DEEPFAKE OR REAL IMAGE PREDICTION USING MESONET

A MAJOR PROJECT REPORT

submitted in partial fulfillment of the requirements

for the award of the degree of

Master of Technology

in

COMPUTER SCIENCE ENGINEEERING

by

Manish Khichi

ROLL NO. 2K19/CSE/11

Under the supervision of

Dr.Rajesh Kumar Yadav

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY,DELHI-110042

June, 2021

CERTIFICATE

I, hereby certify that the Project Dissertation titled **"DEEPFAKE OR REAL IMAGE PREDICTION USING MESONET",**which is submitted by Manish Khichi,Department of Computer Science Engineering, Delhi Technological University, Delhi,in partial fulfillment for the requirement of the award of degree of Master of Technology (Computer Science and Engineering) is a record of a project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Manish

Signature of Candidate

Manish Khichi

Roll No. 2K19/CSE/11

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Yadav

Signature of Supervisor

Date:11 June,2021 **Dr.Rajesh Kumar Yadav,**

Associate Professor

ACKNOWLEDGEMENT

First of all, I express my gratitude to the Almighty, who blessed me with the zeal and enthusiasm to complete this research work successfully. I am extremely thankful to my supervisor Dr.Rajesh Kumar Yadav, Associate Professor, Department of Computer Science Engineering, Delhi Technological University, Bawana Road,Delhi, for their motivation and tireless efforts to help me to get deep knowledge of the research area and supporting me throughout the life cycle of my M. Tech. dissertation work. Especially, the extensive comments, healthy discussions, and fruitful interactions with the supervisors had a direct impact on the final form and quality of M. Tech. dissertation work.

I am also thankful to Dr. Rajni Jindal, Head of the Computer Engineering Department, for his fruitful guidance through the early years of chaos and confusions. I wish to thank the faculty members and supporting staff of Computer Engineering Department for their full support and heartiest co-operation.

This report would not have been possible without the hearty support of my friends. My deepest regards to my Parents for their blessings, affection and continuous support. Also, Last but not the least, I thank to the GOD, the almighty for giving me the inner willingness, strength and wisdom to carry out this research work successfully.

Manish Khichi

ABSTRACT

Advance development in Machine Learning, Deep Learning, and Artificial Intelligence (AI) allow people to exchange the faces and voices of other people in videos so that they look like people did or wanted to say. These videos and photos are called "deepfake" and each day they are more complicated, which worries legislators. This technology uses machine learning technology to provide the computer with real data about the image so that we can falsify it. The creators of Deepfake use artificial intelligence and machine learning algorithms to mimic the work and characteristics of the real human. It differs from traditional fake media because it is difficult to identify. As the 2020 election approaches, 4,444 AI-generated DeepFakes have entered the news cycle. DeepFakes threatens facial recognition and online content. This hoax can be dangerous, because if used incorrectly, you can abuse this technique. Fake video, voice and audio clips can cause enormous damage.We will use Mesonet To make predictions on image data. we will examine four sets of images-correctly identified deepfakes, correctly identified reals, misIdentified deepfakes, misIdentified reals and we will see whether the human eye can pick up on any insights into the world of deepfakes.We will be using the Meso 4 model Trained on the deepfake and real data set.

Contents

List of Figures

List of Tables

List of Abbreviations

ML	Machine Learning
DL	Deep Learning
DNNs	Deep Neural Networks
CNNs	Convolutional Neural Networks
GANs	Generative Adverserial Networks
DFDC	Deepfake Detection Challenge

Chapter 1

INTRODUCTION

1.1 Overview

Figure 1.1: Deepfake image of a person using FaceApp.

The concept of deepfake refers to images, audio, or video that are fakes that is they depict events that never occurred but unlike methods of Manipulating media in the past like Photoshop. These deepfakes are created by Deep neural networks to be nearly indistinguishable from their real counterparts. The advances in the field of deepfakes are equal parts impressive and alarming on the upside. the fidelity with which we can alter media Will certainly lead to some world memes but in The wrong hands. This technology can be used to spread misinformation and undermine public trust almost like a sci-fi type of Identity theft where you can get anyone to say anything and it's the opposite. This means that as we get better at generating deepfakes, we must also get better at identifying them. Mesonet is a convolutional neural network designed for exactly this purpose.Deepfake image of a person using FaceApp is shown in Figure 1.1,FaceApp[1]is one of the best mobile apps for AI photo editing and creating deepfake videos that have a face swap feature.

1.2 Deep Learning

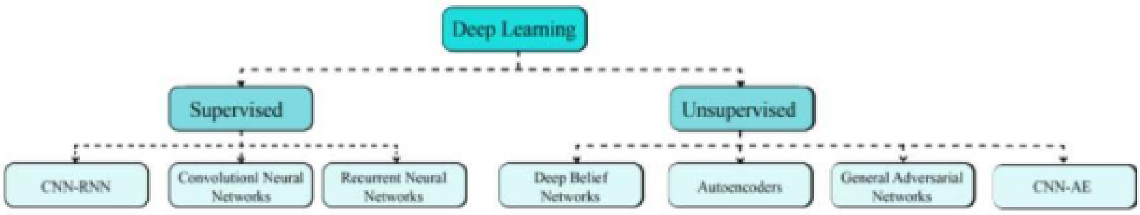

Figure 1.2: Various deep learning methods.

Convolutional Neural Networks (CNNs):CNN is a branch of deep neural network(DNN) in DL, most of which are highly taken in image processing analysis. The multiple layers of the convolutional neural network are designated as an input layer, a convolutional layer, a grouping layer, and a fully connected layer, that is, a dense layer.The input layer takes an image as input. The convolutional layer produces output based on its kernel or filter value (ie feature extractor), used as input to subsequent layers. The grouping layer is used for dimensionality reduction and accelerated calculation.

