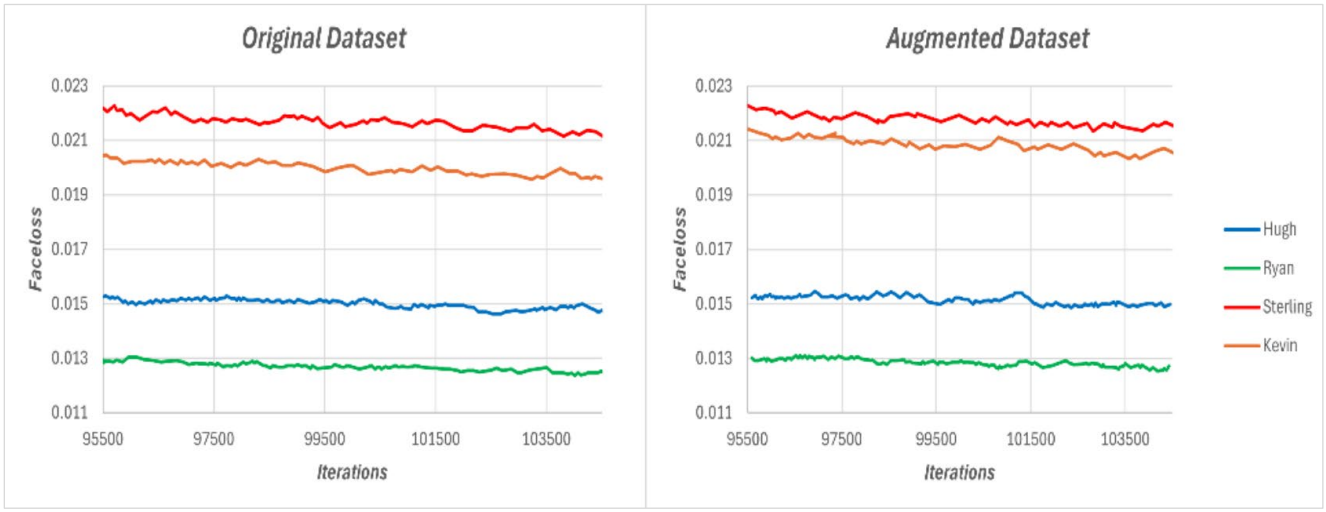

Fig. 9. Hugh Jackman–Ryan Reynolds vs. Sterling K. Brown–Kevin Hart (100K window, Original & Augmented).

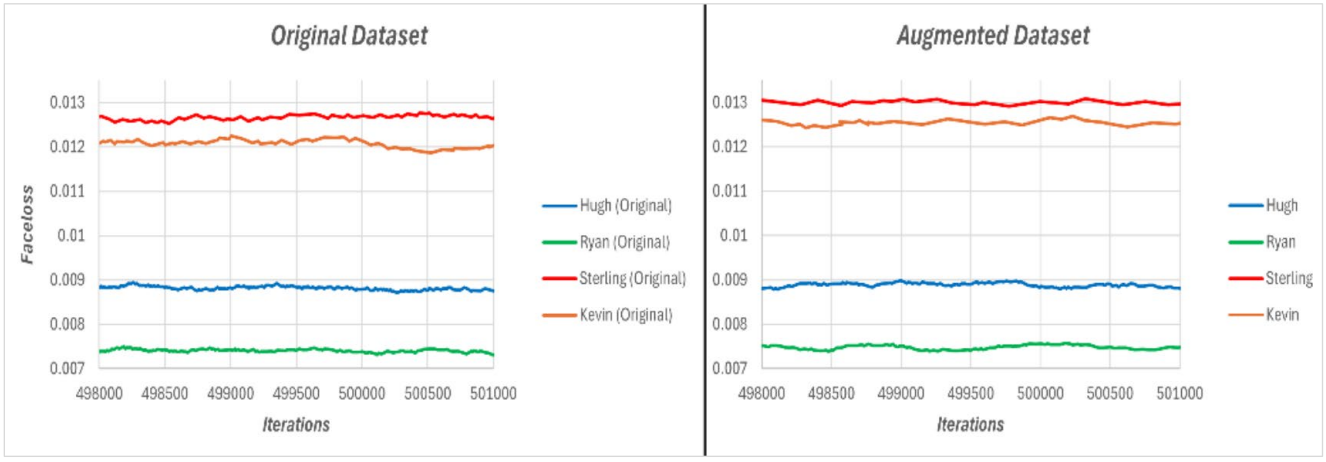

Fig. 10. Hugh Jackman–Ryan Reynolds vs. Sterling K. Brown–Kevin Hart (500K window, Original & Augmented).

