National College of Ireland

# Deepfakes Detection using Convolutional Neural Networks

MSc Research Project
MSc Data Analytics

Tej Rup Sai Munagala
Student ID: x19196628

School of Computing
National College of Ireland

Supervisor: Prof. Cristina Muntean

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | | | |
|---|---|---|---|
| **Student Name:** | Tej Rup Sai Munagala | | |
| **Student ID:** | x19196628 | | |
| **Programme:** | MSc Data Analytics | **Year:** | 2020 |
| **Module:** | MSc Research Project | | |
| **Supervisor:** | Prof. Cristina Muntean | | |
| **Submission Due Date:** | 17/12/2020 | | |
| **Project Title:** | Deepfakes Detection using Convolutional Neural Networks | | |
| **Word Count:** | 6430 | **Page Count**: | 20 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Tej Rup Sai Munagala

**Date:** 17/12/2020

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Deepfakes Detection using Convolutional Neural Networks

Tej Rup Sai Munagala

x19196628

**Abstract**

Recent advancements in deep neural networks have progressed innovative approaches for creating digital content and it is getting very difficult to ascertain between which content is real and what is deepfake. Attackers are using these developed technologies for tampering with videos, and images and they are disclosing them into social media. These actions are impacting many individuals in terms of their reputation, mental health, income, etc. and these are also affecting organizations. In this research paper, the dataset is taken from the deepfake detection challenge by Facebook AI. All the videos are converted into frames and these frames are evaluated by Convolutional Neural Network CNN) model to classify the deepfake content. Hyper-parameter tunning is applied on the CNN model for better accuracy and precision. The model has shown a test accuracy of 80 percent and Area Under Curve (AUC) value of 0.5 from the trained model.

# 1 Introduction

The use of smartphones technology has been skyrocketed in recent years; this rapid development has led to the growth of many sophisticated video editing tools for creating many authentic, realistic videos and it also led to a bloom in social media portals for sharing digital videos in more convenient ways. To date there was a high demand for fake videos, due to a lack of domain knowledge and editing techniques have restricted the realism of fake videos. However, the digital transformation also heightened a new cutting-edge technology called deepfake. Deepfakes are the content generated by artificial intelligence in which the faces are swapped from one person to another person in any given video and these videos look reliable for human eyes.

Deepfake is a creative medium that produces revolutionary applications in fields such as film dubbing, re-animation, virtual shopping, etc. The right use of this technology can bring us our favourite artists and characters into the film. However, the growing rate of this technology also has a terrible scenario in real life. The viral spread of these fabricated videos has caused problems like celebrity humiliation by swapping their face in a few porn videos, propagate false political views by tampering with the videos of political leaders, etc. Mirsky and Lee, 2020 stated that the research publications in this area have been increased from 3 to over 250 between the years 2018 to 2020. Researchers from a cybersecurity firm called

Deeptrace has stated that 96 percent of the fabricated videos are related to pornographic of celebrities and they identified around 15,000 fake videos online which was doubled from the previous year[1].

The machine learning researchers' group has contributed dual attention to the phenomenon of deepfake. On one side they developed tools to create deepfake videos based on expression re-enactment, facial landmark. Techniques such as pix2pix (Isola et al., 2016), CycleGAN (Zhu et al., 2017) and Face2Face (Thies et al., 2016). On the other hand, they have developed various forgery techniques on forensics, behaviour, coherence, and synchronization. Methodologies like detecting inconsistent head pose by using machine learning algorithms like SVM classifier (Yang, Li and Lyu, 2018), capsule neural network which consists of 2D & 1D convolutional neural networks to detect deepfake content of images and videos (Nguyen, Yamagishi and Echizen, 2018), detecting of deepfake on bases of eye blinking by using Long-Term Recurrent CNNs (LRCN) (Li, Chang and Lyu, 2018).
This study is about deepfake detection, for this, all the videos from the dataset are processed into frames. Facial detection techniques like region of interest (ROI) and cascading tools like haarcascade are applied to the extracted frames. All these processed frames are trained by using the CNN model, hyper-parameters like batch normalization, max pooling, and Sigmoid functions are used for fine-tuning the model for better accuracy from the CNN model. The CNN model has shown good accuracy in predicting the deepfake videos.

## 1.1 Research Question

This research paper gives the solution to the following question: "How much accurately and efficiently a deep neural network model can be able to detect a deepfake video."

# 2 Related work

Many Artificial Intelligence (AI) based research has been done on the topic of deepfake. The literature review in this research discusses the recent methodologies that are introduced for a generation of deepfake content and the sophisticated models for the identification of deepfake. This section is divided into the following subsections: the first section gives an outline about deepfake, then the second section continued with the brief technological background related to deepfake. The third and fourth section discusses the advanced neural network models that are available in the creation and defending the deepfake content. Finally, the fifth section talks about the overall conclusion of the literature review.

## 2.1 Overview of Deepfakes

Deepfake is defined as the content that is generated by artificial intelligence that looks genuine to the human eyes. This means that machine learning models are used to swap the face of one person to another person which cannot be identified by humans. However, Mirsky

[1] BBC news: https://www.bbc.com/news/technology-49961089

and Lee, 2020 says “The deepfake content is not only generated to fool the human but also synthesized to deceive machines, these are called adversarial samples”. Deepfake information can be used and mislead in four different areas that are in Hoax, Propaganda, and Entertainment. Hoax is all about tampering of evidence, scams & fraud by spoofing and falsifying the audits and harming credibility by revenge porn. In propaganda, the deepfake is used for misdirection of facts and spread of false conspiracy. In the field of entertainment, deepfake is used in editing and special effects, animating new characters[2].

## 2.2 Technological Background

Even though there is a wide range of neural networks that are developed, most of the deepfake models are created in the combination of encoder-decoder networks and generative neural network (GAN). Most of these models are developed based on the following facial features such as reenactment, replacement, and editing & synthesis. In the reenactment, the following facial factors such as mouth, gaze, body expressions, pose are used to create deepfake content. By using the facial reenactment features it is easy to impersonate one’s identity and helps attackers to tamper the surveillance footage to create fake evidence in real-time. In replacement, the content of one’s identity is swapped with some other identity. This face replacement is like reenacting someone's body features which helps the attackers to circulate fake information through political figures[3]. Finally, editing & synthesis features are used to remove, alter, and add attributes such as clothes, age, hair, weight, ethnicity. Schwartz, 2018 says that the features are used in the field of entertainment for movie editing.