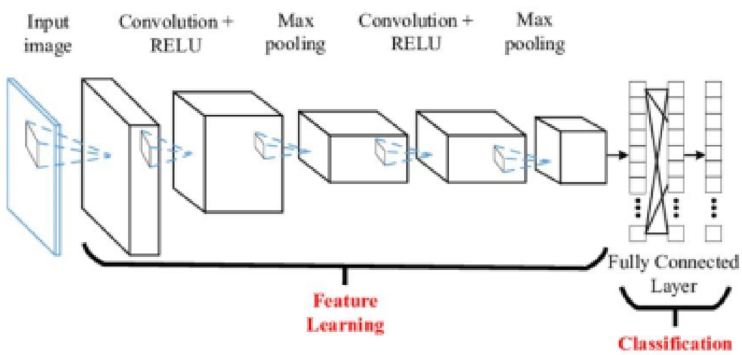

Figure 1.3: CNN Architecture.

Machine learning and deep learning require artificial intelligence to solve datadriven problems by providing a set of algorithms and neural networks . Deepfakes also has a silver lining, using Deepfake face-swapping technology to secure the privacy in medical related videos and help digitally rebuild for the movie Fast and Furious 7 Paul Walker This actor died there in 2015, and no one had at that time Heard of deepfakes before. [2] Jim Carrey played Jack Nicholson in The Shining music video. Picture 1 picture shows a video on YouTube, in which [3] David Beckam speaks nine languages when calling for an end to malaria. Here, deepfake is used to spread awareness. Today, the deepfake technology can bring anyone back to life, even a that left a long time ago, as if there is a video of the Spanish painter Salvador in the Dali Museum in Florida, you can even use it. It takes a selfie. deepfake has a more interesting side, now everyone is dancing. researchers converted professional dance moves into amateur groups.

I use machine learning and deep learning technology to synthesize content from malicious and unethical applications, such as the spread of false information, posing as political leaders and slandering innocent people [4] In April 2018, BuzzFeed demonstrated Deepfake video technology progress by combining the voice of Jordan Peele with the video of Barack Obama.In the video, Jordan Peel warned: "We are entering an era where our enemies can make people appear to be talking anytime, anywhere. Even if they never say those things." Deepfake technology is changing, more and more, changing It's getting more and more complicated and dangerous. This is partly due to the nature of artificial intelligence. There, "traditional" technology requires people's time and energy to improve, and artificial intelligence can learn from it. Therefore, the development capability of artificial intelligence is a double-edged sword. If artificial intelligence was created to do good, that would be great! However, when AI is designed to be malware like Deepfake, the danger of is unprecedented. With the continuous improvement of Deepfake production methods, it is increasingly difficult to find fake videos for . However, here are some points to keep in mind: The lower quality part of the video itself is frame-shaped , the clipping effect around is irregular flickering of the mouth, eyes and neck, skin tone is inconsistent, movement is unnatural, the background changes Or turn it on.

Chapter 2

Deepfake Generation

The change of face and repetition are two profound illusions. When deepfake is used to generate and pass the identity to the target person, the facial expression of the source person can be retained. When exchanging objects, the face of the target person is superimposed on the face of the source. In post-production, overlay face edges blended to match contour of the source face. Faceswaps [5] is a face exchange technology that uses facial landmarks such as eyes,lips,eyebrows,cheeks,nose, chin to turn the target face into a source area, and then process the result, edge polish and color correction. In face playback, it is used to make the subject act and sound like a sound source. Face2Face [6], a facial entertainment tool,used to transform the facial expressions of source object. As long as retains the target's facial features, the position of the target remains unchanged.

2.1 Autoencoders

Deepfakes can be generated by using Autoencoders.At the highest level, Autoencoders work as shown in figure 1. When the data are processed such as image data, Data get compressed by an encoder. The purpose of this compression is to suppress the effect of noise in the data and to reduce computational complexity conversely. the original image can be restored atleast approximately bypassing the compressed version of the image through a Decoder. now suppose we want to create a deepfake that blends Da Vinci's Mona Lisa and Van Gogh's starry nightto do so. When we

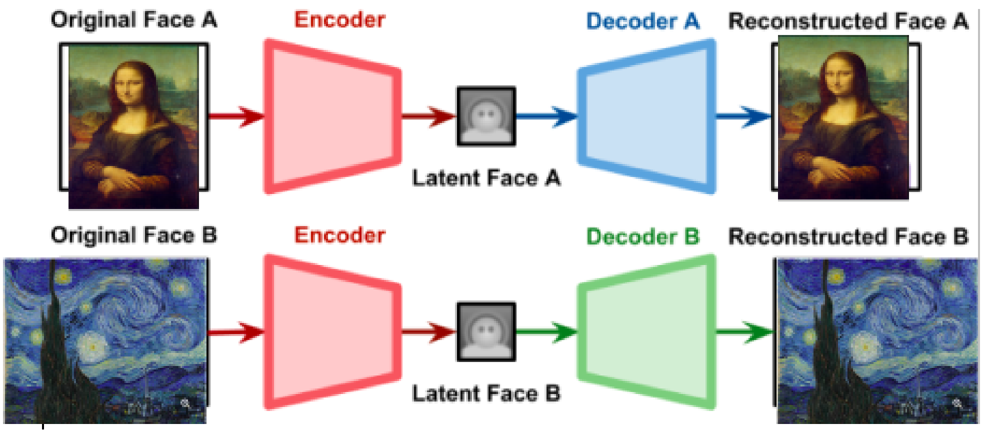

Figure 2.1: Autoencoders

train the autoencoders for different datasets, we allow the encoders to share weights while keeping their decoders separate, that way An image of the Monalisa can be compressed according To general logic. taking into account things like the illumination position and the expression of her face but When it gets restored. This will be done according to the logic specific to starry night which has the effect of overlaying Van Gogh's distinctive style onto da Vinci's Masterpiece although, This is how Deepfakes are often generated.

