National College of Ireland

DEEPFAKE DETECTION: COMPARISON OF PRETRAINED XCEPTION AND VGG16 MODELS

MSc Research Project

Data Analytics

Karthika Nair

Student ID: 22105522

School of Computing

National College of Ireland

Supervisor: Shubham Subhnil

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Karthika Nair		
Student ID:	22105522		
Programme:	MSc. Data Analytics	**Year:**	2023
Module:	MSc. Research Project		
Supervisor:	Shubham Subhnil		
Submission Due Date:	14-12-2023		
Project Title:	Deepfake Detection: Comparison of Pretrained Xception and VGG16 Models		
Word Count:	6479	**Page Count:**	23

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:	Karthika Nair
Date:	14-12-2023

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: COMPARISON OF PRETRAINED XCEPTION AND VGG16 MODELS

Karthika Nair

22105522

Abstract

The detrimental effects of deepfake on the society is the main topic of concern for this research. Deep learning is adopted to battle this issue as it is known for its exceptional proficiency in the effective detection of deepfakes by learning hierarchical features, intricate pattern and spatial relationships from the dataset. The emphasis is on employing CNNs especially Xception and VGG16 for addressing the threat of deepfake technology. The DFDC dataset is extracted from Kaggle aiding in utilising transfer learning to enhance the model efficiency. A comparison is carried out between employing transfer learning that works in collaboration with CNN as the baseline model and Xception and VGG16 are used for finetuning. The architectural design of this research includes the loading pretrained model, fine-tuning, and data augmentation, resulting in an effective system to detect deepfakes. The code execution is done by making use of python libraries such as Keras, Matplotlib, sklearn, Jupyter Notebook and Visual Studio code. The Xception model yields a remarkable accuracy of 84.6% whereas the VGG16 model gives an accuracy of 63%.

Keywords: Deepfake, Convolutional Neural Network (CNN), facial recognition, deep learning, transfer learning, data augmentation, Xception, VGG16

1. INTRODUCTION

1.1 MOTIVATION and OBJECTIVE

Deepfakes are a resultant of high levels of manipulation of video, images or audio using artificial intelligence, especially deep learning. Deepfakes are usually meant to deceive people as it can be highly convincing because it seems authentic even though it is completely synthesised. Using deep neural networks (Dagar & Vishwakumar, 2022), the deepfake technology generates new content by superimposing the facial expressions, voice and/or gestures of actual individuals. This technology has turned out to be a menace to the society since several public figures and celebrities have fallen to be victims of such forged content[1].

This technology has also been associated with multiple cases of identity theft, fraud, and cyberbullying. Every day, the technology seems to get even better, and it has become harder to

[1] https://westernpeople.ie/people-will-end-up-more-sceptical-of-electoral-process-how-ai-and-deepfakes-threaten-elections_arid-4601.html

find stark differences between genuine and forged content, thus deteriorating the trustworthiness of the visual or audio information. It is terrifying to think that through deepfake technology, there is a high potential for malicious intent like spreading misinformation, political manipulation causing reputational damage. Addressing these problems are a necessity in these times, forcing companies to develop tools that can help identify such malicious content.

In support of this research, the dataset undertaken for deepfake detection is obtained from Kaggle. Kaggle is a well-known prominent platform that hosts numerous datasets that deem to be important to the advancement of several technological fields. The dataset utilised here for this project is the "Deepfake Detection Challenge (DFDC)[2]", which feature in a competition hosted by Kaggle which can be considered as a major source for training and evaluation of models committed to distinguishing manipulated content. This dataset is an amalgamation of real and deepfake images and videos, which includes visual content of various individuals, backgrounds, and contextual scenarios. The dataset is a substantial array of labelled images which enabled the exploration and development of detection of deepfake content.

Convolutional Neural Networks have emerged to be pivotal for the detection of deepfake[3] content as they exhibit remarkable proficiency in effective learning of hierarchical features from images, capturing intricate patterns and spatial relationships of the visual content. The key features to be considered while performing deepfake detection are the spatial consistency of facial expressions and features.

This research project will focus on transfer learning, pre-trained CNN models utilizing datasets like ImageNet[4] for the efficient reporting of detecting forged content. Through knowledge transfer the model has an added advantage to distinguish between genuine and manipulated content. Transfer learning with CNNs utilizing architectures like VGG16 and Xception has been researched upon as it can significantly enhance the efficacy of deepfake detection.

Another vastly used CNN architecture is VGG16 which is incorporated with transfer learning to reap its advantages in the detection of deepfake. VGG16 is best known for its efficient extraction of facial and spatial features, thus being a strong building block for recognizing intricate patterns within the visual content.

The model learns the visual characteristics by training itself on large datasets while modifying the top layers of the network for specific detection requirements resulting in optimal performance by fine-tuning the model on target datasets.

The approach of merging transfer learning with Xception (below) and VGG16 helps in speeding up the model convergence as well as facilitating the model to effectively distinguish between genuine and manipulated content. This is done through integrating specialised domain of deepfake detection with pre-existing information from generic image recognition tasks.

[2] https://www.kaggle.com/competitions/deepfake-detection-challenge
[3] https://arxiv.org/abs/2304.03698
[4] https://image-net.org/

1.2 RESEARCH QUESTION

Research Question: "How deepfakes are a threat to cybersecurity and protection of personal data, and why deep learning models are recommended in the detection of deepfakes?"
Sub RQ: "How can the implementation of transfer learning, specifically with Xception model and VGG16 enhance the efficiency of deepfake detection?"

The rest of this paper will include sections in the following format. The subsequent section would be a brief exploration of studies that inspire and support this research. The sections (3 and 4) following that would provide a detailed description of the methodology, acquiring of dataset, preprocessing phases, and the application of deep learning techniques. The implementation of the techniques and evaluation of the model would be explained in the sections 5 and 6. The paper concludes with section 7, articulating the summary of the research in its entirety and providing suggestions in the aspect of future research.

2. RELATED WORK

2.1 DEEPFAKE DETECTION WITH CNN ARCHITECTURES

The research paper (Misra & Singh, 2016) summarizes the challenges faced in detection of faces due to variations in individuals stature, pose, lighting and occlusion by proposing a method which combines CNN with Hough transform which is a computer vision technique that helps in the detection of shapes and sizes. This technique helped the researchers to achieve improvement in face detection accuracy and alignment, with a recognition rate between 90% and 100%, thus surpassing other traditional approaches. Utilizing CNN architecture not only aided with lower failure rates and increased accuracy, but it also showed high resilience to illumination variations.

The study by (Kommalapati, 2021) explains the risks posed by the spread of deepfakes, the influence on media authenticity and integrity. The researchers have emphasised the use of Generative Adversarial Networks (GANs) for the generation of synthetic content like videos in a large scale which include certain public figures. A deepfake detection approach called DefakeHop has also been introduced in this study which adopts PixelHop++ and a feature distillation module. The study was conducted to test the efficacy on CelebDF dataset and the resultant Area Under the Curve (AUC) value was a whopping 94.6%. The idea portrayed through this paper is that the development of advanced detection technologies can help combat the threat of manipulated content by successful generated of refined deepfakes which would help the model learn better.

