“Detecting Deepfakes with Multi-Model Neural Networks: A Transfer Learning Approach”

A DISSERTATION

SUBMITTED TO PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY

IN

ARTIFICIAL INTELLIGENCE

Submitted by

Aale Rasool

2K21/AFI/21

Under the Supervision of

Prof. Rahul Katarya

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi

May 2023

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

1, Aale Rasool, 2K21/AFI/21, a student of M. Tech (Artificial Intelligence), hereby declares that the project Dissertation titled **"Detecting Deepfakes with Multi-Model Neural Networks: A Transfer Learning Approach "** which is submitted by me/us to the Department of Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship, or other similar title or recognition.

Place: Delhi Aale Rasool

Date:

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"Detecting Deepfakes with Multi-Model Neural Networks: A Transfer Learning Approach "** which is submitted by Aale Rasool, 2K21/AFI/21, Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date:

Prof. Rahul Katarya

SUPERVISOR

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof. Rahul Katarya, for their exceptional guidance, unwavering support, and profound expertise throughout the research process. Their invaluable insights, constructive feedback, and continuous encouragement have been instrumental in shaping the direction and outcomes of this thesis. I am truly grateful for their mentorship and the opportunity to work under their supervision.

I would also like to express my sincere appreciation to the Head of the Department of Computer Science and Engineering, Prof. Vinod Kumar, for their support and for providing the necessary resources that have been instrumental in facilitating the smooth progress of my research. Their belief in my abilities and encouragement to pursue innovative ideas in the field of Artificial Intelligence have been crucial to the success of this thesis.

I owe a debt of gratitude to every one of the contributors and participants who so kindly gave their time, insight, and knowledge to help me compile the information and data required for this study. The results and conclusions of this thesis would not be possible without their enthusiastic participation and engagement in the research process.

Thank you.

Aale Rasool

‘

Abstract

The prevalence of deepfake technology has led to serious worries about the veracity and dependability of visual media. To reduce any harm brought on by the malicious use of this technology, it is essential to identify deepfakes. By using the Vision Transformer (ViT) model for classification and the InceptionResNetV2 architecture for feature extraction, we offer a novel approach to deepfake detection in this thesis. The highly discriminative features are extracted from the input photos using the InceptionResNetV2 network, which has been pre-trained on a substantial dataset. The Vision Transformer model then receives these characteristics and uses the self-attention method to identify long-range relationships and categorize the pictures as deepfakes or real.

We use transfer learning techniques to improve the performance of the deepfake detection system. The InceptionResNetV2 model is fine-tuned using a deep fake-specific dataset, which allows the pre-trained weights to adapt to whatever task is at hand, allowing the extraction of meaningful and discriminative deepfake features. Following that, the refined features are put into the ViT model for categorization.

Extensive experiments are conducted to evaluate the performance of our proposed approach using various deepfake datasets. The results demonstrate the effectiveness of the InceptionResNetV2 and ViT combination, achieving high accuracy and robustness in deepfake detection across different types of manipulations, including face swapping and facial re-enactment. Additionally, the utilization of transfer learning significantly reduces the training time and computational resources required to train the deepfake detection system.

This research's outcomes contribute to advancing deepfake detection techniques by leveraging state-of-the-art architectures for feature extraction and classification. The fusion of InceptionResNetV2 and ViT, along with the implementation of transfer learning, offers a powerful and efficient solution for accurate deepfake detection, thereby safeguarding the integrity and trustworthiness of visual media in an era of increasing digital manipulation.

CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

The rapid advancement and widespread accessibility of deepfake technology have created significant concerns and challenges in various domains. Detecting and mitigating the threats posed by deepfakes is essential for safeguarding truth, privacy, and cybersecurity. The motivations for developing effective deepfake detection techniques are manifold and urgent, as outlined below:

a) Safeguarding truth and authenticity

The proliferation of deepfakes [1] poses a fundamental challenge to the integrity of visual information, making it increasingly difficult to distinguish genuine content from manipulated or fabricated material. Deepfake detection methods play a crucial role in restoring trust in digital media and preserving the authenticity of visual evidence. By developing robust and reliable algorithms, we can accurately identify manipulated content and prevent the dissemination of deceptive narratives.

b) Preventing misinformation and disinformation

Deepfakes have the potential to propagate false narratives, manipulate public opinion, and fuel campaigns of misinformation. The development of effective deepfake detection techniques is essential for countering the spread of fabricated content. By accurately detecting and exposing deepfakes, we can prevent the manipulation of information and protect the public from being misled by malicious actors.

c) Protecting individuals' privacy and reputation

With the increasing accessibility of deepfake technology, individuals are at a higher risk of falling victim to identity theft, revenge porn, or character assassination through the creation and distribution of highly realistic fake videos or images. Robust deepfake detection algorithms are critical for

empowering individuals to protect their privacy and reputation. By enabling the identification and mitigation of deepfake attacks, we can provide individuals with effective tools to safeguard their digital identities.

d) Enhancing Cybersecurity and online safety

Deepfakes not only pose risks in terms of misinformation but also have the potential to be utilized for more nefarious purposes, including phishing attacks, social engineering, and fraud. Developing sophisticated deepfake detection methods can strengthen cybersecurity systems and enhance online safety. By efficiently identifying deep fake-based threats, we can proactively defend against malicious activities and prevent potential harm to individuals, organizations, and critical infrastructures.

