Complex & Intelligent Systems
https://doi.org/10.1007/s40747-022-00956-7

ORIGINAL ARTICLE

An explainable deepfake detection framework on a novel unconstrained dataset

Sherin Mathews[1] · Shivangee Trivedi[1] · Amanda House[1] · Steve Povolny[2] · Celeste Fralick[1]

Received: 4 May 2021 / Accepted: 8 December 2022

Abstract
In this work, we created a new large-scale unconstrained high-quality Deepfake Image (DFIM-HQ) dataset containing 140K images. Compared to existing datasets, this dataset includes a variety of diverse scenarios, pose variations, high-quality degradations, and illumination variations, making it a particularly challenging dataset. Since computer vision models learn to perform a task by capturing relevant statistics from training data, they tend to learn spurious age, gender, and race correlations leading to learning biases. To account for AI bias in our proposed DFIM-HQ dataset, we design a simple yet effective image recognition benchmark for studying bias mitigation. Our detection system makes use of an Inception-based network to extract frame-level features and automatically detect manipulated content. We also propose an explainability framework that provides a better understanding of the model's prediction. Such informed decisions provide insights that can be used to improve the model and, thereby, helps to add trust to the model. Our evaluation illustrates that our frameworks can achieve competitive results in detecting deepfake images using deep learning architectures.

Keywords Deepfakes · Deepfake image detection techniques · Digital media forensics · Deep learning · Explainability · AI bias

Arabic Keywords الـذكـاء الاصطنـاعي. رؤيـة الـكمبـيوتـر. عميق. تقـنية

Introduction

Background

Deepfakes correspond to fake data in the form of an image, audio, video, or text. The term "deepfakes" originated after a Reddit user by the same name who created fake videos using an AI-inspired algorithm [1]. Early versions of fake images and video content were created using a combination of deep learning and computer graphics-inspired methods with the intent to make the targeted audience believe that the data are real. Deepfakes often have malicious intent, and, in recent times, deepfakes have been used to serve hostile purposes such as to spread disinformation, tarnish brand reputations, and generate fake news. Examples of a significant number of convincing deepfakes emerged which led researchers to focus on the importance of detecting manipulated content. For example, the infamous video of former US President Barack Obama mouthing words he never spoke convinced several viewers that the altered video was real, until Jordan Peele described the technique behind generating the video

✉ Sherin Mathews
sherin_mathews@mcafee.com; sherinm@udel.edu

Shivangee Trivedi
shivangee_trivedi@mcafee.com

Amanda House
amanda_house@mcafee.com

Steve Povolny
steve_povolny@mcafee.com

Celeste Fralick
celeste_fralick@mcafee.com

[1] Artificial Intelligence Research (AIR), Mcafee, San Jose, USA

[2] Advanced Threat Research (ATR), Mcafee, San Jose, USA

Published online: 17 January 2023

[2]. In another incident, Facebook CEO Mark Zuckerberg appeared to be conveying the power his company possesses and making statements he never made. It was later revealed the video was a deepfake of the original video containing a different speech [3]. Deepfakes can also be abused to commit frauds and scams, as illustrated by an incident that took place in 2019 [4]. Deepfake audio was used to mimic the voice of the CEO of a company to urgently perform a critical business transaction resulting in a loss of close to 240,000 USD. These examples shed light on the urgency and need to develop a strong defense against artificially enhanced content. OpenAI's GPT-2 was utilized to influence public opinion on an Idaho government website that was seeking feedback on changing the Medicaid program[5]. Over half of the of the 1,000 comments on the site were generated by GPT-2 and humans had a difficult time distinguishing human comments from AI generated comments. Fast paced and constantly evolving research in this field, combined with daunting incidents in the wild, makes the detection of deepfakes a pressing issue economically, politically, financially, and socially.

The concept of photograph manipulation dates back to as early as the 1860s [6] and is not particularly new; although over time, underlying techniques to modify photographs have advanced and improved. The initial set of deepfakes were of considerably low quality and could easily be spotted as being manipulated. Corresponding deepfake videos had visible boundaries of face masks superimposed on underlying real faces. Consequently, efforts were made to produce refined deepfakes making them appear highly realistic. Swapping faces between two sources with the help of Generative Adversarial Networks (GANs) and Auto-Encoders (AEs) is the foundation of creating deepfakes. FakeApp [7], still available to download via various 4chan and Reddit forums, was one of the first free tools available to enable users to easily create deepfakes. Faceswap [8] is an open-source tool that creates face swaps with the help of an elaborate pipeline. Dfaker [9] is another example of a freely available tool to perform face-swapping. These tools, albeit free, resulted in unrealistic deepfakes requiring manual effort to make the output appear convincing. DeepFaceLab [10,11], however, uses a stable [12] underlying architecture resulting in compelling deepfakes used by various content creators. Face2Face [13] proposed real-time reenactment resulting in the transfer of expressions from a source image or video to a target image or video. FaceShifter [14] contributed toward further refining face-swapping by developing a two-stage framework specifically overcoming high fidelity and occlusion challenges. The digital manipulation techniques involved in creating deepfakes can be broadly categorized into four major categories, namely entire face synthesis, identity swap, attribute manipulation, and expression swap [15]. Briefly speaking, identity swap refers to swapping the faces of two different persons using either computer graphics methods or deep learning methods. Expression swap is also commonly referred to as facial reenactment where facial expressions and attributes of one person (such as voice) are transferred onto another person. Attribute manipulation is more advanced than swapping techniques as it consists in subtly modifying specific facial features, such as smoothing of the skin or adding glasses. The most advanced method resulting in the generation of highly realistic deepfakes is face synthesis. It refers to creating entirely new fake faces of a non-existent person, supported by deep learning architectures, such as StyleGAN [16], ProGAN [17] and StarGAN [18], to name a few.

Further sections in this paper are arranged as follows: "Deepfake detection: prior art" will cover the prior art of deepfake image research and existing deepfake image datasets. "Proposed deepfake dataset with AI bias and mitigation analysis" will include details of our proposed dataset along with an in-depth AI bias analysis. Lastly, "Proposed detection framework with explainable AI" will provide details of our methods to implement deepfake detection, including a novel adoption of analysis using Explainable Artificial Intelligence (XAI). "Experimental results and analysis" concludes our work and provides directions for further research in this field.