The study (Das, et al., 2022) explores the world of deepfake technology, discussing its positive and negative implications. In a positive point of view, this technology can be used in film-making and virtual reality in an extraordinary fashion. But the negative aspects can be rather threatening to the lives of people, contributing to political manipulation and even extortion. The researchers have worked on a comparative analysis of deepfake detection models in this paper, with emphasis on CNN models like VGG16, EfficientNet, and ResNet, including Recurrent Neural Network (RNN) models such as LSTM. A thorough emphasis is made on the combination of RNN and CNN

which addresses resolution inconsistencies and temporal discontinuity. The result of this study indicates that CNN with Support Vector Machine (SVM) outperforms hybrid CNN and RNN models with a high accuracy value of 98%.

The FaceForensics++ dataset has been put into use in the study (Jolly, et al., 2022) with a different approach on deepfake detection i.e., facial expression detection. This approach has achieved an accuracy of 99% in detecting deepfakes, Face2Face, FaceSwap, and Neural Texture manipulations. The implementation can be seen as a layered process which makes use of facial recognition networks, CNN, long short-term memory (LSTM) layers and Recycle GAN. All these methodologies have been used for spatial and temporal data fusion and indicate promising results in the detection of forged content.

The paper (Ilyas, et al., 2022) addresses various techniques used to detect deepfakes in content that exhibits variations in illumination conditions and ethnicities. These can be considered as challenges and the researchers propose a hybrid deep learning framework called InceptionResNet-BiLSTM that collaboratively combines a customized InceptionResNetV2 for feature extraction with a temporal awareness model called Bidirectional Long-Short Term Memory (BiLSTM). The model evaluation is conducted on FaceForensics++ and FakeAVCeleb datasets and the results demonstrate over 90% accuracy on subsets like FaceSwap, Face2Face and DeepFakes.

The dark potential of deepfakes and the mechanisms for demolishing of the same have been addressed in (Mira, 2023). In this study, it can be seen that deepfakes are generated through a technique called GAN and the paper illustrates more depth in CNNs, RNNs and LSTM for the successful identification of forged and genuine content. Moreover, there is a proposal for deepfake detection which is a combination of YOLO face detector, InceptionResNetV2 CNN, and XGBoost classifier. There is added focus on the recommendation for continuous improvement in detection methods and increasing the quality of deepfakes in the paper.

The adoption of InceptionNet architecture is considered to be of high relevance when it comes to detection of deepfake images in an efficient manner. This is addressed and explored well in (Prasannavenkatesan & Ghouse basha, 2023). This research paper was put into consideration because it conducts a comparative analysis of multiple convolution neural network algorithms, focusing on InceptionNet's role in distinguishing deepfakes. For the purpose of conducting this research, a dataset consisting of 401 training videos is selected and 3,745 images have been generated through augmentation processes. The result of this study gives a promising accuracy of 93% in separating deepfake content from real content. A brief discussion is mentioned in the paper regarding the significance of deepfake in the context of fake news, and it being a threat to national security and also a potential harm to any individual.

Research paper by (Harsh, et al., 2023) indulges in addressing the threats of deepfake technology. This research works by conducting a classification of deepfake human faces using transfer learning. EfficientNetV2 model has been leveraged along with EfficientNetV2B0-B3, S, M and L to evaluate the accuracy of successful distinction of fake and genuine images. The proposed method gives out result which is a whopping 99.97% accuracy. This study has an important emphasis on combating the challenges put forward by deepfake images.

Two approaches have been suggested in this paper, which is the CNN model and transfer learning using pre-trained models including VGG-16, ResNet-50, InceptionV3 in the research paper (Niteesh, et al., 2021). The model in this study achieved a promising accuracy of 87.46%, making it a better approach when compared to standard models. It is consistently proven that logistic regression is the most effective classifier when compared to other pre-trained models.

(Borgalli & Surve, 2022) emphasises on Facial Expression Recognition (FER) using CNN architectures addressing the importance of FER in AI technology and its applications. The study builds its work on real-world scenarios instead of lab-controlled datasets and evaluates performance using standard architectures like VGG and InceptionResnetV. The end resultant of this study reveals an impressive accuracy in distinguishing basic emotions. (R, et al., 2022) also uses the same ideology but uses DenseNet-169 as the foundation network. The key role of this network is to extract and selection of feature in images, especially capturing and recognizing emotions from different angles. The model performs well on the Emotion Recognition dataset with a remarkable 96% accuracy. This study finds applications in medical fields, face unlocking and feedback analysis. A combination of ensemble learning algorithm, VGG16, EfficientNetB0 and InceptionResNetV2 can be seen in (Gwo-Chuan, et al., 2022). The FER-2013 dataset is utilised with data augmentation performed on the dataset. The accuracy achieved in this study is around 70.5% suggesting a 2.81% improvement in the recognition of facial emotions.

2.2 TRANSFER LEARNING WITH XCEPTION MODEL

(Ihsan & AlAsady, 2022) use a deep learning approach which has a modified structure of the Xception net as its base which is a CNN architecture. This altered structure is a separable convolution layer is referred to as the Modified Trimmed Xception (MTXception) forms a dense connection inspired by DenseNet. The researchers have also found that the use of the YCbCr colour system for images has an increased accuracy rate of 99.93% when compared to other colour systems like HSV, Lab and HSV. The dataset used for this research is '140k Real and Fake Faces' dataset which consists of several images including real and fake face generated by StyleGAN. It is observed in this study that MTXception net outperforms the original Xception net and pre-trained Xception net when it comes to parameter efficiency, time of execution and accuracy. The proposed model gives promising effectiveness in detecting forged content in terms of the evaluation metrics like accuracy, precision, recall, F1 measure and loss function.

The effectiveness and accuracy of the model by the adoption of architectural considerations and the number of frames for efficient model performance has been mentioned in (Kusniadi & Setyanto, 2021). The dataset used here for this study is the FaceForensics++ and through MTCNN it was made possible to isolate faces. It was observed that fine-tuning and transfer learning has successfully contributed to the improvement of Xception Net architecture. The testing was carried out on the Celeb-DF dataset resulting in a high accuracy of 83.75%.

A system utilising the Xception model with depthwise separable convolution which is designed specifically for the visual detection of deepfakes generated through AI can be seen in the paper (Muafy, et al., 2023). It leverages a large dataset of 20,000 images that included generated forged

images through StyleGAN dataset and genuine images from CelebA-HQ. This study revealed an impressive score of 98.77% accuracy and showed promising values for the precision, recall and F1-score which indicate high efficiency in the detection of fake faces amongst genuine faces. The methodology works by performing data augmentation on the images for resizing them to 100 pixels, and then consecutively training the Xception model which showcases its depth-wise separable convolution's efficiency in feature extraction through its special architecture containing entry, middle and exit flows. Xception model has proved to be a better performer when compared to other CNN models.

2.3 TRANSFER LEARNING WITH VGG16 MODEL

Transfer learning can work well with limited training data, and this has been addressed in (Tao, et al., 2021). Considering small and light Synthetic Aperture Radar (SAR) images, feature classification has been a studied in (Tao, et al., 2021). The researchers have made use of the feature classification model with the core as VGG16 neural network. A clear distinction of buildings, roads, farms, and trees is the objective for this study.

Stark challenges were observed in obtaining optical images effectively in different weather conditions especially when it is rainy and cloudy. A dataset is constructed with images that have been labelled previously for supporting feature categorisation. This dataset is employed with the VGG-16 framework and the model works specifically by focusing on SAR images. Data augmentation is applied on the dataset to improve the model's capability to predict and identify features in an efficient manner. The study reveals promising experimental results with the model reaching the accuracy till 75% without any parameter adjustments, and with adjustments it increases to 81%. And the additional use of pretrained models through transfer learning improves the accuracy to 87.5%.

