Detection of Deepfake

A Project Report submitted

In partial fulfillment of the requirements for the degree of

Bachelor of Technology in Computer Science

Submitted by

Anwayajyoti Dey(192025004)

Samrat Dey(192025023)

B.Tech CSE, 8th Semester

Royal School of Engineering and Technology

Under the guidance of

Dr. Anupam Das

Associate Professor

Royal School of Engineering and Technology

THE ASSAM ROYAL GLOBAL UNIVERSITY

GUWAHATI: 781035

Session: 2019-2023

ROYAL GLOBAL UNIVERSITY
GUWAHATI

CERTIFICATE OF APPROVAL

It is certified that the work contained in the report entitled **"Detection of Deepfake"** by **Anwayajyoti Dey** bearing Roll No 192025004 and **Samrat Dey** bearing Roll No 192025023 of **Bachelor of Technology, 8th Semester** under the **Computer Science Department**, **Royal School of Engineering and Technology**,The Assam Royal Global University, Guwahati, Assam for the fulfillment of the degree of **Bachelor of Technology** has been carried out under my supervision and that work has not been submitted elsewhere for a degree.

Date:

Place: Guwahati

Project Guide:

Dr. Anupam Das

Associate Professor,

Department of Computer Science and Engineering

Royal School of Engineering & Technology

ROYAL GLOBAL UNIVERSITY
GUWAHATI

FORWARDING CERTIFICATE

It is certified that the work contained in the report entitled "**Detection of Deepfake**" by **Anwayajyoti Dey** bearing Roll No 192025004 and **Samrat Dey** bearing Roll No 192025023 of **Bachelor of Technology, 8th Semester** under the **Computer Scienece Department**, **Royal School of Engineering and Technology** , under the guidance of **Dr. Anupam Das, Associate Professor** has been presented in a manner satisfactory to permit its acceptance as a prerequisite to the degree for which has been submitted.

Date:

Place: Guwahati

Dr. Ishita Chakraborty
HOD, Department of Computer Science & Engineering

DECLARATION

We, **Anwayajyoti Dey** bearing Roll No 192025004 and **Samrat Dey** bearing Roll No 192025023 hereby declare that this project work entitled "**Detection of Deepfake**" was carried out by us under the guidance & supervision of **Anupam Das, Associate Professor**. This project work is submitted during the academic session 2019-2023. This work or no part of it has been submitted elsewherefor any other purpose till date.

Date:
Place: Guwahati

Anwayajyoti Dey	**Samrat Dey**
Roll No. 192025004	**Roll No. 192025023**

ACKNOWLEDGEMENT

I take the opportunity to express my sincere gratitude to all those who supported me throughout this project work. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work.

My special thanks also go to my Faculty Guide and other faculty members for their kind support for the successful completion of my project/dissertation.

Thank you

Anwayajjyoti Dey
Roll No. 192025004

Samrat Dey
Roll No. 192025023

ABSTRACT

The growing computation power has made the deep learning algorithms so powerful that creating an indistinguishable human synthesized video popularly called as deep fakes have become very simple. Scenarios where this realistic face swapped deep fakes are used to create political distress, fake terrorism events, revenge porn, blackmail peoples are easily envisioned. In this work, we describe a new deep learning-based method that can effectively distinguish AI-generated fake pictures from real pictures. Our method is capable of automatically detecting the replacement and reenactment deep fakes. We are trying to use Artificial Intelligence (AI) to fight Artificial Intelligence (AI). Our system uses a MesoNet Convolution neural network to extract the frame-level features and these features and further used to train the Long Short-Term Memory (LSTM) based Recurrent Neural Network (RNN) to classify whether the image is subject to any kind of manipulation or not, i.e., whether the image is deep fake or real video. To emulate the real time scenarios and make the model perform better on real time data, we evaluate our method on large amount of balanced and mixed dataset prepared by mixing the various available dataset like Face Forensic++ [1], Deepfake detection challenge [2], and Celeb-DF [3]. We also show how our system can achieve competitive result using very simple and robustapproach

TABLE OF CONTENTS

LIST OF FIGURES

Chapter-1

Introduction

1.1 Synopsis

Deep fake is a technique for human image synthesis based on neural network tools like GAN (Generative Adversarial Network) or Auto Encoders etc. These tools super impose target images onto source images using a deep learning technique and create a realistic looking deep fake images. These deep-fake video are so real that it becomes impossible to spot difference by the naked eyes. In this work, we describe a new deep learning-based method that can effectively distinguish AI-generated fake images from real images. We are using the limitation of the deep fake creation tools as a powerful way to distinguish between the pristine and deep fake images. Deep fake creation tools leave distinctive artifacts in the resulting Deep Fake images, and we show that they can be effectively captured by Mesonet Convolution Neural Networks. Our system uses a Mesonet Convolution Neural Networks to extract frame-level features. These features are then used to train a Long Short-Term Memory (LSTM) based Recurrent Neural Network (RNN) to classify whether the image is subject to any kind of manipulation or not, i.e., whether the image is deep fake or real image. We tried to make the deep fake detection model perform better on real time data. To achieve this, we trained our model on combination of available datasets. So that our model can learn the features from different kind of images. We extracted an adequate number of videos from Face-Forensic++ [1], Deepfake detection challenge [2], and Celeb-DF [3] datasets.

1.2 Project Idea

In the world of ever-growing social media platforms, Deepfakes are considered as the major threat of the AI. There are many Scenarios where these realistic face swapped deepfakes are used to create political distress, fake terrorism events, revenge porn, blackmail peoples are easily envisioned. It becomes very important to spot the difference between the deepfake and pristine images. We are using AI to fight AI. Deepfakes are created using tools like FaceApp [11] and Face Swap [12], which using pre-trained neural networks like GAN or Auto encoders for these deepfakes creation. Our method uses a LSTM based artificial neural network to process the sequential temporal analysis of the image frames and pre-trained MesoNet CNN to extract the frame level features. MesoNet Convolution neural network extracts the frame-level features, and these features are further used to train the Long Short-Term Memory based artificial Recurrent Neural Network to classify the video as Deepfake or real. To emulate the real time scenarios and make the model perform better on real time data, we trained our method with large amount of balanced and combination of various available dataset like FaceForensic++ [1], Deepfake detection challenge [2], and Celeb-DF [3].

