National College of Ireland

Deepfake Detection Research Project

MSc. Data Analytics
Research Project

Harshpreet Singh
Student ID: 18197396

School of Computing
National College of Ireland

Supervisor: Rashmi Gupta

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Harshpreet Singh
Student ID:	18197396
Programme:	Research Project
Year:	2019-2020
Module:	MSc. Data Analytics
Supervisor:	Rashmi Gupta
Submission Due Date:	17/08/2020
Project Title:	Deepfake Detection Research Project
Word Count:	XXX
Page Count:	22

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:	
Date:	28th September 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):	

Deepfake Detection Research Project

Harshpreet Singh
18197396

Abstract

With the recent developments on the creation of deepfake videos using Generative Adversarial Network (GAN), which can produce realistic photos and videos, the reliability of digital images is becoming more challenging to identify. This research is an approach to develop a deep learning model which can efficiently distinguish between a deepfake and a real video. Research work on transfer learning of computer vision to use the previously build features of the neural network of image categorization and build a new model over it. Deep learning is continuously evolving a lot in both areas of generating and detecting deepfakes. A model developed for detection of deepfake designed with older dataset may expire in time, and a need for new detection technique will always be there. Result of the research is auspicious with more than 90% accuracy and the area of evolvement and advancement.

1 Introduction

Machine learning (ML) has witnessed exponential growth throughout the past decade, and computer science has broadly embraced the machine learning technology in various fields for many practical apps. However, the evolution of ML technology often causes new data protection and security challenges. With the exponential rise in online sites which intake and broadcast videos, the validity of the videos is in desperate need of testing. If any confusion emerges, there must be a solution for different innovative approaches for tackling them.

Deepfakes are videos or pictures altered to look other than their original state with the help of Artificial Intelligence. Initial development of this technology served to fulfil the requirement of generating synthetic videos in the entertainment sector. For animation films and influential science fiction films, deepfake can provide realistic output. With more development in the field, happen the development of applications which allow people to use this face-swapping feature for humour purpose. Such applications were available for everyone to use and were not generating much authentic result. The same technology, when combined with deep learning turned out to be a massive outbreak. With the help of artificial intelligence, deepfakes are the most realistic doctored images and videos.

Big tech companies like Facebook and Google have put together researchers by arranging competitions to help identify the deep-fakes and create a vast volume of a dataset for the same. Google worked with paying and consenting actors to capture hundreds of videos over a year to create it an extensive dataset and created thousands of deep-fakes from

such videos utilising open-source methods of deep-fake creation. With limited effort and simple equipment like smartphones, many new devices accessible on the Web have made it simpler than ever for anybody to create practical "deepfakes." Recent advancement in the development of deep-fake algorithms that generate distorted information has had harmful consequences for anonymity, protection and mass communication.

Everyday technical people are working on advancing technologies. Artificial intelligence is getting so powerful that every industry is using it in some way. Deepfake images may also be detrimental to facial recognition systems by interpreting an individual's facial expressions and manipulating them on someone else's image. Since their first appearance in late 2017 numerous open-source deepfake generation strategies have emerged, leading to a growing amount of synthesised media clippings.

TensorFlow is now for several years the most omnipresent open-source deep machine learning repository. TensorFlow helps in providing high-performance computation with its cloud infrastructure along with it, TensorFlow has a possibly the best-established library network. TensorFlow Cloud is a repository for modular frameworks in deep learning.

A model learned by transfer learning is typically equipped with a reduced dataset relative to the standard training cycle of a neural network framework, other benefits such as better prediction and enhanced rate of training may also be achieved by transfer learning. It usually takes thousands of GPU hours to create a modern computer vision model from scrap.

The size and complexity of the training sample can be significantly decreased by implementing transfer learning on a qualified node; this methodology may be used to tackle the classification and regression analysis with relatively limited datasets. Figure 1 illustrates how transfer learning replacements for the initial layer and generates a separate layer for the grouping of different marks.

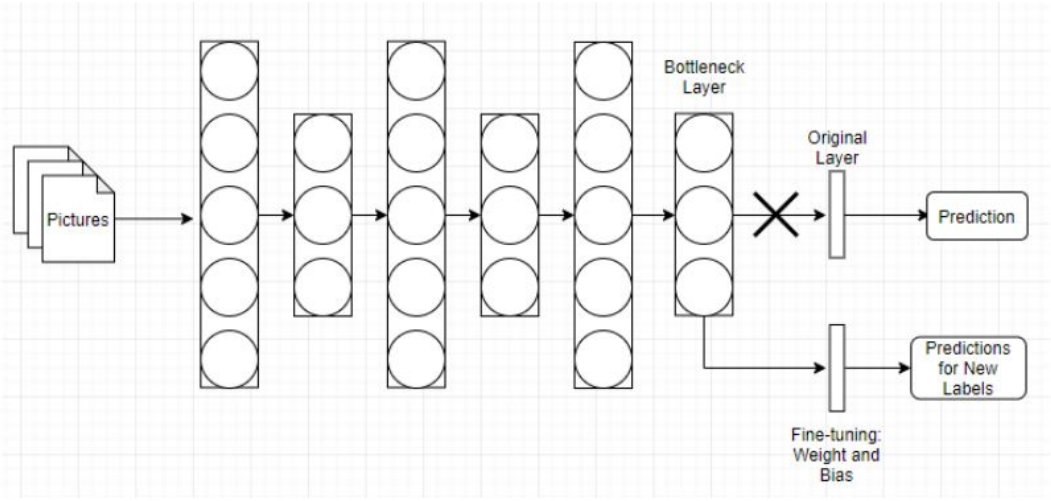

Figure 1: Transfer learning architecture

Deepfakes are new evolving technology which develops compelling AI-generated images and videos. For instance, such a video in which the face of a real person gets altered with some other person that makes it seem like they the original person is doing and saying fictional stuff. One can use such forged videos to violate the identity of a human.