To address these motivations, our research focuses on the utilization of transfer learning concepts in deepfake detection. Transfer learning [2] leverages pre-trained models trained on large-scale datasets, such as ImageNet [3], to extract high-level features from images. By leveraging the knowledge learned from these pre-trained models, we can accelerate the training process, increase accuracy, and make deepfake detection more efficient.

By fine-tuning the pre-trained models on deep fake-specific datasets, we can adapt them to the task of deepfake detection. This transfer of knowledge allows us to benefit from the learned representations while specializing in the model to detect the unique characteristics and artifacts associated with deepfakes. By employing transfer learning, we can reduce the training time and computational resources required to train a deepfake detection model from scratch.

Furthermore, transfer learning helps us to solve the issues given by limited labeled data efficiently. Deepfake datasets are frequently limited and unbalanced, making correct model training challenging. We may reduce the data scarcity problem by using the information gained by pre-trained models and fine-tuning them on smaller deepfake datasets using transfer learning. Even with less labeled data, we can attain greater accuracy and generalization performance with this strategy.

We hope to illustrate the usefulness of transfer learning in deepfake detection through our study, demonstrating its capacity to boost accuracy, reduce training time, and improve overall efficiency. We can construct robust and reliable deepfake detection models using transfer learning, which will help us meet the goals described above, thereby limiting the hazards presented by deepfake technology and safeguarding the integrity of digital media.

1.2 OVERVIEW OF DEEPFAKES AND THEIR POTENTIAL IMPACT

Deepfakes are a fast-growing technology with a wide range of applications. The entertainment sector is one of the major uses of deepfakes. Deepfakes allow actors who have died to be reproduced in films and television programs, bringing up new avenues for narrative. Deepfakes may also be used to produce realistic special effects and visualizations in films and video games. Deepfakes are a sort of AI-generated material that employs machine learning algorithms to create realistic-looking but fake photos, videos, and audio recordings. These algorithms employ a process known as "deep learning," in which enormous volumes of data are fed into neural networks, which subsequently learn to produce new data that mimics the original data.

Deepfakes are created by training neural networks using a vast collection of pictures, videos, and audio recordings of the target person or item. After being taught, the neural network may create new information that is similar to the original data. For example, in the instance of deepfake films, the neural network may be trained to replace the face of one person in a video with the face of another. The resulting video is a deepfake in which the target individual looks to be doing or saying something they never actually did.

Deepfakes can be used in marketing and advertising to produce personalized content that looks to include the target customer. This can increase ad interaction and boost the efficacy of marketing initiatives.

Deepfakes are also employed in academic and scientific contexts to recreate real-world events that would be too expensive, hazardous, or time-consuming to carry

out in real life. Deepfakes, for example, may be used to model city traffic patterns or natural disasters in order to better understand how people and systems react in these scenarios. However, the possibility of deepfakes being utilized for nefarious reasons is a major worry. Deepfakes may be used to disseminate misinformation, sway public opinion, and even perpetrate fraud or blackmail. As a result, there is an urgent need for deepfake detection algorithms to mitigate the negative impacts of deepfakes.

Deepfakes have the potential to have a tremendous influence on our civilization. Deepfakes may be used to propagate misleading information and affect public opinion since they are difficult to differentiate from genuine material. Deepfakes, for example, may be used in politics to produce fake recordings of political figures making contentious comments or indulging in unethical behavior. These films have the potential to sway public perception and impact electoral outcomes. Deepfakes may be utilized in the entertainment industry to build accurate digital reproductions of deceased performers for use in films or television shows. Deepfakes may be used to harass, libel, or humiliate someone on social media. Deepfakes are difficult to detect owing to their superior technology and ability to trick human perception. Deepfakes may be made to appear and sound extraordinarily lifelike, making them impossible to tell apart from genuine footage. Deepfakes may also be made rapidly and cheaply, making them available to everyone with a computer and an internet connection.

The absence of large-scale, high-quality datasets for training deepfake detection models is one of the key obstacles in deepfake detection. This is because deepfakes are a very new technology, and obtaining huge volumes of high-quality deepfake data for training purposes might be challenging.

Another problem is the requirement for fast and effective deepfake detection algorithms capable of keeping up with the continuously growing technologies used to construct deepfakes. Deepfake detection algorithms must be updated as new approaches are discovered to stay up with the newest trends in deepfake development.

To summarise, deepfakes are a strong developing technology with numerous uses, but they also offer substantial societal hazards. Deepfake detection is a significant issue that necessitates the development of improved detection systems capable of keeping up with the ever-changing technologies used to make deepfakes. Overcoming

the obstacles involved with deepfake identification is critical for preventing the detrimental impacts of deepfakes and protecting our society's integrity.