Our paper contributes to the field of deepfake detection in the following ways:

- We curate a high-quality dataset of 70,000 deepfake images and 70,000 real images collected from superior quality sources.
- We propose an end-to-end deep learning-based deepfake detection framework along with a method to include and analyze visual explanations from our models. Our work bridges the gap between detecting deepfakes and understanding visual disparities between deepfake and authentic images.
- We conduct an AI bias assessment to detect and mitigate the presence of bias on DFIM dataset.

Deepfake detection: prior art

This section covers related work in the field of facial artifact detection. Additionally, we also present a table that contains a list of curated datasets with specific information for each.

Deepfake detection methods

Since deepfakes' appearance and continual evolution, there has been significant research surrounding the detection of deepfakes and spotting digitally manipulated content. Preliminary research focused on the detection of inconsistency in eye-blinking present in fake videos and the corresponding

distortion around the mouth [19]. However, this detection method soon became outdated with the onset of new deepfakes which included consistent, natural eye-blinking of human subjects. Various techniques involved leveraging different architectures of Generative Artificial Networks (GANs) and Auto-Encoders (AEs). This paper [20] proposed a solution to identify deepfake videos by learning the distinct features left by warping techniques required to create nearly perfect deepfakes.

Another paper [21] used two CNNs (Convolutional Neural Networks) whose outputs are combined to identify facial tampering. One of the CNNs classifies faces, while the other extracts steganalysis features that are used on an SVM (Support Vector Machine) trained with triplet loss. The resultant scores from both CNNs were used to distinguish between deepfake and real faces. This paper [22] observed that while GANs can generate photo-realistic images, such images contain artifacts that are different from authentic images due to various noticeable constraints, such as non-symmetric facial features. The authors developed an SVM-based method to classify GAN-synthesized images and real images. While not necessarily used to create deepfakes, this paper [23] highlights concerns associated with Deep Generated Networks (DGNs) such as GANs to generate realistic images. To this end, the authors note that there exist differences in color components between images captured through a camera and images generated using a GAN and leverage this fact in their work. This paper [24] uses a forensic technique to observe distinct features associated with an individual in each video, for example, the pattern in which the individual's facial expressions vary over time. Using such traits, the authors utilize a Support Vector Machine (SVM) to build a deepfake detection algorithm.

Government actions to combat disinformation

Due to the malicious impact of deepfakes, various government agencies and companies have taken initiatives to combat disinformation. Defense Advanced Research Projects Agency (DARPA) released a research program called MediFor in mid-2018 [25], specifically dedicated to combat against digital image manipulations and to study media forensics. The program aims to analyze the authenticity of an image or video for informed decisions regarding media usage. In late-2019, Facebook Inc. partnered with Microsoft and several universities to launch the Deepfake Detection Challenge (DFDC) [26] on the popular website Kaggle. The creators of this challenge built and released a curated dataset [27] consisting of 115,000 videos to encourage participants to develop effective and novel detection methods. The results of DFDC were documented and released along with a paper describing methods used to create the dataset [26,27].

Deepfake image datasets

There exist a few deepfake image datasets. FaceForensics++ [29] is a publicly available dataset containing over 1.8M images generated in an automated, supervised fashion. This dataset builds on its predecessor FaceForensics [37] and contains images developed using two computer graphics-based approaches, namely Face2Face and FaceSwap, and two deep learning-based approaches, namely DeepFakes and NeuralTextures. Images generated by StyleGAN are photo-realistic synthetic images produced by combining and borrowing styles from different sets of images. Other researches [32,33] showcased a method to synthetically produce images with varying facial yaw and transformation by using 3-D head pose modeling. The authors released their dataset as well. Another recent paper focuses on spoofing deepfake detectors by removal of specific distinctive properties on fake images produced by GANs [38]. Additionally, they released a dataset by applying their method of evading face detection [31]. This paper used an attention-based mechanism to train binary classifiers to better distinguish between fake and real faces created by using digital manipulation techniques [42]. The authors made their dataset publicly available [40]. The authors of this paper [36] also released their dataset which they collected specifically to evaluate the authenticity of media forensics. This paper [39] also uses the attention mechanism to detect deepfakes, creating 2-D and 3-D Attention-based Deepfake Detection Networks (ADDNets). These datasets have been summarized in Table 1 along with a description of the methods and drawbacks of each dataset.

Proposed deepfake dataset with AI bias and mitigation analysis

In this section, we discuss our dataset highlighting key properties, such as details on bias, ethnicity and gender. We also briefly discuss the suggested steps to mitigate bias in the dataset.

DeepFake Image-high-quality (DFIM-HQ) dataset

Providing benchmark data of high visual quality is key to model a realistic deepfake threat scenario, thereby providing capabilities of future deepfake video synthesis and detection algorithms. DFIM-HQ, therefore, contributes a large amount of manipulated face images with previously unseen deepfake images of high-quality resolution and visual modification quality.

For our deepfake images dataset, we curated StyleGAN images by crawling the website named https://thispersondoesnotexist.com/. Deepfake images in our dataset are of high-quality JPG format having an average resolution of 1024 ×

Table 1 Comprehensive overview of existing image datasets

Dataset name	Number of images	Resolution	Method	Drawback
FaceForensics ++ [28,29]	> 1.8M in total	VGA is 480 × 480, HA is 720 × 720 and FHD is 1080 × 1080	Used four different methods (a) Computer graphics based: Face2Face, FaceSwap and (b) Learning based: DeepFakes, NeuralTextures	More than half of the images have a low resolution of 480p. The rest of the dataset contains images with an average resolution of 720p and a small portion of the images contain a high resolution of 1080p
iFakeFaceDB [30,38]	87000 in total	224 × 224	Removal of fingerprints from images generated from GANs, resulting in a dataset of fake faces with no residue of GAN fingerprints	The images have a low resolution of 224p
Notre Dame Synthetic Face Dataset [32,33]	2M in total	Different resolutions ranging from 100 × 100 to 800 × 600	Dataset is generated by synthesis of fake textures of human faces which are then combined with real 3-D head models at different facial yaws	The images have varied resolutions with the highest resolution being 800 × 600, less than our dataset
Diverse Fake Face Dataset (DFFD) [34,35]	2.6M in total	Real images vary between 200p to 750p. Fake images vary b/w 150p, 200p, 700p and 750p	Real images are obtained from FFHQ and CelebA datasets, along with source frames from FaceForensics ++. Fake images are obtained by combining four different facial manipulation methods. The sources of fake images are: FaceSwap, Deepfake, Face2Face, Deep Face Lab, FaceAPP, StarGAN, PGGAN, StyleGAN	This dataset has different resolutions for real and fake images, with the highest resolution being 750p, less than our dataset
NIST MFC 2018 [36,37]	50,000 (true manipulations and non-true manipulations)	–	Images collected from the Internet are referred to as World Data. These images are then passed through multiple iterations of manipulations to result in a diverse dataset	Not all images are restricted to facial manipulations. There exist many images of natural scenery, which cannot be used for our purpose of detecting deepfakes
FaceSwap (open-source app), SwapMe (iOS app) [21]	4810 in total (Real- 2800, Fake- 2010)	–	Dataset is created using two apps: (a) An open-source app called FaceSwap (b) An iOS app called SwapMe. Post-processing is later applied, including boundary blurring, resizing, and blending processes	(a) Not a large-scale dataset (b)The tampering quality of images created using FaceSwap is less than the tampering quality of images created using SwapMe