This project shows a model designed using the knowledge of adversarial faces to defend against GAN-based Deepfake attacks. The model uses an adversarial face generation approach for shielding the faces of individuals by considering arbitrarily differentiable picture transformations during the Deepfake model preparation. This approach will reliably produce more artefacts in synthesised expressions, rendering it far harder to recognise the produced false videos and photos. Changes in adversarial and edge losses are significant triggers of substantial deterioration of the content of synthesised images. Paper illustrates the reliability and robustness of the model focused on specific metrics.

Motivation and Objective: The Study field will be a distorted grouping. This area is the mechanism by which minor imperfections are found in doctored videos and exposes the false representation of originality. The creation of a doctored image/video needs information to be examined in order to reveal the dishonest, particularly facial expression variables. It is defined as an in-depth image/video study to find minor imperfections such as boundary points, background incoherence, double eyebrows or irregular twitch of the eye

The impetus behind this research is to recognise these distorted media, which is technically demanding and which is rapidly evolving. Many engineering companies have come together to unite a great deal of dataset. Competitions and actively include data sets to counter deepfakes.

Deepfake videos are now so popular that multiple political parties utilise this tool to produce faked images of the leader of their opposing party to propagate hate against them. Fake political videos telling or doing things that have never happened is a threat to election campaigns. These images are the primary source of false media controversies and propagate misleading news.

In order to expose the forgery in extremely detailed facial expression, such details to be investigated frame by frame in the production of a deepfake picture. The goal of this research is to create a deep learning model that is capable of recognising deepfake images. A thorough analysis of deepfake video frames to identify slight imperfections in face head and the model will learn what features differentiate a real image from a deepfake.

Research Question: How well can the deep learning model developed over transfer learning detect deepfake videos generated by AI?

2 Related Work

2.1 Research Trends of Deepfakes

In the era of advanced artificial intelligence, generation and detection methods of deepfakes keeps on getting changed and more advance. The research community is endlessly working on improving deepfake detection algorithms and published various findings. There is an increasing struggle between those who use advanced machine learning to generate deepfakes and those who aim to identify deepfakes from the real videos Nguyen et al. (2019). The consistency in the generation of Deepfakes has been growing, and the detection system efficiency needs to be enhanced accordingly. The premise is, what artificial intelligence has destroyed can also be restored by artificial intelligence Caporusso (2020).

Convolutional Neural Networks have been a massive success in computer vision systems for supervised learning. Radford et al. (2015) presents a new CNN class that rendered arithmetic operations feasible for filters between photos utilizing the latent vector by adding CNN models to GANs, generating cleverer forgeries. DCGANs (Deep Convolutional Generative Adversarial Networks) proved to be the right candidate for unsupervised study in the field of computer vision.

Initial applications of deepfakes were designed for the imitation of celebrities telling funny things. Since the increase of accuracy continues, it has expanded by disrupting peace and prosperity to distribute fake news and generate ruckus of community Hashmi et al. (2020). The word "Deepfake" has several meanings, but we describe a Deepfake is a video that includes a swapped face and is generated using a deep neural network. These get compared with so-called "cheapfakes"-if fake footage has been produced with machine learning, it is a Deepfake, while if it has been made with readily accessible software that has no learning part, and is a cheap fake video.

The theory says that faces or different parts of the body may be synthesized in videos to obtain other people's knowledge deliberately. The authenticity of a video may be established by monitoring major shifts in eye blink patterns in deepfakes using a heuristic approach Jung et al. (2020). With an average intensity of 4.5 blinks per second, with each wink lasting an approx quarter of a second, most video testing samples used for deepfake identification have a small number of faces closed with their heads. Failure to open eyes will also be a positive predictor of an in-depth analysis of deepfake video Pishori et al. (2020).

Studying the most popular deepfake generation algorithms shows, it can produce fake faces with a specific size and resolution. An affine transformation and a blur feature need to be applied to the synthesized faces to balance and suit the source face structure on the original images Younus and Hasan (2020). An algorithm focused on finding inconsistency on the above applications can help in the detection of a synthetically developed deepfake. The process of detecting GAN images produced using colour indications McCloskey and Albright (2018) often added the colour discrepancy between the original images and the images created by GAN. Nevertheless, this approach works for the whole video file and excluding the examination of different areas like in the scenario of deepfake.

2.2 Technical Artifacts

The idea of fake images or images with different person's face is not new, but the recent technical advancements render it more accurate and believable. For the generation of deepfakes, the major roadblock is the quality of the output. Training a confrontational Deepfake model will contribute to a significant deterioration in the quality of the synthesized images Yang et al. (2020).

As with many other state-of-the-art approaches, the computing capacity it requires is far greater. The available models are often comparable when combining deepfakes into actual source video. An approach, depending on the width of the videos to limit the storing of extracted facial features, may stand out Hashmi et al. (2020).

Deep forgery discriminator (DeepFD) Hsu et al. (2018) built to be a total one Convolutional Network maps the features presented in the fake image to find unrealistic information. It is motivated by an utterly convolutional network Long et al. (2015). In the proposed classifier, two channels are added in the last layers, contributing to the learning target of this layer would continue to learn a vigorously active representation of impractical localization of data.

For the detection of deepfakes using neural network work on the extraction of facial features from the video at the level of frames. These can be improved with the addition of more layers to work on the quality of the video output. The main problem that is needed to tackle in this approach is the creation of a model to process a series recursively in a meaningful way Güera and Delp (2018a). Dynamics related to training can have a tremendous effect on the content of the resulting videos.

For high video quality deepfakes, the algorithm based on the measurements of visual consistency in terms of quality, the approach used in the area of display attack detection, turns out to be a better performing approach than algorithms which focused on the inconsistency in the video. Tests in Korshunov and Marcel (2018) show that GAN-generated Deepfake videos are demanding both for face recognition systems and current methods of deepfake detection.

In Neural Texture Synthesis Thies et al. (2019), 3D texture reconstruction is performed under imperfect geometry and generated at real-time speeds. It collects high-level encoding of the presence of the surface and the 3D world. It lets the network exploit the re-rendering of the source voter system on-target candidates with ease. Nirkin et al. (2019) suggested an approach which does not even allow that the goal actor is present inside the training corpus. Paper has used two additional failure functions step by step to failure of continuity and loss of mixing. Stepwise lack of continuity governs the transition of the image of source candidate to goal candidate, and the lack of restoration holds in check the accuracy of facial reconstruction. Poisson mixing tends to smoothly mix with all faces with the surrounding setting with the mind.

