Saahil Sood
ss6358

Efficacy of Deepfake Detection Methods

1. Synopsis

1.1 Overview

As of today, deepfaked images and videos are incredibly realistic to the point that the average person may have trouble distinguishing between them and real images and videos. Although generating the most sophisticated deepfakes takes time, money, and a powerful GPU, the advances in both deep neural networks and convolutional neural networks in recent years has made it easier for the average person to generate less sophisticated, yet realistic deepfakes than before. In fact, generating deepfakes can be done on an iphone. As a result of the rapid advances in deepfake-generation technology, deepfake detectors and scanners have to be constantly updated. This is not to say that current scanners are not good; in fact, they are reliable but they will easily become obsolete if they are not updated constantly. In my project, I aim to test the efficacy of two reputable deepfake scanners, MesoNet and DeepwareAI, with many different types of deepfakes and see the strengths and weaknesses of both. Additionally, for the artificially generated face type of deepfake, I aim to see if a deepfake scanner is best at detecting a certain facial feature over others.

1.2 Novelty

There have definitely been studies regarding deepfake scanners in the past, but these studies involved only a small subset of deepfakes (only faceswap deepfakes, for example). However, deepfake technology advances quickly, and so do the types of deepfakes. In my project, I compare two deepfake scanners using all of the image-based types of deepfakes we have today. These deepfakes include faceswaps, Face2Face, and artificially generated faces.

In addition, since artificially generated faces are the newest type of deepfake, I test which facial features, expressions, or other properties make it easiest for the deepfake detector to detect that the artificially generated face is a deepfake.

1.3 Value to our community

Deepfakes are becoming more realistic and easier to generate than in the past. In fact, some sources suggest that deepfake generation techniques are advancing faster than methods to detect deepfakes. As a result, deepfake detectors remain effective for only limited periods of time; they must constantly be updated to remain effective. For example, the metrics scanners use to detect deepfakes today may become irrelevant in a few years as deepfake generators eliminate the flaws in deepfakes that deepfake detectors rely on. My project will help researchers determine the strengths and weaknesses of current deepfake detectors and get a better understanding of how to improve existing models. That is, if my project indicates that the scanners are not skilled at detecting a particular type of deepfake, researchers can take steps to improve upon them. In addition, my project will compare an image-based scanner with a video-based one. One is not better than the other, but the data from my project can help researchers in understanding the intrinsic strengths and weaknesses of each. By understanding

this, researchers can learn when to use one type of scanner over another in a particular situation.

2. Project

2.1 Research Questions

RQ1: How effectively can deepfake detectors classify real images and videos and different types of deepfaked images and videos?
RQ2: In general, what are the strengths and weaknesses of video-based deepfake detectors and image-based deepfake detectors?
RQ3: Can deepfake detectors readily detect that an artificially generated face/deepfake with a particular facial feature(s) or expression(s) is a deepfake?

To answer all four research questions, I tested the effectiveness of two reputable deepfake scanners, MesoNet and DeepwareAI (also known as Deepfake scanner), at detecting deepfaked images and videos and real (non-deepfaked) images and videos. The following sections explain my methods and findings.

2.2 Project

2.2.1 Background on Deepfakes

There are 4 types of deepfakes I investigated in this project: Lightweight Faceswap Deepfakes, Heavyweight Faceswap Deepfakes, Face2Face, and artificially generated faces. The first two fall under the category of faceswap deepfakes, but they vary in complexity. Lightweight faceswap deepfakes are created using a simple approach where one extracts the face region from one video and places it on the face in the target video. Simple algorithms are used to blend the two faces if necessary. This type of deepfake can be generated on CPUs (as opposed to GPUs) because the approach is so lightweight. Heavyweight faceswap deepfakes are more complex than lightweight faceswaps. These faceswaps involve using two different autoencoders with a shared encoder but distinct decoders. Compared to Lightweight Faceswap Deepfakes, the swapping process is smoother and less error-prone. Face2Face deepfakes are created by transferring the facial expressions of a person from one video to a person in the target video while preserving the identity of the person in the target video. The result is a person with the same identity but different facial expressions. Finally, there are artificially generated face deepfakes. These deepfakes are the most recent type of deepfake. They are generated from Generative Adversarial Networks that have been trained to reproduce very realistic human faces. As a result, these deepfakes appear incredibly realistic, but they still have minor flaws that differentiate them from actual human faces.

2.2.2 Datasets for Research Questions 1 and 2

In order to get the data I needed for this project, I borrowed videos from FaceForensics++ and images from https://this-person-does-not-exist.com/en. More specifically, I got videos for heavyweight faceswap deepfakes, lightweight faceswap deepfakes, Face2face deepfakes, and real (unmodified) videos from FaceForensics++ and got images for artificially generated faces deepfakes from https://this-person-does-not-exist.com/en. I did not necessarily use the data in its original form, however. I had to adjust and modify the data to account for the differences between the datasets and the two scanners. In summary, to test MesoNet, I used images from the folders in my Github (https://github.com/ssood123/Efficacy-of-Deepfake-Detection-Methods):