2.2 Generative Adverserial Networks(GANs)

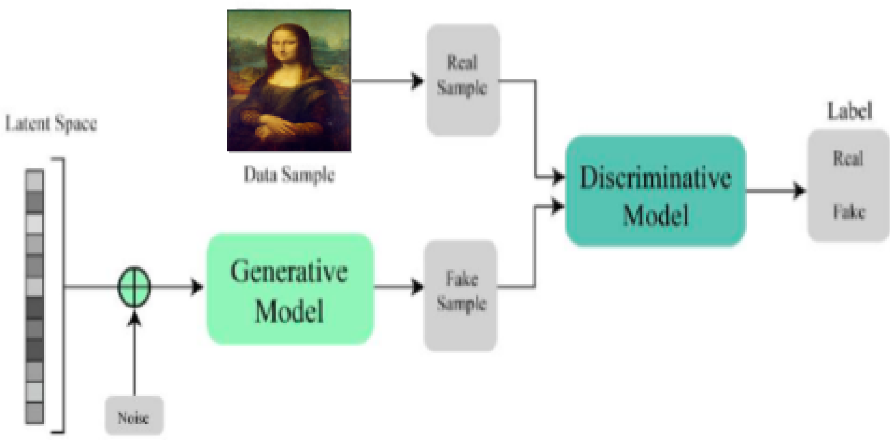

Figure 2.2: Generative Adverserial Networks(GANs)

One of the main problems of deep model training is the limitation of the size of the data set. The generative model with data augmentation solves this problem to a certain extent. It is well known that generative adversarial networks can generate high-quality data. The basic concept in GAN training is a simple minimax game between two models [7]. One of the networks is a generative model that tries to generate data, and the other is a discriminative model to determine whether is True or false. The purpose of the generator is to deceive the discriminator. As competed for victory, they became so proficient in their tasks that the generator finally reached the point where could generate images similar to the images in the original data set.

Chapter 3

Related Work

3.1 Deepfake Detection

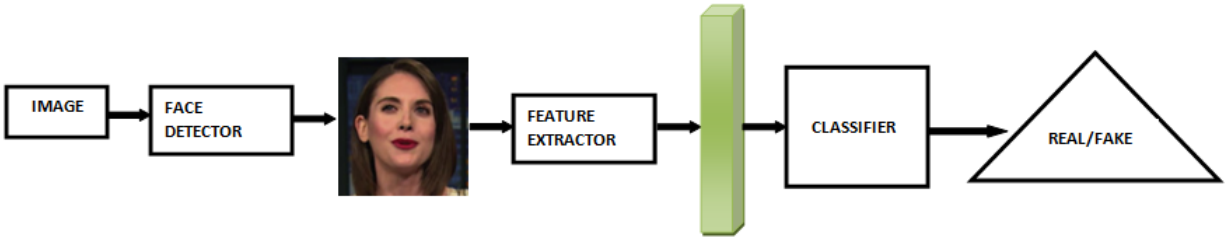

Figure 3.1: Deepfake Image Detection Architecture

Deepfake detection is generally regarded as a binary classification problem, in which a classifier is used to distinguish actual content from actionable content. The problem of fighting against deepfake is to have a ML model,that can said to be real or generated images. Researchers, start-up companies and even the government are already studying tools that can detect deep forgeries, and the US Defense Advanced Research Projects Agency recognizes manual operations. There are at least 2,000 research publications in this field, about , and the number of publications is increasing year by year. Image, video, , and audio are the three main types of composite content.

The function extractor is the core of the key development of the various methods. Several manual element-based algorithms, DL algorithms, recent GAN-based methods being studied. [8] uses 3D head posture difference for distinguish between fake and real images. PRNU (Photo Response Non-uniformity) [9] is the unique noise pattern of the light sensor in each digital camera. This local noise mode is generally used for false image detection because can be changed by any modification of the image. [10] suggests combining RGB spatial features with dual-stream CNN steganalysis features to manipulate face detection. This is a combination of machine learning and manual functions. Classification, segmentation [11] and reconstruction [12] are performed simultaneously in the multi-task learning method.

Chapter 4

Implementation and Results

4.1 Implementation

We are going to explore the topic of deepfakes and examine a neural network designed to identify them called Mesonet.We will use Mesonet to make predictions on image data. We will examine four sets of images-correctly identified deepfakes, correctly identified reals, misidentified deepfakes, misidentified reals and we will see whether the human eye can pick up on any insights into the world of deepfakes.We will be using the Meso 4 model trained on the deepfake data set.Now let's begin how this dataset was assembled and then we will proceed to the model.

4.1.1 Datasets

Table 4.1: Dataset Frames Table

	Face Frames Dataset Quantity
Deepfake	**4258**
Real	**2730**
Total	**6988**

The authors in [13] note that their deepfake dataset was created differently rather than generating deepfakes from scratch which they explained would limit the amount and diversity of the Fake data that they could then feed Mesonet. They choose to extract face images from existing deepfake videos.They used about 175 existing videos pulled from popular deepfake platforms and that created their deepfake dataset. They explained that they extracted the specific frames that contain faces from the deepfake videos. They also note that they followed a similar process for extracting the real image data from real video sources like Tv shows and movies and finally they explain how they satisfied their data so that the various angles of the faces and levels of the resolution were distributed evenly across the real and deepfake datasets.I have used their image dataset of real and deepfake frames extracted from videos online.