2.3 Machine Learning Approaches

Kharbat et al. (2019) uses machine learning algorithms instead of deep learning. An algorithm based on SVM classifier along with a HOG feature point descriptor has accomplished a remarkable result on the topic of deepfake detection. They are showing that improvisation of machine learning algorithms with deep-learning methodologies can show in a remarkable result for the problem of detecting deepfakes.

One of the markers of the bogus video is a combination of DenseNet169 model with an interface of facial warping artefact identification Maksutov et al. (2020). The inference focuses on the assumption that most of the present deepfake algorithms can only synthesize faces with poor quality resolution. So instead, to make the image complete so perfect, they will affine it. This transformation creates recognizable, visible objects.

A novel effort for differentiation of deepfakes from real videos includes manipulating optical flow field dissimilarities as a guide to differentiate between deep and initial images Amerini et al. (2019). The concept addresses potential deviations in the sequence's temporal aspects. For this experiments, motion vectors were represented as a 3-channel picture and then viewed as input for a neural network to solve the question of using a pre-trained network.

Few of previously designed deepfake detection models used a fully convolutional LSTM framework that incorporates a CNN for "object retrieval" and a proposed LSTM for the analysis of "temporal sequence data." Through concentrating primarily on image distortion in each frame of the film, they were able to separate the tasks for identifying deep-fake videos in this pattern Güera and Delp (2018b).

Transfer learning models like Xception uses depth convolutions, enabling the network to extract similarities through cross-channel networks. This transfer learning model takes input from the output of another model, i.e. MTCNN, this model detects and crops the face within a video. To battle classification in low-resolution images, implement a convolutional 3D network will improve the model Kumar et al. (2020).

3 Methodology

The research focuses on a standard data mining process. The Cross-Industry Standard Data Mining method(CRISP-DM) is a prevalent methodology because of its step-by-step approach and general execution. A successful Data Analytics project must have ample business awareness, including analysis, learning and then recognition. Sometimes that is considered a stable and well-planned strategy.

The whole method phase and the relationship between each phase are typically illustrated in Figure 2. The phases essential for the CRISP-DM consists of market knowledge, data comprehension, data preparation, modelling, assessment and implementation.

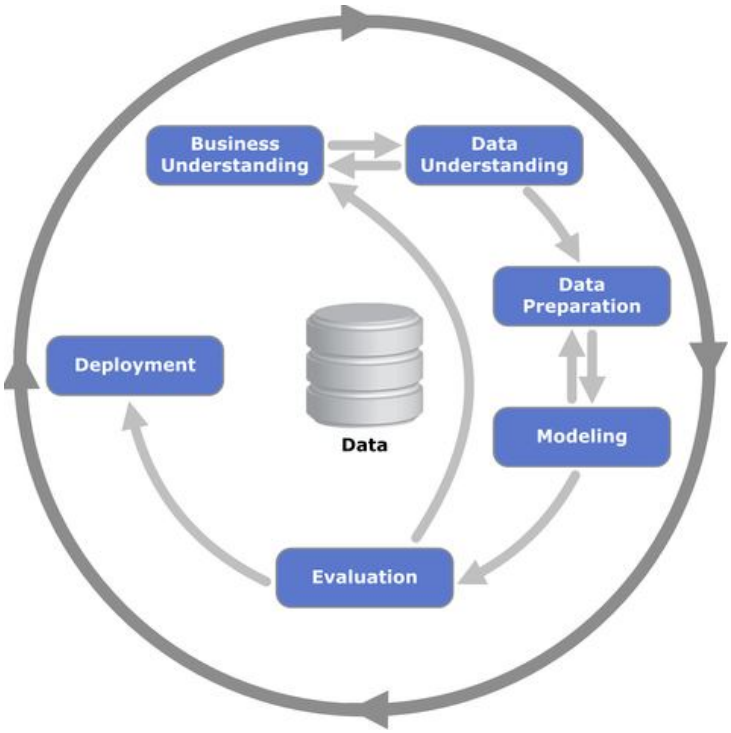

Figure 2: CRISP-DM working cycle.

3.1 Bussiness Understanding

This stage attempts to get a general understanding of the project's requirement. Understanding the nature of the product to be produced for such circumstances is essential. This stage is critical since whole research could fail if this stage is skipped. After the last stage of assessment, CRISP-DM tells to revisit this phase to test how the analysis is moving in an attempt to improve this research's primary purpose.
This research project aims to develop a deep learning model, which can distinguish between a deepfake and a real video.

3.2 Data Understanding

The dataset used in this research is the deepfake detection challenge dataset available on Kaggle. The size of dataset downloaded is 4.13 GB, and the dataset is divided into training and test samples, each containing 400 videos. However, the testing sample does not have the labelling on the videos because the competition was to label them. So this project will be using only the training sample of the original dataset downloaded from Kaggle and divide it into training and test for the analysis and evaluation using the train_test_split library of sklearn.model_selection package.

Exploring the dataset shows that the dataset includes original videos from paid artists that are modified by popular deepfake generator methods. Research continues by checking all the types, counts, labels of the video files. The datset is divided in the ration of 80:20 for fake vs real. After reviewing the types of video files, first concentrate on the metadata information, which is analyzed in depth. In samples of both fake and real videos, some video files have the same original video. Meaning a single real video is recreated in many deepfake videos. In all actual and deepfake videos, one frame is captured analyzed.

After this, the face is highlighted in the captured frame of the videos. This will give us to know how well the model can extract the face features as the majority of the analysis depend on it.

3.3 Data Preparation

The preparation begins with capturing frames from the video. The constructed model will take input of images for the analysis. Many frames are then processed as a new dataset to begin the work as identification of false images. During the initial phase, analysed with a limited number of frames captured per image. However, the result was not appealing, so all frames are captured from the video to get a successful outcome of the study. Then face is highlighted from the frame, and it is cropped. The new dataset with the face images of the individuals in the videos is stored in a new folder.