- ArtificiallyGeneratedFaceImagesAndVideos/ArtificiallyGeneratedFaceImages
- Face2FaceImagesAndVideos/Face2FaceImagesFACE
- HeavyweightFaceswapImagesAndVideos/HeavyweightFaceswapImagesFACE
- LightweightFaceswapImagesAndVideos/LightweightFaceswapImagesFACE
- RealImagesAndVideos/RealImagesFACE

In order to test DeepwareAI, I used videos from the folders in my Github:

- ArtficiallyGeneratedFaceImagesAndVideos/ArtificiallyGeneratedFaceVideos
- Face2FaceImagesAndVideos/Face2FaceVideosFACE
- HeavyweightFaceswapImagesAndVideos/HeavyweightFaceswapVideosFACE
- LightweightFaceswapImagesAndVideos/LightweightFaceswapVideosFACE
- RealImagesAndVideos/RealVideosFACE

The video version of a folder and an image from a folder contain the same data but in different forms. For example, RealVideosFACE and RealImagesFACE contain the same data in different forms. Before reading on, it is suggested to read both Appendix section 6.1 (for a more detailed explanation of how and why I generated the data I used to test the scanners in my project) and the Github repo I provided in the deliverables section (because I refer to folders in my Github repo throughout the project).

2.2.3 Metrics for Research Questions 1 and 2

MesoNet is an image-based scanner that takes in an image and outputs a decimal between 0 and 1. The closer the number is to 1, the more confident MesoNet is that the image is a deepfake. This scale is linear. Therefore, an output of 0 indicates that MesoNet is sure that the image is real while an output of 1 indicates that MesoNet is certain that the image is a deepfake. For each image that MesoNet read and analyzed, I assigned a label (real/fake) and a confidence level (low, medium, high). If MesoNet outputted:

- 0 to 5/30 -> real with high confidence
- 5/30 to 10/30 -> real with medium confidence
- 10/30 to 15/30 -> real with low confidence
- 15/30 to 20/30 -> fake with low confidence
- 20/30 to 25/30 -> fake with medium confidence
- 25/30 to 30/30 -> fake with high confidence

I created a table to organize my results. Additionally, I created a table to depict

- the total number of reals (total number of real with high confidence outputs + total number of real with medium confidence outputs + total number real with low confidence outputs)
- the total number of fakes (total number of fake with low confidence outputs + total number of fake with medium confidence outputs + total number fake with high confidence outputs)

Finally, I made a table depicting:

- total number of high-confidence outputs (total number of real with high confidence outputs + total number of fake with high confidence outputs)
- total number of medium-confidence outputs (total number of real with medium confidence outputs + total number of fake with medium confidence outputs)
- total number of low-confidence outputs (total number of real with low confidence outputs + total number of fake with low confidence outputs)

DeepwareAI is a video-based scanner that takes a video and also outputs a decimal between 0 and 1. Unlike MesoNet, the closer the number is to 1, the more confident DeepwareAI is that the video is *real*. This scale is also linear. An output of 0 indicates that DeepwareAI is sure that the video is real while an output of 1 indicates that DeepwareAI is sure that the video is a deepfake. I analyzed the results of DeepwareAI the same way I analyzed the results of MesoNet except with the consideration that an output of 1 indicates a deepfake prediction and an output of 0 indicates a real prediction. Also, unlike MesoNet, DeepwareAI generates a results.csv file which stores the results of its predictions. I used the .csv files to analyze the results.

2.2.4 Research Question 1: How effectively can deepfake detectors classify real images and videos and different types of deepfaked images and videos?

Please refer to Appendix sections 6.2 and 6.3. Figures 1-10 are relevant for this research question.

The figures relevant to MesoNet are figures 1,2,3,4, and 5.

Figure 1 shows that for images in the "RealImagesFACE" folder, MesoNet classified about 77% of images as real. It was confident that most images were real as about 43% of images fell under the "real with high confidence" category.

Figure 2 shows that for the images in the "HeavyweightFaceswapImagesFACE" folder, most of the time MesoNet thought the images were fake (about 85% of images were classified as fake). It was confident that these images were fake (about 52% of the images fell under the category fake with high confidence).

Figure 3 reveals that for the images in the "Face2FaceImagesFACE" folder, MesoNet thought that the images were real most of the time (about 80% of images were classified as real). It also thought that the images were real with high confidence about 45% of the time. Therefore, it can be concluded that overall it is confident that the images were real.

Figure 4 reveals that for images in the "LightweightFaceswapImagesFACE" folder, most of the time MesoNet thought the images were fake (about 95% of images were classified as

fake). It was also very confident that most of the images were fake (about 77% of the images fell under the fake with high confidence category).

Finally, figure 5 reveals that for the images in the "ArtificiallyGeneratedFaceImages" folder, MesoNet thought that the majority of images were real (about 92% of images were classified as real). It was also very confident that the images were real because about 80% of the images fell under the real with high confidence category.

For the real images, heavyweight faceswap deepfakes, and face2face deepfakes, MesoNet predicted about 50% of images with high confidence, 25% of images with medium confidence, and 25% of images with low confidence. For the lightweight faceswap and artificially generated face deepfakes, Mesonet predicted the vast majority of images with high confidence.

Figures 6,7, 8, 9 and 10 are relevant to DeepwareAI.

Figure 6 reveals that for videos in "RealVideosFACE" (real videos), DeepwareAI classified the videos as real about 81% of the time and with confidence because about 51% of videos fall under the real with high confidence category.