I structure my real and deepfake image data in separate folders underneath a folder called Data. This is important for the flow from the directory method which infers classes from the file structure. The pixel values in our image data exist in the range between 0 and 255.Large integer coefficients like this complicate gradient descent when using typical learning rates. So the next step is to scale our data by a factor of one divided by 255 that way the pixel values fall into the range from zero to one. I instantiate our image data generator that rescales our images then I pass in our data by specifying the directory path to our data folder. I set the batch size to 1 so I process images individually and then I set the class mode to binary for the binary classification task of predicting images as real or deepfake. Now let's check our class indices flow from the directory should have inferred the names of our classes from the names of the sold folders within the data folder and there should be just two classes where our deepfakes are represented by zero and are reals by one. There are two problems that might arise. First, if our classes are reversed such that our deepfakes are represented by one and the reals by zero, we'll have to flip Mesonet predictions by subtracting these values from one, the second issue relates to having more than just two classes as we discussed flow from directory uses file structure to process data and is therefore sensitive to the structure.

Since machine learning is all about finding patterns in data. It is extremely important to understand the nature of the data and how it's collected because you're ultimately going to feed that through a model to understand it as you will see later. With this understanding of the data, we reveal some important insights but stick around for that let's first explore the model.

4.1.2 CNN based Mesonet Model

MESO-4

Meso4 is a convolutional neural network with four convolutional blocks, followed by a fully connected hidden layer, and then an output layer for prediction when we reproduce the model. Convolutional blocks always include a convolutional layer and a maximum detection layer. In Mesonet, these blocks also include a batch normalization layer. Batch normalization is a new technology to improve the speed and stability of neural networks. Its working principle is to normalize the input of each layer of the network to reduce the interdependence between the parameters of a given layer and the input distribution of the next layer. This interdependence is called internal covariate change, which has an unstable effect on the learning process. The last layer of our convolution block is the grouping layer. It is in the pooling layer that we significantly reduce the dimensionality of our data which greatly speeds up computation. Mesonet uses max pooling for this layer which means we reduce a region of pixel values to that region's maximum value. It may seem like we're throwing away too much data in this pooling layer but remember during convolution, our model was able to locate important image features which means we can focus on just the parts that matter most. Think about our visual field right now odds are we are focusing on a very small subset of the available data just like Mesonet. I will be using the essential methods for the network to load weights and to make predictions and next let's create a Meso 4 class. The Meso 4 class takes just one argument and that's the classifier class that just created. I set the gradient descent optimizer and I set its learning rate in the constructor and set the parameters to compile the model.

Now it's time to create the network architecture. First I create our input layer and assign it to the variable x for the input layer that just needs to pass the three dimensions of our image data then create our four convolutional blocks. The convolutional layer represented by Conv2D is the trickiest part here I set the size and the number of filters that will use in convolution. Each filter represents a distinct image

