Deep Fake Detection using Deep Neural Architectures

Pruthivi Raj Behera
MT20037
pruthivi20037@iiitd.ac.in

Prabal Jain
MT20115
prabal20115@iiitd.ac.in

T G Narayanan
MT20027
narayanan20027@iiitd.ac.in

ABSTRACT

Deep Learning Techniques have been applied to various complex real-world problems, from big data analytics to computer vision and human-level control. Generative adversarial networks using deep learning lead to creating very realistic deepfake videos by playing with the digital content of images and videos, deep learning advances; however, they have also been employed to create software that can cause threats to privacy, democracy, and national security. DeepFake algorithms are used to create fake images and videos which are indistinguishable by humans. This paper presents an efficient approach to detect deep fake videos with the help of state-of-art Models. This study provides a comprehensive way to deal with different Deep Fake Detection Techniques.

Many approaches are available in the recent literature. We used a similar kind of approach with a few modifications to the Architecture of state-of-art models, which are being used for the Detection of Fake Videos. Our approach Defines a simple way of detecting DeepFake videos with the help of frame by frame classification. We have presented Baseline models, which Comprise CNN and MesoNet on a subset of DFDC Dataset; for the proposed approach, we have used a combination of MesoNet and Inception Module architecture on 5000 videos with 32 frames per video. We achieved an accuracy of 80.48 on Test Data. Thus obtained results proposed new architecture with relevant accuracy.

1. INTRODUCTION

Deepfake is a technique used to create convincing images/videos by replacing the face of a source person with the face of the target person. The results produced using deep fake models are quite advanced. Public figures such as politicians or celebrities are the ones targeted the ones. The deepfake results with such influential figures as a subject can prove to be lethal to society. Moreover, with a vast collection of their images already available online, the deepfake results using them are near indistinguishable. The only way to distinguish it from the actual video is to look for the slight inconsistency in the hair of the person, asymmetry in the face structure, morphed teeth, and the background of the face. This brings about the dire need for deep fake detection tools that are well trained with consistent deep learning techniques. Deepfake involves superimposing images of the face of the target person to a video with another person who is doing or saying something. This causes trouble as the generated video looks real enough to cause confusion among the viewers.

There are certain useful applications of deepfake, such as shooting stunt scenes in movies. The stunt artists' face is replaced with that of the main actor's face. This technology is further extended to create fake scenes to complete certain scenes that cannot be shot in real. But as the demerits of this technology weigh upon the merits, it makes it necessary to have deep fake detection mechanisms. Our work focuses on deep fake detection in the videos. The detection technique is thus required to analyze facial expressions and movements. Certain cues used to detect a fake video are observing the eye movement. Due to less availability of celebrities' photos online with their eyes shut, it becomes difficult to mimic the eye closing movements present in the real videos. Observing the duration and irregularities in the eye pattern hence proves to be a great cue for achieveing deep fake detection goals.

During the initial years of the development of deepfake detection methods, the obviously visible inconsistencies were used as cues to detect a fake video. The recent methods, on the other hand, apply deep learning techniques to get the features-weights. As a video can either be fake or real, this makes it important to train the dataset on a large set of instances. This makes the DFDC dataset so popular for achieving genuine models that are capable of detecting deepfakes with more precision than the models developed using smaller datasets. The fake detection in the videos is not as simple as it is for an image. Fake image detection cannot be applied directly to the frames of the video because of its format. Instead of storing every frame as an image, compressed videos are formed using complex compression algorithms that store the minute changes in pixels from frame to frame instead of storing the complete frame data. This makes it necessary to have a different data processing approach for videos.Deepfake videos are generated with nominal resolution using appraoches such as sharing, scaling and rotation. This creates changes in the surrounding areas of the target image when added to the video. These inconsistencies can hence be exploited using deep learning models. MesoNet method such as MesoInception-4 are few of the many prevalent methods that can be employed to achieve our goal. This brings us to exploring the potential of MesoInception-4 on the highly renowned DeepFake Detec-

tion Challenge dataset. The dataset was available in two versions, one of which had 5000 videos which we have experimented with. Each one of them comprises of 32 frames. The DFDC dataset was created by unanimous efforts of teams from AWS, Facebook, Microsoft, the Partnership on AI's Media Integrity Steering Committee, and various academics. The dataset was created with the goal to encourage deep learning enthusiasts all over the world to develop deepfake detection models.