1.3 Motivation of the project

The increasing sophistication of mobile camera technology and the ever-growing reach of social media and media sharing portals have made the creation and propagation of digital images more convenient than ever before. Deep learning has given rise to technologies that would have been thought impossible only a handful of years ago. Modern generative models are one example of these, capable of synthesizing hyper realistic images, speech, music, and even video. These models have found use in a wide variety of applications, including making the world more accessible through text-to-speech, and helping generate training data for medical imaging. Like any trans-formative technology, this has created new challenges. So called "deep fakes" produced by deep generative models that can manipulate video and audio clips. Since their first appearance in late 2017, many open-source deep fake generation methods and tools have emerged now, leading to a growing number of synthesized media clips. While many are likely intended to be humorous, others could be harmful to individuals and society. Until recently, the number of fake images and their degrees of realism has been increasing due to availability of the editing tools, the high demand on domain expertise. Spreading of the Deep fakes over the social media platforms have become very common leading to spamming and peculating wrong information over the platform. Just imagine a deep fake of our prime minister declaring war against neighboring countries, or a Deep fake of reputed celebrity abusing the fans. These types of the deep fakes will be terrible, and lead to threatening, misleading of common people. To overcome such a situation, Deep fake detection is very important. So, we describe a new deep learning-based method that can effectively distinguish AI generated fake images (Deep Fake images) from real images. It's incredibly important to develop technology that can spot fakes, so that the deep fakes can be identified and prevented from spreading over the internet

1.4 Objectives

Our project will reduce the Abuses' and misleading of the common people on the world wide web.

Our project will distinguish and classify the video as deepfake or pristine.

1.5 Statement of the scope

There are many tools available for creating the deep fakes, but for deep fake detection there is hardly any tool available. Our approach for detecting the deep fakes will be great contribution in avoiding the percolation of the deep fakes over the world wide web. This project can be used to detect and spot differences between real and fake images.

Chapter-2

Literature Review

2.1 Background Study

Face Warping Artifacts [15] used the approach to detect artifacts by comparing the generated face areas and their surrounding regions with a dedicated Convolutional Neural Network model. In this work there were two-fold of Face Artifacts. Their method is based on the observations that current deepfake algorithm can only generate images of limited resolutions, which are then needed to be further transformed to match the faces to be replaced in the source image. Their method has not considered the temporal analysis of the frames. Detection by Eye Blinking [16] describes a new method for detecting the deepfakes by the eye blinking as a crucial parameter leading to classification of the image as deepfake or pristine. The Long-term Recurrent Convolution Network (LRCN) was used for temporal analysis of the cropped pixels of eye blinking. As today the deepfake generation algorithms have become so powerful that lack of eye blinking cannot be the only clue for detection of the deepfakes. There must be certain other parameters must be considered for the detection of deepfakes like teeth enchantment, wrinkles on faces, wrong placement of eyebrows etc. Capsule networks to detect forged images [17] uses a method that uses a capsule network to detect forged, manipulated images in different scenarios, like replay attack detection and computer-generated image detection.

2.2 Discussion

In their method, they have used random noise in the training phase which is not a good option. Still the model performed beneficial in their dataset but may fail on real time data due to noise in training. Our method is proposed to be trained on noiseless and real time datasets. Recurrent Neural Network [18] (RNN) for deepfake detection used the approach of using RNN for sequential processing of the frames along with ImageNet pre-trained model. Their process used the HOHO [19] dataset consisting of just 600 videos. Their dataset consists small number of images and same type of images, which may not perform very well on the real time data. We will be training out model on large number of Realtime data. Synthetic Portrait Images using Biological Signals [20] approach extract biological signals from facial regions on pristine and deepfake portrait image pairs. Applied transformations to compute the spatial coherence and temporal consistency, capture the signal characteristics in feature vector and photoplethysmography (PPG) maps, and further train a probabilistic Support Vector Machine (SVM) and a Convolutional Neural Network (CNN). Then, the average of authenticity probabilities is used to classify whether the image is a deepfake or pristine. Fake Catcher detects fake content with high accuracy, independent of the generator, content, resolution, and quality of the image. Due to lack of discriminator leading to the loss in their findings to preserve biological signals, formulating a differentiable loss function that follows the proposed signal processing steps is not straight forward process.

Chapter-3

FEASIBILITY STUDY AND REQUIREMENT ANALYSIS

3.1 Feasibility Study

A feasibility study considers various constraints within which the system should be implemented and operated. In this stage the resource needed for implementation such as computing equipment, manpower and costs are estimated. The estimated are compared with available resources and a cost benefit analysis of the system is made. The main objectives of the feasibility study are to determine whether the project would be feasible in terms of the following categories:

- Economic feasibility
- Operational feasibility

3.1.1 Economic Feasibility

Economic feasibility attempts to weigh the costs of developing and implementing a new system, against the benefits that would gather from having the new system in place. This feasibility study gives the top management the economic justification for the new system. A simple economic analysis which gives the actual comparison of costs and benefits are much more meaningful in this case. In addition, this proves to be a useful point of reference to compare actual costs as the project progresses.

3.1.2 Schedule Feasibility

The dateline of a software system can be easily estimated if the proper team and achievable goals are formed