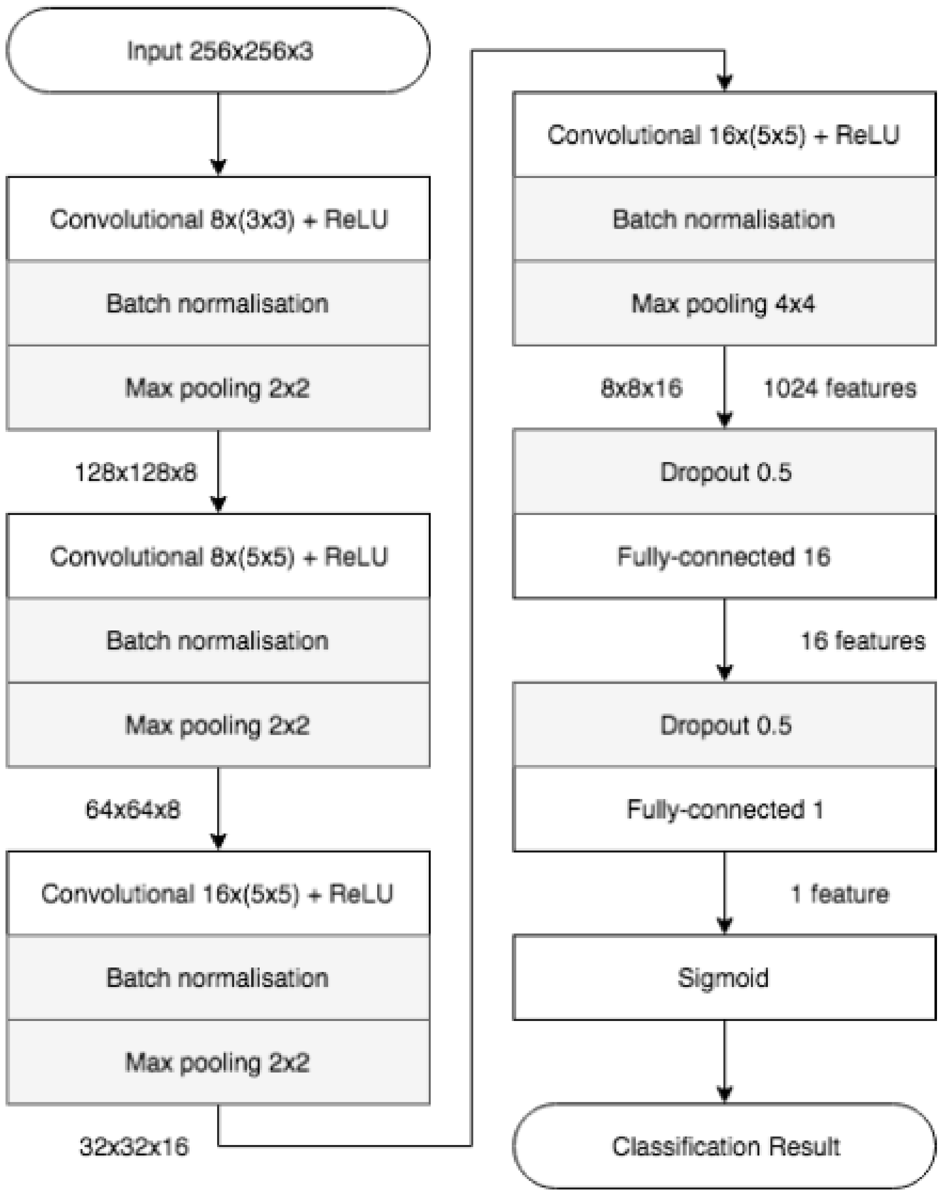

Figure 4.1: Meso-4 Network Architecture

feature for example a horizontal line during convolution. This filter is passed over an image to assess the degrees to which specific regions of that image corresponds with the filter. If you're a math whiz, this is done by calculating the dot product of the filter. With each filter, the size region of the image for each of the color channels for the rest of us. The important thing to know is that the filter identifies the existence and location of specific image features like horizontal or vertical lines after the convolutional layer comes the batch normalization layer.

With successive blocks in the convolutional base, CNNs proceed to higher-order feature representations from lines to corners to shapes to faces.Meso-4 is a convolutional neural network with four convolutional blocks followed by one Fully Connected Hidden Layer then the Output Layer for prediction. Now that I've got the network architecture established, I need to instantiate the model and load the weights. I download the weights online and then save the file meso4 DF in a folder called weights. I load the weights using the load method of our classifier by specifying the file path. Mesonet is now ready to make predictions on image data.

4.2 Results

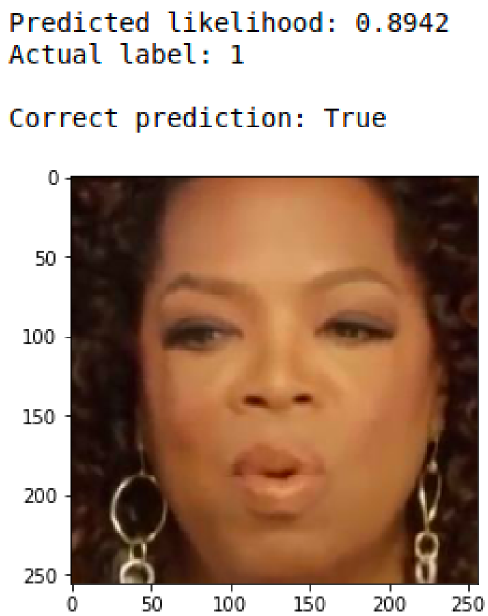

Figure 4.2: Example of one of the image that model predicted

Now we're ready to pass in an image through Mesonet. The generator.next method returns two items, the pixel data of a given image and the actual label for it whether it's real or deepfake. So let's set variables x and y equal to generator. next. let's write three print statements to evaluate the prediction. For the first, let's show Meson that's predicted output for the image rounded to four digits, and for the second, let's show the actual label and for the third, let's see whether Mesonet's prediction is accurate that is whether the actual label corresponds with mesonet's predicted output after rounding to 0 or 1. Now we can see how Mesonet performed on a particular image in Figure 4.2,when predicted outputs are nearly 0 or 1.This corresponds with a high degree of confidence in the prediction but when the predicted output approaches 0.5, the confidence in the model prediction approaches a random guess, for that run this a few more times and see any patterns.