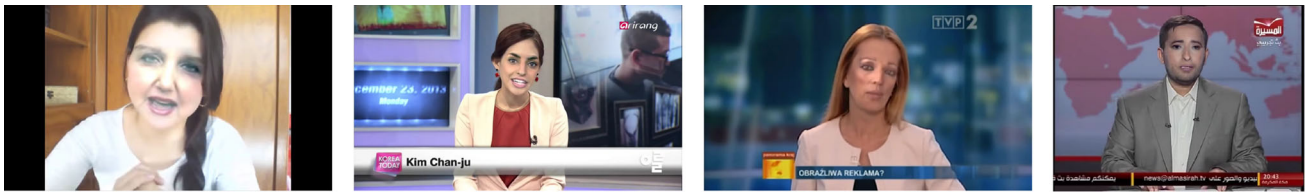


Fig. 1 Illustrates sample deepfake images from face forensics dataset

Fig. 2 Illustrates sample deepfake images from our proposed DFIM-HQ Dataset

Table 2 DFIM-HQ Dataset split into train, validation, and test images

Images	Real	Deepfake
Total	70, 000	70, 000
Train	52, 500	52, 500
Validation	10, 500	10, 500
Test	7000	7000

1024 pixels. The real images from our dataset are also high-quality PNG images at 1024 × 1024 resolution and contain considerable variation in terms of age, ethnicity, and image background. For our real images dataset, we used images from the Flicker-Faces HQ dataset (FFHQ). The images were taken from the Flickr dataset, thus inheriting all the biases of that website, and automatically aligned and cropped using dlib. Thus, in total, our dataset contains 70,000 images of fake content synthesized with STYLEGAN2 and 70,000 real images. A visual illustration comparing the image quality deepfake images from Face Forensics Dataset and the proposed DFIM-HQ dataset is presented in Figs. 1 and 2. We can clearly observe the improved higher quality of deepfake images in DFIM-HQ dataset (Fig. 2) as compared to Face Forensics Dataset (Fig. 1).

Additionally, we also present a breakdown of our dataset for training, validation, and testing indicated in Table 2. The total dataset size is 140,000 images which is the largest highest quality deepfake image dataset produced to date. This dataset serves as a baseline to promote future research in deepfake rendering and detection approaches for forensics study. We conducted bias and race analysis in this dataset which shows that there is significant variation in terms of age, race, and ethnicity. Though deepfake research immensely benefits from datasets, most of these deepfake databases, including ours, have been created under controlled conditions using parameters such as position, pose, lighting, expression, background, camera quality, occlusion, and gender. While there are many applications for deepfake technology in which one can control the parameters of image acquisition, there are also many realistic applications in which the practitioner has little or no control over such parameters. Thus, a dataset that exhibits natural variability in a pose, lighting, focus, resolution, facial expression, age, make-up, occlusions, background, and photographic quality is needed.

Evaluation of AI bias in images

A key concern related to datasets consisting of human subjects is if the images are representative of a diverse population. Under-representation of certain protected classes can result in models biased toward a particular age, race, or gender. For example, a model trained on 70% men and 30% women might perform more favorably toward men. If this model has an impact on decisions such as college admission, then the result could be fewer women receiving admittance. We analyzed the images in the DFIM-HQ dataset using the Deepface framework [44]. Deepface is a python framework for facial recognition and facial attribute analysis (age, gender, emotion, and race). The framework relies on several state-of-the-art models such as VGG-Face, Google FaceNet, OpenFace, Facebook Deep-Face, DeepID, ArcFace, and Dlib. We focused on facial attribute analysis for age, gender, and race to determine if any bias toward certain attributes was present. A subset of real and deepfake images was analyzed using the Deepface framework and the results are displayed in Figs. 3, 4, and 5.

The results of the bias analysis show that some protected classes are not equally represented. Age is concentrated between 10 and 60 years old. This leaves individuals with

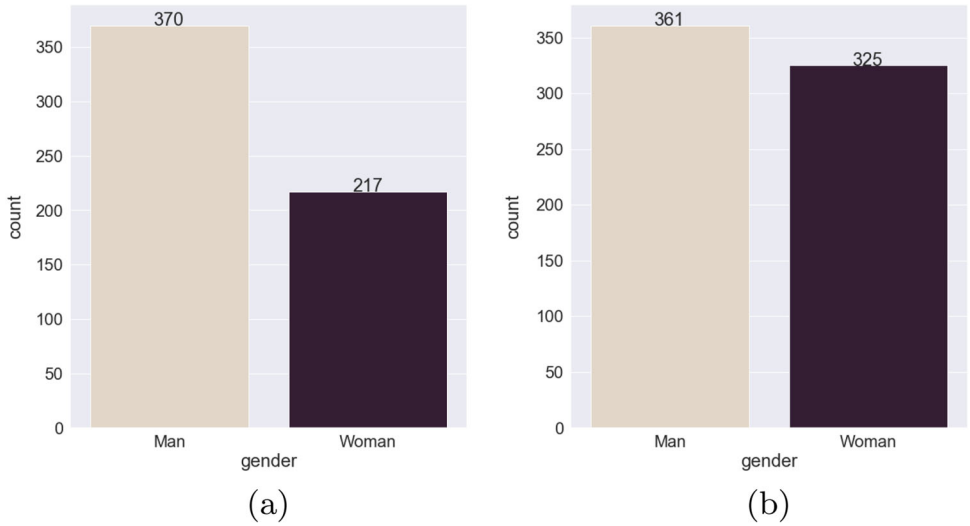


Fig. 3 **a** Indicates DFIM-HQ real images gender bias analysis. **b** Indicates DFIM-HQ Deepfake images gender bias analysis