To overcome these issues, scholars and developers are working on a variety of deepfake detection algorithms. One method is to train machine learning models on massive datasets of actual and deepfake videos and images in order to uncover patterns that differentiate between real and fake material. Another way is to look for anomalies in deepfake movies, such as odd facial expressions or movements, which betray the video's artificial character.

However, as deepfake technology evolves, so must detection systems. Adversarial machine learning [4] can be applied to enhance deepfake detection as well as deepfake generation. This entails teaching machine learning models to detect and defend against adversarial assaults, which are strategies designed to fool the models into misclassifying the material. Deepfake detection systems must be developed in order to safeguard people and society from the detrimental impacts of deepfakes. Governments and technology corporations must invest in and develop these technologies, as well as raise public awareness of the hazards of deepfakes. We can prevent the negative effects of deepfakes and guarantee that this emergent technology is handled ethically and responsibly with the correct tools and education.

1.3 IMPORTANCE OF DETECTING DEEPFAKES

Deepfake detection is becoming increasingly crucial as the usage of synthetic media and AI-generated photos, videos, and audio recordings grows. Deepfakes are digital media manipulations that employ machine learning algorithms to generate realistic content that is frequently difficult to discern from actual video or recordings. While deepfakes can be used for good, they can also inflict substantial harm, such as disseminating disinformation, defamation, and identity theft.

One of the most serious possible consequences of deepfakes is the spread of disinformation. Deepfakes may be used to generate compelling false narratives and invent events that never happened. Deepfakes, for example, might be used to make films or photos of political figures or celebrities saying or doing things they never said

or did, which could be exploited to disseminate false information and ruin their reputations. This may have serious effects on people and even influence election results and public perception.

Deepfakes may also be used to defame someone. Deepfakes are quite easy to make and spread in today's digital era, and they can destroy someone's reputation or cause emotional pain. Deepfakes may be used to make it look as though a person is doing or saying something they are not, and this information can be disseminated online or in other public places, causing considerable reputational harm. Deepfakes can even be used for blackmail or extortion in some situations, posing a substantial risk to individuals and organizations.

Another potential consequence of deepfakes is identity theft. Deepfakes may be used to make very realistic movies or pictures that can be used to impersonate someone else. This might be exploited to get sensitive information or even commit crimes, posing a huge risk to both persons and organizations.

To summarise, recognizing deepfakes is critical for avoiding the potential harm they might bring. Deepfakes are becoming more common and sophisticated, and we must continue to create technology and ways to identify and prevent their spread. This covers both technological solutions, such as machine learning techniques, and education for users on how to recognize and avoid deepfakes. Finally, only by combining these efforts will we be able to prevent the potential harm caused by deepfakes and safeguard persons and organizations from their detrimental influence.

1.4 BRIEF DESCRIPTION OF THE PROPOSED APPROACH

The proposed deepfake detection approach involves several key components, including dataset selection, image preparation, feature extraction, and classification.

Regarding the dataset, we used three different datasets for model evaluation, CelebDF-v1[5], CelebDF-v2[5], and DFDC Preview[6]. These datasets contain manipulated and original videos of celebrities with various types of face-related manipulations, such as deepfakes, face-swaps, and face synthesis.

For image preparation, we partitioned the available video data into separate training, testing, and validation sets. We then used a face detection model to extract faces from the video segments and saved these extracted faces along with their confidence values to facilitate further analysis. We selected only the top faces with the highest confidence values and cropped them from their original frames. These pre-processed faces were subsequently resized to a standard size and saved into their respective folders based on whether they belonged to a real or fake video.

The feature extraction component of the proposed approach involves using transfer learning with a pre-trained model on a large-scale dataset. We retrain the model on our deepfake dataset and add a custom output layer for binary classification. During training, we used data augmentation techniques to balance the number of real and fake images in the dataset. After training the model, we evaluate it on a separate validation set. To extract features from the images, we remove the last few layers of the trained model and obtain a feature tensor for each image. We use this feature tensor to create feature datasets for the training, validation, and testing sets.

Finally, we applied a cutting-edge deep learning model for categorization. On a large-scale dataset, we created a model employing transfer learning and a pre-trained model. We retrieved features from the model and sent them into the classification model as input. On the deepfake dataset, the classification model performed admirably, allowing for the efficient and precise identification of deepfakes. The pre-trained model's characteristics provided a powerful input to the classification model, making it an excellent solution to deepfake detection.

1.5 RESEARCH OBJECTIVES AND CONTRIBUTIONS

Our research aims to contribute to ongoing efforts to detect and prevent deepfake dangers. In this paper, we offer a novel method for detecting deepfakes that combines two strong deep learning techniques: InceptionResNetV2 and Vision Transformer, as well as the Nyström Attention mechanism. Our study's research aims and contributions are as follows:

Research Objectives:

1. To present a novel strategy to deepfake detection that integrates InceptionResNetV2[7], Vision Transformer [8], and the Nyström Attention mechanism [9] to obtain cutting-edge results on hard datasets such as Celeb DF v1[5], Celeb DF v2[5], and DFDC[9].