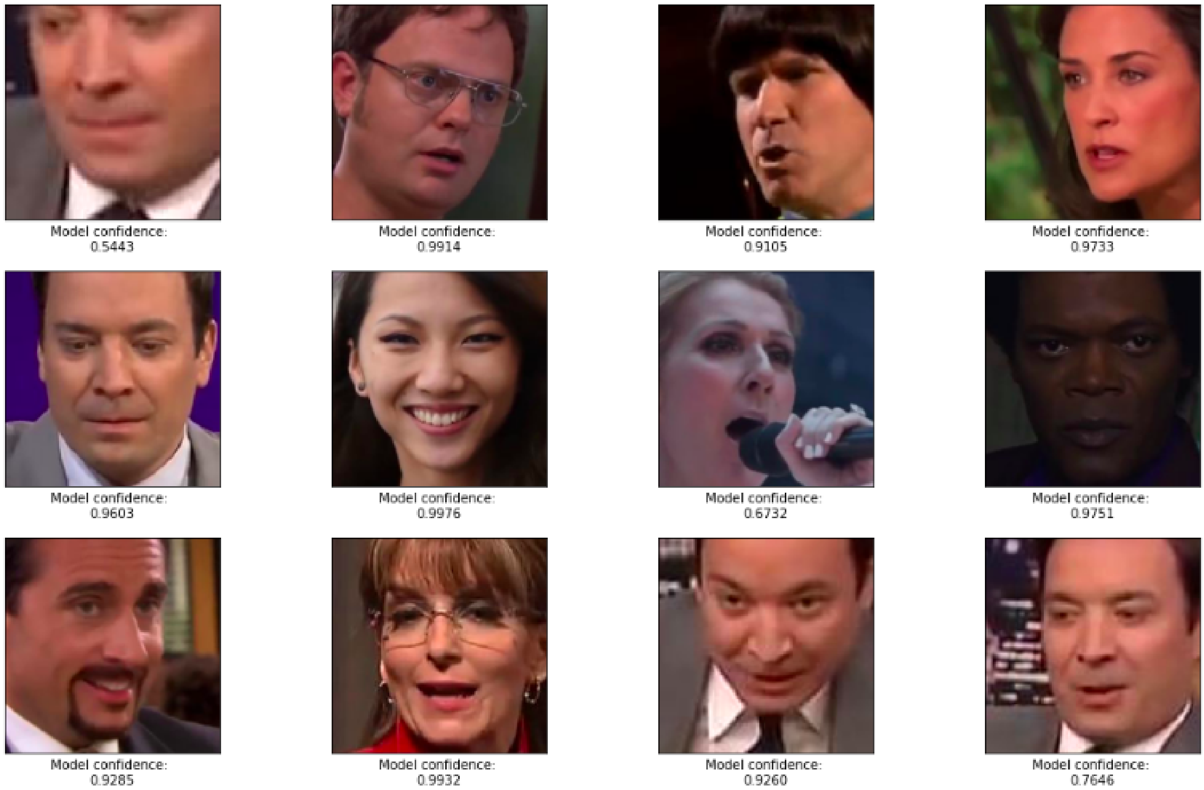

Figure 4.3: Correct Real vs Correct Real Predicted Output

Let's organize our predictions into four categories correctly predicted deepfakes, correctly predicted reals, misclassified deepfakes, and misclassified reals. We create lists to keep track of which images fall into each category then we write a for loop to iterate through our dataset and sort each observation into one of these four categories. We create two lists for each category, one to store the image data and the other to store the corresponding prediction value. Now let's create our plotter function which we'll use to show random batches of images in each category. The plotter takes two arguments a list of image data and the list of corresponding predictions. We use an f string for the x label so we can show Mesonet's predicted output for the image in question.

Let's check out a collection of correctly identified real images, we pass correct real and correct real predicted(Figure 4.3) into our plotter function.Checking out the output images look pretty good, you might see some characters from TV or movies

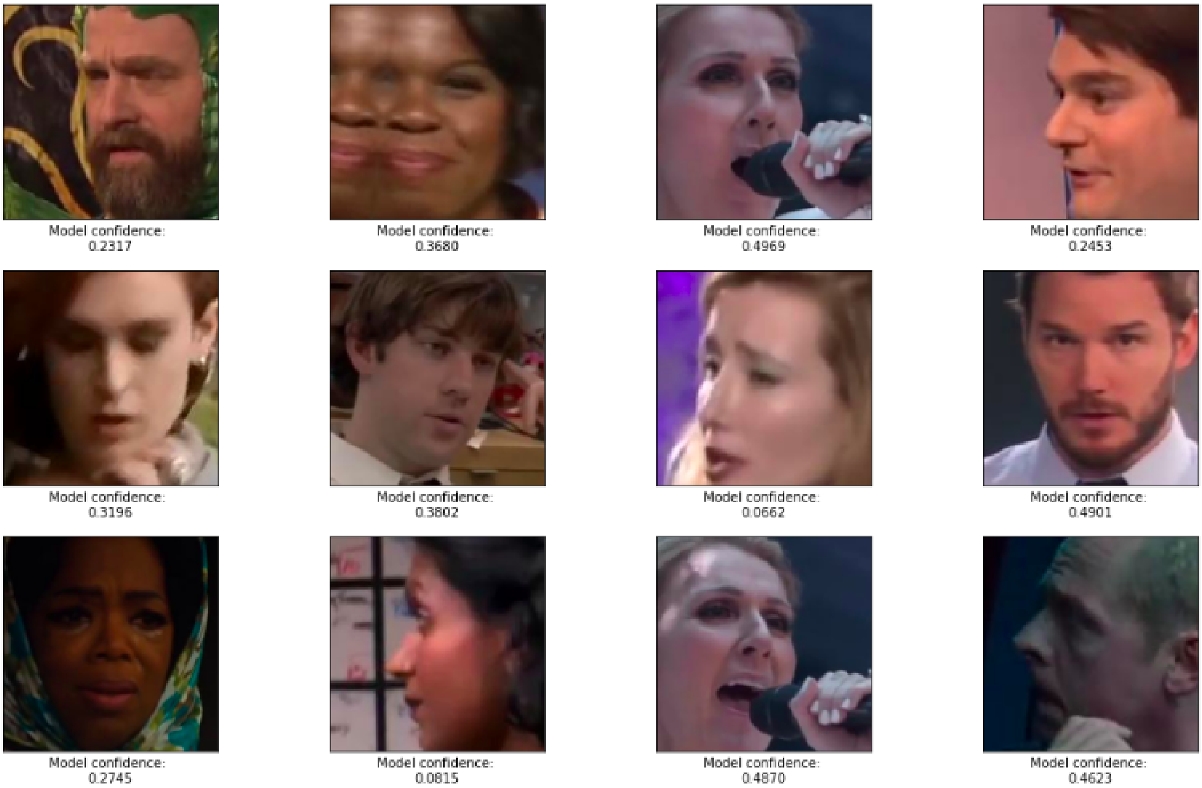

Figure 4.4: Misclassified Real vs Misclassified Real Predicted Output

that you recognize, here look at the predicted output for the images, the closer these values are to one the more confidence Mesonet has that the image is real. Notice that most of these outputs vary between a range pretty close to one. Okay for contrast let's look at real images that were misclassified as deepfakes, let's again use our plotter function but this time we pass it as misclassified real and misclassified real predicted(Figure 4.4). Although these are all representing errors,it is reassuring to note that the model's confidence for these predictions tend to be closer to 0.5 and that's to say that it's less confident in these predictions that turned out to be wrong, so it's pretty much a guess and I guess that's okay. Now let's check our correct deepfakes and correctly identified deepfakes(Figure 4.5),that's an odd pattern in our data there. Let's also check the deepfakes misclassified and misclassified deepfakes predicted(Figure 4.6).