Deep learning techniques have been known to give outstanding performance for face detection owing to the depth of the layers and large-scale training datasets. But for the same reason, they require extensive computational resources. (Perdana & Prahara, 2019) observes the difficulties in recognizing facial features by employing a light-CNN network with modified VGG16 model as the base for the task of face recognition. The methodology follows four stages: face detection, face alignment, feature extraction, and classification.

Handcrafted features like Eigenfaces and Local Binary Patterns have been put into use for extraction of features. But the catch with this method is that in the conditions of unconstrained environments, there is a chance for performance degradation because of certain factors like variations in illumination, occlusion, poses and complicated backgrounds. This model works with limited datasets as certain layers are removed from the VGG16 model making it more simplified yet highly effective. This study focuses on a dataset with 7,250 images and 30 labels with the model achieving an accuracy of 94.4% on the test dataset thus outperforming other modified VGG16 models with an indication that light-CNN has the potential to give remarkable outcomes in comparison to deeper alternatives.

A hybrid approach for the detection of facial features with high resolution RGB images can be observed in the study (Aung, et al., 2021). This approach is an amalgamation of the VGG16 pretrained CNN with the You Only Look Once (YOLO) algorithm resulting in an improved technique for face detection in real-time live video content. The study addresses key features like facial positioning, colours, and illumination variations and provides an apt solution to them. Functional foundation of this model is obtaining an FDDB dataset with VGG16 model to extract facial features in them and using YOLOv2 to detect faces. To address the issues that could lead to overfitting, data augmentation is employed and with this along with the removal of unnecessary layers to simplify the architecture, the model achieves an outstanding precision value of 95% on the test image set.

3. METHODOLOGY

This section will explain procedure to create a model that detects deepfakes efficiently by following certain key steps that come under a systematic framework known as Knowledge Discovery in Databases (KDD)[5]. Data selection is one of the initial steps which includes choosing diverse and relevant datasets comprising of real as well as fake or forged images. The KDD process also mentions cleaning and preprocessing of data followed by transforming the data to make it compatible for the models to learn.

The chief goal is to unveil patterns and features in the images of the dataset while keeping the business objectives in mind with a thorough consideration to following the ethical requirements. The final step would be to obtain visualisations of the discovered knowledge. The following sections will explain each of these principles in detail:

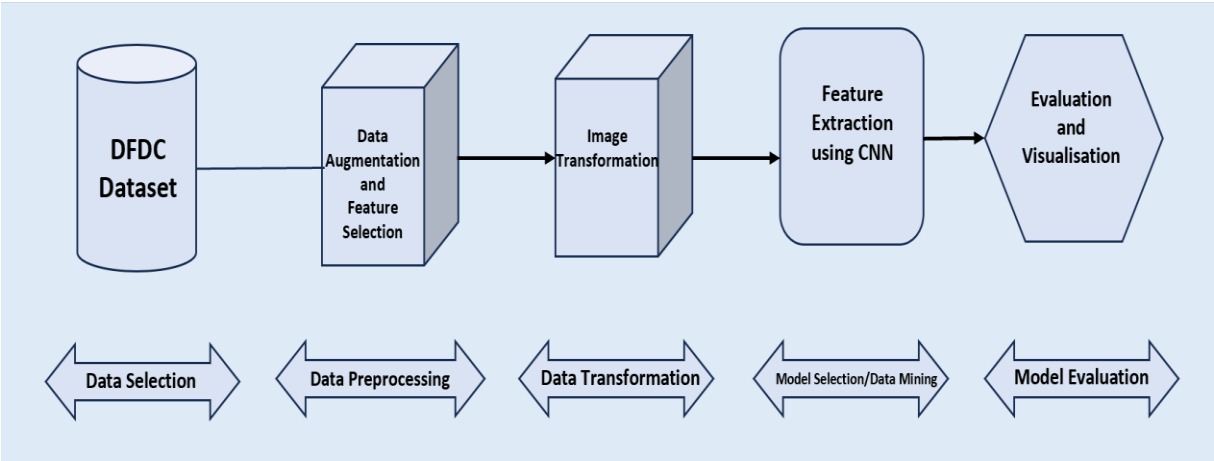

Fig. KDD process for deepfake detection

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7924527/

3.1 DATA SELECTION

Deepfake detection has been a rising challenge in the world of artificial intelligence and because of this, multiple datasets have been made publicly available to help mend the problem of deepfakes. One such dataset has been obtained for this research from the Kaggle DeepFake Detection Challenge[6].

For external model training, the training set has been made accessible through a GCS bucket, enabling researchers to employ various methodologies in order to develop models that successfully detect deepfakes outside the scope of Kaggle notebooks. This large dataset has been distributed among 50 files to make it easier for researchers to develop models by working with dataset sizes that are suitable for their environment.

A public validation set is also available as a subset of 400 videos, to help validate the developed models and finally the model is fitted on the test set for generating comparative insights for the performance of the model.

3.2 DATA PREPROCESSING

Preprocessing of dataset is a crucial step in the development of a deepfake detection model to prevent generation of irregular outcomes due to attributes values that are completely irrelevant and/or due to false information. The '`pull_dataset`' function works on the preprocessing by fetching the dataset from a specified location with information on the video names and labels marking the video as real or fake and converting it into JPG format. To facilitate the model training process, the function performs encoding on these labels and converts them to numerical format by denoting them as 1 for 'FAKE' and 0 for 'REAL' and generates NumPy arrays with images and labels which is later used for model training and evaluation.

3.3 DATA TRANSFORMATION

Raw image data needs to be transformed to a format that is ideal and compatible for the training models especially for the training stages. Incorporation of necessary data transformation techniques is crucial for the model to perform effectively.

The TensorFlow model construction and training process is accountable for the process of data transformation especially during conversion of the NumPy arrays that represent training, validation, and test sets to generate TensorFlow Datasets.

The pixel values of the images are made compatible with the pretrained models (Xception and VGG16) through preprocessing and standardising using the preprocess function of the models on each of the images in the datasets. For the purpose of preparing these generated datasets for efficient training, they are shuffled, batched and prefetched.

[6] https://www.kaggle.com/competitions/deepfake-detection-challenge/data

3.4 MODEL SELECTION

The generalisation and the performance of the model depends entirely on the model that is selected with rigorous thought and decision process, making it a crucial step in the deep learning pipeline. Every model would have a different architecture, parameters, and complexities but in the end, the performance of the model is determined by how well they can capture the intricate patterns in the data. A suitable model should be chosen based on the problem at hand, the characteristics of the dataset and other factors like computational resources. Overfitting or underfitting would be completely avoided by a model which is chosen wisely as it makes a stark balance between the complexity and the simplicity of the model's architecture.

CNN Model: CNNs are known for handling image classification tasks effectively because they possess the capability to capture spatial and hierarchical features from any visual data. For this reason, they are accepted extensively for detection of deepfakes. CNN's special architecture comprising of convolutional layers and pooling operations allows it to learn relevant features on its own while remarkably obtaining intricate patterns that indicate manipulation in the content. CNNs can effectively distinguish between real and fake content owing to its ability to detect both low-level features, such as textures and edges, and high-level features, like facial expressions.

The adoption of transfer learning can contribute to better prediction since large image datasets, such as ImageNet, are used to pretrain CNN as they have consistently familiarised with generic features for visual detection and recognition tasks. As suggested in (Rahman, et al., 2022) in deepfake detection, the use of CNN with transfer learning promotes better recognition and identification as it learns the details and nuances of the content through the knowledge acquired from multiple datasets. When CNN is built as the base model, it mainly employs itself as a feature extractor and through transfer learning, the model is fine-tuned on target dataset.