The MesoInception-4 contains an architecture that comprises of four convolution layers and two dropout layers. We have modified the architecture a bit by employing three instead of four convolution layers and instead replacing the fourth layer with three additional convolution layers, followed by dropout layers and a softmax layer.

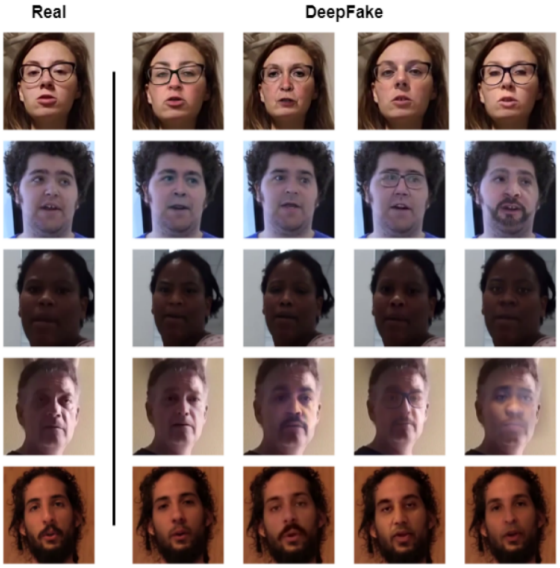


Figure 1: Sample Deepfakes (Top), Sample real (Bottom)

2. RELATED WORK

Some of the existing works are already done in this area. One of the first works in this field was done by [6] with recurrent neural networks. In this paper, the authors created a dataset with 300 videos from video-streaming websites. Additional 300 videos were taken from HOHA dataset to increase the size of dataset. For the method, a Conv-LSTM model was implemented with 80 frames for every video. The channel mean was subtracted from each channel and every frame was resized to 299x299.

Some more strides were made by [8] in the field of recurrent network. In the paper, attempts are made to distinguish the videos prepared using Deepfake, Face2Face or FaceSwap techniques. FaceForensics++ dataset was used for evaluation and a 4.55% increase in accuracy was obtained.

Another recent work was done by [2] but with CNN which comprised of optical flow based techniques. In this paper, they have used optical flow fields to distinguish deepfakes from original ones. PWC-Net is a forward flow technique to extract optical flow using the CNN model. A semi-trainable CNN named Flow-CNN is given a input of computed optical flow. They have tested VGG16 and ResNet50 as backbones. The initial layers of network were fixed while last layers including dense one were trained and adam optimizer was used.

Alternative methods such as different ensembling methods have also been tried. One such study was done by [7] to address the challenges posed by deepfake multimedia. The methods such as stacking ensemble and randomized weighted ensemble were used where the former uses the output of base-learners to train meta learner for better mapping of results and latter optimizes the weights and take its average.

Most recent work done is done by [5] in this field with convolutional traces. In this paper, a novel deepfake detection method focused on images representing human faces (convolutional traces) was introduced. Different GAN's like STYLEGAN2, FACEFORENSICS++, etc. were used to generate deepfake and classification tests were carried out on the obtained feature vectors with highest accuracies in range of 92-97%.

We will be experimenting the DFDC dataset [3] since it is by far the most detailed dataset which complies with the ethics of the society. Each person featured in the videos are volunteers or paid actors who are aware of the terms of use of their recordings. The DFDC dataset comprises of 128,154 clips sourced from 3,426 paid actors. Out of this, 104,500 are unique fake videos. Augmentations such as Gaussian blurring, grayscale conversion, horizontal flipping, resolution alteration, rotation, etc were applied to the test set. Various methods such as DFAE, MM/NN Face Swap, NTH, FSGAN and StyleGAN were used. DFAE produced the best results while StyleGAN produced the worst overall results, both at the frame level and at the video level.