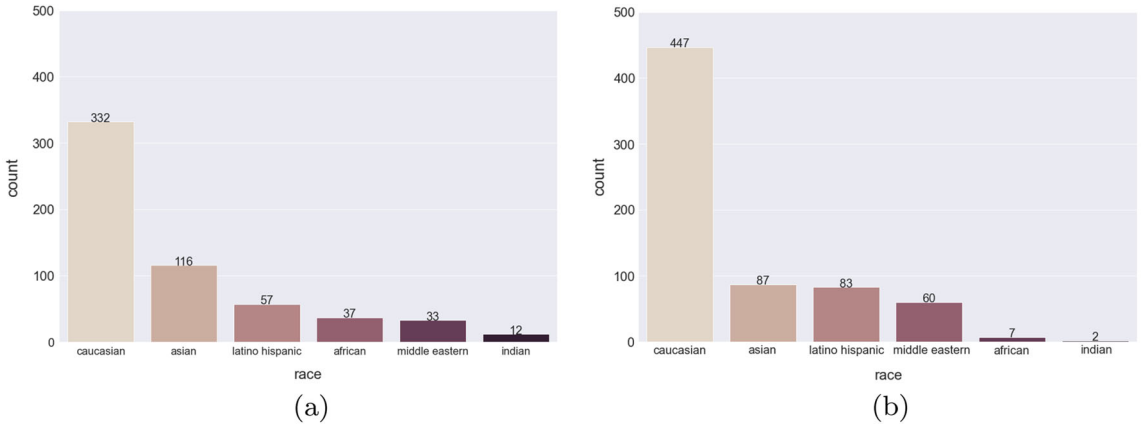


Fig. 4 **a** Indicates DFIM-HQ real images race bias analysis. **b** Indicates DFIM-HQ deepfake images race bias analysis

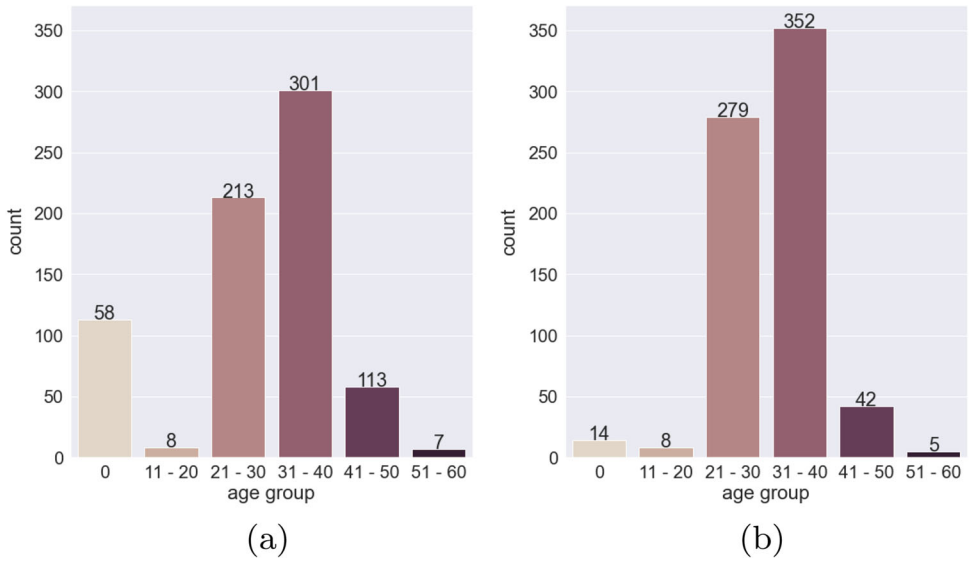


Fig. 5 **a** Indicates DFIM-HQ Real Images Age Bias Analysis. **b** Indicates DFIM-HQ Deepfake Images Age Bias Analysis

age less than 10 or greater than 60 not represented in the dataset. However, this could be a result of the Deepface tools prediction algorithm for age. Further exploration is needed to determine if the tool has an impact on age predictions and under-representation. Men are the dominant class in both deepfake images and real images. The biggest gap between men and women is present in the real images dataset with women only representing 31% of the images in the tested subset. Concerning race, the prominent race for both real images and deepfake images appears to be Caucasian population. This leaves the racial groups of Asian, Latino Hispanic, Middle Eastern, African, and Indian under-represented. Indians and Africans are the two least represented classes.

AI bias mitigation analysis

Mitigating bias present in a dataset is essential to ensure the resulting model does not contain a bias toward a specific age, race, or gender. Several metrics help calculate bias as detailed in IBM's AI 360 Fairness tool [45]. The metrics used to assess bias are statistical parity difference, equal opportunity difference, average odds difference, disparate impact, and thiel index. Three of these metrics were used to calculate age, race, and gender results on the DFIM-HQ dataset and the results are shown in Tables 3 and 4.

For gender, the privileged class was man; for race, the privileged class was Caucasian; and for age, the privileged class was 31 - 40. The acceptable fair range for statistical parity difference is between −0.1 and 0.1. The acceptable fair range for disparate impact is between 0.8 and 1.25. The acceptable fair range for equal opportunity difference is between −0.1 and 0.1, respectively. Steps can be taken to mitigate bias using techniques such as reweighting or adversarial debiasing [45]. Some methods are applied directly to the training dataset, while others are applied to the trained model or model predictions. The proposed DFIM-HQ dataset falls under acceptable fair range for all the three metrics in the category of age and gender. For the DFIM-HQ dataset, we propose using adversarial debiasing to reduce the bias present in age, race, and gender. This technique helps create a fair model without discrimination toward certain protected classes that an adversary could exploit. This technique can also be utilized when removing the attribute causing bias from the dataset is difficult.

In our case, race cannot be removed from the dataset since it is an inherent property of the image. Adversarial debiasing [46] is applied directly to the model where the model is trained to predict if the image is a deepfake or not while conversely preventing a jointly trained adversary from predicting the race attribute. We recommend performing mitigation on the trained model instead of the dataset to preserve the integrity of the images. This means that reweighting is not an applicable technique that can be applied to our dataset to mitigate bias. Reweighting works by creating a weights vector for our protected attribute and output variable and, as a result, does not apply to our image dataset. Reject option-based classification is a mitigation technique that is directly applied to the model predictions [45]. This technique changes the prediction probabilities by adjusting them to remove bias. This technique can be applied to our model since the bias mitigation occurs at the prediction step. However, since the prediction probabilities are changed, the result may increase false positives or negatives for our deepfake model. Therefore, we recommend first applying adversarial debiasing and measuring the performance on reduced bias. An additional step that could be taken to mitigate bias before training is balancing the dataset by under-sampling images for specific ages, races, or genders. This would work best on the gender classes. Removing a small percentage of men could help balance the dataset between men and women. We propose using adversarial debiasing before taking this step as undersampling reduces the size of training data available.