Chapter 4

SYSTEM DESIGN

4.1 Use Case View

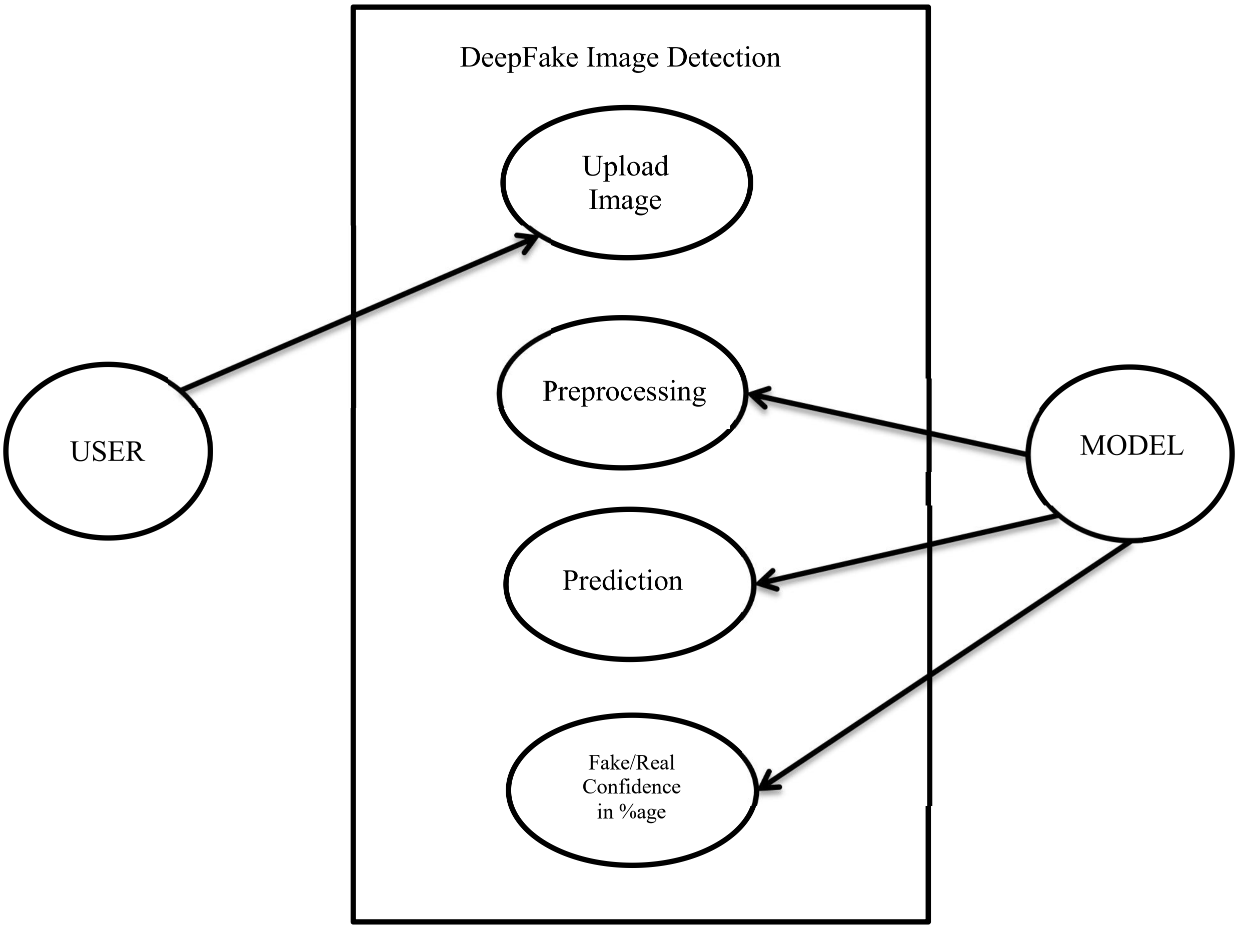


Figure 4.1: Use Case Diagram.

4.2 Data Flow Diagram

4.2.1 DFD level 0

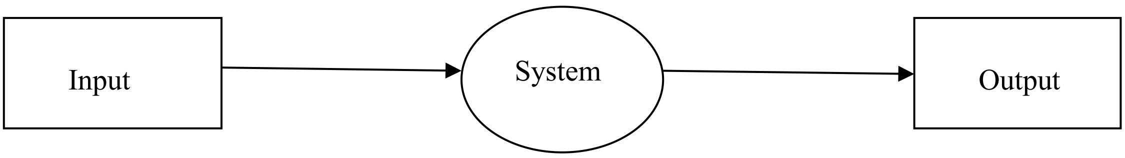


Figure 4.2.1 : DFD level 0.

DFD level – 0 indicates the basic flow of data in the system. In this System Input is given equal importance as that for Output. Input: Here input to the system is uploading image. System: In system it shows all the details of the Image. Output: Output of this system is it shows the fake image or not. Hence, the data flow diagram indicates the visualization of system with its input and output flow.

4.2.2 DFD Level 1

DFD Level 1 gives more in and out information of the system. Where system gives detailed information of the procedure taking place.

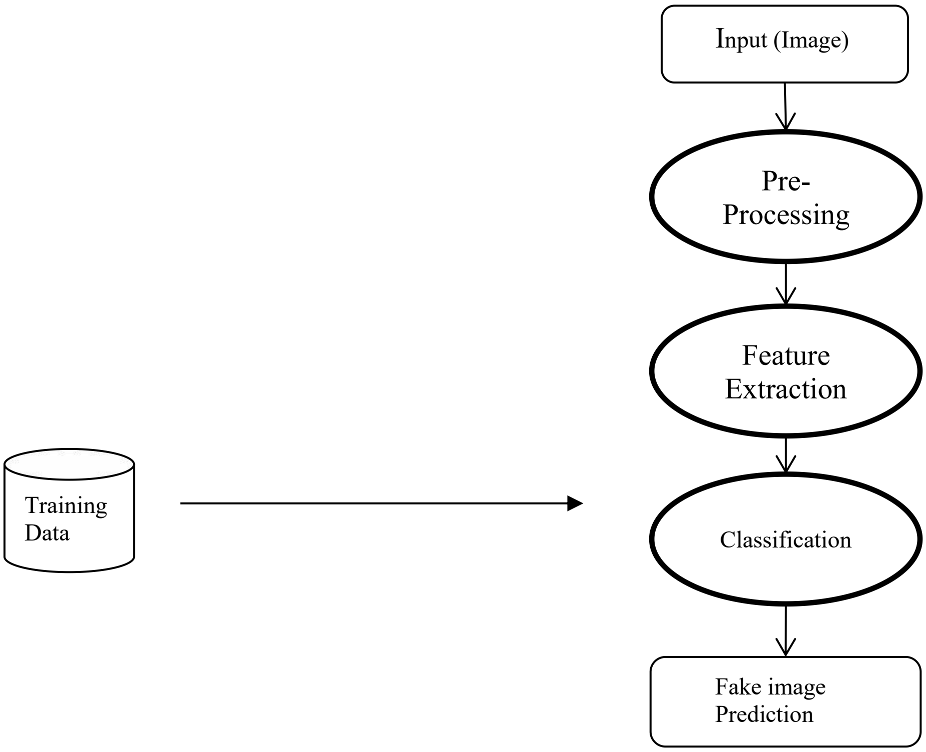


Figure 4.2.2: DFD level 1

4.2.3 DFD Level 2

DFD level 2 enhances the functionality used by user.