Here CV2 package is used to explore the videos and getting frames from the videos. CV2 is a package from OpenCV, a cross-platform library which helps to create applications for real-time computer vision. This package focuses mainly on image recognition, video capturing and interpretation, including functionality such as face identification and object detection.

The latest dataset consisting of fake and actual content derived face range will be used to train the model. Using flatten images with less noise has been shown to improve the algorithm's performance significantly.

3.4 Modelling

In this stage, the execution of modelling will be conducted. The model will take input from the last stage's result. Model for deepfake videos detection will conduct an image categorisation analysis of each frame in the video, Discussion in Jiao et al. (2019) on deepfake image is helpful in this research project. This project has succeeded to identify a modern way of identifying the deepfake picture. Inadequate data leads to poor performance. So no matter how successful the simulation is, it ends up getting back to this point so performing it over again if there are any mistakes while predicting the tests.

For developing a model for this research project, a pre-trained model is used. Pre-trained models are designed to tackle a particular problem, and this is called transfer learning. Transfer learning is the idea of removing the free form of thought by utilising learned experience of one question to overcome related ones. It provides the advantage of using pre-trained models to solve diverse and complex tasks in computer vision as effective extractors for new pictures.

Transfer Learning
In transfer learning, research is using InceptionResNet-V1 model from the Keras. Inception-ResNet-v2 is a CNN model, trained from the ImageNet database with more than one million images. The model has 164 layers and can identify artefacts into 1000 types of items, like the screen, cursor, pen, and several species.

A last layer of the pre-trained model is added, which will check the outcome of the analysis of the other hidden layers and use it to determine the category of the input.

The model is designed to learn the usual indicators of Deepfake videos, obvious things that can tell us about photo fakeness, which can be founded by learning the images. The model will be learning below mentioned aspects, including many others to learn categorizing deepfake video from a real one.

Too smooth skin, lack of skin information – these indicators are the product of one deep-fake algorithm problem: low synthesised face resolution. However, detection can sometimes be challenging, primarily due to makeup on one of the two faces. The initial deep-fake algorithm produces 64x64 pixel images, and we need to resize them. Many of the algorithms could now generate 128x128 or even 256x256 faces, but even those sizes may not suffice for good deep-faced video.

The colour difference between the synthesised image and the initial skin-this measure may be used in deep identification of humans. These mismatches, though, may also be very difficult to spot through the hand, but this will not explain a unique algorithm.

Seen sections of the first face or temporal flickering-we can see objects of the first face or even entire first face flickering when face swapping algorithm got the odd option of the face field. It is just only one frame of the whole one-hour film. However, we should take a more precise look at this framework. Head location indicator can also help determine deepfakes accurately.

Activation Function
The softmax activation function is a function which transforms a vector of real K values into a vector of real K values which sums up to 1. The input values may be positive, negative, zero, or greater than one, but the softmax converts them into ranging from 0 to 1, and they can be represented as odds. If the input is low or negative, the softmax will make it a slight likelihood, and if the input is high, it will make it a high likelihood, but it would still stay between 0 and 1.

Softmax method is often also called the softargmax method. The softmax is a logistic regression analysis of several classes which could be used for classification problems. Its methodology is very analogous with the sigmoid function used for logical regression methods.

Multi-layer neural networks may result in a penultimate layer that delivers real-valued ratings that are not efficiently scaled and can be tough to negotiate. The softmax is very helpful here as it transforms the scores into a weighted distribution of probabilities that can be presented to a consumer or used as feedback to specific applications.

Loss Function
The loss function is a measure to decide how the algorithm is modelling data efficiently. If the research prediction is incorrect, the outcome will be higher, and if the estimate is genuine, the effect will be lower. In this analysis, there are different ways of loss function from which to introduce the system Log Loss. This is a simple modification of the probability function.

In this research, Binary cross-entropy function is used. Binary cross-entropy is a property of failure and is used for binary classification tasks. Those are tasks which answer with only two choices to a query (yes or no, A or B, 0 or 1, left or right). Multiple separate questions of this type may be asked concurrently, such as in the multi-label grouping or the binary picture segmentation. The representation of loss function in the neural network is shown in Figure 3.

The binary cross-entropy needs to calculate the natural logarithm of input only from 0 to 1. The softmax activation function is the one to ensure the product remains into this spectrum.

Optimisation Function
Choosing the deep learning model's optimisation algorithm will mean the difference between positive outcomes in minutes, hours, and days. Adam is an optimisation method used in the testing phase to adjust weights in a neural network after each epoch.
The Adam optimisation algorithm is an advancement to stochastic gradient descent, which is adopted widely for images related machine learning and NLP modelling.
Comparison of various optimisation functions with Adam is shown in Figure 4.

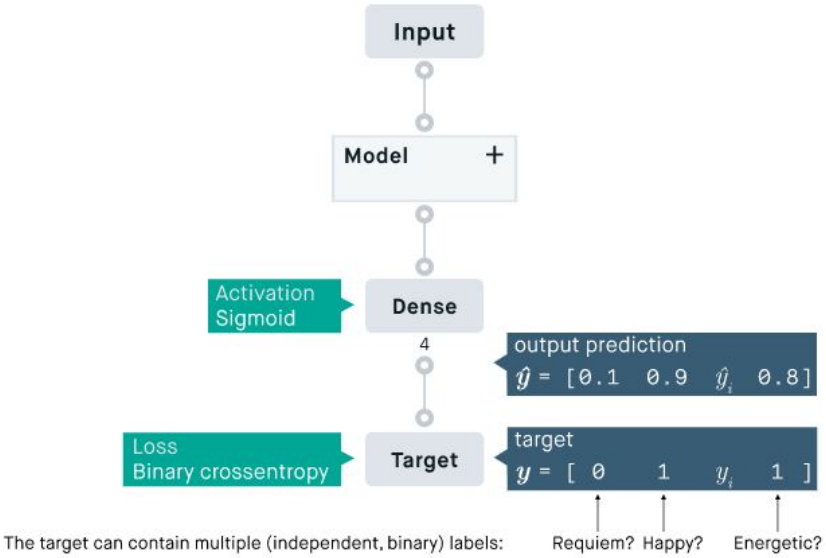

Figure 3: A model with Binary crossentropy loss function.