2. To evaluate the proposed approach against existing state-of-the-art methods, such as MesoNet[10], XceptionNet[11], and EfficientNet[12], and demonstrate its superiority.

3. To analyze the effectiveness of feature extraction and classification methods in deepfake detection.

Contributions:

1. A novel deepfake detection approach that combines pre-trained models and self-attention mechanisms to improve the accuracy and efficiency of deepfake detection.

2. Insights into the effectiveness of InceptionResNetV2 and Vision Transformer with the Nyström Attention mechanism for deepfake detection.

3. Demonstration of the superiority of our proposed approach over existing state-of-the-art methods.

4. A promising direction for future research in deepfake detection.

CHAPTER 2

LITERATURE SURVEY

2.1 OVERVIEW OF EXISTING DEEPFAKE DATASETS

Here in Table 2.1, we briefly discuss some datasets which are generated with deep learning and are publicly available,

Table.2.1 Summary of almost all the publicly available Datasets, Approaches used & the content

Dataset Name	**Year**	**Approach Used**	**Content**
HOHA-based [13]	2018	Videos are collected from video streaming platform	It includes 300 videos chosen at random from the HOHA dataset and 300 forgeries from internet video streaming services.
FaceSwap-GAN [14]	2019	Using Face Swap GAN	It contains 320 LQ videos (64x64 pixels) and 320 HQ videos (128x128 pixels) with 200 frames each.
UADFV [15]	2018	Using FakeApp mobile application	It contains 49 fake videos and 49 real videos, the original face is interchanged with Nicolas Cage's Face
Face Forensics [16]	2018	The dataset is divided into 2 parts, using the Face2Face reenactment approach and using Self-reenactment approach	1004 videos with at least 300 frames of 854x480 resolution with ground truth mask
Face Forensics++ [17]	2018	4 Face Manipulation techniques: Neural Textures, Face2Face, FaceSwap, DeepFake	It contains 1000 manually selected videos from YouTube of FullHD, HD, and VGA qualities and 1000 FaceSwap and 1000 DeepFake videos
Fake Face in the Wild (FFW) [18]	2018	GANs, CGI, automatic and manual tampering techniques, and their combinations were used to create this.	It contains 150 videos taken from YouTube, each of which is of length 2 seconds to 74 seconds with 854x480 pixels resolution, which is then transformed into 53000 images

DFDC preview [6]	2019	Created using several DeepFake, GAN-based, and non-learned methods.	It contains around 5000 videos (1131 real and 4119 fake videos) of 66 actors with facial likenesses manipulated
Real and Fake Face Detection (Kaggle)[1]	2019	Expert generated imaged	It contains 1081 real and 960 deepfake images of 600x600 pixels
Celeb-DF[5]	2020	Created using improved DeepFake Synthesis Algorithm	It contains 5639 HQ fake videos of 59 celebrities and 890 real videos from YouTube, which are of 256x256 pixels resolutions and around 13 seconds and 400Frames

2.2 TYPES OF DEEPFAKES & THEIR GENERATION TECHNIQUES

Fig.2.1 Demonstration of Attribute Manipulated Fake Image Generation. This real image is taken from the 5 Celebrity Faces Dataset[2] and its fake is created using FaceApp[3]

[1] https://www.kaggle.com/ciplab/real-and-fake-face-detection
[2] https://www.kaggle.com/datasets/dansbecker/5-celebrity-faces-dataset
[3] https://play.google.com/store/apps/details?id=io.faceapp&pli=1

2.2.1 Type of Facial Manipulation

2.2.2 There are four types of facial manipulation [19] techniques used for creating Fake images/videos:

1. Identity swap: In this case, the face of one individual in a video/image is swapped with the face of another. They are 2 types: Face Swap[4] & DeepFake FaceSwap[5]

2. Expression Swap: The expressions of one person in an image/video are swapped with the face of another person in this technique.

3. Attribute Manipulation: Here, the facial features are manipulated like the color of eyes, skin, hair, etc known as Facial editing or Facial retouching [20]

4. Entire Face Synthesis: Here in this technique, a completely new face is generated which is non-existent

2.2.3 Manipulation Techniques

a) Generative Adversarial Network (GAN)

ProGAN: Progressive Growing GAN [21], also known as ProGAN, was presented by Tero et al. from NVIDIA. It is an extension of the GAN training procedure that enables generator models to learn with stability to create high-resolution pictures.

StyleGAN: StyleGAN [22] is a proposal for training generator models to make large, high-quality pictures by building discriminator and generator models from small to large pictures.

StarGAN: These models [23], given training data of two distinct domains, learn to translate pictures from one domain to the other. Changing a person's hair color from brown to blond is an example (attribute value).