Table 3 Bias metrics analysis on DFIM-HQ dataset using Meso-Net model

Meso-Net	Age	Race	Gender
Statistical parity difference	0.0376	0.1615	−0.0873
Equal opportunity difference	−0.0164	−0.0006	0.0167
Disparate impact	1.0787	1.3791	0.8254

Table 4 Bias Metrics Analysis on DFIM-HQ Dataset using MesoInception-Net model

MesoInception-Net	Age	Race	Gender
Statistical parity difference	0.0531	0.16611	0.0728
Equal opportunity difference	0.00411	−0.0026	−0.0008
Disparate impact	1.1125	1.3872	1.1391

Proposed detection framework with explainable AI

In this section, we present our detection framework to spot digital manipulation in high-quality deepfake images. We also apply Explainable AI to generate visual explanations for our models' decisions and understand its predictions.

Deepfake detection methodology

We employ two deep learning frameworks for obtaining a baseline on our DFIM-HQ image dataset. One of our networks consists of a sequence of four layers of successive convolutions, batch normalization, and max pooling. The framework is called the Meso-Net model and is a dense

Table 5 A table showing hyperparameter values used for Meso-Net and MesoInception-Net

Hyper-parameter	Value
Number of epochs trained	50
Number of steps/training epoch	250
Number of steps/validation epoch	125
Learning rate	0.001
Activation function between layers	ReLU
Activation at last layer	Sigmoid

Table 6 Training and validation metrics for MesoInception-Net and Meso-Net

Model	MesoInception-Net		Meso-Net	
	Train	Validation	Train	Validation
ACC (%)	99.65	99.87	94.88	94.52
Precision	0.9971	1	0.9393	0.9045
Recall	0.9963	0.9974	0.9615	0.9940
BCE	0.0214	0.01653	0.1395	0.1341
MAE	0.0062	0.0013	0.0821	0.07667
RMSE	0.0062	0.196	0.0821	0.07667
R^2	0.9895	0.9961	0.8352	0.8376
F-1	0.9966	0.9986	0.9484	0.9453

network with one hidden layer [32]. The convolutional layers make use of ReLU activation functions to introduce non-linearities. Our framework uses batch normalization to regularize their output and to prevent the vanishing gradient effect. Additionally, l_2 regularization of 0.01 and drop out of 0.5 are included to regularize and improve the network robustness. Our other baseline framework is an alternative structure of the Inception model. It replaces the first two convolutional layers of the Meso-Net model with a variant of the Inception model. The underlying concept is to stack the output of several convolutional layers with different kernel shapes, thereby increasing the function space in which the model is optimized. The network makes use of dilated convolution to deal with multi-scale information. Both our networks are trained for 50 epochs each, with 250 training steps per epoch, and 125 validation steps per epoch. We used sigmoid as our activation function and optimized the binary cross-entropy loss using the Adam optimizer. Since the output of the sigmoid function is restricted between [0,1], we used 0.5 as the threshold value to determine the classification result. Any image that results in an output of greater than or equal to 0 but less than or equal to 0.5 gets classified as a deepfake image, whereas an image that results in an output of greater than 0.5 but less than or equal to 1 gets classified as a real image. Table 5 summarizes our hyper-parameters used for both our networks.

We present an overview of our deepfake detection framework in Fig. 6. The input to our detection framework can be an image in any format. We preprocess the image to a resolution of 256x256 pixels and pass it through our convolution networks MesoInception-Net and Meso-Net. We record the prediction output from the last layer of our CNNs. Parallelly, we employ an explainable framework using Grad-CAM that extracts visual explanations from the last convolution layers of both MesoInception-Net and Meso-Net. The explanations are generated in the form of heatmaps, indicating what our models visualize to produce a prediction outcome. Finally, we analyze the prediction output from our CNNs along with corresponding explainable heatmaps to generate a final output of whether an image is authentic or a deepfake.

Explainability in deepfakes

Machine learning models are known to operate as a black box. A black box is a system performing input and output operations without providing true knowledge of its internal workings. The cause and reason for decisions provided by machine learning models are often hard to explain, creating multiple concerns, one of them being the trust issues of using the model in the real world. This creates a need for "Explainability," a technique that provides an understanding of how an AI-assisted system or model generated an output. AI models arriving at conclusions must be interpretable to ensure they perform without any unexpected bias, provide algorithmic fairness, and help users understand when and why a model generated an incorrect prediction [46–48].

Explainability can be implemented by using Gradient Weighted Class Activation Mapping (Grad-CAM) [49] and Local Interpretable Model-Agnostic Explanations (LIME) [50], among other explainability techniques [51]. Grad-CAM is a method to produce visual explanations to outputs of CNNs. This method generates a heat map from the second-last layer of a CNN to create a class-discriminative localization map. This map highlights pixels whose intensity should be increased to increase the chances of classification of the desired class. Hence, the results of Grad-CAM highlight important pixels relevant to a specific class. Grad-CAM is a generalization of CAM [51], as the latter is also a technique to view visualizations; however, it can only be applied to CNNs of a particular kind, whereas the latter can be applied to any deep neural network performing image or even textual classification in general.