Figure 7 reveals that for videos in "HeavyweightFaceswapVideosFACE", DeepwareAI predicted the videos to be fake about 86% of the time and with high confidence (70% of videos fall under fake with high confidence category).

Figure 8 reveals that for videos in "Face2FaceVideosFACE", DeepwareAI predicted the videos to be real about 63% of the time and with moderate confidence (only about 37% of the videos fall under the real with high confidence category).

Figure 9 reveals that for videos in "LightweightFaceswapVideosFace", DeepwareAI predicted the videos to be fake about 61% of the time and with moderate confidence (only about 32% of the videos fall under the fake with high confidence category).

Figure 10 reveals that for videos in "ArtificiallyGeneratedFaceVideos", DeepwareAI predicted the videos to be real exactly 100% of the time and with very high confidence (about 99% of images fell under the real with high confidence category).

For the heavyweight faceswap and artificially generated faces deepfake types, DeepwareAI predicted the majority of videos with high confidence. For face2face deepfakes, DeepwareAI predicted about 50% of the videos with high confidence, 25% of the videos with medium confidence, and 25% of the videos with low confidence. For the lightweight faceswap deepfake type, DeepwareAI predicted the results with slightly lower confidence overall than face2face. For the real videos, DeepwareAI predicted the results with slightly higher confidence overall than face2face.

Based on these statistics, it can be concluded that the deepfake detectors (both video-based and image-based) are skilled at detecting faceswap-based deepfakes but not skilled at detecting other types of deepfakes. The most likely reason for this is that faceswap-based deepfakes are easier to detect than face2face and artificially generated deepfakes. Even though heavyweight faceswap deepfakes are more realistic than their lightweight counterparts, both types involve swapping a face with another face. Although the swapped face can appear realistic, the swapping process is error prone and introduces flaws that are not present in real images. face2face images are less flawed than faceswap-based

deepfakes because they may change the expression of a person but still preserve the person's original identity unlike faceswap-based deepfakes. As a result, face2face images may appear more realistic than faceswapped images. Both scanners classified more images or videos in the face2face category as real than fake. Although the results are far from ideal, one reason for this is that the transformations I applied to the raw videos and images from FaceForensics++ and https://this-person-does-not-exist.com/en somehow decreased the quality of the images/videos, making is more difficult for both scanners to classify the images or videos accurately. Artificially generated face images were the hardest type for both scanners to detect. In fact, both scanners thought that most, if not all, images or videos in this category were real. This makes sense because styleGAN is able to generate incredibly realistic-looking faces that are sometimes indistinguishable from real images. Unlike the other 4 datasets, I did not apply as many transformations to the images in the "ArtificiallyGeneratedFaceImages" folder, so I believe that the classification results for this deepfake category are highly accurate.

Also, if we look at the results of both scanners for the real images and videos, we can see that both predicted correctly about 80% of the time. DeepwareAI predicted the videos with slightly higher confidence overall than MesoNet did with images, however. These statistics indicate that deepfake detectors are good at classifying real images as real. However, I believe that both of these scanners should've done better (classify a higher percent of images or videos as real). The reason is that both detectors classified a higher percent of artificially generated face images and videos as real than real images or videos. This shouldn't be the case because even though artificially generated face images are realistic, they sometimes contain features that give away the fact that they are a deepfake. A possible explanation for this discrepancy is that the videos downloaded from FaceForensics++ had lower resolution than the images downloaded from https://this-person-does-not-exist.com/en. Another explanation is that many transformations were applied to the videos from the Zenodo datasets while only FFMPEG was applied to the images from https://this-person-does-not-exist.com/en. Each transformation might have the effect of decreasing the resolution of an image or video. In either case, the real images and videos likely had worse resolution than the artificially generated face images and videos, resulting in both scanners having more difficulty in classifying real images and videos accurately.

2.2.5: Research Question 2: In general, what are the strengths and weaknesses of video-based deepfake detectors and image-based deepfake detectors?

We can use the results and analysis done in 2.2.4 to answer this research question. Based on the MesoNet results, the order of how well this scanner can detect different deepfake types from best to worst is: Lightweight faceswap, Heavyweight faceswap, face2face, and artificially generated faces. For DeepwareAI, the order is (from best to worst once again): Heavyweight faceswap, Lightweight faceswap, face2face, and artificially generated faces. However, this doesn't mean that DeepwareAI is better at accurately classifying Heavyweight faceswap deepfakes than MesoNet. In fact, DeepwareAI is just barely better at detecting heavyweight faceswap deepfakes than MesoNet. Overall, MesoNet performed better than DeepwareAI. It seems that MesoNet had more accurate predictions and predicted with higher confidence most

of the time. There are some cases where DeepwareAI performed better than MesoNet (For example, it detected face2face deepfakes more accurately than MesoNet), but MesoNet had better results overall. Still, we can't conclude that MesoNet is better than DeepwareAI. DeepwareAI's GitHub page suggests that inputting the raw videos from FaceForensics++ to the scanner produces detection accuracies of 90% on average. Since both scanners were given the same data (but in different forms), it can be concluded that video-based classifiers like DeepwareAI need more data than image-based classifiers like MesoNet to be effective. However, when given enough data, video-based classifiers can perform equally as well as or even outperform image-based classifiers. Therefore, in contexts with limited amounts of data, image-based classifiers may be more effective. Otherwise, in contexts with abundant amounts of data, video-based classifiers may be more effective.