Pretrained Xception Model: Xception model is often chosen for image classification tasks as when it is bridged with CNN through transfer learning, it has given remarkable results. Being a deep CNN architecture, it extends multiple benefits for detection of deepfakes. Computational efficiency is ideally maintained while this model extensively captures features that are complex owing to the depthwise separable convolutional layers that it possesses. Pretraining itself through a wide range of visual content dataset like ImageNet helps it learn complex hierarchical features.

The dependency on labelled dataset specifically for deepfake detection is generally reduced as pre-trained Xception model is introduced, making it more compliant to identifying forged content.

Pretrained VGG16 Model: The architectural complexity of VGG16 model is somewhat uniformly distributed and simple, with repetitive blocks of convolutional filters with max-pooling layers which decide the depth of the model. These convolutional filters are responsible for scanning the entire input volume. Due to the simplicity of the model architecture, VGG16 is used for several image classification tasks.

The model has a strong foundation that learns spatial and hierarchical features from the dataset owing to the convolutional filters along with the pooling layers. Integration with CNN layers during transfer learning makes the process of detection and classification more seamless and effective. Pretraining on large datasets such as ImageNet and then fine-tuning it for capturing intricate patterns generates knowledge for the model to use it later on for efficient detection of facial features. Since the model pretrains on large dataset, the difficulty of working with smaller and concise datasets have been eliminated as the model specialises in detecting unique patterns because it already has the knowledge from the pretraining.

3.5 MODEL EVALUATION

The model's performance will deem to be successful if it efficiently and effectively identifies the fake images from the genuine images. The accuracy of the predictions, the precision values and the recall value measure the model's ability to identify intricate patterns.

In this study, the evaluation is made through visual representations with a strong emphasis on accuracy and loss values. Training accuracy and validation accuracy is depicted in line graphs denoting training epochs with its corresponding accuracy values. With the aim of this visualisation being a clear visualisation of the model's learning dynamics, the fluctuations might imply any potential issues like underfitting or overfitting, meanwhile the convergence of the accuracy curve might imply efficient model learning.

4. DESIGN ARCHITECTURE

The aim of this design architecture is to create an adaptable, robust and efficient system for deepfake detection which can accurately distinguish real and forged content. The dataset is extracted from DFDC through Kaggle as it is available publicly. The architecture works by employing a deep learning approach with CNN as its base model and utilizing pre-trained Xception and VGG16 model.

These models are known for enhanced feature extraction and following this, data augmentation is performed which proactively improves the generalisation of the model. The performance of the model is judged and depicted through data visualisation techniques. Moreover, fine-tuning of the pretrained models (Xception and VGG16) with additional CNN layers is essential for the process of detection.

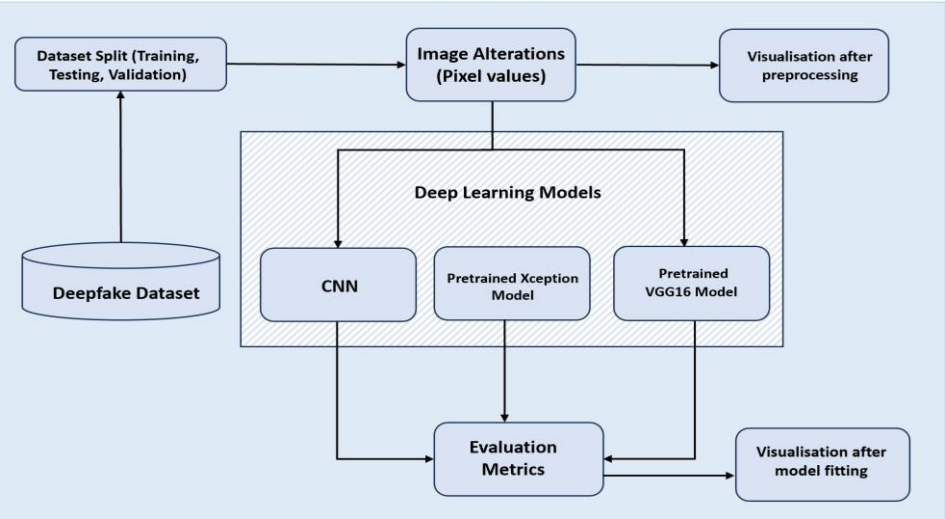

Fig. 2 Design Architecture

5. IMPLEMENTATION

5.1 DATA LOADING, PREPROCESSING, and EXPLORATION

One of the fundamental steps in the development of an efficient deepfake detection system is to have a proper structure for data loading and preprocessing as the attributes of the dataset decides the performance of the model immensely. The dataset undertaken for this research is utilized from the DFDC which comprises of both forged and genuine images which is used to train the model and evaluate its ability to distinguish real content from manipulated ones.

The data loading process is carried out by reading the metadata from '`metadata.csv`' file which comprises of all the information about each video and images in the dataset with their labels as an indication of the authenticity of the data. A dataframe is generated by combining the variables that were created while splitting the metadata into variables that have distinct real and fake content information.

This is followed by a split of the dataset into training, validation and test sets which facilitates the model to performing training and consequently, evaluate its performance. The '`train_test_split`' function is used for this purpose. To prevent a class imbalance during training, it is important to perform a stratified split which in turn ensures a balanced representation of the datasets. The Pandas library is used to perform data manipulation and Scikit-learn for splitting of the dataset.

	videoname	original_width	original_height	label	original
0	aznyksihgl.mp4	129	129	FAKE	xnojggkrxt.mp4
1	gkwmalrvcj.mp4	129	129	FAKE	hqqmtxvbjj.mp4
2	lxnqzocgaq.mp4	223	217	FAKE	xjzkfqddyk.mp4
3	itsbtrrelv.mp4	186	186	FAKE	kqvepwqxfe.mp4
4	ddvgrczjno.mp4	155	155	FAKE	pluadmqqta.mp4

Fig. 3 Content in metadata.csv

The preprocessing is done on the dataset to make it ready for model training. A balanced subset is created by generating certain samples that address potential biases that could come into picture from an imbalanced sheet. Sampling ensures an equal number of samples of real and fake data during training.