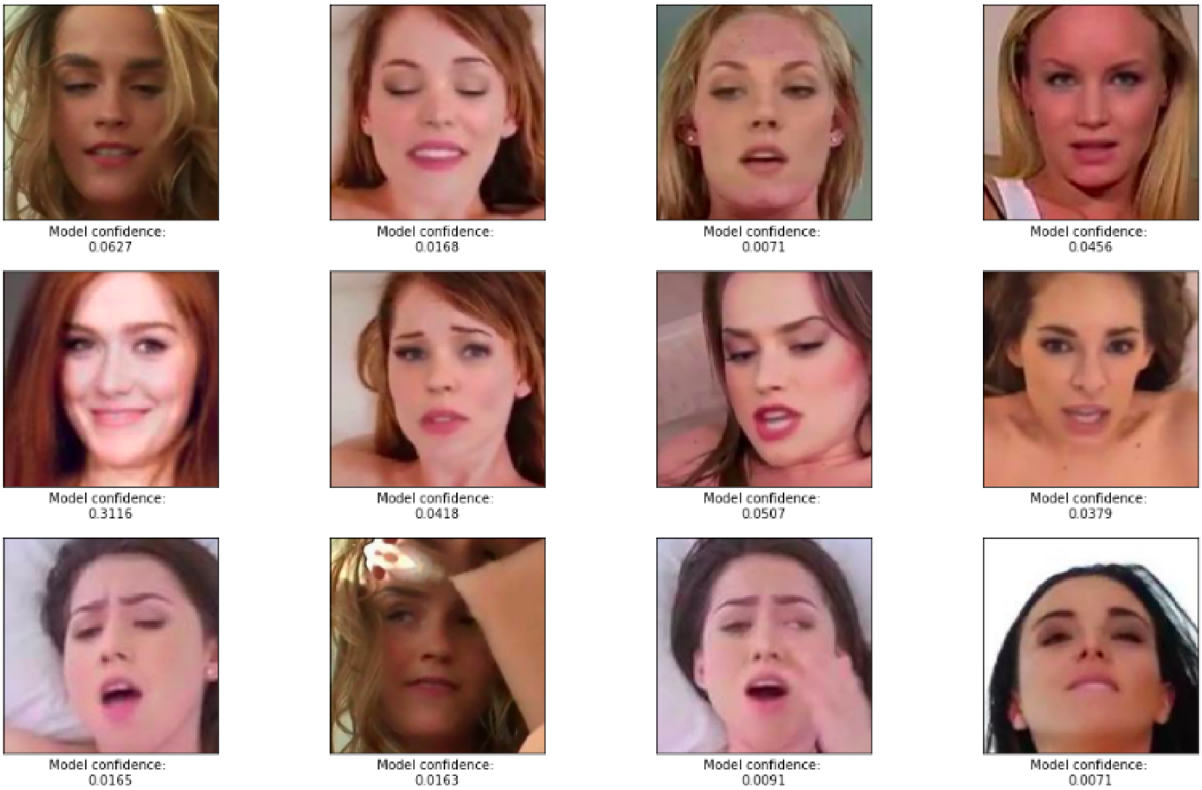

Figure 4.5: Correct Deepfake vs Correct Deepfake Predicted Output

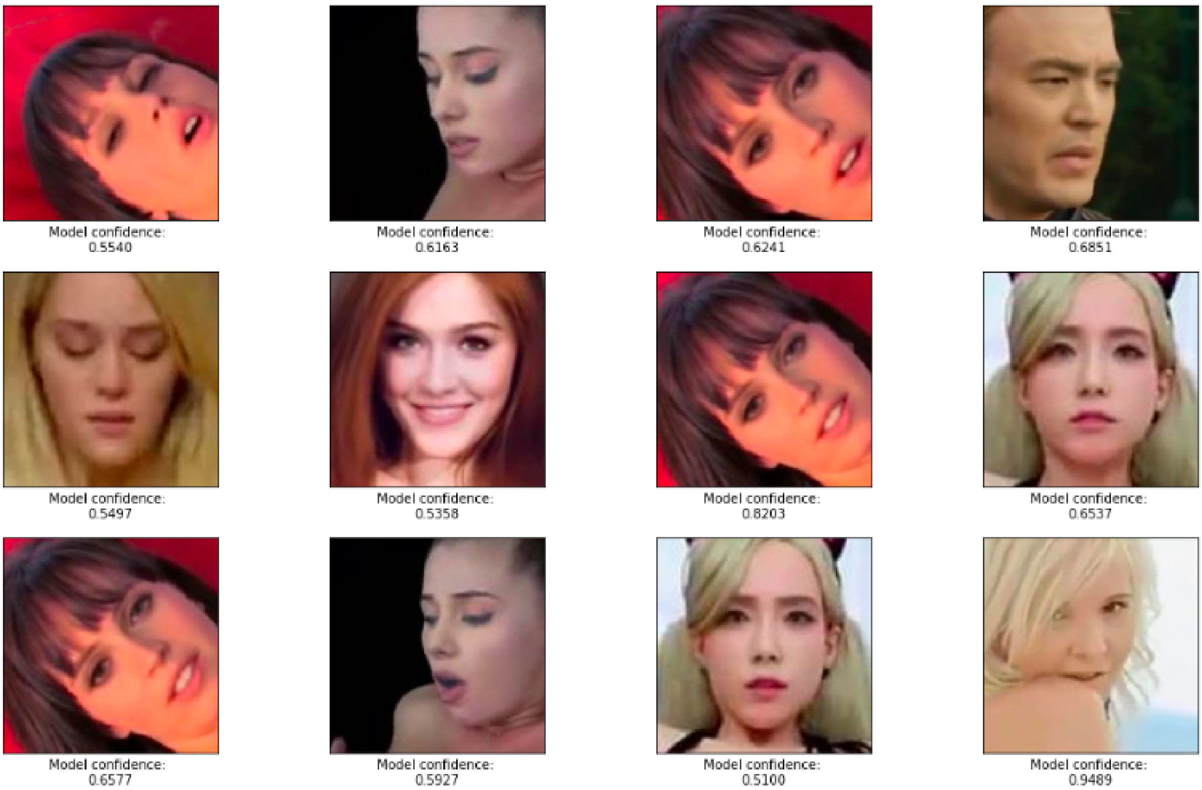

Figure 4.6: Misclassied Deepfake vs Misclassified Deepfake Predicted Output

Chapter 5

Discussion

Mesonet's method of collecting deepfake data,they acquired deepfake image data from popular deepfake video platforms that are online. According to the September 2019 report called the state of deepfakes conducted by deep trace labs,96 percent of deepfake media is pornographic. Deepfake pornography is among the worst abuses in the world of deepfakes and besides being a significant social issue this also complicates the technical side of our deepfake classifier. There is why Mesonet was trained on real data acquired from TV and movie sources that offer a great variety of facial expressions and settings. However, the deepfake data acquired was from popular deepfake platforms on the internet sources that are overwhelming dominated by pornographic content therefore it's expected that the model took advantage of a data artifact the statistical reality that deepfakes tend to be pornographic and reals tend to be non-pornographic and used it as a heuristic for its predictions. The model makes predictions under real conditions of diffusion on the internet which explains their use of popular deepfake platforms including pornographic websites. One wonders then whether we can neutralize this effect and force the model to recognize deepfakes without the aid of statistical accidents by using some different data. Since deepfake data is limited in supply and overwhelmingly pornographic. We could acquire our real data from pornographic sites as well rather than just TV shows and movies and unlike deepfake data pornographic is relatively easy to find on the internet. I examined a model designed to identify deepfake images called Mesonet and I implemented it to explore how it works in doing so. I did encounter a problematic data artifact that an overwhelming amount of deepfake images are indeed pornographic. I explained why this matters and I suggested how we could potentially neutralize this effect.

Chapter 6

Conclusion & Future Scope

Deepfakes is a synthetic face generation technology that uses a GAN. Due to the continuous advancement of video processing technology and the improvement in quality of videos, deepfake detection becomes more difficult. The current methods for detecting deep forgery can identify signals such as borders, shadows and uniform artifacts or double eyebrows, but the technology that produces deep forgeries is developing rapidly, so we must devote ourselves to building more technologies and detection tools. One of the biggest limitations when a video is tampered with is the lack of real data set that can be used to test new detection technologies, so Facebook is using paid actors to commission the first data set of this