2.2.6: Dataset for Research Question 3

For research question 3, I used images I generated from StyleGAN. More specifically, I used stylegan.ipynb on my github repo. StyleGAN is a generative adversarial network that allows a person to generate very realistic-looking faces from an input to seed. These artificially generated faces are a more recent type of deepfake. More specifically, a seed is used to generate a latent vector (a 512-number long vector), which is then used to generate the face. Any slight change to any of the elements in the latent vector will change the generated face ever so slightly. Because of the method styleGAN uses to generate faces, one can use styleGAN to generate faces with particular facial expressions or features. My method of doing so was first generating a long list of generated faces from a list of consecutive seeds. From all of the images in that list, I would select the images that contained my desired facial expression or feature. To generate more of my preferred images without having to create a longer list of images from consecutive seeds, I would select two random images with my desired expression or features; get their latent vectors; create a new latent vector by averaging the two latent vectors; and add a small, randomized offset to it. This method works because the average of two latent vectors that generate a facial image with a certain expression or feature is very likely to generate a new facial image with the same expression or feature. The small offset adds a slight touch of uniqueness without changing the face too much. The randomized offset is also necessary in case my algorithm selects two particular latent vectors to average more than once. Using my method, I created 10 sets of facial images, each set having images of people with particular facial expressions or features. These image sets can be found under "StyleGANImages" in my Github repo. Each image set has 600 images, so there are a total of 6000 images in all of the 10 image sets. While all images in an image set share a particular facial expression or feature, that is the only constant. For example, in AngryImages, we can have 3 faces with angry expressions, but all having different hair lengths or ages. Here is a brief description of each image set found in the "StyleGANImages" folder

- AngryImages (Faces with angry expressions)
- FrightenedImages (Faces with Frightened expressions)
- HappyImages (Faces with happy expressions)
- LongHairImages (People with long hair)
- OldImages (Faces of people in old age)

- ReadingGlassesImages (People wearing reading glasses)
- ShortHairimages (People with short hair)
- SunglassesImages (People wearing Sunglasses)
- SurprisedImages (Faces with surprised expressions)
- YoungImages (Faces of people who are very young)

2.2.7. Metrics for Research Question 3

I will be giving the images stored in all 10 folders mentioned in 2.2.6 to MesoNet. The methods described in section 2.2.3 will be used to analyze the results.

2.2.8. Research Question 3: Can deepfake detectors readily detect that an artificially generated face/deepfake with a particular facial feature(s) or expression(s) is a deepfake?

Please refer to Appendix section 6.4. Figures 11-20 are relevant for this research question.

In general, the answer to this research question based on my results is a resounding no. For all 10 image sets I used, MesoNet predicted most, if not all images in the dataset to be real even though the images were artificially generated faces. However, the results presented in figures 11-20 can still reveal insights about what types of facial features or expressions MesoNet has the most potential to detect. Here is list of how well MeosNet performed for each dataset in terms of classifying images correctly (from worst to best):

- "ReadingGlassesImages"
- "AngryImages"
- "ShortHairImages"
- "LongHairImages"
- "HappyImages"
- "YoungImages"
- "SurprisedImages"
- "FrightenedImages"
- "SunglassesImages"
- "OldImages"

Figures 16, 14, 17, 11, and 13 all reveal that MesoNet performed extremely similarly for the images from the "ReadingGlassesImages", "LongHairImages", "ShortImages", AngryImages", and "HappyImages" folders in the sense the all or nearly all images in each dataset were predicted to be real. In fact, MesoNet performed equally well for images from the "ShortHairImages", "LongHairImages", and "AngryImages" folders in that MesoNet only predicted 1 image from each of the three image sets to be fake. In this case, they were ordered based on how many images fell under the category "real with high confidence". MesoNet performed slightly better for images from the "YoungImages", "SurprisedImages", and "FrightenedImages" image sets. This is revealed by figures 20, 19, and 12. MesoNet performed better still with images from the "SunglassesImages" dataset. This is revealed by figure 18. Finally, MesoNet performed exceptionally well with images from the "OldImages" folder relative

to other datasets. This is shown by figure 15. Although the trend is loose, my results suggest that MesoNet may have more potential to detect very animated expressions compared to more tame ones. An explanation for this is that StyleGAN has a harder time generating perfectly realistic animated/expressive facial expressions compared to more tame types. If we take an example, a comparison between the images in "HappyImages" and "SurprisedImages" reveals that surprised expressions may be slightly less realistic than happy expressions. In fact, some of the images in the "SurprisedImages" folder contain a distinct brown-colored spot of people's tongues when people's mouths are open. This would rarely, if ever, appear in images in the "HappyImages" folder as people in those images are smiling and showing their teeth rather than opening their mouths. This is why MesoNet performed slightly better with the "SurprisedImages" dataset than the "HappyImages" dataset. MesoNet performed best with the last two datasets, the images from "SunglassesImages" and "OldImages". MesoNet might've performed relatively well with the "SunglassesImages" because sunglasses obscure a person's eyes, meaning that MesoNet couldn't use a person's eyes to determine whether the image is real or fake. Lastly, StyleGAN seemed to have the most trouble generating realistic images of old people. While I was generating these images, I noticed that some images contained faces with an unusually high amount of wrinkles compared to what one would expect. As a result of these flaws, MesoNet might've had an easier time detecting that these images were deepfakes.