## 2.3 Deepfake Creation

Deepfake contents are mostly created on five types of neural networks namely Encoder-Decoder Networks (ED), CNN, GAN, and Recurrent Neural Networks (RNN). Deepfakes content generation models are mostly created based on driving a specific identity based on a one-to-one, many-to-one, and many-to-many. In the ED networks, modeling methods like autoencoders, variational auto render (VAE) models are developed for deepfake creation. Under GAN models’ techniques like Pix2Pix and cycle GAN are introduced and the videos that are resulted from GAN models. Many advancements are also done in RNN models for deepfake detection like LSTM, GRU (gate recurrent units) parameters, and Pix2Pix methodologies are introduced and these improved models resulted in accurate deepfake videos compared to other neural network models.

## 2.4 Deepfake Detection

Deepfake detection models are used to detect a given video or image is deepfake content or not. These deepfake detection methodologies can be segregated into two divisions namely artifact-based and undirected based approach. In artifact-based models, the deepfake content

[2] The Facebook: https://about.fb.com/news/2018/05/inside-feed-facing-facts/#watchnow

[3] Theguardian: https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth

is identified on basis few artifacts that are very difficult to be analyzed by humans, but the machines can easily identify the deepfake content. In the undirected approach, instead of depending on the artifacts, the researchers developed a generic neural network classifier model that will decide a given content is deepfake or not.

### 2.4.1 Artifact-based Detection

In this approach, the deepfake detection model is developed on the bases of the following spatial artifacts such as blending, environment, forensics, behavior, physiology, synchronization, and coherence.

#### 2.4.1.1 Blending

The target facial features are spliced with some familiar person's facial features; it is done by blending the boundaries of the target person. Authors like Rossler et al. (2019) and Nguyen, Fang, Yamagishi, and Echizen (2019) have advanced their deep learning approaches on the basis of blending artifact features. Rossler progressed their work on XceptionNet CNN, they proceed their dataset in such way that it has wide verities of facial features and the dataset was trained on many deepfake algorithms like Face2Face, Faceswap, and NeuralTextures. The XceptionNet neural network has shown a good precision score on the test dataset however the CNN had less precision with low-quality images.

Nguyen and their team had developed a Y-shaped autoencoder decoder neural network, here the blended artifact features were sent to the encoder part of the neural network and the decoder part consists of detection and segmentation of manipulated frames. The neural net was trained on four different types of datasets and the Y-shaped architecture was structured with different forensics manipulated techniques in different layers. The overall architecture had shown good results in detecting the blended boundaries; however, these models have lesser accuracy in detection of deepfake on bases other artifact features.

#### 2.4.1.2 Environment

The fake content is identified based on features like hair and background under environment artifact features. Researchers like Ciftci et al. (2020) and Yang et al. (2019) had used features like biological signals and facial landmarks in the training dataset for developing the deepfake detection models. Ciftci had used particularly hidden features like biological signals from the video dataset, all these extracted features from the videos are processed through temporal consistency and spatial coherence. All these pre-processed signals are transformed into pairwise selection signals that are used as input for the CNN classifier. The neural model had shown an accuracy of about 91.7 percent for deepfake detection.

Author Yang had used facial landmarks such as head poses, facial gestures, and faces from the video dataset to create synthesized frames. All these synthesized frames were sent through the machine learning algorithm support vector machine (SVM) classifier algorithm and the model had been evaluated on the different test datasets. For the UADFV dataset, it had shown an Area Under ROC (AUROC) value of 0.89 and for the DARPA dataset, the model had resulted in about 0.843 AUROC value, which means the machine learning classifier able to detect the replicated head poses on any given video dataset.

### 2.4.1.3 Forensics

Under the forensics artifact, wide verities of facial features and patterns were used as parameters for the model. Lots of research work is being done on this forensics section as it includes most of the facial features in the identification of deepfake content. Research works by Li, Yang, Sun, Qi, and Lyu (2020) and Amerini et al. (2019) had shown how the forensics artifacts had been used for developing a fake detection neural model. In the paper, Li and their team had taken a dataset contains diverse videos from low resolution to high resolution. As the dataset contains all the subtle facial features, they trained the dataset through different methodologies such as DeepFaceLab and faceswap for better prediction results. The trained dataset had processed through various models like two-stream and XceptionNet CNN. These models had shown good accuracy for prediction however additional factors like enhanced image quality, accurate anti-forensics methodologies will be helpful for better prediction.

Research work by Amerini et al. (2019) had considered all the facial features from the dataset and then these videos were pre-processed into RGB (Red Green Blue) frames and 3D optical frames before they were sent into the modeling. All these frames were used as input for the VGG16 and ResNet50 CNN models, both these models had shown an accuracy of 81.6 percent and 75.4 percent, respectively. However, these subtle facial features help in the identification of deepfake, but these do not generalize all the evolved artifact features.

### 2.4.1.4 Synchronization

Synchronization is also an artifact factor that helps in revealing a given video is deepfake or not. Authors like Li and Lyu (2018), and Afchar and Nozick (n.d.) few others have researched deepfake detection models based on complete video synchronization. Li and Lyu had used a modified version of DLIB packages to extract all the distinctive artifacts from each video clip of the dataset. After extracting all these features, the gaussian blur algorithm was applied to find inconsistency in all the clips. Finally, all these processed frames were applied on the Resnet CNN and got an accuracy of about 97.4 percent.

Afchar and Nozick (n.d.) had used various neural networks on different layers of the entire architecture. During the pre-processing inception model was applied to all the videos that are in the dataset to segregate the videos that were not fully synchronized. All these filtered video clips are sent through the ensemble neural network model to detect the forged videos and the model had shown an accuracy of about 97 percent on the Face2Face dataset. These researchers had concluded that models fail to detect the deepfake videos if the mouth is closed completely.

### 2.4.1.5 Coherence

The coherence is an artifact feature that is used to check any jitter and flickers in a clip to identify the deepfake content. Research work by Nguyen, Yamagishi and Echizen (2019), Ivanov et al. (2020) had detected the deepfake video based on the flickers and jitters features. Author Nguyen had configured a capsule neural network that consists of three layers of capsules. All the videos were split into frames by using the latent features of VGG19 CNN, all these splatted frames were sent into the three-layered capsule model. There are 2D CNN, 1D CNN, and statistical polling respectively in the three layers of the capsule model to identify the coherence features in each video. The trained model had shown a good percentile

of accuracy about 97 percent on the test dataset on the detection of jitters and flickers in any given video.