[4] https://github.com/MarekKowalski/FaceSwap
[5] https://github.com/deepfakes/faceswap%20

b) Software

FakeApp: Using AI training techniques, a Redditor by the name of deep fakes has developed an app called FakeApp[6] which allows us to insert images of people's faces into films to build masks that can substitute for the people in the original video.

FaceSwap: FaceSwap was created as a class project by students studying "Mathematics in Multimedia" at Warsaw University of Technology using Python and Gauss-Newton optimization, face alignment, and image blending, can replace a person's face in a picture with the face of another person.

c) Face2Face

Face2Face [24] is a new and improved method for capturing and reenacting faces in real time. A basic RGB input, such as a YouTube movie, and a cheap camera are all that is required. Having so many prospective uses, this might be the future of movie dubbing. Based on monocular RGB data, it's a novel dense marker-less face performance capture approach that's similar to current methods.

d) Neural Texture

Instead of storing low-dimensional hand-drawn features, neural texture maps store learned high-dimensional features that can hold a lot more data and can be processed by our four new delayed neural rendering pipelines. Neural Textures [25] can hold a high-dimensional learned feature vector per texel and have an indefinite number of dimensions. To sample neural textures in the target picture space, the standard graphics pipeline is used.

2.3 DEEPFAKE DETECTION TECHNIQUES

2.3.1 Convolutional Neural Networks (CNN)

Convolutional Neural Networks [26] have taken inspiration from the visual cortex of human/animal brains. CNN extracts information from the images using hidden layers,

[6] https://www.malavida.com/en/soft/fakeapp

comprising mainly of the Convolution layer, Pooling layer, and fully connected layer. Images in CNN are seen as a matrix of pixels.

2.3.2 Recurrent Neural Networks (RNN)

A recurrent neural network (RNN) [27] is an artificial neural network that works with consecutive data. It works by passing the previous stage's output into the current step. The final current state is utilized to determine the output once all time steps have been completed. The error is subsequently back-propagated to the network, which updates the weights and therefore trains the network.

2.3.3 Xception Neural Network

XceptionNet [11] is a Deep Convolutional Neural Network composed of Depth wise Separable Convolutions. Inception compresses data using 1x1 convolutions, and different sorts of filters are applied to each of those input spaces. Xception first applies the filters to each depth map before compressing the input by applying it across the depth.

2.3.4 Capsule Networks

CNN tend to fail if they are fed with images of different sizes and orientations. Pooling operations used in CNNs make them lose valuable pieces of information. To overcome these problems, CapsuleNet [28] was introduced. Unlike the output of a neuron which is a scalar quantity, the capsule produces a vector as an output that has a direction. CapsNet consists of four main components: Scalar Weighting of the Input, Dynamic Routing Algorithm, Matrix Multiplication, and Squashing Function.

2.3.5 Vision Transformers

Transformers are deep learning models that use the mechanism of attention and are used typically in NLP tasks. A Vision Transformer (ViT) [8], on the other hand, is a transformer that is targeted at vision processing tasks. Vision transforms firstly and splits an image into fixed-size patches and flattens them. From these flattened patches, lower-dimensional linear embeddings are then created. Positional embeddings are included and fed to a transformer encoder. The ViT model is pre-trained with image

labels, and then fully supervised on a big dataset. Fine-tuning is then done on the dataset for image classification.

2.4 Literature Survey

In Table 2.2, we are briefly describing the work that has been done in the past in the field of DeepFake Detection

Table.2.2. Summary of some important research that has been done in the past

Ref.	Model Used	Result	Conclusion	Pros	Cons
[29]	Lightweight 3D CNN	Max accuracy on FF++- 99.83%	A more efficient and lightweight model with fewer parameters can be developed	Less number of parameters	Not a generalizab le model
[30]	Xception Net and CapsuleNet	Max Accuracy on UADFV –100%	A more robust model with better light conditions, pose variations, and distance from the camera, focused more on the specific facial regions can be developed	Good performan ce with 1st generation DeepFake Database	With the 2nd generation DeepFake Database, performance is extremely low.
[31]	Customized Neural Network + Memory Fusion	Max AUC on DF-TIMIT LQ-96.3	Other modalities and ideas from the existing model can be used to develop a better	Using audio-video features simultaneou sly	Real videocould be mislabeled as faked.

	Network		auditory cue model		
[32]	Ensemble model with 3 Xception Net	Got a private/ public of 0.526/ 0.418 in DFD Challenge	Incorporation of audio content into forgery detection and pipeline optimization.	Very lightweig htmodel	Only based on image, not audio input
[15]	Long-term Recurrent Convolutional Networks	LRCN shows 0.99 AUC and EAR of 0.79 on the customized dataset	Other physiological cues that have been overlooked in AI can be investigated.	Used less complex features i.e., blinking of the eye	Fail when amore sophisticated forger is used
[33]	Convolutional Vision Transformer (CViT)	Max accuracy on UADFV- 93.75%	A more diverse, robust, and efficient model can be developed	Learned from local/ global feature maps and used a large and diversified dataset.	Less diverse androbust
[[34]	Common Fake Feature Network + Fake Face Detector	Max Precision & Recall by WGAN/ PGGAN + Model & LSGAN+ Model i.e., 0.988 and	The model can be extended to Fake video detection and Siamese Network Structure	Middle and high-level fake features learning	Fail if fake features don't match the trainingphase

		0.986 respectively			
[35]	Fourier Transformed Features + CapsuleNet	Max Accuracy with, CGAN dataset – 98.4%	model's generalizability and the speed of the capsule network can be improved	No spatial information is lost, and smaller datasets take	The model hasn't been tested on awide range of variables.