3. Deliverables

On Zenodo, you can find the videos I downloaded directly from FaceForensics++. Each dataset consists of 300 videos.

- Lightweight Faceswap Video Dataset: https://zenodo.org/record/6476243#.YnSSPtrMI2w
- Heavyweight Faceswap Video Dataset: https://zenodo.org/record/6430356#.YnSSudrMI2w
- Original Video Dataset: https://zenodo.org/record/6476482#.YnSSydrMI2w
- Face2Face Video Dataset: https://zenodo.org/record/6430891#.YnSS2trMI2w

Additionally, I used the website https://this-person-does-not-exist.com/en to get the images that can be found in "ArtificiallyGeneratedFaceImagesAndVideos/ArtificiallyGeneratedFaceImages" on Github.

Most of my other scripts and datasets are stored on Github. A description of the files I put on Github are in the README.md. Here is a link to my Github repository for this project: https://github.com/ssood123/Efficacy-of-Deepfake-Detection-Methods

4. Self-Evaluation

In this project, I was able to transform datasets from FaceForesncis++ and https://this-person-does-not-exist.com/en into forms that were acceptable for my experiment. After applying the transformations to the raw datasets, I was able to input them into MesoNet and DeepwareAI to not only test their individual efficacies, but compare the two scanners as well. I was also able to generate realistic-looking faces using StyleGAN. This project taught me how to efficiently work with large datasets given limited resources. For example, datasets from Zenodo have 1000 videos each, but I had to figure out how many videos I could get from each to make my experiment effective while still considering the storage limits of my SSD. This was a challenging procedure and it took some trial and error, but I eventually narrowed it down to 300 videos from each dataset. This project also taught me how to organize large amounts of data into presentable forms. That is, I had to figure out how to organize the data I got from MesoNet and DeepwareAI into a way that would allow me to extract meaningful data from it. I decided to use predictions and confidences to get the most meaningful results from my data. The most difficult task for me was learning how to use StyleGAN. I was very daunted by how complex the setup was as well as how it required a GPU, but breaking the code down into sections and understanding how StyleGAN worked on a high level helped me use StyleGAN effectively.

5. References

[1] Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niebner. 2019. “FaceForensics++: Learning to Detect Manipulated Facial Images.” 2019 IEEE/CVF International Conference on Computer Vision (ICCV). https://arxiv.org/abs/1901.08971
[2] Afchar, Darius, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. “MesoNet: a Compact Facial Video Forgery Detection Network.” 2018 IEEE International Workshop on Information Forensics and Security (WIFS). https://ieeexplore.ieee.org/document/8630761.

6. Appendix

6.1 Generation of datasets used for research questions 1 and 2

Due to the various types of deepfakes, I used data from both FaceForensics++ and https://this-person-does-not-exist.com/en. I got 4 of the 5 datasets from FaceForensics++: *dataset 1* consisted of 1000 short videos of new speakers and journalists (these are unmodified), *dataset 2* consisted of the 1000 original videos from dataset 1 modified using lightweight faceswap techniques, *dataset 3* consisted of the 1000 original videos from dataset 1 modified using heavyweight faceswap techniques, and *dataset 4* consisted of the 1000 original videos from dataset 1 modified using Face2Face techniques. However, due to hardware limitations, I could not work with all 1000 videos from each dataset. Instead, I chose to work with 300 videos from each dataset. I randomly selected the 300 videos from dataset 1 mentioned

earlier and selected the corresponding 300 videos from datasets 2,3, and 4. Altogether, I got 1200 videos, which I uploaded on Zenodo. On Zenodo:

- **Lightweight Faceswap Video Dataset** (consists of 300 videos from dataset 2)
- **Heavyweight Faceswap Video Dataset** (consists of 300 videos from dataset 3)
- **Original Video Dataset** (consists of 300 unmodified/real videos from dataset 1)
- **Face2Face Video Dataset** (consists of 300 videos from dataset 4)

Additionally, I got 1225 images from https://this-person-does-not-exist.com/en. All the images on this website are StyleGAN-generated images. These images can be found in the **"ArtificiallyGeneratedFaceImagesAndVideos/ArtificiallyGeneratedFaceImages"** folder in my Github repo.

Due to the differences between the datasets from Faceforensics++ and https://this-person-does-not-exist.com/en as well as the differences between MesoNet and DeepwareAI (the two scanners I investigated for this project), I had to modify each of the datasets to allow for a fair comparison between the two scanners. First, I took the 300 videos from each of the four datasets from Faceforensics++ and extracted frames from them using extractFramesFromVideos.py in my github repo (located in the DatasetManipulationScripts folder) . For all 1200 videos in all of the four datasets, a frame/image was extracted every .5 seconds of the video. For example, the first video from the original video dataset on Zenodo could be split into 7 frames/images named "video1_frame1","video1_frame2"..."video1_frame7". The second video would be split into images named "video2_framex" where x is a number, and so on. These images/frames I got from the original video dataset were stored in the "RealImages" folder on Github. I applied the same procedure to the other 3 datasets (not including the artificially generated face dataset). You can find these frames/images on Github in the folders named:

- **"HeavyweightFaceswapImagesAndVideos/HeavyweightFaceswapImages"** (this folder stores heavyweight faceswap images/frames extracted from the heavyweight faceswap video dataset on Zenodo),
- **"Face2FaceImagesAndVideos/Face2FaceImages"** (this folder stores face2face images/frames extracted from the face2face video dataset on Zenodo)
- **"LightweightFaceswapImagesAndVideos/LightweightFaceswapImages"** (this folder stores lightweight faceswap images/frames extracted from the lightweight faceswap video dataset on Zenodo)
- **"RealImagesAndVideos/RealImages"** (this folder stores real/unmodified images/frames extracted from the original video dataset on Zenodo)..