[4] A novel DeepFake detection method focused on images representing hu-man faces (convolutional traces) was introduced. To Extract the Convolu-tional Traces Expectation-Maximization Algorithm was used which capturedthe correlation among the pixels to discriminate between the real ones and Fakeones. EM algorithm consists of two steps Expectation step and Maximiza-tion step. For Experiments, different GANs like STYLEGAN, STYLEGAN2, FACEFORENSICS++, etc. were used to generate the Deep Fake images withthe help of pre-trained models. Different GAN's like STYLEGAN2, FACE-FORENSICS++, etc. were used to generate DeepFake, and classification testswere carried out on the obtained feature vectors with the highest accuracy inthe range of 92-97 percent.
An Effective and Fast way of DeepFake DetectionMethod was proposed by M.A. Younus and T.M. Hasan [9] in which they haveused Haar Wavelet Transform and Edge Detection to detect DeepFakes. ForPreprocessing faces were extracted from Original images after that they per-form Haar Wavelet transform on the extracted image and also on the rest ofthe Frame dis-jointly then E-type analysis, E sharpness analysis is performedwhich are the Edge maps constructed in each scale, For E-Type analysis If theimage is not blurred then it is classified as Real otherwise Blur Ex-

tent (A) andBlur Extent(B) from E-Type Sharpness Analysis of Extracted Face and Rest ofthe Frame are taken into account respectively. Here Blur Extent represents the3 image blur Coefficient, So if the blur is found compare the Blur Coefficient withthe rest of the image, if Blur Extent of the rest of the frame is greater than theBlur Coefficient of the Extracted image then it is classified as Real Otherwiseit is Fake. They have used the DeepFake Video Dataset which is "UADFV" inwhich a total of 98 videos are there out of which 50 percent are fake and 50percent are real and contains total frames of about 35280. On Performing overthe Dataset their Proposed model achieved an accuracy of 90.5 percent which is greater than the Existing models like Meso-4, MesoInception-4, TwoStreamNN, and Head Pose models. Thus their model effectively Captured DeepFakesby detecting the Difference between Extracted image and the Rest of the Image using Haar Wavelet transformation.

3. METHODOLOGY

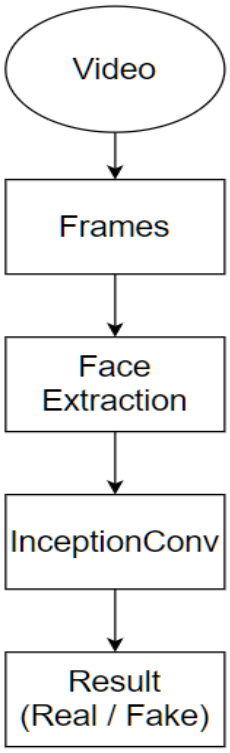


Figure 2: Black Box Architecture

In this section, we will describe our proposed DeepFake Detection Algorithm. During the literature survey, we found out that a pre-trained model like EfficientNets significantly outperforms other encoders and many people used it to score well on the leaderboard. Our approach is simple and sequential. We used frame by frame classification approach, i.e., for each of the frames extracted from the video, our model will classify it as real or fake based on the model and a majority vote.

We came across a huge data which contains 124k videos with eight facial modification algorithm. Data was further divided into 50 batches. As per our Computational Threshold, we decided to extract 100 videos from each batch to make our data normalized and not biased towards the count of fake and real video. We decided to use 100 videos from each batch in which 50 were fake, and the remaining 50 were real.

We used fake videos corresponding to the real video for efficient learning, so our final data contains a total of 5000 videos in which 2500 videos were fake, and 2500 videos were real corresponding to the fake videos. During the Literature Survey of various frame-by-frame approaches, we found that 32 frames per video perform well on Deep Convolutional Neural Network models. The face is essential in detecting deepfakes, so we used a simple MTCNN detector to extract faces from the images. During face extraction, we found out that some images extracted from the videos by MTCNN were not required faces. So, in order to provide correct data to our model, we removed extracted images in which the face was not extracted.