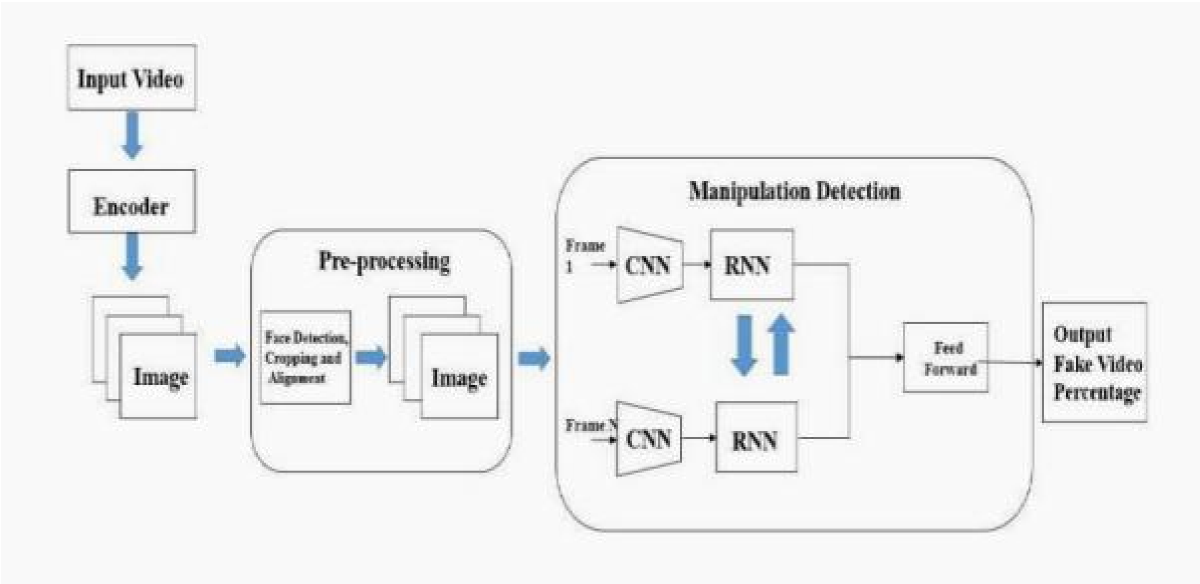


Figure 4.2.2: DFD level 2
Source: https://www.studocu.com/row/document/tribhuvan-vishwavidalaya/csit/deepfake-detection-my-project-report/42650432

4.3 Workflows

4.3.1 Training Workflow

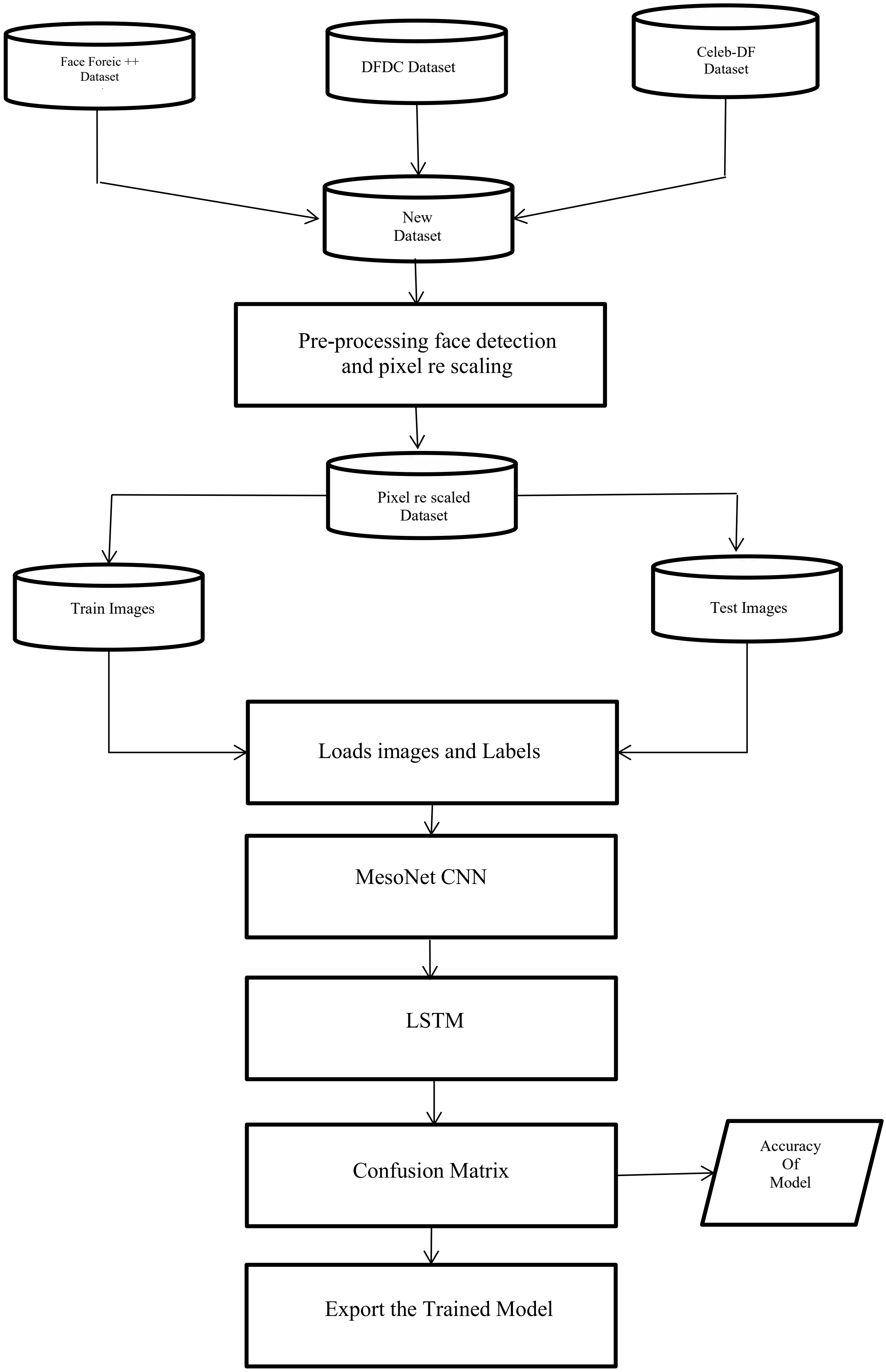


Figure 4.3.1 Training Workflow.