However, using these images was not enough. The photos from the artificially generated faces dataset (which are in the folder **"ArtificiallyGeneratedImagesAndVideos/ArtifiicallyGeneratedFaceImages"** on Github) showed faces close-up while all the other images from the other 4 folders ("RealImages","LightweightFaceswapImages", "Face2FaceImages", "HeavyweightFaceswapImages") didn't. Therefore, I used the scripts in the

Create-Face-Data-from-Images folder in my github repo (located in the DatasetManipulationScripts folder) to go through the images in the 4 aforementioned folders to create a close-up of the face by extracting the faces in the images. The reason I did this was to ensure a fair comparison between all 5 datasets. You can find the modified images under the following folders on my Github repo:

- **"HeavyweightFaceswapImagesAndVideos/HeavyweightFaceswapImagesFACE"** (contains images of the zoomed-in faces of the images from "HeavyweightFaceswapImages")
- **"Face2FaceImagesAndVideos/Face2FaceImagesFACE"** (contains images of the zoomed-in faces of the images from "Face2FaceImages")
- **"LightweightFaceswapImagesAndVideos/LightweightFaceswapImagesFACE"** (contains images of the zoomed-in faces of the images from "LightweightFaceswapImages")
- **"RealImagesAndVideos/RealImagesFACE"** (contains images of the zoomed-in faces of the images from "RealImages")

A comparison between the images in "Face2FaceImagesFACE" and "Face2FaceImages" reveals that the images in "Face2FaceImagesFACE" are zoomed-in versions of the images in "Face2FaceImages". In some cases, my python script could not extract the faces from the images; this is why there are less images in "HeavyweightFaceswapImagesFACE" than "HeavyweightFaceswapImages", for example.

All of the images in the "HeavyweightFaceswapImagesFACE", "Face2FaceImagesFACE", "LightweightFaceswapImagesFACE", "RealImagesFACE", and "ArtificiallyGeneratedFaceImages" folders were used to test MesoNet since it is an image-based deepfake detector. However, since DeepwareAI is a video-based classifier, I had to create new videos from the images stored in each of the five aforementioned folders. Since the folders "HeavyweightFaceswapImagesFACE", "Face2FaceImagesFACE", "LightweightFaceswapImagesFACE", and "RealImagesFACE" contained (modified) frames of the original videos on Zenodo, I decided to combine the frames into videos using FFMPEG. For example, there are 7 frames/images for the first video in the "LightweightFaceswapImagesFACE" folder. They are labeled "video1_frame1",...,"video1_frame7". Using these 7 images, I used the script createVideosFromImages.py in my github repo (located under the DatasetManipulationScripts folder) to create a 7-second long video where each image was shown for exactly one second in the video. The script uses FFMPEG. I chose 1 second for each image to give DeepwareAI the chance to process each of the 7 images shown in the video. This process was done for all frames/images stored in all of the 4 folders (all the folders except "ArtificiallyGeneratedFaceImages"). The process was different for the "ArtificiallyGeneratedFaceImages". For each of the 1200 videos I got from FaceForensics++, I averaged the number of frames I was able to extract from each video (which turned out to be 9). I then grouped each of the images in the "ArtificiallyGenerateFaceImages" folder into groups of 9 images each and created videos for each group. Like before, each image was shown in its respective video for 1 second exactly, resulting in 9-second long videos. These videos can be

found in the "ArtificiallyGeneratedFaceImagesAndVidoes/ArtificiallyGeneratedFaceVideos" folder on Github. In summary, the video data can be found in the following folders in my Github repo:

- **"HeavyweightFaceswapImagesAndVideos/HeavyweightFaceswapVideosFACE"** (contains videos created from the frames in "HeavyweightFaceswapImagesFACE")
- **"Face2FaceImagesAndVideos/Face2FaceVideosFACE"** (contains videos created from the frames in "Face2FaceImagesFACE")
- **"LightweightFaceswapImagesAndVideos/LightweightFaceswapVideosFACE"** (contains videos created from the frames in "LightweightFaceswapImagesFACE")
- **"RealVideosAndImages/RealVideosFACE"** (contains videos created from the frames in "RealImagesFACE")
- **"ArtificiallyGeneratedFaceImagesAndVideos/ArtifiicallyGeneratedFaceVideos"** (contains videos created from the images in "ArtificiallyGeneratedFaceImages")

The videos in each of the five folders were given to DeepwareAI for testing. Although I could've inputted the videos I downloaded from FaceForensics++ directly into DeepwareAI, I decided to use my custom-generated videos to ensure a fair comparison between the DeepwareAI and FaceForensics++ scanners. That is, each scanner would get the same exact data in either image or video form. Inputting the videos I downloaded from FaceForensics++ into DeepwareAI would give the scanner an unfair advantage.