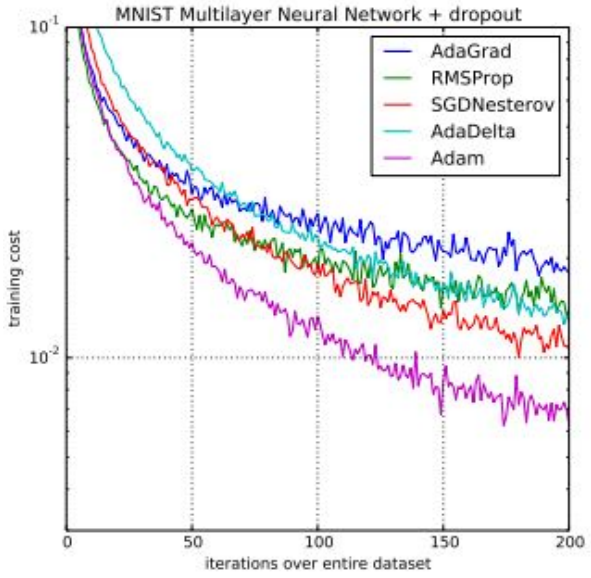

Figure 4: Comparison of Adam Optimisation fumction with other optimisation functions training a multi-layer perceptron

3.5 Evaluation

Confusion Matrix

A confusion matrix is a beneficial technique for illustrating a classifier model; This matrix will observe the relationship between the classified and the actual values. A matrix of ambiguity will summarise quickly how accurate a classification process is. In real or fake related problems in which categorisation is binary, it is very relevant, because it predicts two factors positive or negative.
The confusion matrix has four sections: true positive, false positive, false negative and true negative, as represented in the Figure 5.

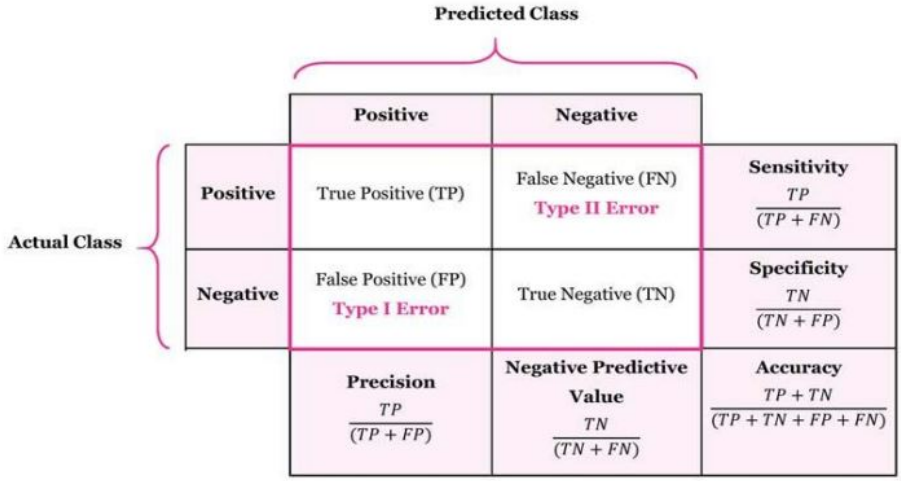

Figure 5: A confusion matrix representation for evaluation of model

Train and Validation Accuracy and Loss

Accuracy is one parameter used to test models of classification. Imprecisely, accuracy is the proportion of observations that the process has got correct.
Some of the most significant issue in a machine learning algorithm will be overfitting. That is when the model suits well with the dataset, but it cannot make assumptions and make accurate assumptions about the general data.
This graph will represent the ratio between the accuracy observed when training the data, and the accuracy with the validation data.

Train and Validation Loss

Training a model involves learning (determining) positive values from designated instances with both the weights and the bias. A supervised deep learning model is shaped by looking at several instances and trying to come up with a solution that reduces the loss; this method is called computational risk minimization.
Loss is retribution for poor judgment. It is a measure that describes the forecasting of the model. If the forecasting of the approach is perfect, then the loss has to be 0; else the loss is higher.

4 Design Specification

Transfer Learning

Transfer learning is a form of machine learning in which a process built for a specific task is used as a preliminary step for a proposed task. It is a standard method in deep learning where pre-trained models are used as a starting point for computer vision and natural language processing tasks. Due to the immense computing and time, this method provides a significant capacity leap in resources required to build neural network models on these issues. The flow of transfer learning is defined in Figure 6.

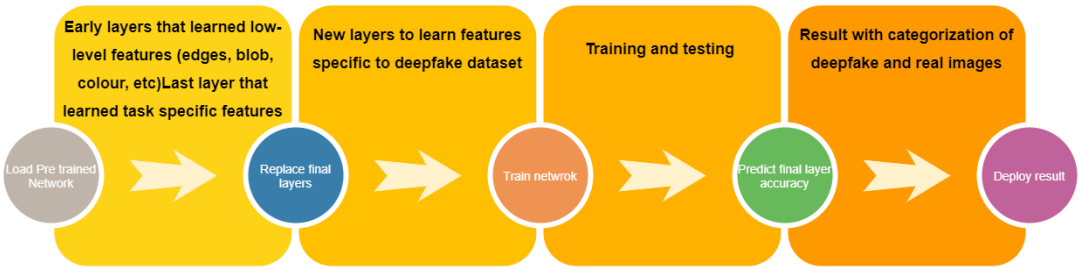

Figure 6: Transfer learning architecture schema).

Inception-ResNet-V2

Inception-ResNet-V2 is a robust deep convolutional network essential to the most significant developments in image recognition in recent years. The Architecture has shown to produce very higher efficacy at a reasonably low computational expense.
As described in Szegedy et al. (2016), Inception-ResNet-V2 carries the design of the hybrid Inception with vastly enhanced identification performance compared to ResNet-V1 and Inception-V4. Figure 7 shows the architecture of Inception-ResNet-V2.