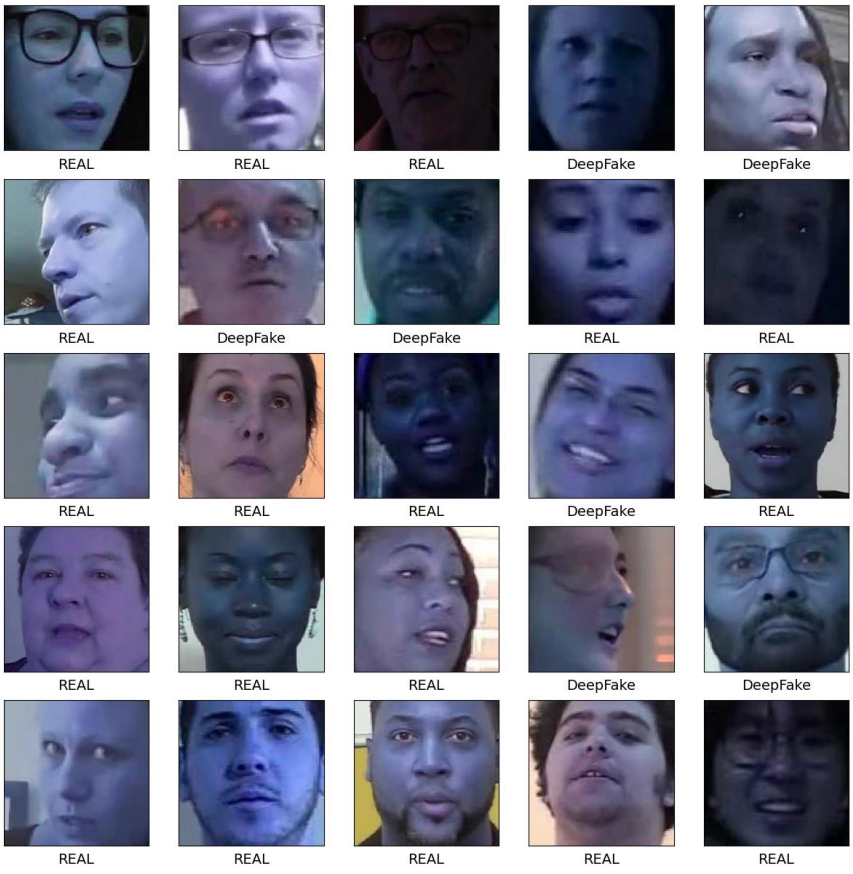

Fig. 4 Data exploration of real and deepfake images

The following figure shows the total number of classes per set in train, validation and test sets.

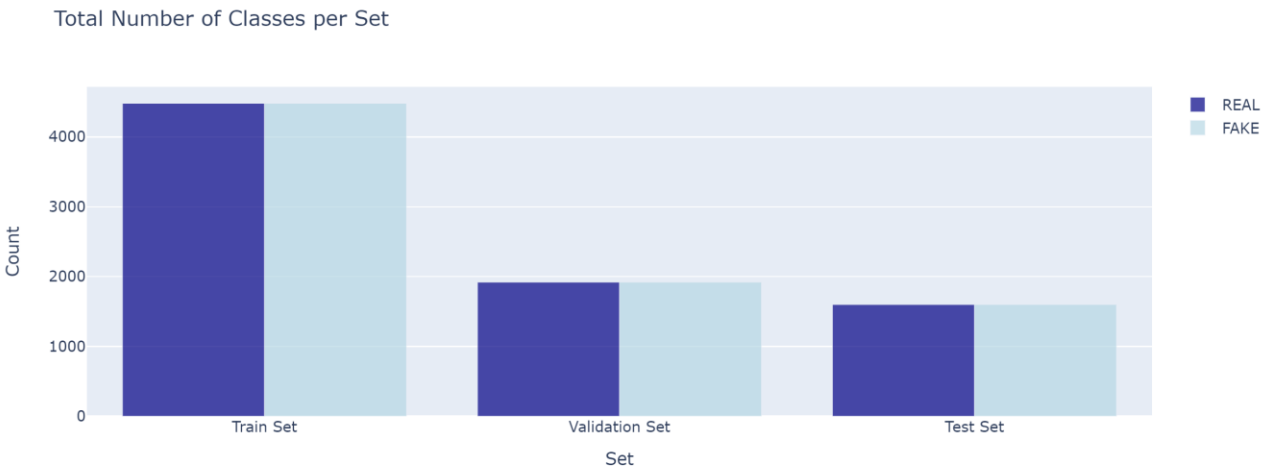

Fig. 5 Total number of REAL and FAKE classes in each set

The bar graph above illustrates the total number of classes that exist in the training, validation, and test sets providing a visual representation of the distribution of real and fake classes within each of these subsets. A balanced distribution is mandatory for an unbiased and reliable model behaviour. If the bars are of equal height, it is indicative of a balanced distribution of the dataset.

5.2 MODEL IMPLEMENTATION and EVALUATION

The model implementation is carried out with a simple Convolutional Neural Network (CNN) as a baseline model. The model is responsible for learning intricate patterns and features from images. For finetuning, this study has included pretrained models such as Xception and VGG16. The following sections will explain in detail how the implementation has been employed and finally a brief discussion on which model serves best result will be explained.

5.2.1 CNN as Baseline Model

The model's foundation is built upon TensorFlow's Keras API. And the network's layers are responsible for learning different features initiating with learning simpler features and then slowly progressing to complex patterns and objects.

Model compilation is done using binary cross-entropy as the loss function and Nadam optimizer. The dropout value chosen for this model is 0.5 as there should be a threshold for relying too intensely on specific features. The dropout layers disable some neurons during training which aids in solving this issue. The model compilation is done using binary cross-entropy as the loss function with accuracy as the metric, and Nadam optimizer.

The model is trained through 10 epochs with 140 cycles in each epoch, generating an accuracy rate of 50% with an F1 score of 0.66. This rate of accuracy is good but not great and thus, it is decided to use Xception and VGG16 models to see if finetuning with these models can help achieve better results.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_8 (Conv2D)           (None, 224, 224, 64)      9472

 max_pooling2d (MaxPooling2  (None, 112, 112, 64)      0
 D)

 conv2d_9 (Conv2D)           (None, 112, 112, 128)     73856

 conv2d_10 (Conv2D)          (None, 112, 112, 128)     147584

 max_pooling2d_1 (MaxPoolin  (None, 56, 56, 128)       0
 g2D)

 flatten (Flatten)           (None, 401408)            0

 dense_2 (Dense)             (None, 128)               51380352

 dropout (Dropout)           (None, 128)               0

 dense_3 (Dense)             (None, 64)                8256

 dropout_1 (Dropout)         (None, 64)                0

 dense_4 (Dense)             (None, 1)                 65

=================================================================
Total params: 51619585 (196.91 MB)
Trainable params: 51619585 (196.91 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Fig. 6 CNN Model compilation results