6.2 Results of the MesoNet experiment with the 5 datasets from FaceForensics++ and https://this-person-does-not-exist.com/en

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	169	2962	5.71%
fake with medium confidence	195	2962	6.58%
fake with low confidence	312	2962	10.53%
real with low confidence	463	2962	15.63%
real with medium confidence	522	2962	17.62%
real with high confidence	1301	2962	43.92%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	2286	2962	77.18%
Fake	676	2962	22.82%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	775	2962	26.16%
medium confidence	717	2962	24.21%
high confidence	1470	2962	49.63%

Figure 1 (above): MesoNet results for images in the "RealImagesFACE" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	1527	2966	51.48%
fake with medium confidence	562	2966	18.95%
fake with low confidence	447	2966	15.07%
real with low confidence	270	2966	9.1%
real with medium confidence	121	2966	4.08%
real with high confidence	39	2966	1.31%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	430	2966	14.5%
Fake	2536	2966	85.5%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	717	2966	24.17%
medium confidence	683	2966	23.03%
high confidence	1566	2966	52.8%

Figure 2 (above): MesoNet results for images in the "HeavyweightFaceswapImagesFACE" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	122	2959	4.12%
fake with medium confidence	165	2959	5.58%
fake with low confidence	289	2959	9.77%
real with low confidence	418	2959	14.13%
real with medium confidence	630	2959	21.29%
real with high confidence	1335	2959	45.12%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	2383	2959	80.53%
Fake	576	2959	19.47%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	707	2959	23.89%
medium confidence	795	2959	26.87%
high confidence	1457	2959	49.24%

Figure 3 (above): MesoNet Data for images in the "Face2FaceImagesFACE" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	1830	2403	76.15%
fake with medium confidence	289	2403	12.03%
fake with low confidence	161	2403	6.7%
real with low confidence	62	2403	2.58%
real with medium confidence	33	2403	1.37%
real with high confidence	28	2403	1.17%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	123	2403	5.12%
Fake	2280	2403	94.88%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	223	2403	9.28%
medium confidence	322	2403	13.4%
high confidence	1858	2403	77.32%

Figure 4 (above): MesoNet Data for images in the "LightweightFaceswapImagesFACE" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	39	1225	3.18%
fake with medium confidence	31	1225	2.53%
fake with low confidence	29	1225	2.37%
real with low confidence	39	1225	3.18%
real with medium confidence	105	1225	8.57%
real with high confidence	982	1225	80.16%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	1126	1225	91.92%
Fake	99	1225	8.08%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	68	1225	5.55%
medium confidence	136	1225	11.1%
high confidence	1021	1225	83.35%

Figure 5 (above): MesoNet Data for images in the "ArtificiallyGeneratedFaceImages" folder

6.3. Results of the DeepwareAI experiment with the 5 datasets from FaceForensics++ and https://this-person-does-not-exist.com/en

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	17	300	5.67%
fake with medium confidence	18	300	6.0%
fake with low confidence	23	300	7.67%
real with low confidence	36	300	12.0%
real with medium confidence	53	300	17.67%
real with high confidence	153	300	51.0%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	242	300	80.67%
Fake	58	300	19.33%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	59	300	19.67%
medium confidence	71	300	23.67%
high confidence	170	300	56.67%

Figure 6 (above): DeepwareAI Data for videos in the "RealVideosFACE" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	212	300	70.67%
fake with medium confidence	27	300	9.0%
fake with low confidence	19	300	6.33%
real with low confidence	29	300	9.67%
real with medium confidence	7	300	2.33%
real with high confidence	6	300	2.0%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	42	300	14.0%
Fake	258	300	86.0%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	48	300	16.0%
medium confidence	34	300	11.33%
high confidence	218	300	72.67%

Figure 7 (above): DeepwareAI Data for videos in the "HeavyweightFaceswapVideosFACE" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	33	300	11.0%
fake with medium confidence	41	300	13.67%
fake with low confidence	36	300	12.0%
real with low confidence	41	300	13.67%
real with medium confidence	36	300	12.0%
real with high confidence	113	300	37.67%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	190	300	63.33%
Fake	110	300	36.67%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	77	300	25.67%
medium confidence	77	300	25.67%
high confidence	146	300	48.67%

Figure 8 (above): DeepwareAI Data for videos in the "Face2FaceVideosFACE" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	96	300	32.0%
fake with medium confidence	36	300	12.0%
fake with low confidence	50	300	16.67%
real with low confidence	51	300	17.0%
real with medium confidence	33	300	11.0%
real with high confidence	34	300	11.33%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	118	300	39.33%
Fake	182	300	60.67%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	101	300	33.67%
medium confidence	69	300	23.0%
high confidence	130	300	43.33%

Figure 9 (above): DeepwareAI Data for videos in the "LightweightFaceswapVideosFACE" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	0	137	0.0%
fake with medium confidence	0	137	0.0%
fake with low confidence	0	137	0.0%
real with low confidence	1	137	0.73%
real with medium confidence	0	137	0.0%
real with high confidence	136	137	99.27%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	137	137	100.0%
Fake	0	137	0.0%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	1	137	0.73%
medium confidence	0	137	0.0%
high confidence	136	137	99.27%