During literature survey, we found out that MesoNet models performs well on Facial Forgery Detection, so we followed the combined architecture of MesoNet and Inception Modules from scratch (architecture is shown in figure 6) in which we used three blocks of Inception Modules which consist of the Input layer, 1x1 convolution layer, 3x3 convolution layer, 5x5 convolution layer, Max pooling layer, Concatenation layer, Pooling downsamples the input data to create a smaller output with a reduced height and width. Within an Inception module, we add padding(same) to the max-pooling layer to ensure it maintains the height and width as the other outputs(feature maps) of the convolutional layers within the same Inception module. By doing this, we ensure we can concatenate the outputs of the max-pooling layer with the outputs of the conv layers within the concatenation layer, within a convolutional layer there is ReLU(rectified linear unit) activation function which is utilized. The Inception module allows for the utilisation of varying convolutional filter sizes to learn spatial patterns at different scales. So in our Proposed architecture we added 3 different blocks of Inception Module.

Our data contains 159116 frames with the corresponding label. We resized our images to 256x256, so to train our model on this data, we decided our train data 80 Percent, out of which 20 Percent is validation data and the remaining 20 Percent for test data, So we train our model on 101833 frames and perform predictions on 25459 frames we were able to achieve an accuracy of **80.4684**. Respective accuracy and loss plots are shown in figure 9. For our model, we tried with epochs [10,20,30], and we found out that after 30 epochs model starts to vary a lot in terms of loss and accuracy on validation data.

We proposed to detect Fake Videos generated by various GAN's on MesoScopic level of Analysis as Analysis based on Noise in an image in a Compressed Video would not be efficient So we adopted an intermediate approach using a deep Convolutional Network with Less number of Layers comparing to Very deep Neural Network like ResNet, InceptionNet, etc. to make it Computationally Effective In this Model

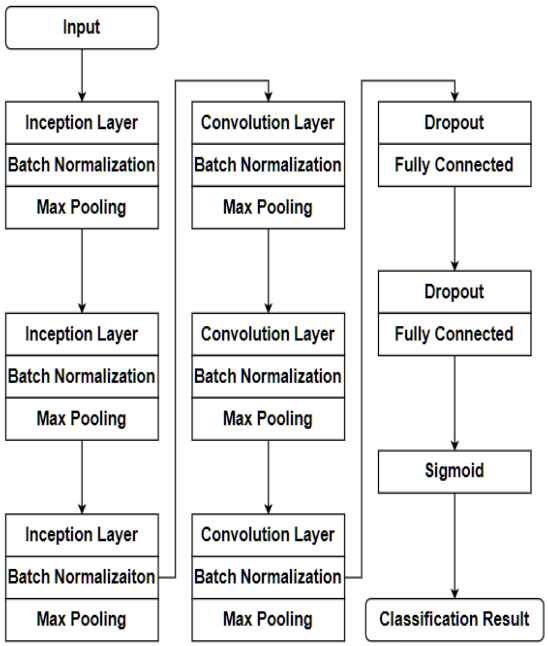


Figure 3: Final Architecture

architecture we used 3 Inception Blocks in which Each Inception Blocks consisting of Several Convolutional Layers after every Convolutional Layer a Dilated Convolution was applied, Thus it Provides better Feature Extraction.

4. DATASET

We partnered with other industry leaders and academic experts in September 2019 to create the Deepfake Detection Challenge (DFDC) in order to accelerate development of new ways to detect deepfake videos. In doing so, we created and shared a unique new dataset for the challenge consisting of more than 100,000 videos. The DFDC has enabled experts from around the world to come together, benchmark their deepfake detection models, try new approaches, and learn from each others' work.

The DFDC dataset consists of two versions, one is Preview dataset having 5k videos that features two facial modification algorithms. The second one is Full dataset consisting of 124k videos that features eight facial modification algorithms. Some generic observations for fake faces are as follows:

- Two eyes do not match, iris colors are different
- Asymmetrical face
- Misaligned teeth, stain/ mark on teeth, gap between teeth
- Difference between two ears
- Something odd about shape of earlobe
- There is something odd about earrings, both earrings do not match
- Stain or patch in skin
- Hint of presence of other unexpected objects on skin/neck
- Something odd about other people in background
- Some asymmetry between two sides of collar/ sleeve of dress

Figure 4: Face Extracted Dataset

5. BASELINES

We have implemented the following 2 variants for each of the 2 baselines models as mentioned below. For implementation, we worked on a subset of the dataset performing undersampling with 20 frames for fake videos and oversampling with 150 frames for real videos to combat class imbalance. We have also tabulated the results as shown in Table 1 and generated plots as shown in Figure 7.