type for AI community use. It is part of the DFDC [14],that was also established in in cooperation with AWS, Microsoft, the AI Media Integrity Association Committee and academia. The mission of DFDC, to encourage researchers all around the world to develop fresh cutting-edge technologies,which will enable to spot counterfeit tampered media [15]. We hope that by helping the AI community come together, we can foresee the challenges of this emerging technology. The research community is committed to developing the Deepfake detection algorithm to solve the disturbing deepfake problem, and has published a number of research results. This report implements Mesonet using Meso 4 to make predictions on image data. We will be using the Meso 4 model Trained on the deepfake Data set. we will examine four sets of images-correctly classified deepfakes, correctly classified reals, misclassified deepfakes, misclassified reals.

Obviously,a war is brewing between the people who use advanced machine learning to generate advanced deep forgeries and those who try to detect them. Because authenticity provides a secure environment, deepfake technology must be integrated by applying a real layer across the Internet to provide trust metrics in social media and other networks. The detection of distorted vision content has become a hot topic in the scientific community. We will continue to study successful defense strategies. We will explore the new network model to more accurately discover the content of Deepfake , which may continue to be part of our project.

Bibliography

[1] Faceapp, https://www.faceapp.com/

[2] https://www.youtube.com/watch?v=Dx59bskG8dc

[3] https://www.youtube.com/watch?v=U-mg7a1vwkw

[4] https://www.youtube.com/watch?v=bE1KWpoX9Hk

[5] M.Kowalski, "faceswap," https://github.com/MarekKowalski/FaceSwap.

[6] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner,"Face2Face: Real-Time Face Capture and Reenactment of RGB Videos," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp. 2387–2395, November 2016.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair,A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680.

[8] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 8261–8265.

[9] N. Bonettini, D. Güera, L. Bondi, P. Bestagini, E. J. Delp, and S.Tubaro, "Image anonymization detection with deep handcrafted features," pp.2304–2308, Sep. 2019.

[10] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2017, pp. 1831–1839.

[11] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," arXiv preprint arXiv:1906.06876, 2019.

[12] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection," arXiv preprint arXiv:1812.02510, 2018.

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

[14] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detec- tion challenge (DFDC)preview dataset.

[15] "facebook deepfake detection dataset," https://deepfakedetectionchallenge.ai/