Figure 10 (above): DeepwareAI Data for videos in the "ArtificiallyGeneratedFaceVideos" folder

6.4: Results of the MesoNet experiment with the 10 image sets of images generated from StyleGAN

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	0	600	0.0%
fake with medium confidence	0	600	0.0%
fake with low confidence	1	600	0.17%
real with low confidence	1	600	0.17%
real with medium confidence	5	600	0.83%
real with high confidence	593	600	98.83%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	599	600	99.83%
Fake	1	600	0.17%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	2	600	0.33%
medium confidence	5	600	0.83%
high confidence	593	600	98.83%

Figure 11 (above): MesoNet Data for images in the "AngryImages" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	0	600	0.0%
fake with medium confidence	2	600	0.33%
fake with low confidence	9	600	1.5%
real with low confidence	0	600	0.0%
real with medium confidence	11	600	1.83%
real with high confidence	578	600	96.33%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	589	600	98.17%
Fake	11	600	1.83%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	9	600	1.5%
medium confidence	13	600	2.17%
high confidence	578	600	96.33%

Figure 12 (above): MesoNet Data for images in the "FrightenedImages" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	0	600	0.0%
fake with medium confidence	1	600	0.17%
fake with low confidence	1	600	0.17%
real with low confidence	2	600	0.33%
real with medium confidence	18	600	3.0%
real with high confidence	578	600	96.33%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	598	600	99.67%
Fake	2	600	0.33%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	3	600	0.5%
medium confidence	19	600	3.17%
high confidence	578	600	96.33%

Figure 13 (above): MesoNet Data for images in the "HappyImages" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	0	600	0.0%
fake with medium confidence	0	600	0.0%
fake with low confidence	1	600	0.17%
real with low confidence	4	600	0.67%
real with medium confidence	11	600	1.83%
real with high confidence	584	600	97.33%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	599	600	99.83%
Fake	1	600	0.17%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	5	600	0.83%
medium confidence	11	600	1.83%
high confidence	584	600	97.33%

Figure 14 (above): MesoNet Data for images in the "LongHairImages" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	4	600	0.67%
fake with medium confidence	17	600	2.83%
fake with low confidence	12	600	2.0%
real with low confidence	6	600	1.0%
real with medium confidence	0	600	0.0%
real with high confidence	561	600	93.5%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	567	600	94.5%
Fake	33	600	5.5%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	18	600	3.0%
medium confidence	17	600	2.83%
high confidence	565	600	94.17%

Figure 15 (above): MesoNet Data for images in the "OldImages" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	0	600	0.0%
fake with medium confidence	0	600	0.0%
fake with low confidence	0	600	0.0%
real with low confidence	0	600	0.0%
real with medium confidence	1	600	0.17%
real with high confidence	599	600	99.83%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	600	600	100.0%
Fake	0	600	0.0%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	0	600	0.0%
medium confidence	1	600	0.17%
high confidence	599	600	99.83%

Figure 16 (above): MesoNet Data for images in the "ReadingGlassesImages" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	0	600	0.0%
fake with medium confidence	0	600	0.0%
fake with low confidence	1	600	0.17%
real with low confidence	2	600	0.33%
real with medium confidence	5	600	0.83%
real with high confidence	592	600	98.67%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	599	600	99.83%
Fake	1	600	0.17%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	3	600	0.5%
medium confidence	5	600	0.83%
high confidence	592	600	98.67%

Figure 17 (above): MesoNet Data for images in the "ShortHairImages" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	4	600	0.67%
fake with medium confidence	3	600	0.5%
fake with low confidence	9	600	1.5%
real with low confidence	4	600	0.67%
real with medium confidence	0	600	0.0%
real with high confidence	580	600	96.67%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	584	600	97.33%
Fake	16	600	2.67%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	13	600	2.17%
medium confidence	3	600	0.5%
high confidence	584	600	97.33%

Figure 18 (above): MesoNet data for images in the "SunglassesImages" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	0	600	0.0%
fake with medium confidence	0	600	0.0%
fake with low confidence	9	600	1.5%
real with low confidence	0	600	0.0%
real with medium confidence	1	600	0.17%
real with high confidence	590	600	98.33%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	591	600	98.5%
Fake	9	600	1.5%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	9	600	1.5%
medium confidence	1	600	0.17%
high confidence	590	600	98.33%

Figure 19 (above): MesoNet Data for images in the "SurprisedImages" folder

Category	Number of images in this category	total number of images	percent of total
fake with high confidence	1	600	0.17%
fake with medium confidence	1	600	0.17%
fake with low confidence	3	600	0.5%
real with low confidence	2	600	0.33%
real with medium confidence	8	600	1.33%
real with high confidence	585	600	97.5%

Predicted Label	Number of Images for this label	Total number of Images	Percent of total
Real	595	600	99.17%
Fake	5	600	0.83%

Confidence	Number of images for this confidence interval	total number of images	percent of total
low confidence	5	600	0.83%
medium confidence	9	600	1.5%
high confidence	586	600	97.67%

Figure 20 (above): MesoNet Data for images in the "YoungImages" folder