4.3.2 Testing Workflow

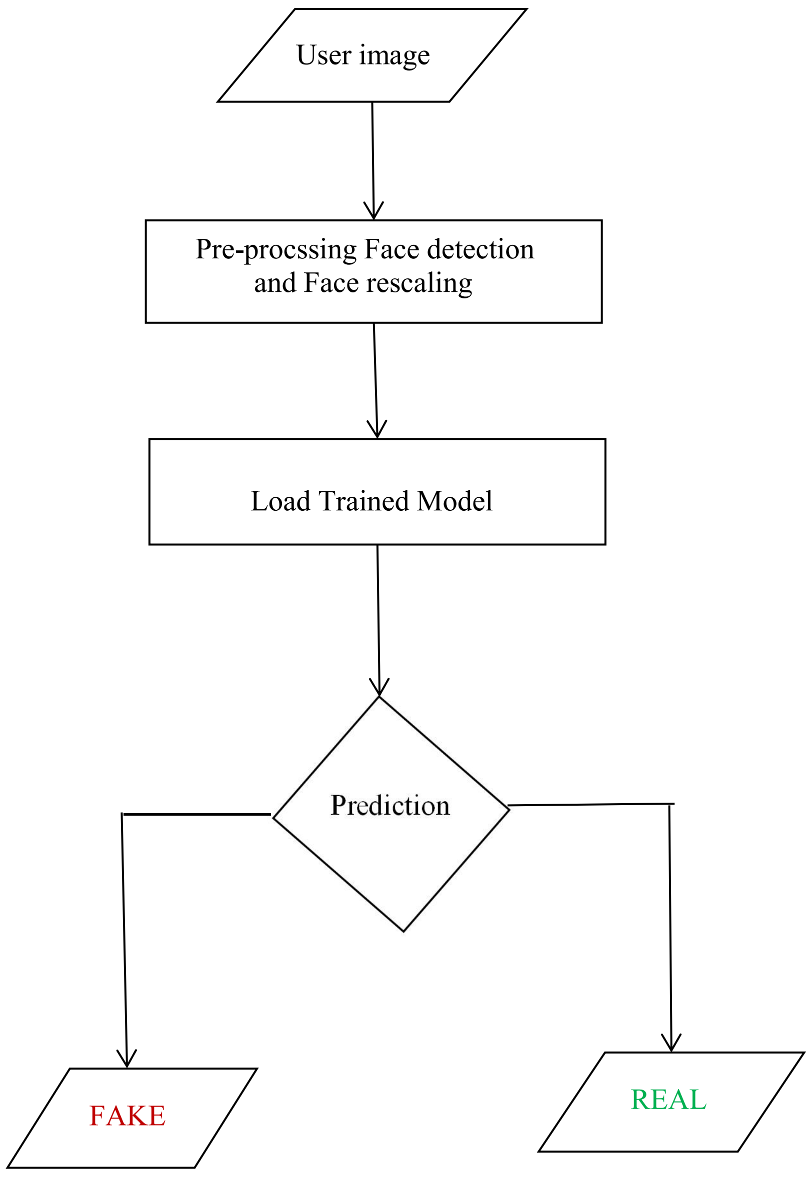


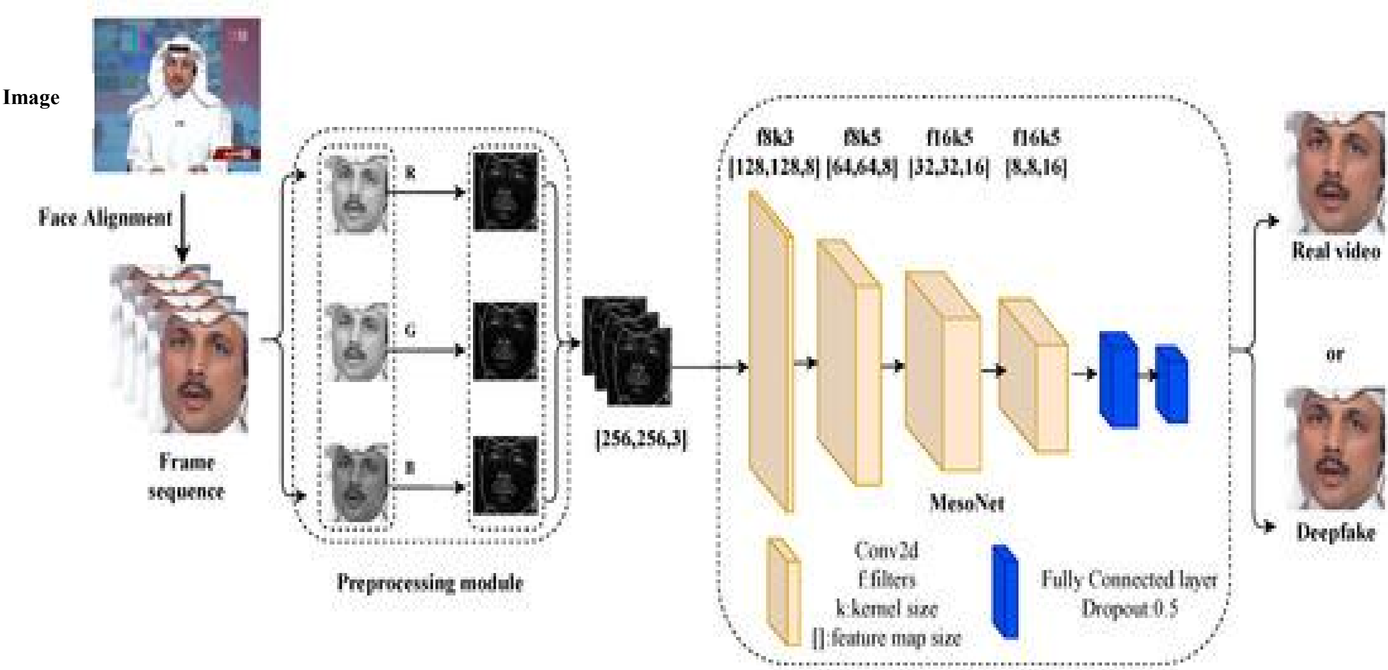
Figure 4.3.2: Testing Workflow

Chapter-5

METHODOLOGY

5.1 System Architecture

In this system, we have trained our deepfake detection model on ample number of real and fake images. The system architecture of the model is showed in the figure. In the development phase, we have taken a dataset, preprocessed the dataset and created a new processed dataset which only includes the face cropped images.



`Figure 5.1 System Archutecture.
Source: MesoNet: a Compact Facial Video Forgery Detection Network.
https://arxiv.org/pdf/1809.00888.pdf

5.2 Creating Deepfake Videos

The process to generate Deepfake images is to gather aligned faces of two different people **A** and **B**, then to train an auto-encoder **EA** to reconstruct the faces of **A** from the dataset of facial images of **A**, and an auto-encoder **EB** to reconstruct the faces of **B** from the dataset of facial images of **B**. The trick consists in sharing the weights of the encoding part of the two auto-encoders **EA** and **EB**, but keeping their respective decoder separated. Once the optimization is done, any image containing a face of **A** can be encoded through this shared encoder but decoded with decoder of **EB**. This principle is illustrated in the below Figure.