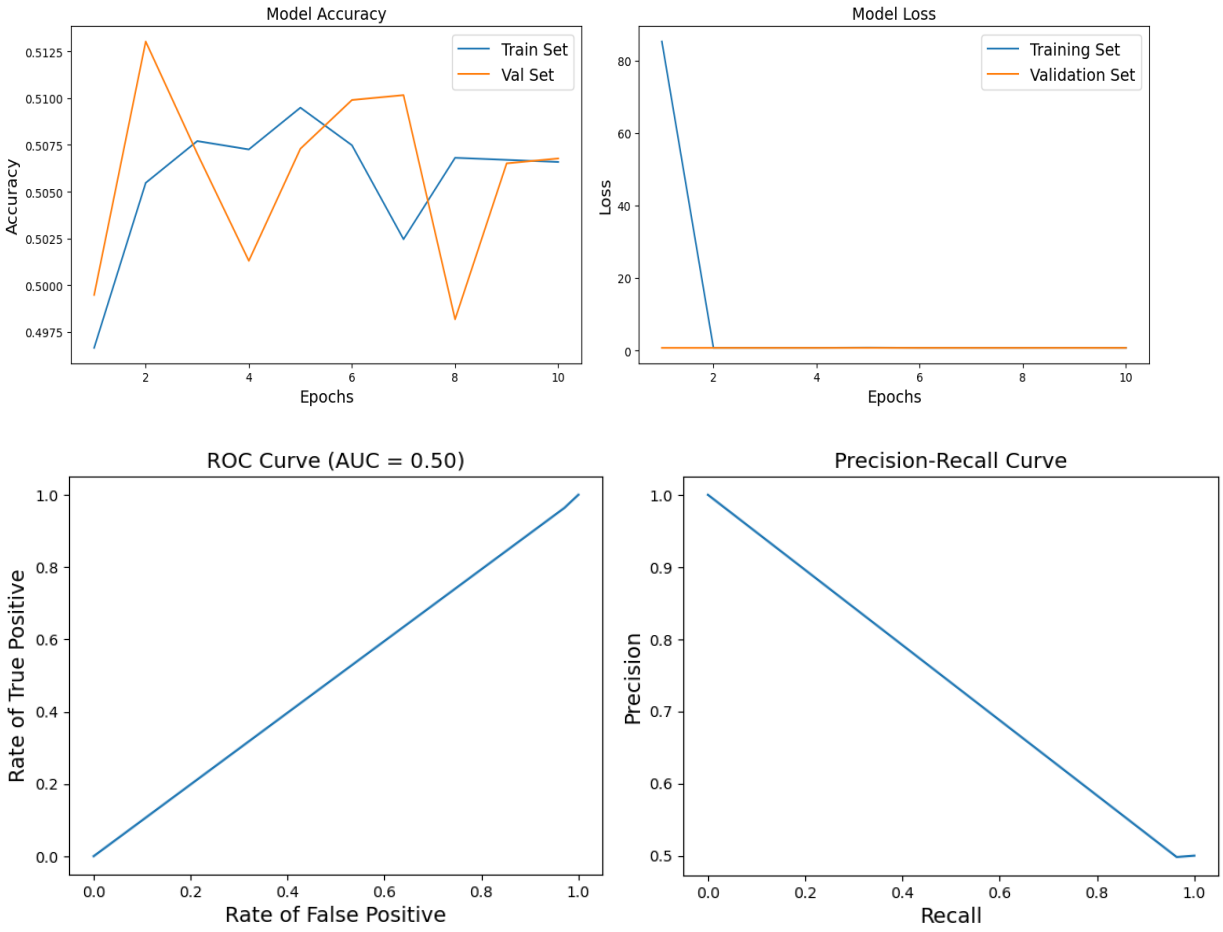

Fig. 7 CNN Model Evaluation results

5.2.2 Pretrained Xception for Finetuning

Being a sophisticated pretrained model, Xception model is used for fine-tuning which involves the adjustment of its weights to make it compatible for the task of deepfake detection. Data augmentation is performed in order to have all the images in the dataset to be of the same size for batching.

The '`preprocess_input()`' function is used for the same purpose making it prepared for the model by shuffling and prefetching the images in the training set. The top part of the Xception model is fine tuned to make it perform ideally as the weights of the model's top layers are adjusted in such a manner that it fits our task better.

The pretrained model is first loaded without the inclusion of its top layers and replacing it with the task that is required to be done. Once it is loaded, the model is trained for a few epochs while keeping the weights of the base model fixed. The fine tuning of the top layer makes the accuracy of this model to hike up to 63%. The following step undertaken is to make the top layers of the base model trainable again with a rather lower learning rate of 0.05. This makes the accuracy of the final model increase up to 84.6%.

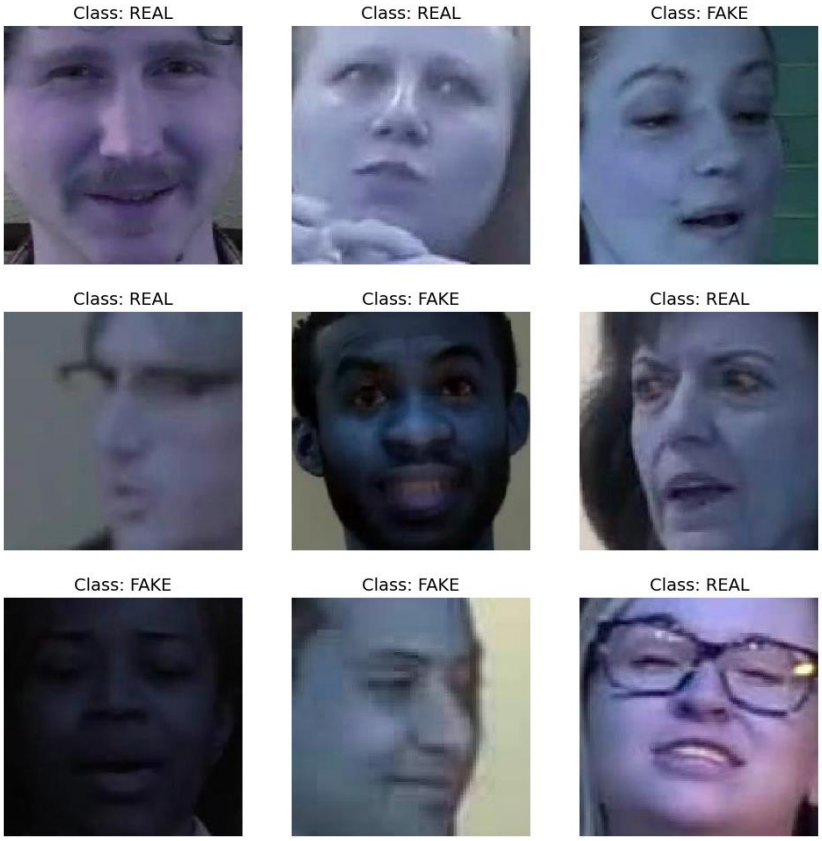

Fig. 8 Data augmentation results

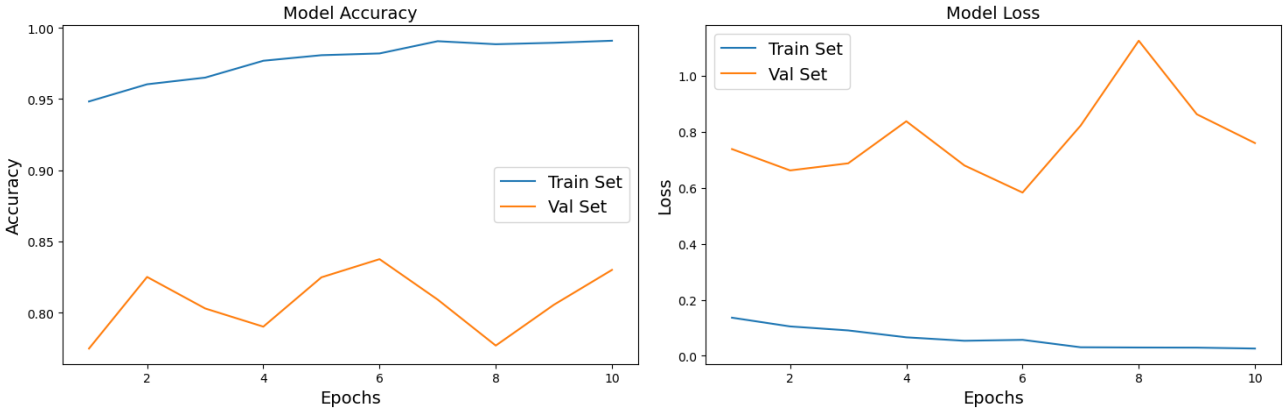

Fig. 9 Model Evaluation using pretrained Xception model

5.2.3 Pretrained VGG16 for Finetuning

The same ideology as that of Xception model is used here except for the fact that we utilise the pretrained VGG16 model for finetuning. VGG16 has been trained on vast datasets like ImageNet, contributing to its effectiveness in learning intricate patterns in the visual content. Finetuning of the pretrained model is done by altering the weights of the top layer to make it work on a specific task at hand extrinsically.