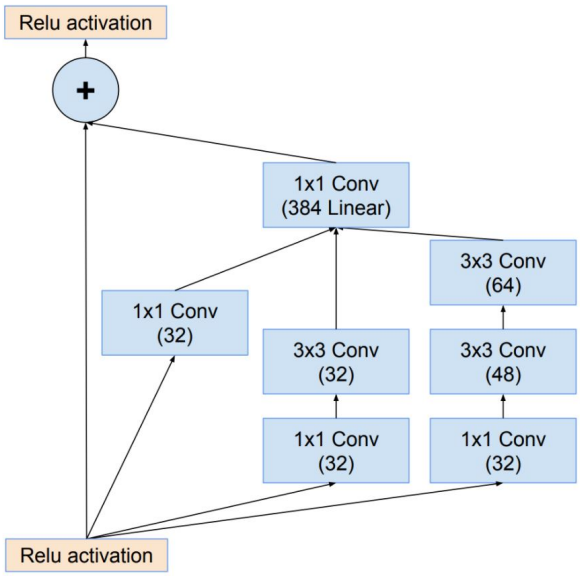

Figure 7: The architecture schema of Inceptions ResNet V2 neural network (35 x 35 grid module).

Sigmoid Activation Function
The Sigmoid Activation function (Figure 8) condenses input values in the range of 0 to 1.

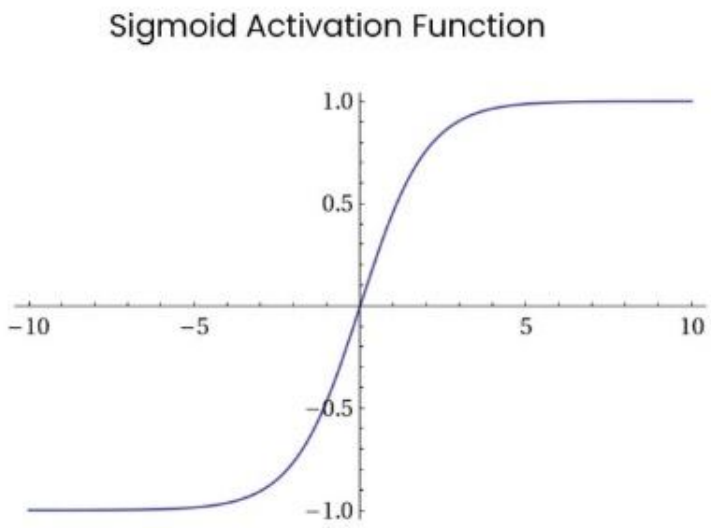

Figure 8: Sigmoid Activation Function.

Global Average Pooling 2d layer
An average pool takes a kernel size and offers the average value, so the stride value can pass. Global Average Pool, therefore, is the kernel scale of H x W lengths. It takes the global average over height and width and gives a 1xC tensor for the input of H x W x C

Binary Cross-entropy Loss Function
Known as Loss of the Sigmoid Cross-Entropy activation function. It is a Sigmoid activation, and as well as a loss of Cross-Entropy. It is unique to every vector variable, which implies that other vector variables do not influence the Loss measured for each segment of the CNN output matrix.

5 Implementation

Prelimary Data Exploration
Implementation begins with the exploration of the videos dataset. Loading the dataset in the python and exploring it. Determining the type of files present in the dataset. It is said that data analytics is all about understanding the data. The dataset contains video files and a JSON file with the labelling of fake or real video with their names.

After this, check if any data is missing in the dataset, like video filename not present in the JSON. This problem has many solutions to deal with, but the dataset used in this research did not have any missing value. Research progresses with checking for unique values, most many originals and missing video.

Initially, images are extracted from the videos. A folder is created which will store frames captured from the video. In the code, it can be defined how many frames are required to capture. This project is taking all possible frames from the videos. Quickly, it can be said that if a short, low-resolution clip includes too many smudgy pixels, across the head, this indicates a high likelihood that this video could be a fake one. Nevertheless, this approach cannot work correctly as with the evolving technology; more clear deepfakes are getting produced. Also, few real and fake videos are played to see any visible differentiation between the two categories.

It is required to dig further into the understanding of deepfake videos for determining the source of flaws in them. Here CV2 package is used to explore the videos and getting frames from the videos. CV2 is a package from OpenCV, a cross-platform library which helps to create applications for real-time computer vision. This package focuses mainly on image recognition, video capturing and interpretation, including functionality such as face identification and object detection.

Face Extraction
Next, step is the face extraction, two new folders are created to store face extracted from each frame of the video. Each frame of the video is captured using the CV2 library, and then facial landmark is defined using the get_frontal_face_detector of dlib package. Detection of face images is a subcategory of the topic of form analysis. A pattern determinant tries to locate key places of interest all along with the structure of an image given in input.

All the extracted face images are stored in the newly generated folders of real and deepfake face. These faces contribute to the dataset, which will be sent to the model for the analysis.

Next, pre-process the new dataset before sending it into the model for training. Define the input shape of the image, normalize the images and reshaping is performed at this stage.

After the pre-processing, the dataset is divided into testing and training sample using the train_test_split method of sklearn package.

Transfer Learning
Transfer learning is a machine learning method in which a framework generated and learned for a general analysis like image categorization and can be used again as the preliminary step for another analysis. Transfer learning is helpful if the data is inadequate for the analysis or the available computation for the modelling is less than required for a neural network to be trained appropriately.

By applying transfer learning on a trained server, the size and difficulty of the training sample may be significantly reduced; this approach can be used to tackle the classification and regression analysis of reasonably small data sets.

This project is using Inception-ResNet-V2 transfer learning model. Inception-ResNet-V2 is a combination of the two most popular transfer learning networks, i.e. Inception block and ResNet blocks, to boost the performance. Architecture has shown to produce very higher efficacy at a reasonably low computational expense.