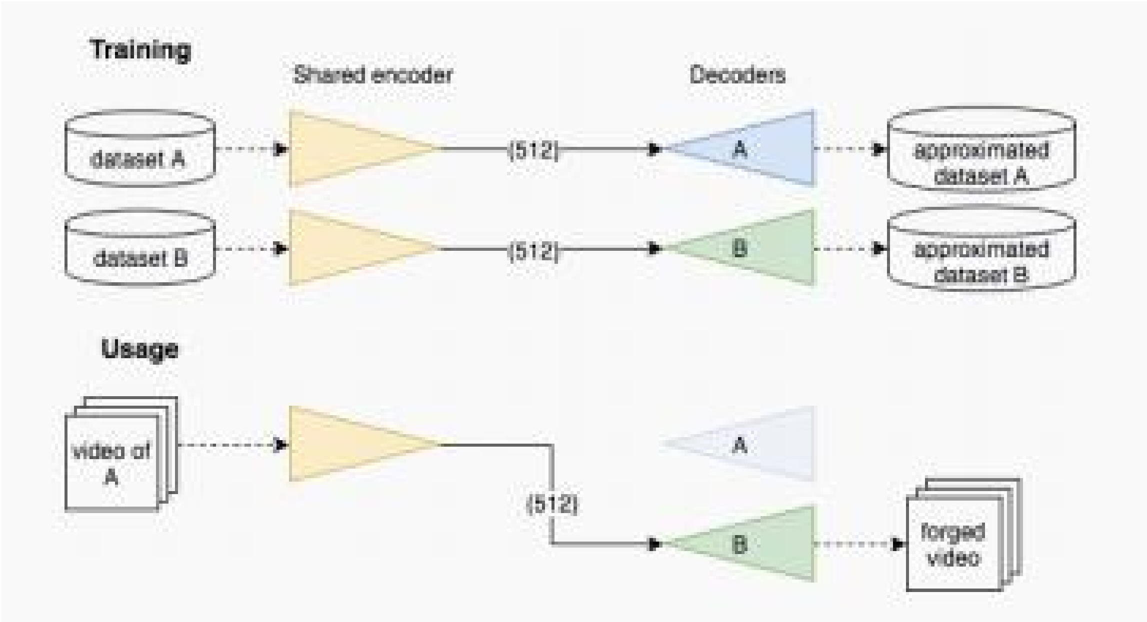


Figure 5.2 Creating Deepfake
Source: MesoNet: a Compact Facial Video Forgery Detection Network.
https://arxiv.org/pdf/1809.00888.pdf

5.3 Architecture Design

5.3.1 Dataset Gathering

For making the model efficient for real time prediction. We have gathered the data from different available datasets like FaceForensic++(FF) [1], Deepfake detection challenge (DFDC) [2], and Celeb-DF [3]. Further we have mixed the dataset the collected datasets and created our own new dataset, to accurate and real time detection on different kind of images.

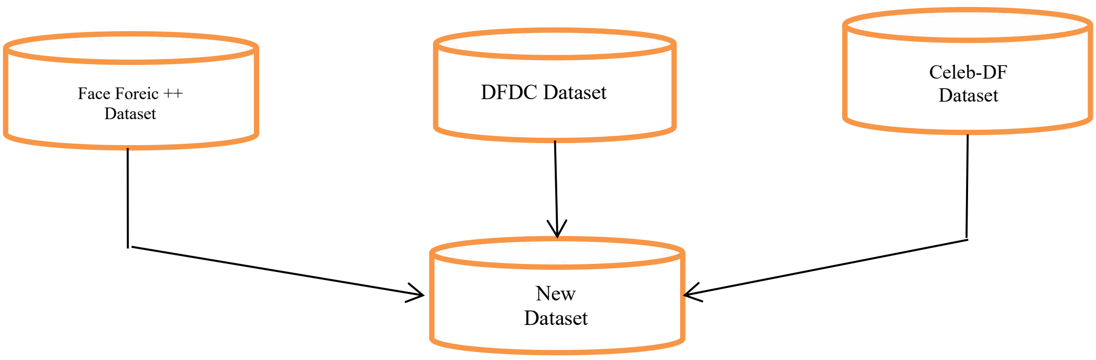


Figure 5.3.1 Dataset Gathering

5.3.2 Datasets

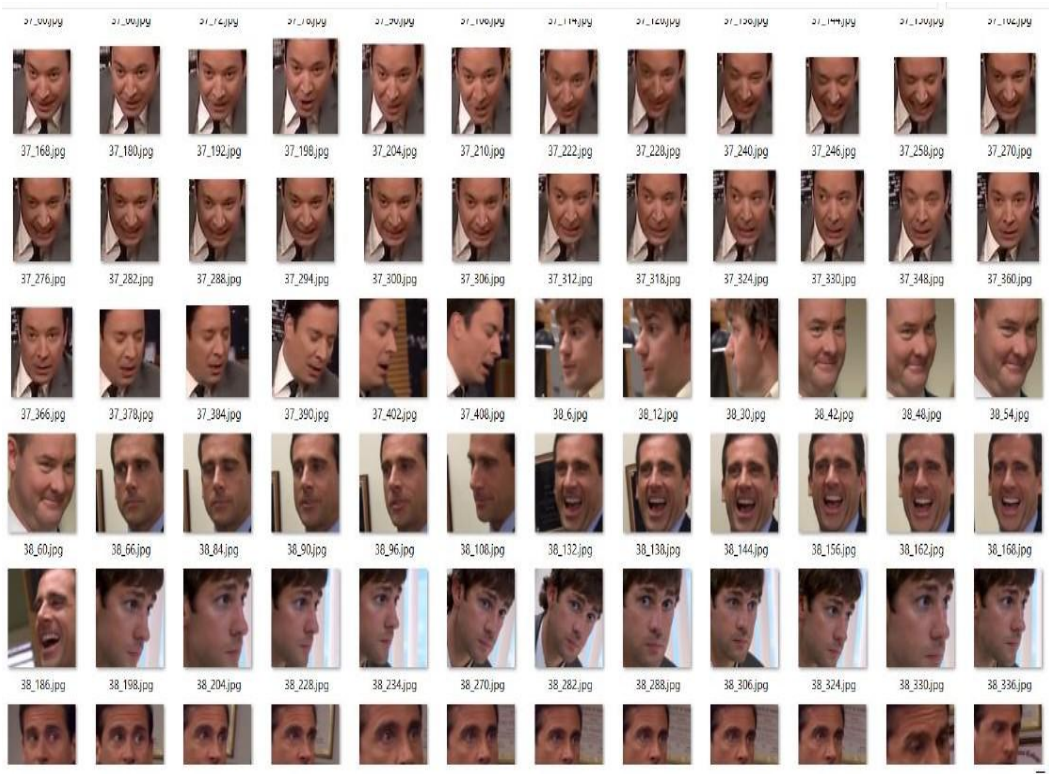


Figure 5.3.2.1 Real Dataset.
Source: https://github.com/kiteco/python-youtube-code/tree/master/Deepfake-detection