Data augmentation is carried out to ensure that the data is stable across the platform and produces uniform results. The top layers of the model is replaced with the specific detection task that is required for this study, and then the model is trained for 10 epochs with 280 cycles in each epoch. The accuracy of the model at this stage is 64%. The base model's top layer is made trainable again with lower learning rate of 0.05, making the accuracy 63%.

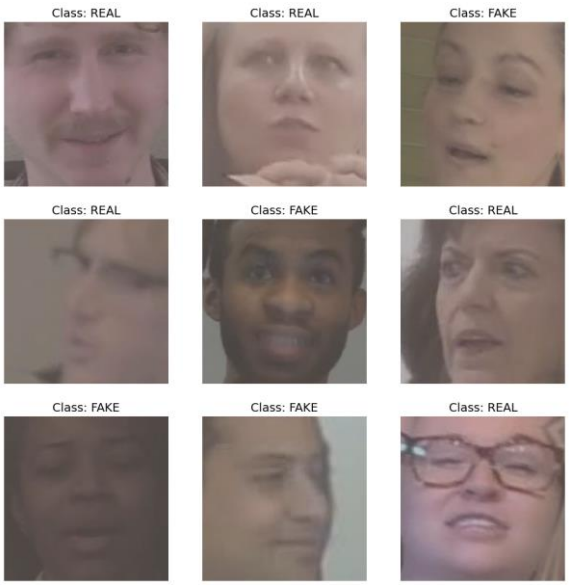

Fig. 10 Data augmentation results for VGG16 training model

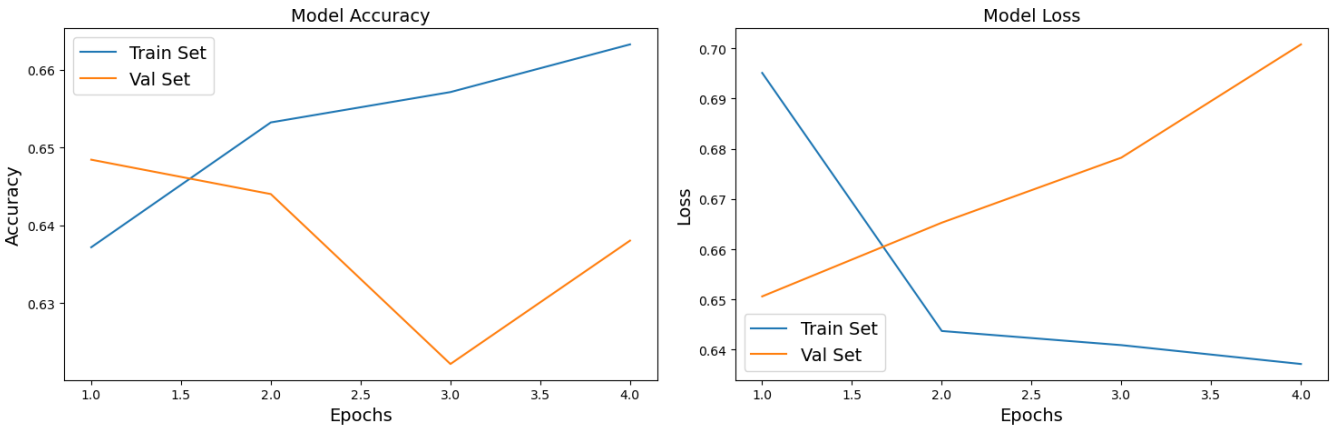

Fig. 11 Data visualisation results for VGG16 training model

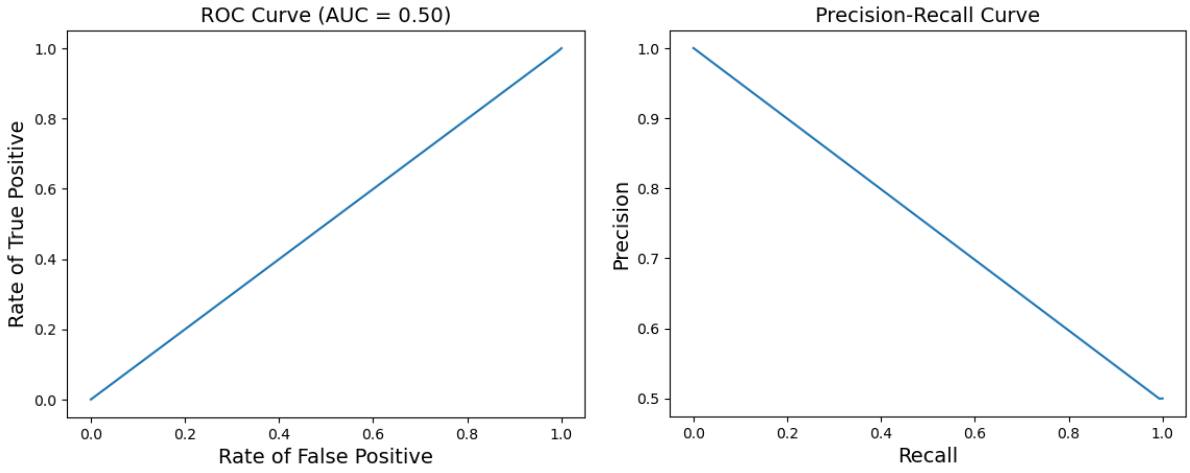

Fig. 12 Data visualisation results – ROC and Precision-Recall Curve

5.3 MODEL COMPARISON

The study reveals that the performance of the models using CNN with pretrained architecture like Xception and VGG16 indicate differences in identifying relevant features between real and fake images.

Through continuous training and finetuning, it is observed that using the Xception pretrained model generates far better model accuracy results owing to the depthwise separable convolutional layers, contributing to its ability to effectively distinguish between manipulated and real content. The model's accuracy is noted to be 84.6% for Xception model, whereas for VGG16, the accuracy is only 64%. This shows that the model can work efficiently by detecting deepfakes with a possibility of 84.6% which is a good measure.

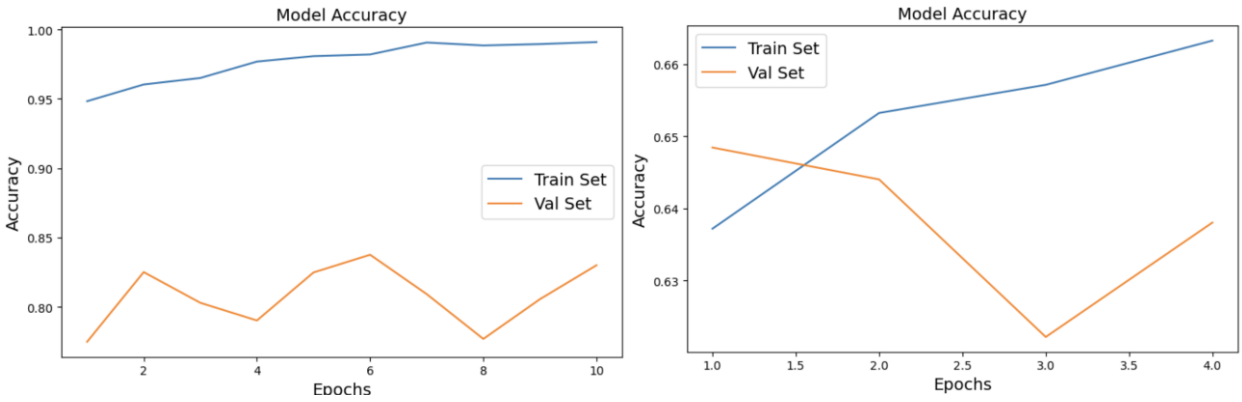

Fig. 13 Xception Model Accuracy vs. VGG16 Model Accuracy

```
100/100 [==============================] - 144s 1s/step - loss: 0.6804 - accuracy: 0.8462

[0.6803543567657471, 0.8462499976158142]
```

(a)

	precision	recall	f1-score	support
0	0.60	0.72	0.65	1600
1	0.65	0.51	0.57	1600
accuracy			0.62	3200
macro avg	0.62	0.62	0.61	3200
weighted avg	0.62	0.62	0.61	3200

(b)

Fig. 14 (a) Xception Model Evaluation Metrics vs. (b) VGG16 Model Evaluation Metrics

6. CONCLUSION and FUTURE WORK

To conclude this study, the detection of deepfake content is of utmost importance at this day and age as it can be the centre of dilemma for several individuals' reputation and has the potential to sabotage their privacy.