For our use case, when a deepfake classification model classifies media files as "real" or "deepfake," explainability outputs a mapping to illustrate important features used by the model to perform the classification. In this manner, our methods use explainability to further understand and correct misclassifications generated by the deepfake classification model, thereby increasing the model's robustness along with increased trust

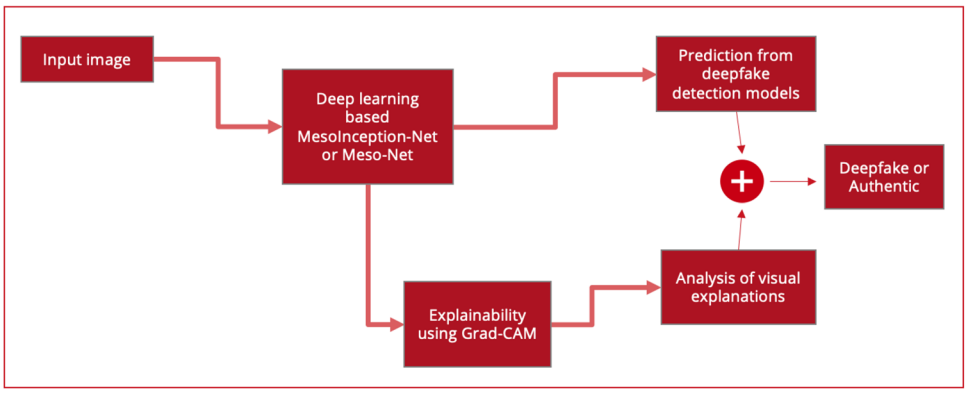


Fig. 6 An overview of our proposed deepfake detection framework showing how we combine Explainable AI and predictions from deep learning models to conclude if an image is a deepfake or an authentic image

in classifier predictions. Grad-CAM provides visual explanations for deep learning models. Specifically, in our case, Grad-CAM highlights areas of an image that MesoInception-Net and Meso-Net consider important when classifying the image as a deepfake image or a real image. Explainability adds value to the decisions of our models by helping us understand what factors influenced them to distinguish deepfake images from real images. By carefully observing the heat map generated, we can avoid misclassifications and create robust deepfake detection models.

Experimental results and analysis

This section covers a discussion of results from our experiments. We include detailed analysis from our results and provide evaluation metrics.

Benchmark results and analysis

We trained the models Meso-Net and MesoInception-Net on the training dataset as given in Table 2 for 50 epochs. Based on the various runs, we chose the best-performing weights of our models to be at epoch 43 for MesoInception-Net, and at epoch 44 for Meso-Net. Our best-performing weights are validated on the validation set. Additionally, we calculated other metrics of evaluation such as Binary Cross-Entropy loss (BCE), Accuracy (ACC), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2), Recall, Precision, and F-1 score. Our metrics are documented in Table 6.

Comparing our models, we see that MesoInception-Net is our best-performing model, giving a validation accuracy of 99.87%. In comparison, Meso-Net generated a validation accuracy of 94.52%. The corresponding ROC curves and Precision–Recall (PR) curves for both MesoInception-Net and Meso-Net are shown in Figs. 7 through 12. The Accuracy Under the Curve (AUC) of MesoInception-Net is observed to be 0.99973 and that of Meso-Net is observed to be 0.99333 for validation data.

Explainability (XAI) results

To better understand our models' predictions, we randomly sampled high-quality real and deepfake images and the explainability framework allows us to observe activations for each class. Figure 13 illustrates Grad-CAM activation maps on two real images using MesoInception-Net model. Similarly, Fig. 14 showcase two StyleGAN generated fake image and corresponding Grad-CAM visual activations. The results in Fig. 13 are significantly different from those in Fig. 14. For the fake image of a man with spectacles shown in Fig. 14a, the Grad-CAM map highlights the hair and spectacles, indicating the MesoInception-Net model considers them to be the most important features when classifying the image as fake. Similarly, consider the fake image shown in Fig. 14b: the Grad-CAM map highlights the eyes, the lips, and the hair of the woman indicating they are seen as significant features to classify the image as a fake image. Comparing facial feature activations with the Grad-CAM maps shown for real images in Fig. 13a, we notice no such activations. In contrast, for the real image of the man with spectacles in Fig. 13a, the Grad-CAM map shows an outline of the spectacles, nose, and the background as important features while classifying the image as a real image. Similarly, for the real image of a girl in Fig. 13b, the corresponding Grad-CAM map shows a similar outline of her nose and cheeks.

Conclusions

We introduced a large-scale challenging dataset for the development and evaluation of deepfake detection methods. There are, in total, 140,000 high-resolution images in the DFIM-HQ dataset. The DFIM-HQ dataset reduces the gap in the visual quality of deepfake datasets and the actual deepfake images circulated online on social media platforms. The