5.1 Deep Convolutional Neural Net

Convolutional neural network (CNN) has been traditionally proven to be effective in fields such as image classification and recognition. The advantage of using CNN is that it can automatically learn complex feature utilizing massive simple neurons and back-propagation. Figure 5 shows the architecture for our deep convolutional neural network in which we used ReLu activation Function with kernel size 3x3, we have used our CNN Model on input frame 64x64 on SGD and Adam optimiser, Also last hidden layer has the keep probability of 0.5 and our output Layer consist of 2 units each units will give the probability of image belonging to that particular class.

5.2 MesoNet

In MesoNet [1], the authors presented a method to detect face tampering in videos via deep learning approach. Before processing with Meso-4, the images are resized to the size of 256x256. In Meso-4 architecture as shown in figure 6, the network begins with a sequence of 4 convolution layers followed by a pooling layer. The convolution layers use ReLU activation to improve generalization. Batch normalization layers are used to regularize the outputs and prevent vanishing/exploding gradients. Dropout layers are also used to prevent overfitting and prevent complex computations due to the immense amount of data.

5.3 Baseline Results

The CNN model with Adam performed the best by giving a training loss of 0.066 and testing loss of 0.134. The training accuracy was about 97.58% and the testing accuracy was about 95.08%. Meso-4 performed better than Meso-1, reuslting in an training accuracy of 93.38% and a testing accuracy of 92.52%, giving training loss of about 0.07% and testing loss of 0.055%.

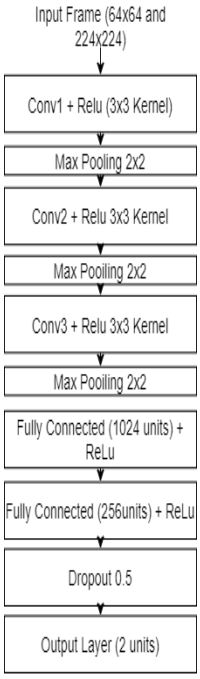


Figure 5: CNN Architecture

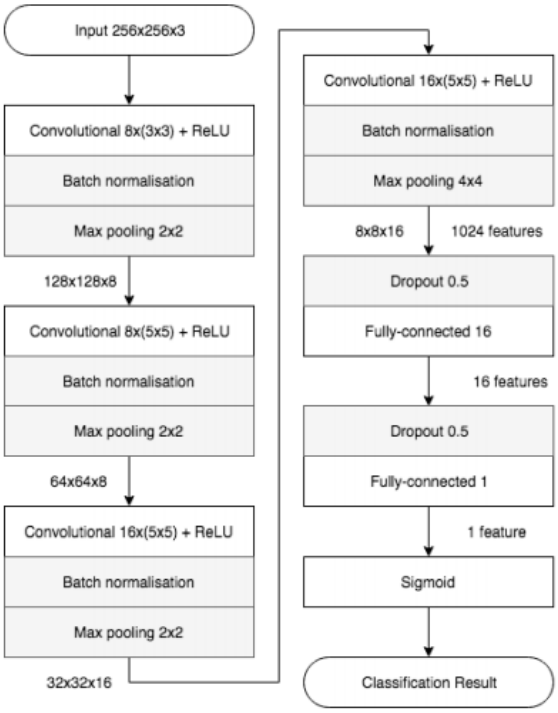
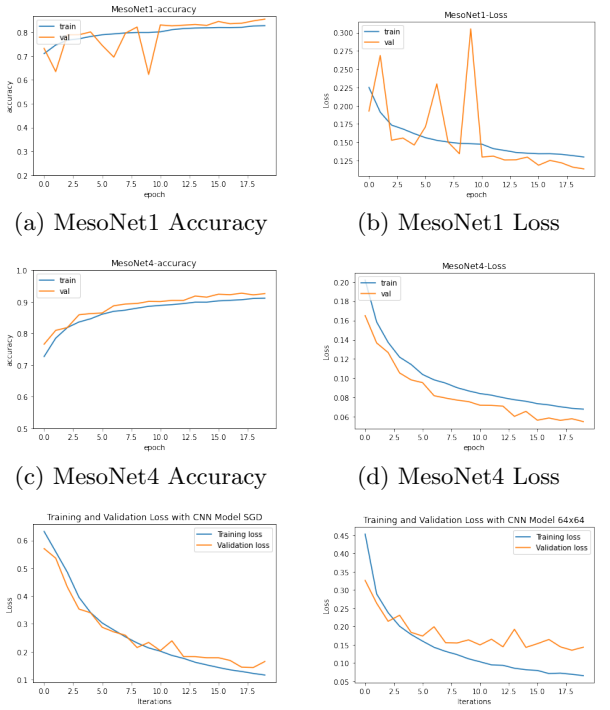