Figure 5.3.2.2 Deepfake Dataset
Source: https://github.com/kiteco/python-youtube-code/tree/master/Deepfake-detection

5.3.3 Pre-Processing

In this step, the images are preprocessed and all the unrequited and noise is removed from images. Only the required portion of the image i.e., face is detected and cropped. The first steps in the preprocessing of the image are to split the image into pixels. After splitting the image into pixels, the face is detected in each of the pixel and the pixel is cropped along the face. Later the cropped frame is again converted to a new video by combining each frame of the video. The process is followed for each video which leads to creation of processed dataset containing face only videos. The frame that does not contain the face is ignored while preprocessing.

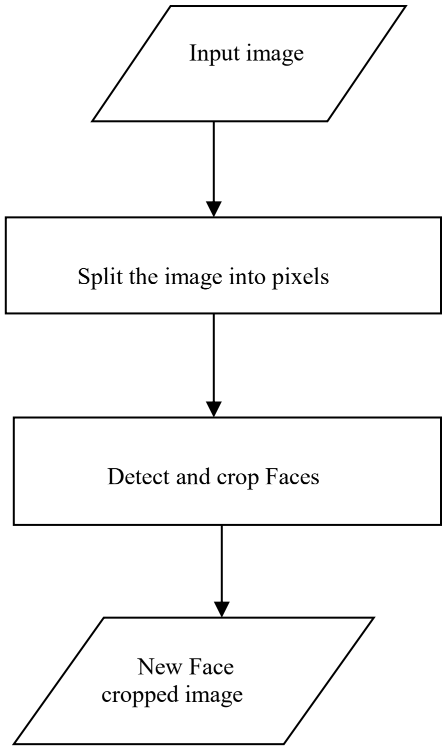


Figure 5.3.3: Pre-processing

5.3.4 Model Architecture

Our model is a combination of CNN and RNN. We have used the Pre- trained Mesonet CNN model to extract the features at pixel level and based on the extracted features a LSTM network is trained to classify the image as deepfake or pristine. Using the Data Loader on training split of images the labels of the images are loaded and fitted into the model for training.

Meso 4

We have started our experiments with rather complex architectures and have gradually simplified them, up to the following one that produces the same results but more efficiently. This network begins with a sequence of four layers of successive convolutions and pooling, and is followed by a dense network with one hidden layer. To improve generalization, the convolutional layers use ReLU activation functions that introduce non-linearities and Batch Normalization to regularize their output and prevent the vanishing gradient effect, and the fully-connected layers use Dropout to regularize and improve their robustness.

MesoInception-4

An alternative structure consists in replacing the first two convolutional layers of Meso4 by a variant of the inception module introduced by Szegedy et al [16]. The idea of the module is to stack the output of several convolutional layers with different kernel shapes and thus increase the function space in which the model is optimized. Instead of the 5×5 convolutions of the original module, we propose to use 3×3 dilated convolutions [17] in order to avoid high semantic. This idea of using dilated convolutions with the inception module can be found in as a mean to deal with multi-scale information, but we have added 1×1 convolutions before dilated convolutions for dimension reduction and an extra 1×1 convolution in parallel that acts as skip-connection between successive modules.

LSTM for Sequence Processing

We are using 1 LSTM layer with 2048 latent dimensions and 2048 hidden layers along with 0.4 chance of dropout, which is capable to do achieve our objective. LSTM is used to process the pixels in a sequential manner so that the temporal analysis of the image can be made, by comparing the pixel at ‘t’ second with the pixel of ‘t-n’ seconds. Where n can be any number of pixels before t.

5.4 Algorithm Details

5.4.1 Preprocessing Details

- ❖ Dataset preparation: A diverse and balanced dataset is essential for training an effective deepfake detection model. The dataset should include a mixture of real and manipulated images, covering various types of deepfake techniques. Each image in the dataset is labeled as genuine or fake.

- ❖ Image resizing: Resizing the images to a consistent resolution is a common preprocessing step. It helps ensure that all input images have the same dimensions, which is necessary for feeding them into the neural network. Typically, deepfake detection models like MesoNet accept square input images of a specific size, such as 256x256 pixels.

- ❖ Data augmentation: Data augmentation techniques are often employed to increase the diversity and robustness of the training data. Augmentation techniques may include random rotations, translations, flips, or changes in brightness and contrast. These variations help the model learn to generalize better and improve its performance on real-world data.

- ❖ Normalization: Normalizing the pixel values of the images is crucial for effective training. It involves scaling the pixel values to a standardized range, such as [0, 1] or [-1, 1].

Normalization helps in reducing the impact of variations in pixel intensity across different images and ensures more stable training.

- ❖ Pretrained model initialization: MesoNet may utilize a pretrained convolutional neural network (CNN), such as VGGNet or ResNet, as a backbone. Pretrained models are typically trained on large-scale image classification datasets like ImageNet. The weights of the pretrained model are loaded, and further training is performed on the deepfake detection dataset to fine-tune the model for the specific task.

- ❖ Input representation: The input images are converted into a suitable format for feeding into MesoNet. This typically involves transforming the images into tensors, which are multi-dimensional arrays compatible with deep learning frameworks. The shape of the input tensor is determined by the network architecture and input requirements of MesoNet.

- ❖ Dataset preparation: A diverse and balanced dataset is essential for training an effective deepfake detection model. The dataset should include a mixture of real and manipulated images, covering various types of deepfake techniques. Each image in the dataset is labeled as genuine or fake.