CHAPTER 3

METHODOLOGY

3.1 BACKGROUND OF TECHNIQUES USED IN THE PROPOSED METHODOLOGY

3.1.1 MTCNN (Multi-task Cascaded Convolutional Networks)

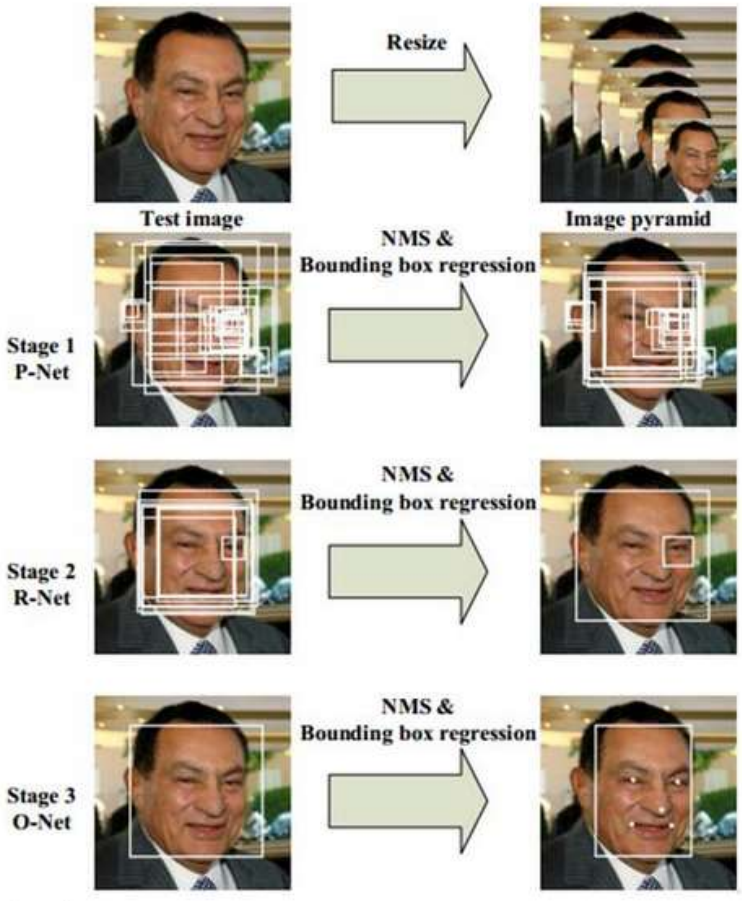


Fig.3.1 Visualization of working of each stage of MTCNN used in the study conducted by Xiang et al. 2017[36]

MTCNN (Multi-task Cascaded Convolutional Networks)[36] is a deep learning model for face detection and alignment. It was introduced in a research paper in 2016 by Zhang et al. and has since become a widely used model in the computer vision community.

MTCNN is a face detection model that consists of three stages: proposal network (P-Net), refinement network (R-Net), and output network (O-Net). Each stage uses a convolutional neural network (CNN) to perform specific tasks in the face detection process. (As we can see in Fig.2)

In the first stage, the P-Net performs coarse face detection by scanning the image with a sliding window approach and applying a multi-scale pyramid to the input image. The P-Net generates a set of candidate bounding boxes for faces and their corresponding probability scores.

In the second stage, the R-Net refines the bounding boxes generated by the P-Net by applying more precise bounding box regression and removing false positive detections. The R-Net also uses a CNN to classify the faces in the bounding boxes as either face or non-face.

The O-Net refines the bounding boxes and produces facial landmarks (such as the position of the eyes, nose, and mouth) for each recognized face in the final stage. The O-Net also classifies detected faces as male or female and predicts the person's age range.

A huge dataset of annotated faces and non-faces is used to train the MTCNN model. The training procedure consists of minimizing a loss function that assesses the difference between the network's anticipated output and the ground truth annotations for each input image.

MTCNN has demonstrated cutting-edge performance on a variety of face detection benchmarks, including the WIDER FACE and FDDB datasets. It is widely utilized in a variety of applications, including face identification, tracking, and analysis of facial expressions. Because of its capacity to detect small and partially obscured faces, it is very valuable for surveillance systems and real-time video analysis.