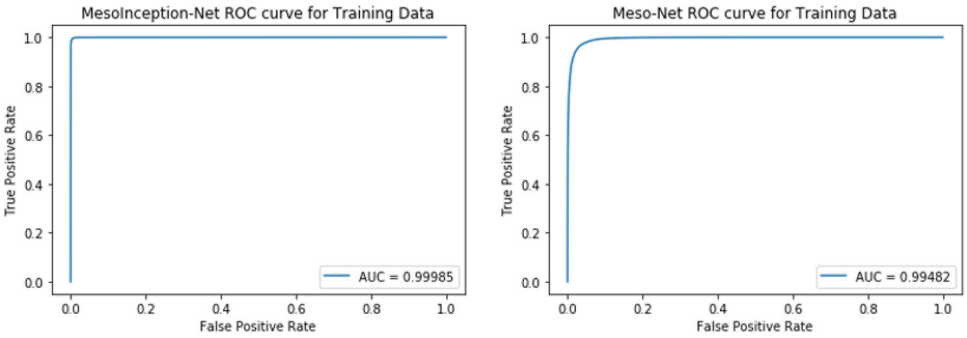


Fig. 7 ROC curves on training dataset

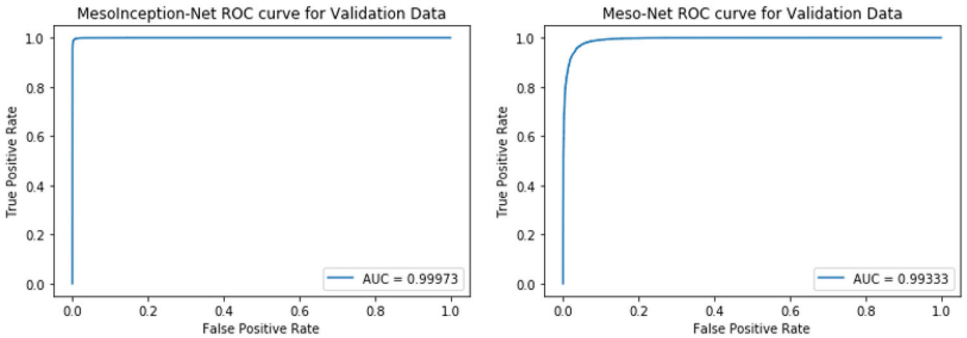


Fig. 8 ROC curves on validation dataset

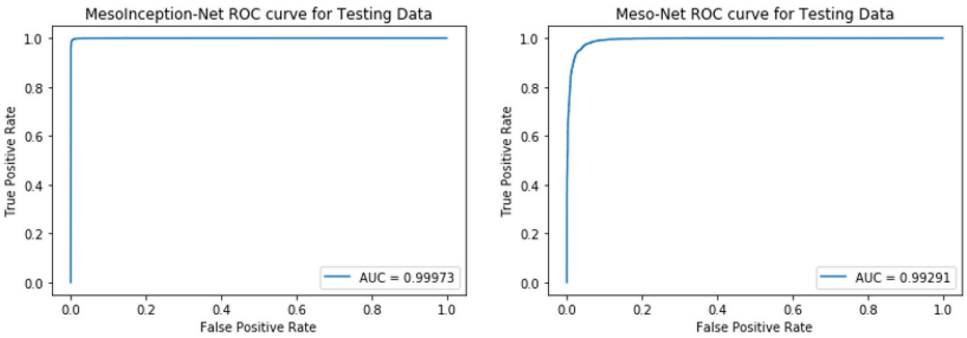


Fig. 9 ROC curves on testing dataset

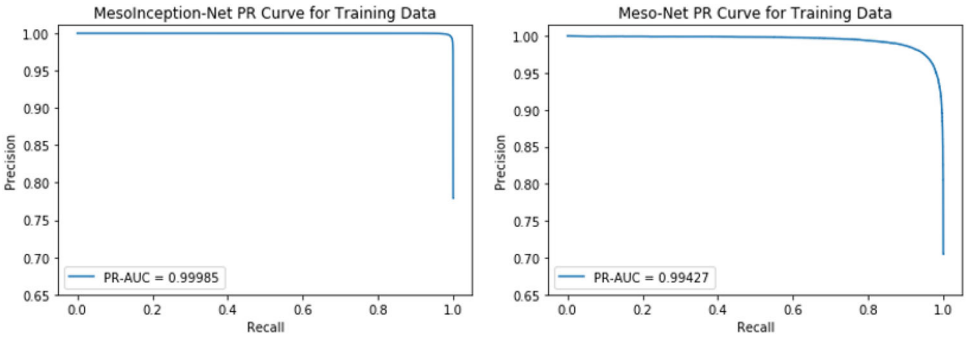


Fig. 10 Precision–Recall curves on training dataset

overall visual quality of these high-resolution images in the DFIM-HQ dataset is superior when compared to existing datasets, with significantly fewer notable visual artifacts.

With recent work on AI Bias, we acknowledge that the presence of spurious bias in the training data can severely degrade the accuracy of current models, even when the biased dataset contains more information than an unbiased dataset. Based on our proposed DFIM-HQ dataset, we set out to provide an in-depth look at the problem of training visual classifiers to ensure model fairness and study AI bias mitigation. In order to detect and mitigate AI bias, we perform a thorough analysis of bias mitigation techniques in visual recognition models and draw several important algorithmic conclusions. Our work helped bridge the gap, proposing an avenue for exploring mitigating bias in deep learning models within a simpler and easier-to-analyze setting.

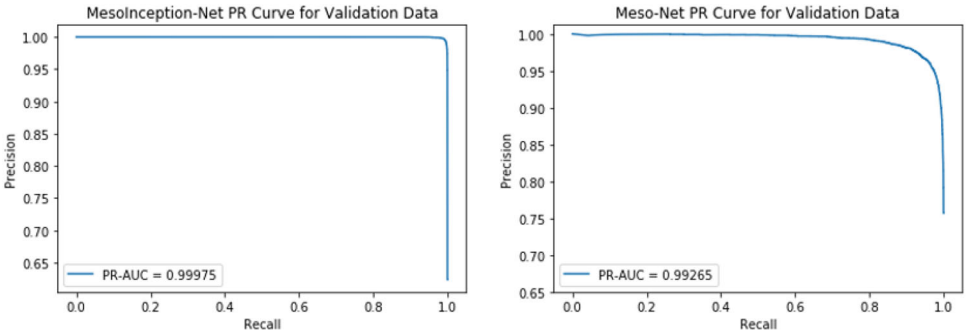


Fig. 11 Precision–Recall curves on validation dataset

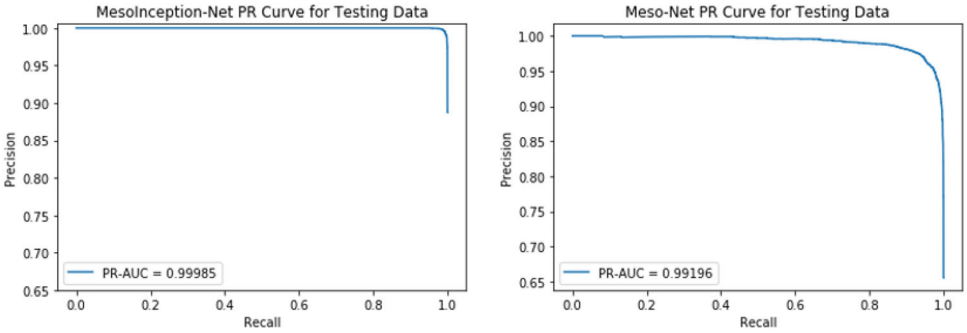
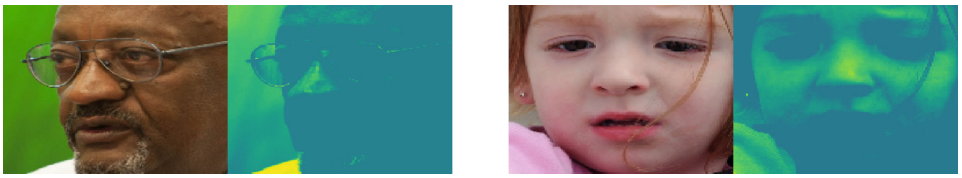
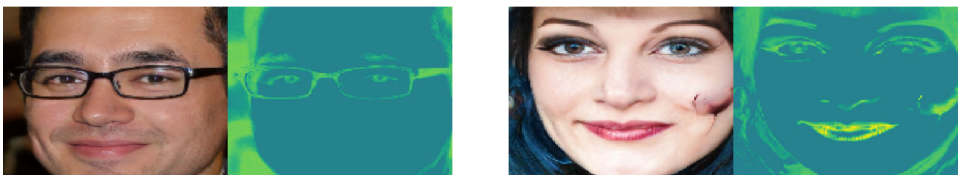