Researcher Ivanov and his team also derived an architecture that uses both the CNN and Long Short-Term Memory (LSTM) neural network to find the deepfake videos. All the videos from the training dataset were sliced into the sequence of frames and these sequential frames are inputted into the CNN to extract few vectorized features. These trained vector features were sent into the LSTM RNN model, by using softmax as activation function and with a dropout rate of 0.5 to distinguish the deepfake vectors. This model had shown a higher accuracy even on the large dataset, even with varying vectors.

### 2.4.2 Undirected Approaches

In this approach, the deepfake detection is performed by creating a generic neural network identifier, this detection is categorized based on two approaches namely - anomaly detection and classification detection.

#### 2.4.2.1 Classification

This classification approach uses deep learning algorithms and performs very well than the traditional image-based forensics detection models. Various researchers have developed complex neural network models for the identification of deepfake content and most of the advancements happened on the CNN and GAN models.

Yu et al. (n.d.) worked on an architecture called the GAN model for the identification of deepfake. They designed the architecture in such a way that leaves special fingerprints on the processed videos for separation of real and deepfake videos. The GAN model had resulted in numerous frames with special fingerprints which were used to identify the fake videos. This model has evaluated on the Celeb A dataset, which contains many video files and various conditions of videos. After evaluation, the model had shown better accuracy, however, the model failed in a few scenarios when the videos had noise, blur, and compression factors.

To rectify the above-unseen distortion issues author Marra et al. (2018) had used a methodology called image-to-image translation in the modeling part. They developed a GAN model that can be used for both generating the deepfake and can be used for the detection of deepfake. The dataset was pretrained based on the Leave-one-manipulation-out technique and this pretrained dataset was modeled through the neural network. After training the model, the model was tested with numerous scenarios of dataset and the model has shown good accuracy in detecting the deepfake content in all the conditions.

Few other research works are being done for the development of the CNN model for deepfake detection. Research work from Mo and Chen (2018) and Tariq et al. (2018) had developed a complex CNN architecture. Researcher Tariq had developed a three-layered Shallow convolutional neural network (ShallowNet). There are dropout functions, max pooling, and normalization in each layer of the ShallowNet. This ensemble model was inputted with Generated Synthetic Images (GAN), which were derived from the training dataset. After training the model had shown a better AUROC value, however, the model prediction was decreased when the model was compiled with lower-dimensional images.

To rectify the image dimensionality issues with the neural network model, author Mo had introduced progressive three-layered CNN. The input dataset was pre-processed and RGB color images were extracted, then these images were passed through high filters before sending it to the model. The layered CNN had Leaky Rectified Linear Activation (LReLU) function and max pooling to process the images with a dimension of 128*128 frames. Finally, the outputted frames are classified with a probability value of 0.5 with the use of the softmax function.

Author Chan et al. (2019) used encoder-decoder neural network video-to-video classifier for identification of a video is fake or not. The encoder part of the neural network is inputted with all the facial poses and pose sticks of the person. All these poses of frames were processed to generate synthetic GAN images, these images were transferred through the decoder part of the model that detects the deepfake content by looking at the motion sequence of the GAN images. Overall, this model has lesser accuracy for the consecutive frames with extreme poses.

### 2.4.2.2 Anomaly Detection:

In the anomaly detection methodology, the outliers from the dataset are included so that the model will perform well even when there are distortions in the videos. Researchers had included many inconsistencies in the videos during the training process to make sure the model detects the deepfake even with the irregular videos.

Author Li, Li, Tan, and Huang (2020) had processed the videos into synthetic images by including the mismatching colour profiles between the consecutive images from the camera. All the synthesized images are sectioned off based on colour patterns and then these images were sent into the first-order differential operation to extract any residuals from the images. After dropping out these residual images, they were sent into the different neural network models for the classification of images that are fake or not. Finally, the neural model with one class classification has shown good accuracy in detecting fake videos.

Researcher G¨uera and Delp (2018) included video clips that had inconsistent head poses. All the video clips with irregular landmarks are segregated in a sequence of frames. These inconsistent frames were pre-processed and sent through the Resnet50 CNN model with higher epochs of the modeling process. After evaluation of the model with the UADFV dataset, it had shown an accuracy of about 94 percent in detecting the deepfake content.

A triplet mining methodology was used by the author Schroff et al. (2015), he considered sequence of frames which are matching and mismatching. The image frames were taken from the Labelled Faces in the Wild (LFW) dataset, all these frames were sent into the CNN model with L2 optimizer before embedding. After evaluation all the frames sent into the Triplet loss function and the model had shown an accuracy about 95 percent even after considering 1000 frames during the evaluation process.

The machine learning model was used by researcher Agarwal et al. (2019) by considering the irregularities in facial expressions and speaking patterns. The video clips were converted into frames based on the region of interest (ROI) with the help of OpenFace2 package. All these filtered frames were sent into the support vector machine (SVM) classifier and the model was evaluated with different hyperparameters. The model had shown good accuracy even for every 10 seconds of frames.

Author Li et al. (2018) implemented a complex CNN neural net architecture, he considered the inconsistent eye blinking videos for training the model. During the pre-processing stage, the misaligned continuous eye frames were cropped out. All these filtered frames were sent to the LRCN model to memorize the sequence pattern to the model for better prediction. Inside the LRCN model, there are three subcategories of functions like sequence learning, feature extraction, and predictions in a sequenced form. Overall, the model had shown good performance accuracy.

### 2.5 Summary of Literature Review

In the overall/ literature review related to deepfake models we can see that initially the deepfake detection was done bases mainly by looking over the facial features of the video and mostly the computation was done by using machine learning algorithms. In the early 2000s as the computational resources were improved, many modeling methodologies were also developed in the deep learning algorithm. From most of the papers, we can infer that CNN outperformed other deep learning neural networks. The prediction accuracy of the models is increased when the entire facial region is considered rather than individual artifact facial features. By considering all the parameters from previous work, a novelty approach has been implemented in the detection of deepfake.

## 3 Research Methodology

The research work is done by applying using an ensemble model that has both the CNN and RNN models for detection of deepfake content. The Knowledge Discovery in Databases (KDD) methodology is used for the implementation of this ensemble model. As this project has lots of knowledge-driven business applications, KDD methodology is used instead of Cross-Industry Standard Process for data mining (CRISP-DM). From figure 1, it is illustrated how the whole process of KDD methodology was applied for this research. The KDD methodology is explained in the following sections like dataset selection, target dataset creation, pre-processing and cleaning dataset, data transformation, choosing and applying data mining models, evaluation of model results.