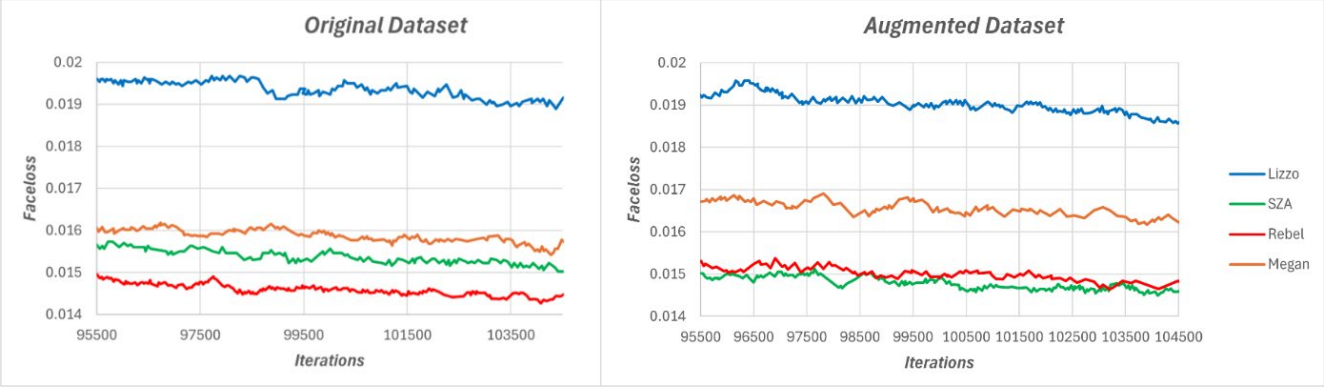

Fig. 11. Lizzo-SZA vs Rebel Wilson-Megan Fox (100K window, Original & Augmented).

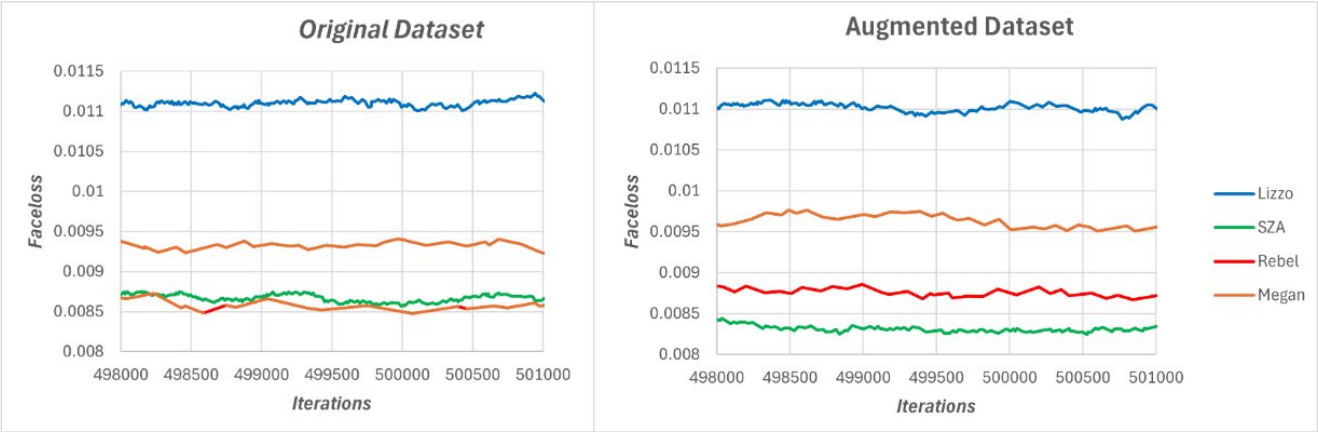

Fig. 12. Lizzo-SZA vs Rebel Wilson-Megan Fox (500K window, Original & Augmented).