Fig. 12 Precision–Recall curves on testing dataset



(a) XAI results on real image 1 (b) XAI results on real image 2

Fig. 13 XAI results on sample real images



(a) XAI results on deepfake image 1 (b) XAI results on deepfake image 2

Fig. 14 XAI results on sample deepfake images

Preliminary results with the proposed deep learning detection network and explainability framework are very promising and show that this kind of system seems to point out some existing inhomogeneities between the two analyzed classes. This evidence paves the way for many possible future works: firstly, to evaluate the forensics benefits that can be obtained by addition of explainability frameworks for deepfake content identification by testing against more datasets and with other deep learning networks. Secondly, it would be prudent to further study how this approach could exploit inconsistencies on the temporal axis and be combined with well-known state-of-the-art frame-based methodologies to improve their performances. Lastly, we believe that our work paves the way to encourage researchers to develop more powerful detection techniques, anti-forensic techniques, and countermeasures against real-world deepfakes.

Data availability The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Eyerys, An AI Capable In Creating Fake Porn, Is Starring Gal Gadot And More: A Terrifying Implication https://www.eyerys.com/articles/news/ai-capable-creating-fake-porn-starring-gal-gadot-and-more-terrifying-implication
2. Cnet, Jordan Peele turns Obama into foul-mouthed fake-news PSA. https://www.cnet.com/news/jordan-peele-buzzfeed-turn-obama-into-foul-mouthed-fake-news-psa/
3. Motherboard, This Deepfake of Mark Zuckerberg Tests Facebook's Fake Video Policies, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
4. The Wall Street Journal, Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
5. Knight W. AI-powered text from this program could fool the government [Internet]. Wired. Conde Nast; 2021 [cited 2021Apr29]. Available from: https://www.wired.com/story/ai-powered-text-program-could-fool-government/
6. Photo Tampering Throughout History. https://web.archive.org/web/20150908155915/, http://www.cc.gatech.edu/~beki/cs4001/history.pdf
7. FakeApp. https://www.malavida.com/en/soft/fakeapp/
8. Faceswap. https://github.com/deepfakes/faceswap#deepfakesfaceswap
9. Dfaker. https://github.com/dfaker/df
10. Petrov I, Gao D, Chervoniy N, Liu K, Marangonda S, Umé C, Jiang J, RP L, Zhang S, Wu P, Zhang W (2020) Deepfacelab: a simple, flexible and extensible face swapping framework. arXiv preprint arXiv:2005.05535
11. DeepFaceLab. https://github.com/iperov/DeepFaceLab
12. Introducing the new SAEHD model. https://www.reddit.com/r/SFWdeepfakes/comments/dfgv68/introducingthenewsaehdmodel/
13. Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M (2016) Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2387-2395)
14. https://lingzhili.com/FaceShifterPage/
15. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inform Fus 1(64):131–48
16. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4401-4410)
17. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196
18. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8789-8797)
19. Li Y, Chang MC, Lyu S (2018) In ictu oculi: exposing ai created fake videos by detecting eye blinking. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-7). IEEE
20. Li Y, Lyu S (2018) Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656
21. Zhou P, Han X, Morariu VI, Davis LS (2017) Two-stream neural networks for tampered face detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 1831-1839). IEEE
22. Yang X, Li Y, Qi H, Lyu S (2019) Exposing gan-synthesized faces using landmark locations. In: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security (pp. 113-118)
23. Li H, Li B, Tan S, Huang J (2020) Identification of deep network generated images using disparities in color components. Signal Process 1(174):107616
24. Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H (2019) Protecting world leaders against deep fakes. InCVPR Workshops (pp. 38-45)
25. https://www.darpa.mil/program/media-forensics
26. https://ai.facebook.com/datasets/dfdc/
27. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Ferrer CC (2020) The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397
28. https://niessnerlab.org/projects/roessler2019faceforensicspp.html
29. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. (2019) Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1-11)
30. https://github.com/socialabubi/iFakeFaceDB
31. iFakeFacesDB Dataset. https://github.com/socialabubi/iFakeFaceDB
32. Notre Dame Synthetic Face Dataset. https://cvrl.nd.edu/projects/data/
33. Banerjee S, Scheirer WJ, Bowyer KW, Flynn PJ (2019) Fast face image synthesis with minimal training. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 2126-2136). IEEE
34. http://cvlab.cse.msu.edu/dffd-dataset.html
35. Dang H, Liu F, Stehouwer J, Liu X, Jain AK (2020) On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5781-5790)
36. Guan H, Kozak M, Robertson E, Lee Y, Yates AN, Delgado A, Zhou D, Kheyrkhah T, Smith J, Fiscus J (2019) MFC datasets: large-scale benchmark datasets for media forensic challenge evaluation. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW) Jan 7 (pp. 63-72). IEEE
37. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2018)Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179
38. Neves JC, Tolosana R, Vera-Rodriguez R, Lopes V, ProençSa H, Fierrez J (2020) Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. IEEE J Select Top Signal Process 14(5):1038-1048
39. Zi B, Chang M, Chen J, Ma X, Jiang YG (2020) WildDeepfake: a challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM International Conference on Multimedia (pp. 2382-2390)
40. http://cvlab.cse.msu.edu/project-ffd.html
41. LFW dataset. http://vis-www.cs.umass.edu/lfw/
42. Dang H, Liu F, Stehouwer J, Liu X, Jain AK (2020) On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5781-5790)
43. FFD Dataset. http://cvlab.cse.msu.edu/project-ffd.html
44. Serengil SI, Ozpinar A (2020) LightFace: a hybrid deep face recognition framework. In: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU) Oct 15 (pp. 1-5). IEEE

45. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, Nagar S (2018) AI Fairness 360: an extensible toolkit for detecting. Understanding, and Mitigating Unwanted Algorithmic Bias
46. Ribeiro MT, Singh S, Guestrin C (2016) Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386
47. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) Oct 1 (pp. 80-89). IEEE
48. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592
49. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (pp. 618-626)
50. Ribeiro MT, Singh S, Guestrin C (2016)"Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining Aug 13 (pp. 1135-1144)
51. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2921-2929)
52. Afchar D, Nozick V, Yamagishi J, Echizen I (2018) Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS) Dec 11 (pp. 1-7). IEEE
53. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1–9)
54. Mathews SM (2019) Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. In: Intelligent computing-proceedings of the computing conference. Springer, Cham, pp 1269–1292