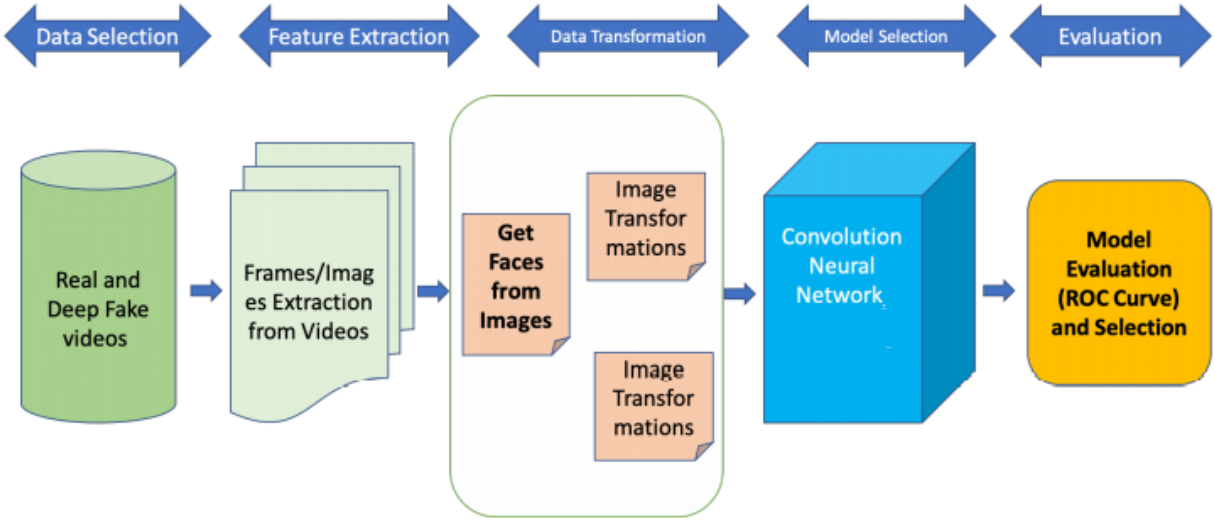


**Figure 1: KDD Methodology for Deepfake Video Detection**

## 3.1 Dataset Selection and Understanding

Deepfake videos are taken from deepfake detection challenge dataset, which is organized by Facebook AI. There are about 1,24,000 videos that are available in the dataset and the memory size of this dataset is about 500 gigabytes (GB). Facebook has created all these videos under different backgrounds, lightning shades and various ethnicity of genders acted in the videos. However, for this research, a total of 800 sample videos are chosen from the entire deepfake detection dataset. The reason behind choosing 800 videos is due to the computational problems within the resources of Jupyter notebook and compilation time of deep learning algorithms.

## 3.2 Dataset Creation

The entire 800 videos are sub divided into two equal divisions training dataset and testing dataset. There are total of 400 videos in the training dataset and from figure 2 illustrates there are about 323 fake videos and 77 real videos. There about 400 videos in the testing dataset which is used for testing the deep learning models. Each video in the training dataset has a length of 10 seconds.

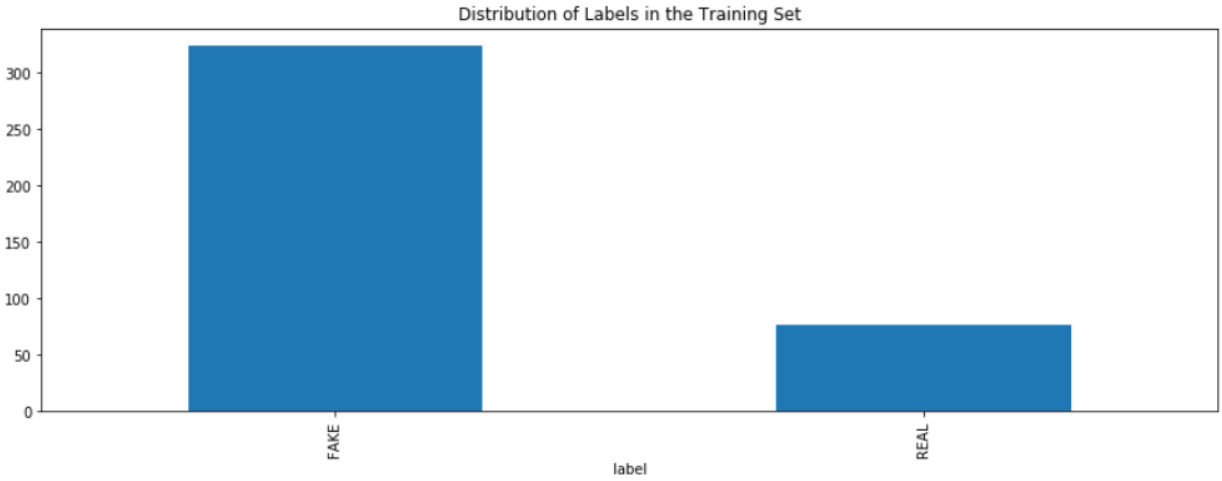


**Figure 2: Distribution of labels in training set**

## 3.3 Data Preparation

Before providing data to the neural network model the following pre-processing steps are included like extraction of frames from videos and extraction of region of interest (ROI) from frames. Few more parameters like batch sizing, frame segmentation, resizing, and face extraction are applied within the model.

### 3.3.1 Extraction of Frames from Videos

The generation of frames is mandatory from each and individual video that is in the training dataset. Generation of frames is a mandatory pre-processing step for the learning model and the extracting of frames also helps in running the model in less computational time. Figure 3 gives an overall description of individual videos like shape, frames per second (fps), and duration. We can understand that duration of each video is 10 seconds, most of the videos have 30 fps and the dimension of the videos are 1080*1920 and 1920*1080.

| | frame_shape | fps | duration |
|---|---|---|---|
| aagfhgtpmv.mp4 | (1080, 1920, 3) | 30 | 10 |
| aapnvogymq.mp4 | (1080, 1920, 3) | 30 | 10 |
| abarnvbtwb.mp4 | (1080, 1920, 3) | 30 | 10 |
| abofeumbvv.mp4 | (1080, 1920, 3) | 30 | 10 |
| abqwwspghj.mp4 | (1080, 1920, 3) | 30 | 10 |

**Figure 3: Calculated frame shape, fps, and duration**

### 3.3.2 Face Detection

The face detection process is done by using the OpenCV library called haarcascade, this library helps in detecting the faces on bases of region of interest. Haarcascade tool filters the faces in three ways, features like cascade in red circle, facial gestures are filtered in green box and the whole face is seen in the green box. Figure 4 shows how the cascading tool is filtering out the extracted frames.