Figure 6: Meso-4 Architecture

6. ANALYSIS

6.1 Results

We have Implemented Different Architecture : MesoNet and a CNN based Model for our Baselines, Due to computational BottleNeck we performed Classification on Subset of Videos



(a) MesoNet1 Accuracy (b) MesoNet1 Loss

(c) MesoNet4 Accuracy (d) MesoNet4 Loss

(e) CNN Model SGD Loss (f) CNN Model Adam Loss

Figure 7: Accuracy and Loss Curves

Table 1: Classification accuracies and Loss for baseline models performed on DFDC dataset.

Model	Train Loss	Test Loss	Train Acc	Test Acc
CNN-SGD	0.119	0.169	95.09	93.19
CNN-Adam	**0.066**	**0.134**	**97.58**	**95.08**
Meso-1	0.130	0.113	86.03	85.85
Meso-4	**0.070**	**0.055**	**93.38**	**92.52**

Table 2: Classification Report (InceptionConv)

	precision	recall	f1-score
0	0.845444	0.746878	0.793110
1	0.772668	0.863033	0.815354
accuracy	0.804864	0.804864	0.804864
macro avg	0.809056	0.804956	0.804232
weighted avg	0.809113	0.804864	0.804215

Provided in the batch 'dfdc_train_part_13.zip' Due to Class Imbalance for Fake Videos we decided to increase the number of Frames for Real Videos by the Factor of 3, and we achieved Significantly Good Acuracy on our Test Set using CNN Model with Adam Optimizer and Meso-4 Model with Testing Accuracy of 97.58 and 93.38 respectively. Furthermore we analyze that videos in the particular match were not shuffled and contains more videos of a the same actor which is where we found the bottleneck of our CNN and Meso-4 Based Model. Results for the Baseline are shown in

table 2. After Implementing Our Proposed Method we get the a Final Accuracy of 80.48 percent as given in figure 9 . We analysed on our Loss plots and Accuracy Plots on Test Data and after epoch 10, Loss on Validation Data starts to Fluctuate as training Loss Decreases Smoothly.