The study meticulously explores the design implementation, and evaluation of a deepfake detection system, utilizing advanced convolutional neural networks (CNN) with Xception and VGG16 pretrained models. The highlighted features of the proposed architecture include its robustness and adaptability with an emphasis on the importance of transfer learning, data augmentation and model fine tuning in achieving high accuracy in identifying real from manipulated content.

The foundational strength of the model, incorporating hierarchical and spatial features through convolutional filters and pooling layers, was useful in efficient detection and classification. The integration of CNN layers on large datasets provided a strong foundation for intricate pattern detection.

The evaluation criteria focused on accuracy, precision, and recall values, with visual representation depicting learning dynamics. The design architecture aimed at adaptability, robustness, and efficiency, employing a deep learning approach with CNN as the base model and integrating pretrained models for enhanced feature extraction. Data augmentation improved model generalization.

Proper data loading and preprocessing structures were crucial, utilizing publicly available DFDC datasets and ensuring balanced subset creation. Data exploration visualisations demonstrated a balanced distribution of real and manipulated classes, crucial for unbiased model behaviour. Model implementation involved a baseline CNN, followed by fine-tuning with pretrained Xception and VGG16 models. Xception outperformed VGG16 significantly, achieving an accuracy of 84.6% compared to 64% accuracy of the VGG16 model.

Although this study proves to be of high effectiveness, there is always room for improvement. Divulging into accommodating advanced model architectures can be included, expanding the scope of the study. An inclusion of ensemble learning techniques can also provide better insights into making the model perform better. Another addition that could be made is to adopt real-time deployment considerations contributing to a better and resilient system that prevents such adversarial attacks. While the adoption of real-world applications can help improve the development of detection systems, it is also equally important to commit to the ethical standards.

ACKNOWLEDGEMENT

I would sincerely like to thank my supervisor, Shubham Subhnil, who helped me immensely throughout this research. With his technical expertise and insights, I was able to work my way through this study successfully.

REFERENCES

Aung, H., Bobkov, A. V. & Tun, N. L., 2021. *Face Detection in Real Time Live Video Using Yolo Algorithm Based on Vgg16 Convolutional Neural Network.* Sochi, Russia, IEEE, International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM).

Borgalli, R. A. & Surve, S., 2022. *Deep Learning Framework for Facial Emotion Recognition using CNN Architectures.* Tuticorin, India, IEEE, International Conference on Electronics and Renewable Systems (ICEARS).

Dagar, D. & Vishwakumar, D., 2022. A literature review and perspectives in deepfakes: generation, detection, and applications. *IEEE, Int J Multimed Info Retr 11,* pp. 219-289.

Das, A., K. S. A., V. & Sebastian, L., 2022. *A Survey on Deepfake Video Detection Techniques Using Deep Learning.* Kottayam, India, IEEE, Second International Conference on Next Generation Intelligent Systems (ICNGIS).

Gwo-Chuan, L., Zi-Yang, L. & Tsai-Wei, L., 2022. *Ensemble Algorithm of Convolution Neural Networks for Enhancing Facial Expression Recognition.* Hualien, Taiwan, IEEE 5th International Conference on Knowledge Innovation and Invention (ICKII), Hualien, Taiwan.

Harsh, V., Nitin, Y., Ayush, R. & Sushant, J., 2023. *Detecting Deepfake Human Face Images Using Transfer Learning: A Comparative Study.* Bangalore, India, IEEE International Conference on Contemporary Computing and Communications (InC4).

Ihsan, S. & AlAsady, T. A. A., 2022. *Deep fake Image Detection based on Modified minimized Xception Net and DenseNet.* Al-Najaf, Iraq, IEEE, 5th International Conference on Engineering Technology and its Applications (IICETA).

Ilyas, H., A., I., A., J. & K. M., M., 2022. *Deepfakes Examiner: An End-to-End Deep Learning Model for Deepfakes Videos Detection.* Lahore, Pakistan, IEEE, 16th International Conference on Open Source Systems and Technologies (ICOSST).

Jolly, V. et al., 2022. *CNN based Deep Learning model for Deepfake Detection.* Ravet, India, IEEE, 2nd Asian Conference on Innovation in Technology (ASIANCON).

Kommalapati, A., 2021. *Comparative study of state of the art deepfake detection models.* Dublin, Ireland, IEEE, School of Computing, National College of Ireland.

Kusniadi, I. & Setyanto, A., 2021. *Fake Video Detection using Modified XceptionNet.* Yogyakarta, Indonesia, 4th International Conference on Information and Communications Technology (ICOIACT).

Mira, F., 2023. Deep Learning Technique for Recognition of Deep Fake Videos. *IEEE IAS Global Conference on Emerging Technologies (GlobConET),* pp. 1-4.

Misra, O. & Singh, A., 2016. *An approach to face detection and alignment using hough transformation with convolution neural network.* Bareilly, India, IEEE,International Conference on Advances in *Computing*, Communication, & Automation (ICACCA) (Fall).

Muafy, M. B. I., Sthevanie, F. & Ramadhani, K. N., 2023. *Generated AI Face Detection using Xception Model.* Bandung, *Indonesia*, IEEE, International Conference on Data Science and Its Applications (ICoDSA).

Niteesh, K., Pranav, P., Vishal, N. & Geetha, V., 2021. *Deepfake Image Detection using CNNs and Transfer Learning.* Pune. India, *IEEE*, International Conference on Computing, Communication and Green Engineering (CCGE).

Perdana, A. B. & Prahara, A., 2019. *Face Recognition Using Light-Convolutional Neural Networks Based On Modified Vgg16 Model.* Medan, Indonesia, IEEE, International Conference of Computer Science and Information Technology (ICoSNIKOM).

Prasannavenkatesan, T. & Ghouse basha, N., 2023. *Deepfake Face Detection Using Deep InceptionNet Learning Algorithm, Bhopal*, India: IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS).

Rahman, A. et al., 2022. *Short And Low Resolution Deepfake Video Detection Using CNN.* Hyderabad, India, IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC).

R, S., G, S. & V, A., 2022. *Facial Emotion Recognition using Deep Learning Approach.* Pudukkottai, India, IEEE, International Conference on Automation, Computing and Renewable Systems (ICACRS).

Tao, J. et al., 2021. *Research on vgg16 convolutional neural network feature classification algorithm based on Transfer Learning.* Shanghai, China, IEEE, 2nd China International SAR Symposium (CISS).