Steps followed to generate model using transfer learning are as follows. First, remove the default loss layer, the output layer used to render forecasts, and substitute that one with a deepfake detection loss output layer. This failure feedback is a fine-tuning network to decide whether preparation restricts variations from either the data identified or the performance forecasted. Retention vigorously throughout the whole network does not necessarily deliver the most efficient result.

In this research the model is run for 20 epochs to learn the training dataset. The model uses Sigmoid activation function. The sigmoid activation function, sometimes known as the logistic function, is a prominent activation function for neural networks. The function information is translated to a value within 0 and 1.

Summary of the model shows that the output shape of Inception_Resnet_V2 model layer is (None, 2, 2, 1536) with 54336736 parameters, output shape of Global_Average_Pooling2D average pooling layer is (None, 1536) with 0 parameters and of the dense layer is (None, 2) with 3074 parameters.

6 Evaluation

Accuracy
The accuracy value is the reliability of the evaluation (test dataset) on the other hand, validation accuracy, a measure of the reliability of the model on the data collection not used to train the model.

In the research, the accuracy value achieved by the model after 20 epochs is 0.9923. This accuracy means that the model is 99.2% accurately categorising deepfake videos from the real one. Also, the validation accuracy achieved by the model is 0.9092, meaning 90.9% correct categorisation of deepfakes from the validation dataset, the dataset which model has not seen before. The ratio between accuracy and validation accuracy is shown in Figure 9.
The less difference between the accuracy and the validation accuracy means the model trained in the research is more general in usage.

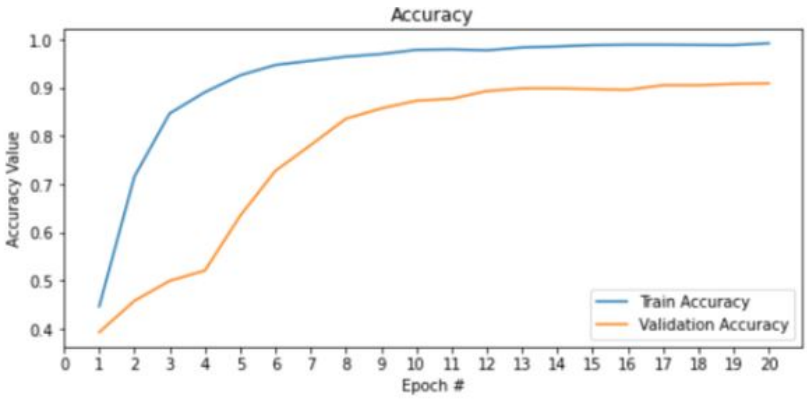

Figure 9: Ratio of loss vs validation accuracy

Loss
Loss value is the loss measured throughout a model preparation on the training data, whereas val_loss value is the loss calculated on the validation data. Loss is compared with val_loss, and if the loss is substantially lower, the data is likely to be overfitted meaning that the model is trained exhaustible on the training data and is not suitable on general data.

In the research, the loss value of the model result is 0.0309. This is a really good value in terms of image data training model. Suggesting that the model is good at determining deepfakes.

A validation loss is a great predictor of whether the lack of experience with more dataset will mitigate more. The research yields a validation loss value of 0.2572, this value is suggesting the model is making some bad decisions with the general data as compared to the training data but the difference is not that huge. The comparison of value changes between loss and validation loss is shown in Figure 10.

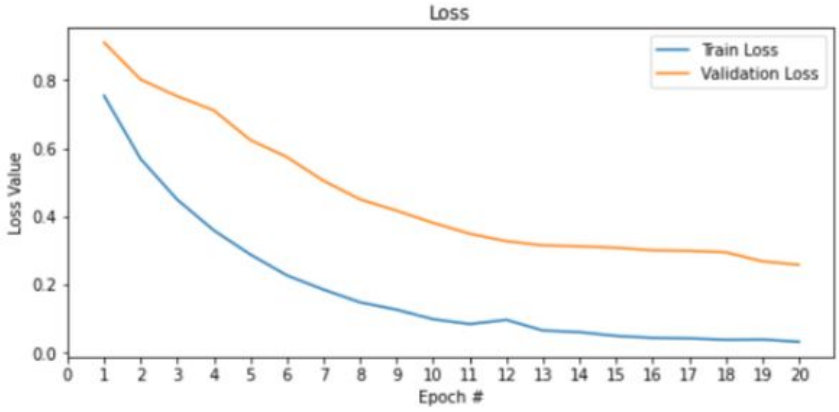

Figure 10: Ratio of loss vs validation loss

It is observed that on epoch number 12, the difference between loss and validation loss is reduced, but the loss value is increased; hence the research does not stop on epoch 12 and continues till defined number of epochs.

Confusion Matrix
A confusion matrix is a description of the outcomes of analysis over a classification issue. The count of observations that are accurate and inaccurate is listed and subdivided by class. It reveals how uncertain the model of grouping is as it creates assumptions. This offers one visibility in not only to point the mistakes produced by a classifier but, most specifically, the kinds of mistakes created.

Following are the describing classes of a confusion matrix:
Positive (P): Positive conclusion (for example: yes a deepfake).
Negative (N): Note is not optimistic (for instance: not a deepfake).
True Positive (TP): Observation is good, so it is predicted optimistically.
False Negative (FN): Observation is good but adversely expected.
True Negative (TN): Perception is negative and optimistic is expected.
False Positive (FP): Observation is negative but positively predicted.

The confusion matrix of the research is shown in Figure 11.