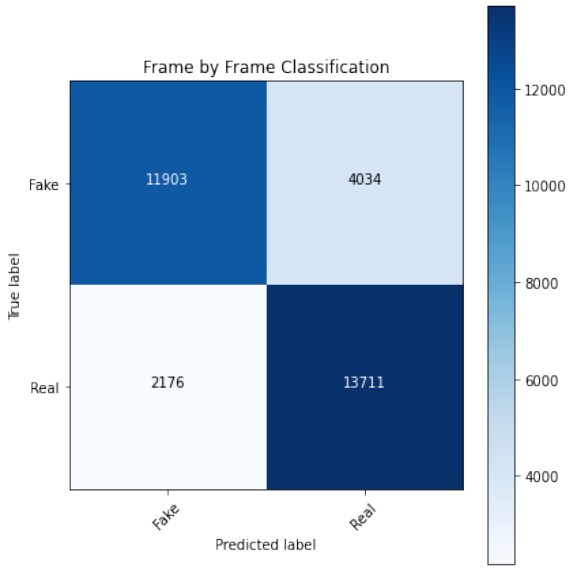


Figure 8: Confusion Matrix

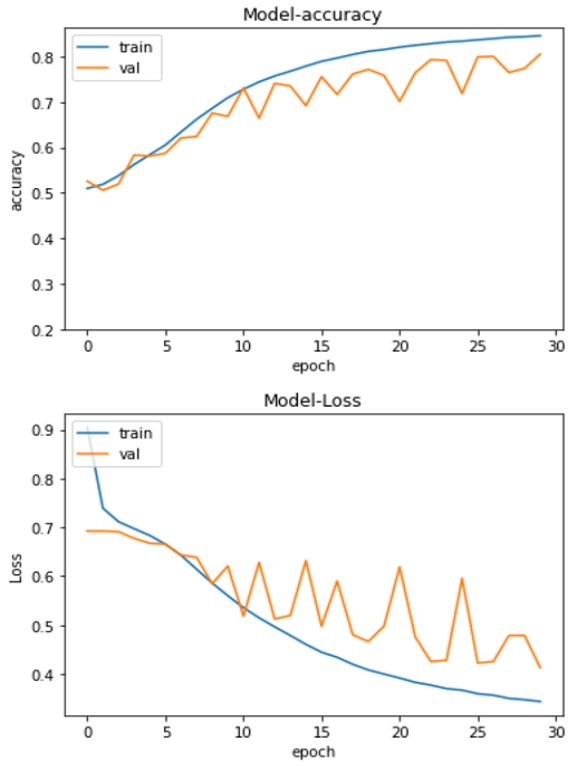


Figure 9: Accuracy and Loss Plots

6.2 Error

During Implementation of Baseline Models we Discarded the Architecture of MesoNet-1 Model as we can see Loss Plot of MesoNet1 Performs inefficiently as validation loss increases abruptly on epoch 9 Providing a decline in Accuracy, On the Other hand MesoNet 4 Performs Smoothly to fit the Image data to Classify it as Fake or Real we performed 20 iteration on the Subset of Deep Fake Data and achieved an Accuracy of 92.52 But we realised the Limitation of MesoNet4 as Model was biased Towards the Data and Hence it Performs inefficiently on the whole data. Figure 7 contains Loss Plots for MesoNet-1 and MesoNet-4. So to reduce further error we proposed a Deep Architecture for Classification Problem as stated in Methodology.

6.3 Explainability

CNNs typically consist of multiple convolution and pooling layers which help the deep learning model in automatically extracting relevant features from visual data like images. Due to this multi-layered architecture, CNNs learn a robust hierarchy of features, which are spatial, rotation, and translation invariant. Any image can be represented as a tensor of pixel values. The convolution layers help in extracting features from the image (forms feature maps). Shallower layers (closer to the input data) in the network learn very generic features like edges, corners and so on. Deeper layers in the network (closer to the output layer) learn very specific features pertaining to the input image.

Inner working of our Model states extracting important very fine features to perform classification of a given video frame by frame to real or fake class, Our model can be reuse and Provided more Computational Resources a Deep Combination of Convolutional Neural Network can be deployed. Inception Modules can be Utilized by Performing Tuning for the Number of Layers and Inception Modules to be integrated into the Network, As we were able to achieve 80 Percent Accuracy which is reliable and Trustworthy.

6.4 Visualization

Deep Neural Networks are one of the most powerful class of machine learning models, It has only one drawback these networks are completely black-box. We still have very little knowledge as to how deep Convolutional Neural networks learn their target patterns so well, especially how all the neurons work together to achieve the final result , we now have the ability to visualise the filters that Convolutional Neural Networks (CNNs) learn from their training By visualising the learned weights we can get some idea as to how well our network has learned. For example, if we see a lot of zeros then we'll know we have many dead filters that aren't going much for our network which is a great opportunity to perform pruning for model compression. The feature maps of a CNN capture the result of applying the filters to an input image that is at each layer, the feature map is the output of that layer. The reason for visualising a feature map for a specific input image is to try to gain some understanding of what features our CNN detects perhaps it detects some parts of our desired object and not others or the activations die out at a certain layer, Also early layers of the network detect low-level features (colours, edges, etc) and the later layers of the network detect high-level features (shapes and objects). So in this Section we visualized some

of the Convolutional Feature Maps.