5.4.2 Model Details

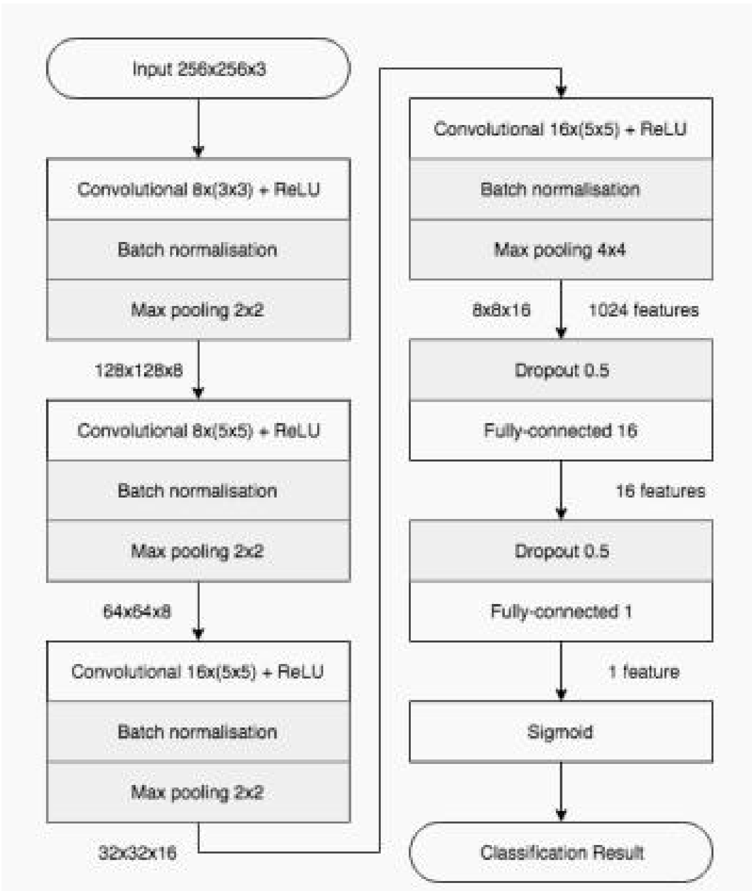


Figure 5.4.2 Meso4 Architecture.
Source:https://paperswithcode.com/paper/mesonet-a-compact-facial-video-forgery

5.5 Tools and Technologies Used

5.5.1 Planning

Open Project

5.5.2 Programming Languages

Python3

5.5.3 IDE

Google Collab

5.5.4 Libraries

- NumPy
- Matplotlib
- TensorFlow
- Keras

Chapter-6

Results

6.1. Image classification results

Classification scores of both trained network are shown in Table 3 for the Deepfake dataset. Both networks have reached fairly similar score around 90% considering each frame independently. We do not expect a higher score since the dataset contains some facial images extracted with a very low resolution.

Network	Deepfake classification score		
Class	Forged	Real	Total
Meso-4	0.882	0.901	0.891
mesoInception-4	0.934	0.900	0.917

Table 6.1.1. Classification scores of several networks on the Deepfake dataset

Table 4 presents results for the Face2Face forgery detection. We observed a notable deterioration of scores at the strong video compression level. The paper that introduces the FaceForensics dataset used in our tests [20] presents better classification results using the state-of-the-art network for image classification Xception [4]. However, with the configuration given by the latter paper, we only managed to fine-tune Xception up to obtain a 96.1% score at the compression level 0 and 93.5% score at level 23. It is therefore unclear how to interpret the results.

Network	Face2Face classification score		
Compression level	0	23 (light)	40 (strong)
Meso-4 mesoInception-4	0.946 0.968	0.924 0.934	0.832 0.813

Table 6.1.2 Classification scores of several networks on the Face Forensics dataset

Chapter-7

Conclusion and Future Scope

7.1 Conclusion

These days, the dangers of face tampering in video are widely recognized. We provide two possible network architectures to detect such forgeries efficiently and with a low computational cost. In addition, we give access to a dataset devoted to the Deepfake approach, a very popular yet underdocumented topic to our knowledge. Our experiments show that our method has an average detection rate of 98% for Deepfake videos and 95% for Face2Face videos under real conditions of diffusion on the internet. One fundamental aspect of deep learning is to be able to generate a solution to a given problem without the need of a prior theoretical study. However, it is vital to be able to understand the origin of this solution in order to evaluate its qualities and limitations, which is why we spent a significant time visualizing the layers and filters of our networks. We have notably understood that the eyes and mouth play a paramount role in the detection of faces forged with Deepfake. We believe that more tools will emerge in the future toward an even better understanding of deep networks to create more effective and efficient ones.

7.2 Future Scope

The future scope of deepfake detection using MesoNet and similar techniques is likely to involve continuous advancements in model architectures, diverse training datasets, adversarial defense mechanisms, real-time detection systems, collaborations with platforms and regulation, and user education. These collective efforts can help mitigate the potential harms posed by deepfake videos and promote a safer and more trustworthy digital environment.

References

[1] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images” in arXiv:1901.08971.

[2] Deepfake detection challenge dataset: https://www.kaggle.com/c/deepfake detection challenge/data Accessed on 26 March 2020

[3] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi and Siwei Lyu “Celeb-DF: A Large scale Challenging Dataset for DeepFake Forensics” in arXiv:1909.12962

[4] Deepfake Video of Mark Zuckerberg Goes Viral on Eve of House A.I. Hearing: https://fortune.com/2019/06/12/deepfake-mark-zuckerberg/ Accessed on 26 March 2020

[5] TensorFlow: https://www.tensorflow.org/ (Accessed on 26 March 2020)

[6] Keras: https://keras.io/ (Accessed on 26 March 2020)

[7] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. arXiv:1702.01983, Feb. 2017

[8] J. Thies et al. Face2Face: Real-time face capture and reenactment of rgb videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2387–2395, June 2016. Las Vegas, NV.

[9] Face app: https://www.faceapp.com/ (Accessed on 26 March 2020)

[10] Face Swap: https://faceswaponline.com/ (Accessed on 26 March 2020)

[11] Yuezun Li, Siwei Lyu, “ExposingDF Videos by Detecting Face Warping Artifacts,” in arXiv:1811.00656v3.

[12] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.

[13] ResNext Model: https://pytorch.org/hub/pytorch_vision_resnext/ accessed on 06 April 2020

[14] https://www.geeksforgeeks.org/software-engineering-cocomo-model/ Accessed on 15 April 2020

[15] MesoNet: a Compact Facial Video Forgery Detection Network

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, et al. Going deeper with convolutions. Cvpr, 2015.

[17] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015