For simplicity, this confusion matrix is shown with the darkness of colour in each class. More bright values mean large numbers; dark values mean smaller numbers. The confusion matrix has a very dark colour in the section of False-positive (41) and False-negative (49), this means that the model is making very fewer mistakes and predicting more precisely.

```
True positive =  2945
False positive =  41
False negative =  49
True negative =  710
```

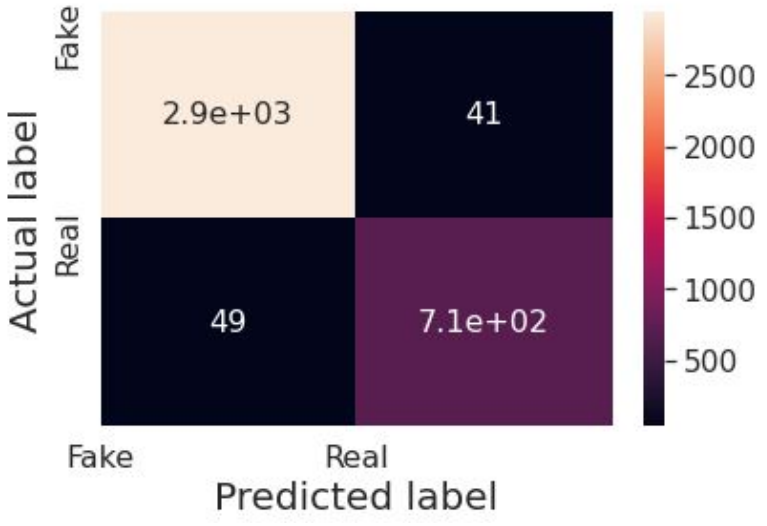

Figure 11: Confusion matrix with the result of model

7 Conclusion and Future Work

This research is conducted to determine how well can the deep learning model developed over transfer learning perform in precisely distinguishing deepfake videos generated by AI from real video shot on camera.

The research uses InceptionResNet-V2 transfer learning model, which is trained on massive image datasets as the pre-trained model. This model is comparatively better performing than other models in the area of image categorisation. After the evaluation of the model, it can be concluded that deep learning model developed over a pre-trained model can very effectively determine deepfake videos from a real one and has much future scope.

The evaluation of the research shows that the model trained on 200 short video dataset can achieve 90.02% accuracy with loss of 25%. This result is a promising result of the analysis to answer it is the research question.

Limitation and Future Work
The main limitation dealt in this research was the lack of computation power. As the research is dealing with video dataset, a lot of RAM is required to store data which is under the analysis. There were many options used to directly train the model on video dataset, like optimising the model to run minimum analysis, reducing the size and factors in videos. However, the model was running out of memory in every scenario.

Due to the above limitation, the model had to be trained on the images of frames captured from videos. This lack of computation caused a decrease in the efficiency of detecting deepfake videos for a model which could learn categorisation of videos directly. This research holds a great opportunity for future work. The dataset used in this research is a small portion of data from a challenge organised for generating a deepfake detection model. The entire dataset was huge of 950GBs which could not be trained on a local machine, or even in Google Colab.

After trying various approaches, it can be deduced that deepfake videos hold many challenges in detection. Moreover, training with limited videos is not enough to develop a very general model.
With more high computation power and memory resources, more advance model can be developed, which can take videos in the input.

References

Amerini, I., Galteri, L., Caldelli, R. and Del Bimbo, A. (2019). Deepfake video detection through optical flow based cnn, pp. 1205–1207.

Caporusso, N. (2020). Deepfakes for the good: A beneficial application of contentious artificial intelligence technology, *Advances in Intelligent Systems and Computing* pp. 235–241.

Güera, D. and Delp, E. J. (2018a). Deepfake video detection using recurrent neural networks, pp. 1–6.

Güera, D. and Delp, E. J. (2018b). Deepfake video detection using recurrent neural networks, pp. 1–6.

Hashmi, M. F., Ashish, B. K. K., Keskar, A. G., Bokde, N. D., Yoon, J. H. and Geem, Z. W. (2020). An exploratory analysis on visual counterfeits using conv-lstm hybrid architecture, *IEEE Access* **8**: 101293–101308.

Hsu, C.-C., Lee, C.-Y. and Zhuang, Y.-X. (2018). Learning to detect fake face images in the wild, *2018 International Symposium on Computer, Consumer and Control (IS3C)* pp. 388–391.

Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z. and Qu, R. (2019). A survey of deep learning-based object detection, *IEEE Access* **7**: 128837–128868.
URL: *http://dx.doi.org/10.1109/ACCESS.2019.2939201*

Jung, T., Kim, S. and Kim, K. (2020). Deepvision: Deepfakes detection using human eye blinking pattern, *IEEE Access* **8**: 83144–83154.

Kharbat, F. F., Elamsy, T., Mahmoud, A. and Abdullah, R. (2019). Image feature detectors for deepfake video detection, pp. 1–4.

Korshunov, P. and Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection.

Kumar, A., Bhavsar, A. and Verma, R. (2020). Detecting deepfakes with metric learning, pp. 1–6.

Long, J., Shelhamer, E. and Darrell, T. (2015). Fully convolutional networks for semantic segmentation, pp. 3431–3440.

Maksutov, A. A., Morozov, V. O., Lavrenov, A. A. and Smirnov, A. S. (2020). Methods of deepfake detection based on machine learning, pp. 408–411.

McCloskey, S. and Albright, M. (2018). Detecting gan-generated imagery using color cues.

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T. and Nahavandi, S. (2019). Deep learning for deepfakes creation and detection: A survey.

Nirkin, Y., Keller, Y. and Hassner, T. (2019). Fsgan: Subject agnostic face swapping and reenactment, *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* pp. 7183–7192.

Pishori, A., Rollins, B., van Houten, N., Chatwani, N. and Uraimov, O. (2020). Detecting Deepfake Videos: An Analysis of Three Techniques, *arXiv e-prints* p. arXiv:2007.08517.

Radford, A., Metz, L. and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks.

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A. (2016). Inception-v4, inception-resnet and the impact of residual connections on learning.

Thies, J., Zollhöfer, M. and Nießner, M. (2019). Deferred neural rendering, *ACM Transactions on Graphics (TOG)* **38**: 1 – 12.

Yang, C., Ding, L., Chen, Y. and Li, H. (2020). Defending against gan-based deepfake attacks via transformation-aware adversarial faces, *ArXiv* **abs/2006.07421**.

Younus, M. A. and Hasan, T. M. (2020). Effective and fast deepfake detection method based on haar wavelet transform, pp. 186–190.