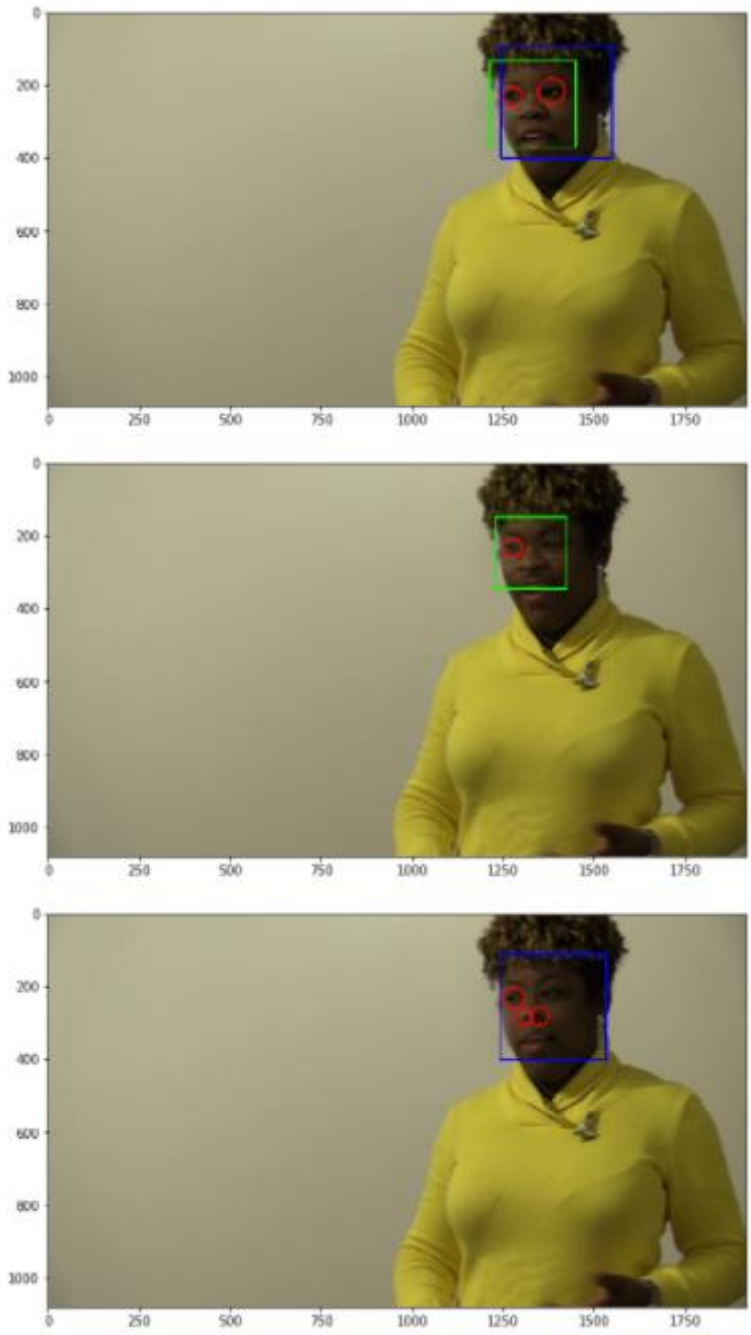


**Figure 4: Haarcascade implementation**

During pre-processing stage, all the frames are resized to a lesser dimension of about 299*299 frames and all the frames are segmented into two categories like the real and fake frame.

## 3.4 Modelling

As this research deal with a lot of image processing, computer vision, and artificial intelligence, the CNN model is utilized for model prediction. Before initializing weights to model, pre-processing packages like Blazeface library from Open CV are used for better utilization of computational power and it also helps for better prediction accuracy.

# 4 Design Specification

The design specification that is required for implementing this model is divided into two levels. The first level consist of the presentation layer and the second level is the business logic layer.

## 4.1 Design flow

The design process is demonstrated in figure 5, under the presentation layer the transformed frames that resulted during the pre-processing stage are presented and the evaluation results after the modelling are also represented in the presentation layer. In the business logic layer includes about the pre-processing and post processing steps the included for the modelling.