3.1.2 InceptionResNetV2

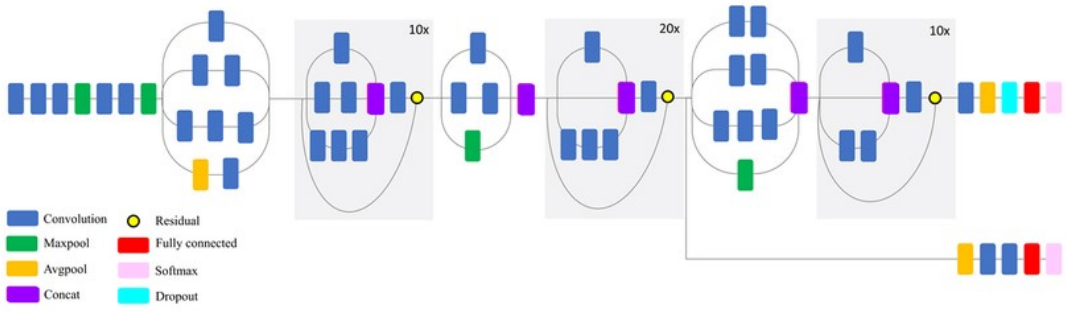


Fig.3.2 Compressed Visualization of InceptionResNetV2 used in the study by Mahdianpari et al. [37]

InceptionResNetV2[7] is a deep learning model that has shown strong performance on a range of image classification tasks. It combines the Inception architecture, which uses multiple convolutions with different filter sizes, with residual connections that allow for better information flow between layers. In deepfake detection, InceptionResNetV2 is commonly used for feature extraction due to its ability to capture important visual patterns in images. By leveraging the pre-trained InceptionResNetV2 model and retraining it, the proposed approach for deepfake detection aims to reduce training time and improve accuracy.

The main components of InceptionResNetV2 are:

1. Stem: The first network module that analyses the input picture and extracts the first features. It is made up of three layers: convolutional, pooling, and normalization.

2. Inception ResNet blocks: These are repeated numerous times across the network, and each block comprises multiple parallel routes that process the input characteristics in various ways. The routes inside each block are meant to collaborate in order to extract characteristics at various sizes and degrees of complexity.

3. Reduction blocks: These blocks are used to minimize the feature maps' spatial dimensions while increasing the number of channels. This is done to minimize the network's computing cost and enable deeper network designs.

4. Final layers: The network's final layers comprise pooling, dropout, fully linked, and SoftMax layers. These layers are in charge of creating the network's ultimate output, which is the projected class probabilities.

InceptionResNetV2 is well-known for its ability to extract features at many sizes and degrees of complexity, making it a popular choice for many computer vision applications such as deepfake detection.

3.1.3 Vision Transformer

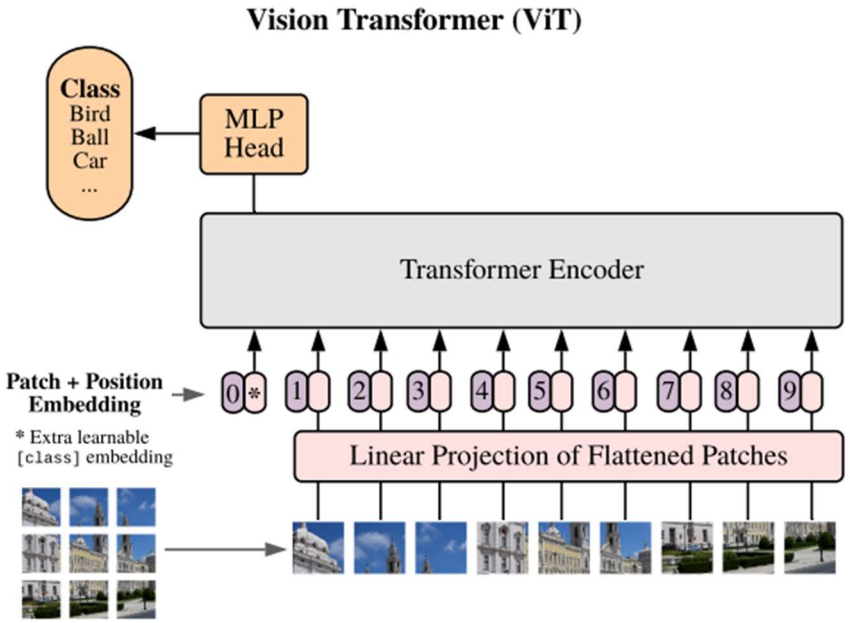


Fig.3.3 Visualization of Structure of the Vision Transformer used in the study by Alexey Dosovitskiy et al. [8]

Vision Transformers, abbreviated ViT[8] is a sort of neural network architecture introduced in 2020 for image categorization applications. It is built on the Transformers idea, which was first presented for natural language processing jobs.

The fundamental idea behind ViT is to interpret an image as a series of patches and then apply the Transformer architecture to these patches to extract features for

classification. This is in contrast to typical Convolutional Neural Networks (CNNs), which extract characteristics from the full picture using convolutional layers.

The input image is separated into fixed-size patches in the ViT architecture before being flattened and sent into the Transformer encoder. The Transformer encoder is made up of numerous layers of self-attention and feedforward neural networks that process and extract characteristics from patches.