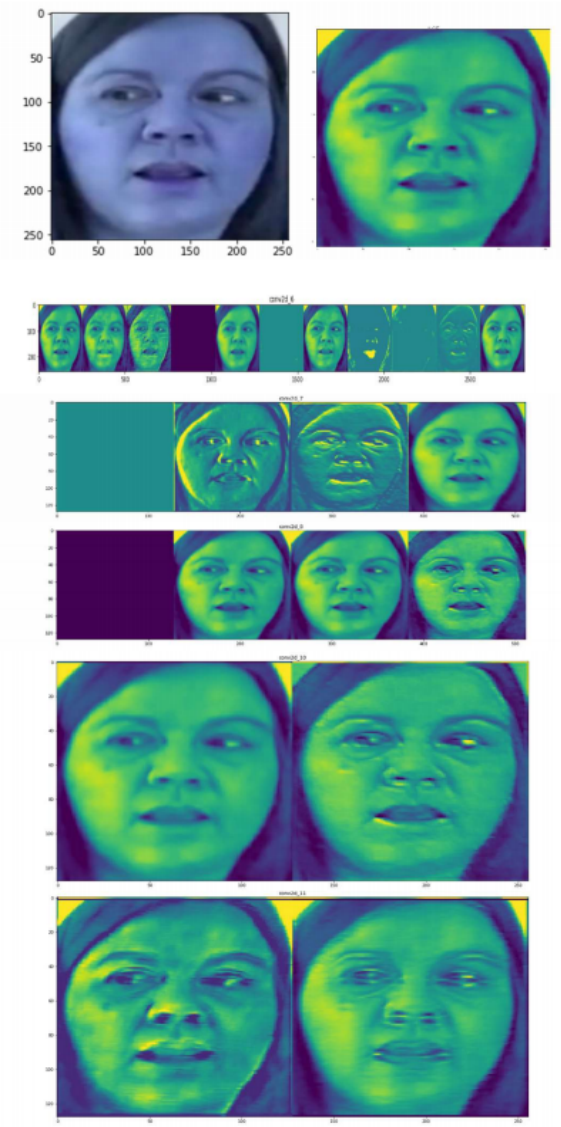


Figure 10: Feature Map Visualization

7. CONTRIBUTION

For parallel processing on huge video data, we divided our work as –

Pruthivi Raj Behera: Preprocessing of Video Batch [0-17], Baseline Meso1, Proposed Architecture, Implementation of InceptionModule, Training Model.

Prabal Jain : Preprocessing of Video Batch [18-35], Baseline Meso4, Analysis, Hyperparameter Tuning on InceptionConv, Training Model.

T G Narayanan : Preprocessing of Video Batch [36-49], Data Cleaning, CNN-Adam, CNN-SGD Implementation, Hyperparameter Tuning InceptionConv, Training Model Report.

All contributed equally towards the literature review.

8. CONCLUSIONS

In this paper, the Frame by Frame Classification method was used to classify the frames as real and fake as it reveals its importance in the detection of DeepFake Videos. The results are exciting and show that a combination of architecture of MesoNet with inception blocks provides better accuracy on a frame by frame classification. Our approach was applied to 5000 videos available in the datasets. Due to computational complexity, 100 videos from each batch were considered. Our approach should be applied to the whole dataset of 124k videos to achieve the robustness of the model. Following our results, we can create a simple Video classifier based on the features extracted by the Deep Convolutional Neural Network. This method detects fake videos independent of the generator and the type of video. Authenticity is the need for a habitable world. For future work, our model should be trained on the full dataset. Data augmentation like image compression, noise, blur, resize with different interpolations, color jittering, scaling, and rotations must be performed to achieve robustness. Time sequence models such as RNN with a pre-trained model will perform well if provided more computing resources. Variants of Efficient Nets (Pretrained) were effective on the DFDC dataset. Our method provided significant accuracy on the test set.

References

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

I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 1(2), 2020.

L. Guarnera, O. Giudice, and S. Battiato. Fighting deepfake by exposing the convolutional traces on images. *IEEE Access*, 8:165085–165098, 2020. doi: 10.1109/ACCESS.2020.3023037.

L. Guarnera, O. Giudice, and S. Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 666–667, 2020.

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

M. S. Rana and A. H. Sung. Deepfakestack: A deep ensemble-based learning technique for deepfake detection. In *2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, pages 70–75. IEEE, 2020.

E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3 (1), 2019.

M. A. Younus and T. M. Hasan. Effective and fast deepfake detection method based on haar wavelet transform. In *2020 International Conference on Computer Science and Software Engineering (CSASE)*, pages 186–190, 2020. doi: 10.1109/CSASE48920.2020.9142077.