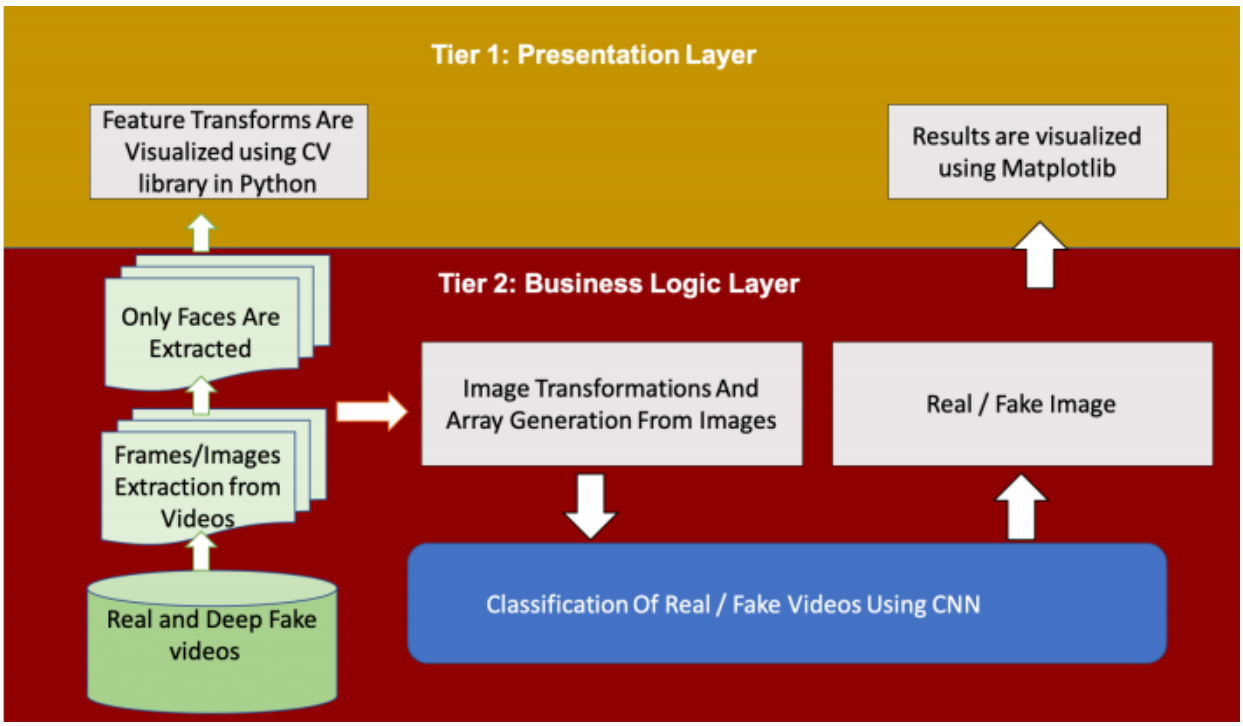


**Figure 5: Design process for deepfake detection**

# 5 Implementation

This section discusses the overall implementation of the CNN model for the detection of deepfake videos.

## 5.1 System Setup

CNN is a deep learning model that requires a lot of time for processing the video files during the training stage. Jupyter notebook is used for executing the CNN model and 16 GB CPU RAM (Random Access Memory) required for processing the model. Python libraries like TensorFlow, OpenCV, and Keras are required for CNN model implementation. Model coding is done based on python version 3 libraries. Generation of frames from videos, face extraction

is applied by using PIL, NumPy, panadas, and Keras libraries. The processed frames are sent to the CNN model as shown in figure 5.

## 5.2 Data Handing

There is a total of 800 videos that are available in the dataset and these videos are divided equally for the train and test dataset. A user-defined function like frame extraction is written for the extraction of frames from the training dataset. Face extraction function is used for extracting facial features by using haarcascade library, all these facial features like an eye, smile, profile, and frontal are stored in a data frame. ROI function is used for updating the extracted frames with more frames that have only the facial features. All these extracted frames are inputted into the model for evaluation, during the modeling max pooling, batch normalization and Dropout are used for resizing the frames to lower dimension and these parameters also increase the computational time of the model.

## 5.3 CNN Architecture

CNN is a sequential neural network that has multiple layers within the model. For this research, a CNN model with four levels of continuous convolutions, batch normalization, and pooling is used. ReLU activation functions are used at each level of convolution to remove any irregularities that are inputted into the neural network. The batch normalization is used to regularize the output from the model and Dropout used for improving the accuracy of the model. Figure 6 illustrates the overall architecture of the CNN model that is used for the prediction.

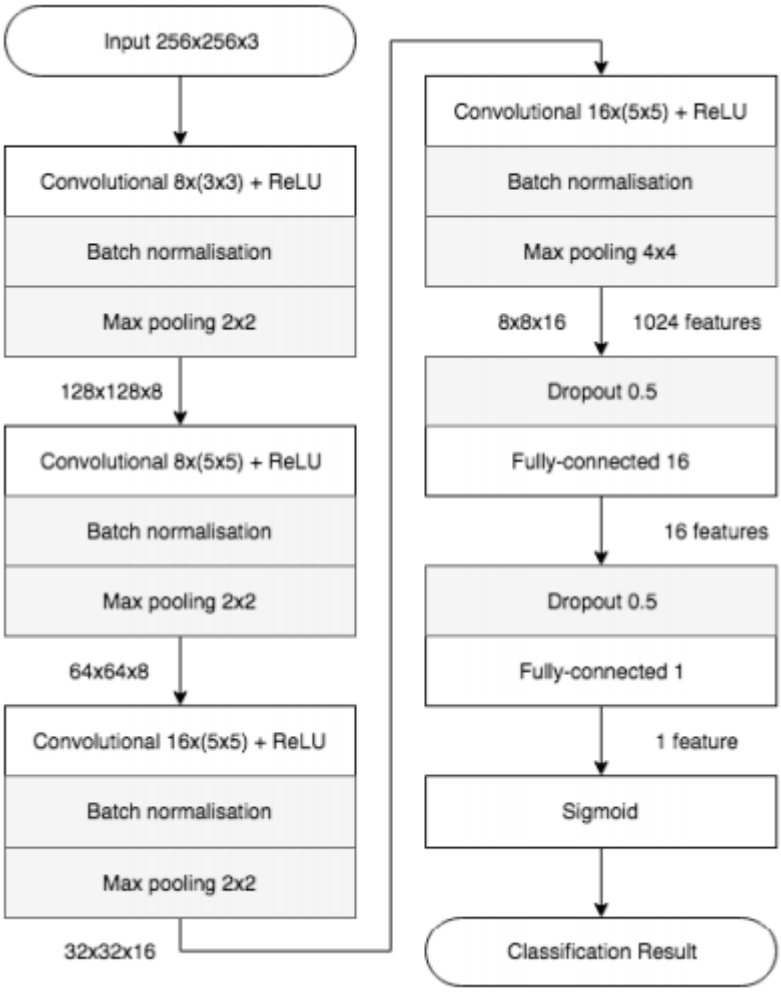


**Figure 6: CNN Architecture**[4]

# 6 Evaluation

This section discusses the results of the CNN model, the model is evaluated based on parameters like accuracy, loss, and validation loss. The plot between the model loss value and the number of epochs tells us about the performance of the model. For a deeper analysis of the model accuracy plot for both the training and test, datasets are plotted. The confusion matrix for the predicted model is evaluated based on true positive, true negative, false positive, and false negative values. Finally, the resultant values of the test dataset are plotted to illustrate how many videos of them are real and fake.

[4] CNN architecture: https://arxiv.org/pdf/1809.00888.pdf

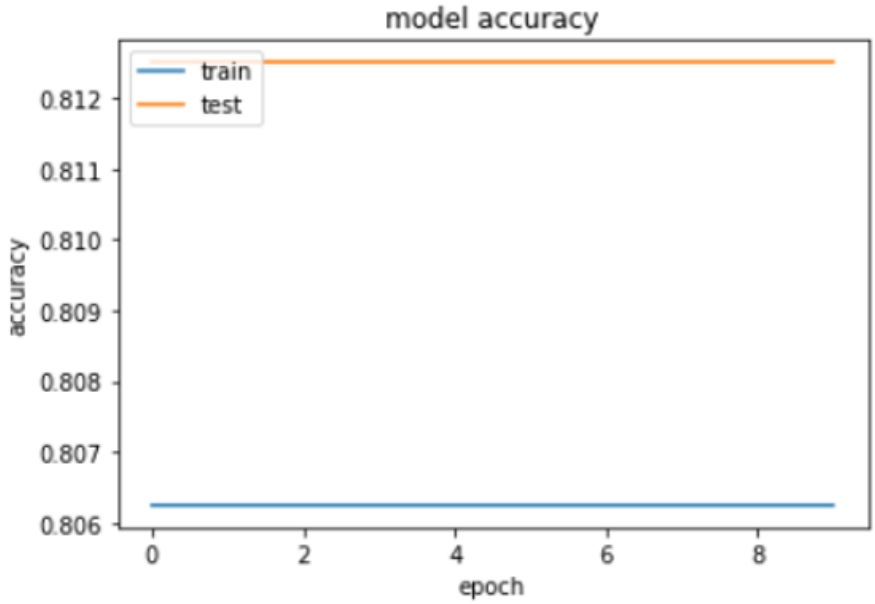


**Figure 7: Model Accuracy Plot**

Figure 7 illustrates the model accuracy for the training and test dataset for epoch values. It can be inferred that the model accuracy is content for both datasets even at different epoch values. The model has shown 80% accuracy in detecting the fake content.

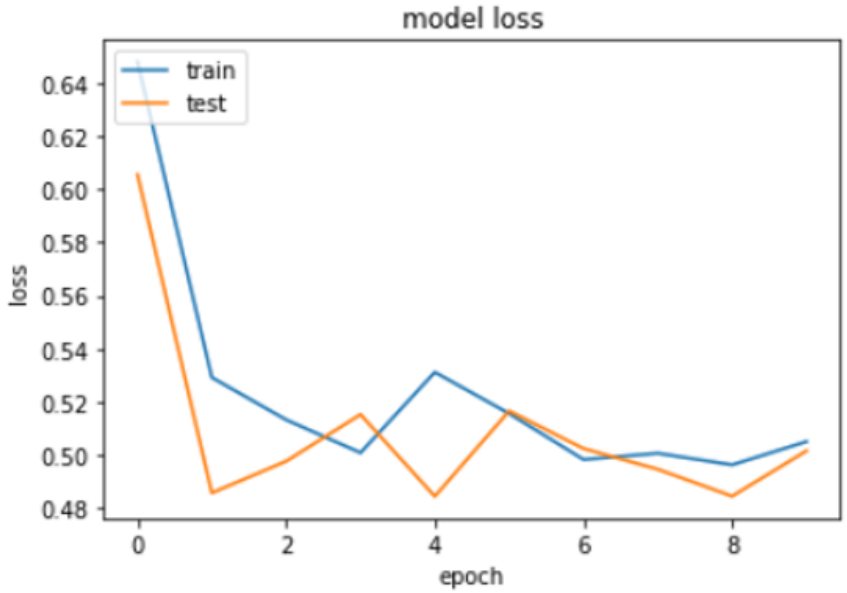


**Figure 8: Model Loss Plot**

Figure 8 shows how the model loss values are changing for different epochs, it can be inferred that for model loss value is declined dramatically for the continuous increase in the epoch values. This means that model is good at classifying the fake videos from the whole dataset. The initial iteration of the CNN model has shown an AUC value of about 0.4 and a loss value of 0.7, these values can be seen in table 1.

| | loss | auc | accuracy | val_loss | val_auc | val_accuracy |
|---|---|---|---|---|---|---|
| 0 | 0.725248 | 0.496187 | 0.80625 | 0.693147 | 0.5 | 0.8125 |

**Table 1: Results from CNN Model**

The hyperparameters of the CNN model are tuned for better prediction accuracy. In the previous model, there are about 3,745,89 parameters are used for training the model and the new model has a total of 20,905 parameters.

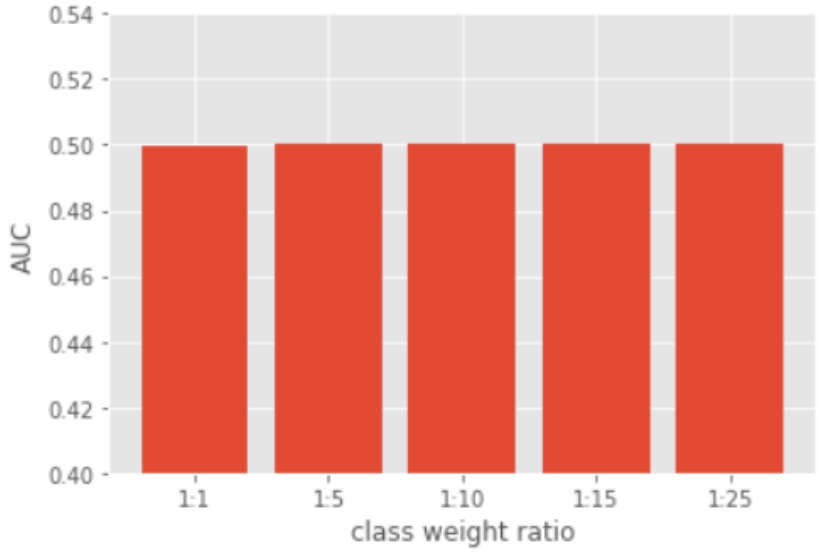


**Figure 9: Class Weight Ratio**

The new hyper parameterized model was executed for five more iterations by including the class weights for each iteration. Even after evaluating the model five time the AUC value is 0.5, which means the model has reached a threshold point and the model are not overfitted.

| | Name | actual | predicted |
|---|---|---|---|
| 0 | test_videos\aassnaulhq.mp4 | 1 | 0.960093 |
| 1 | test_videos\aayfryxljh.mp4 | 0 | 0.945127 |
| 2 | test_videos\acazlolrpz.mp4 | 0 | 0.988654 |
| 3 | test_videos\adohdulfwb.mp4 | 0 | 0.465923 |
| 4 | test_videos\ahjnxtiamx.mp4 | 1 | 0.990661 |
| .. | ... | ... | ... |
| 395 | test_videos\ztyvglkcsf.mp4 | 0 | 0.987158 |
| 396 | test_videos\zuwwbbusgl.mp4 | 0 | 0.121230 |
| 397 | test_videos\zxacihctqp.mp4 | 0 | 0.690090 |
| 398 | test_videos\zyufpqvpyu.mp4 | 0 | 0.959646 |
| 399 | test_videos\zzmgnglanj.mp4 | 0 | 0.952335 |

**Table 2: Predicted values from CNN**

Table 2 illustrates the predicted values for the complete test dataset and most of the prediction values are in the range of 0.6 to 0.9. As the AUC value is 0.5, so this value is chosen as a cut-off point for the real and fake video. The figure 9 illustrate the predicted real and fake video by the CNN model.

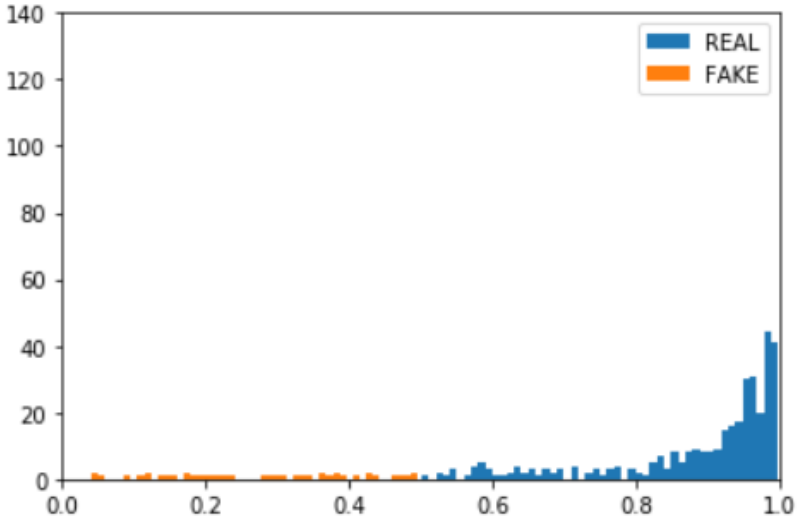


**Figure 10: Predicted real and fake video distribution**

### 6.5 Discussion

Overall, CNN model has shown 80 percent of accuracy which is higher than other benchmark models. As the model evaluation was done on the CPU, it took around 8,700 seconds to run each iteration of the model. The number of layers in the CNN model is reduced and the number of parameters is also reduced from 3,745,89 to 20,905 to improve the model accuracy of the CNN model. Even after five iterations of the CNN model the model accuracy has not improved much, it is around 80 percent. The model is trained a smaller number of epochs around 10 because of the limitation in hardware.

## 7 Conclusion and Future Work

The aim of the research is about deepfake detection in each video. After carefully understanding the pre-processing techniques and methodologies from the literature review, the CNN model is chosen for evaluation. The model is configured with various hyperparameters to under the model limitations for the various situation and it also helped in seeing the change in model accuracy for different parameters. Overall, the model has shown an average accuracy of about 80 percent for detecting the deepfake.

The future work for the deepfake would be by modeling the techniques not only based on images instead features like audio files and biological signals from videos would be more reliable for deepfake detection. The reverse engineering approaches like cycle GAN and encoder-decoder neural networks will help detect the flaws in the frames and these advanced

techniques also consider edge detection and blurriness of frames. Focusing on the micro-level of facial features will be more resourceful in detecting the deepfake.

# 8 Acknowledgment

I acknowledge Facebook AI for providing data for the deepfake detection research project.

I would like to express sincere gratitude to my supervisor Prof. Cristina Muntean for the guidance and encouragement offered throughout the research module. Her expertise in the machine learning space has assisted me during the toughest situation of the project and made research simple and achievable.

I would also thank my family and friends for believing in me to reach my goal.

# References

Afchar, D., Nozick, V., Yamagishi, J. and 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.

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K. and Li, H., 2019, June. Protecting World Leaders Against Deep Fakes. In CVPR Workshops (pp. 38-45).

Amerini, I., Galteri, L., Caldelli, R. and Del Bimbo, A., 2019. Deepfake video detection through optical flow based cnn. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 0-0).

Anantrasirichai, N. and Bull, D., 2020. Artificial Intelligence in the Creative Industries: A Review. arXiv preprint arXiv:2007.12391.

Ciftci, U.A., Demir, I. and Yin, L., 2020. Fakecatcher: Detection of synthetic portrait videos using biological signals. IEEE Transactions on Pattern Analysis and Machine Intelligence.

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

Ivanov, N.S., Arzhskov, A.V. and Ivanenko, V.G., 2020, January. Combining Deep Learning and Super-Resolution Algorithms for Deep Fake Detection. In 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus) (pp. 326-328). IEEE.

Isola, P., Zhu, J.Y., Zhou, T. and Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. and Aila, T., 2020. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8110-8119).

Li, H., Li, B., Tan, S. and Huang, J., 2018. Detection of deep network generated images using disparities in color components. arXiv preprint arXiv:1808.07276.

Li, Y., Chang, M.C. and Lyu, S., 2018, December. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-7). IEEE.

Li, Y. and Lyu, S., 2018. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656.

Li, Y., Yang, X., Sun, P., Qi, H. and Lyu, S., 2020. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3207-3216).

Marra, F., Gragnaniello, D., Cozzolino, D. and Verdoliva, L., 2018, April. Detection of gan-generated fake images over social networks. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 384-389). IEEE.

Mirsky, Y. and Lee, W., 2020. The Creation and Detection of Deepfakes: A Survey. arXiv preprint arXiv:2004.11138.

Mo, H., Chen, B. and Luo, W., 2018, June. Fake faces identification via convolutional neural network. In Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security (pp. 43-47).

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

Nguyen, H.H., Yamagishi, J. and Echizen, I., 2019, May. Capsule-forensics: Using capsule networks to detect forged images and videos. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2307-2311). IEEE.
Oliver, N., FATEN: A framework for governance in the era of data-driven decision-making algorithms1.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M., 2019. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1-11).

Schroff, F., Kalenichenko, D. and Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C. and Nießner, M., 2016. Demo of Face2Face: real-time face capture and reenactment of RGB videos. In ACM SIGGRAPH 2016 Emerging Technologies (pp. 1-2).

Tariq, S., Lee, S., Kim, H., Shin, Y. and Woo, S.S., 2018, January. Detecting both machine and human created fake face images in the wild. In Proceedings of the 2nd international workshop on multimedia privacy and security (pp. 81-87).

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

Yu, N., Davis, L.S. and Fritz, M., 2019. Attributing fake images to gans: Learning and analyzing gan fingerprints. In Proceedings of the IEEE International Conference on Computer Vision (pp. 7556-7566).

Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).