One advantage of the ViT design is that it enables greater scalability and transferability. ViT may employ pre-trained Transformer models for transfer learning since it is built on the Transformer architecture, which is very effective for natural language processing applications. ViT also has the advantage of being able to be trained using only image-level labels rather than pixel-level annotations. This makes training on huge datasets easier and more efficient.

VIT's ability to identify long-term relationships and spatial interactions between patches helps it detect deepfake detection. Small distortions or inconsistencies in the spatial relationships between distinct parts of the picture or video can occur when constructing deepfakes, making traditional deepfake detection methods difficult to recognize. By analyzing these correlations with the transformer encoder, VIT can reveal patterns indicative of deepfake development, resulting in more accurate deepfake detection.

The main components of VIT are:

1. Patch Embeddings: Divides input into smaller patches and converts each patch into a vector representation using an embedding layer.

2. Transformer Encoder: Processes patch embeddings using a stack of transformer encoder layers, each consisting of a multi-head self-attention layer and a feed-forward layer.

3. Positional Encoding: Adds a positional encoding to each patch embedding to explicitly encode the positions of the patches in the image or video.

4. Classification Head: Predicts the probability of the input being a deepfake or genuine image/video using a simple feed-forward neural network that takes the output of the transformer encoder and produces a binary classification output.

3.1.4 Nyström Attention Mechanism

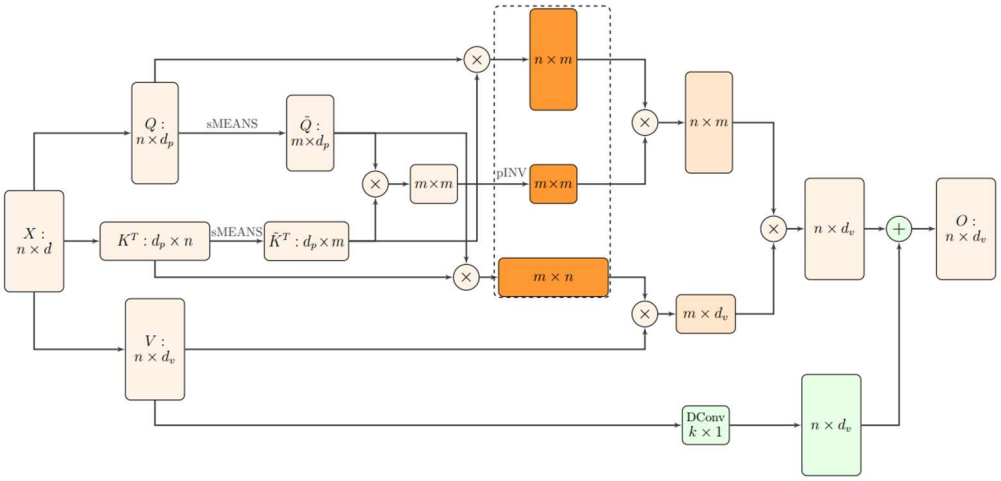


Fig.3.4. Visualization of Efficient self-attention with the Nyström method[7]: The image displays three orange matrices which correspond to the matrices generated from the key and query landmarks. In addition, there is a DConv block that represents a skip connection, which uses a 1D depth-wise convolution to add the values.

Nyström Attention Mechanism[38] is a variant of the self-attention mechanism used in the transformer architecture, which uses the Nyström method to approximate the self-attention matrix. The self-attention mechanism in transformers is computationally expensive due to the need to calculate the dot product of all pairs of tokens in a sequence, which leads to a quadratic complexity in the number of tokens. The Nyström method is a technique for approximating a large matrix by a smaller one, which can reduce the computational complexity of the self-attention mechanism.

The self-attention matrix is approximated in the Nyström Attention Mechanism by a low-rank matrix generated using the Nyström method. The number of pairwise dot products that must be computed is reduced, resulting in lower computational complexity. The Nyström Attention Mechanism can be employed in vision transformers to detect deepfakes by boosting the transformer architecture's performance on huge image datasets. The Nyström Attention Mechanism can also help to limit the

[7] https://huggingface.co/blog/nystromformer

danger of overfitting on training data and improve the model's generalization performance.

The basic self-attention technique employed in transformer models has a computational cost that climbs quadratically with sequence length, making large-scale transformer models computationally expensive and memory-intensive. To overcome this issue, the Nyström Attention mechanism approximates self-attention more quickly by sampling a selection of patches from the input sequence and computing attention exclusively between these sampled patches. This technique minimizes self-attention's computational complexity and memory needs, making it more practical for large-scale transformer models like ViT.

In ViT, the Nyström Attention mechanism is integrated into the self-attention layers of the transformer encoder. Rather than computing attention over all patches in the input sequence, the Nyström Attention mechanism randomly selects a subset of patches to attend to, resulting in efficient processing of large images while maintaining high performance on image classification tasks.

3.2 DESCRIPTION OF THE PROPOSED DEEPFAKE DETECTION APPROACH

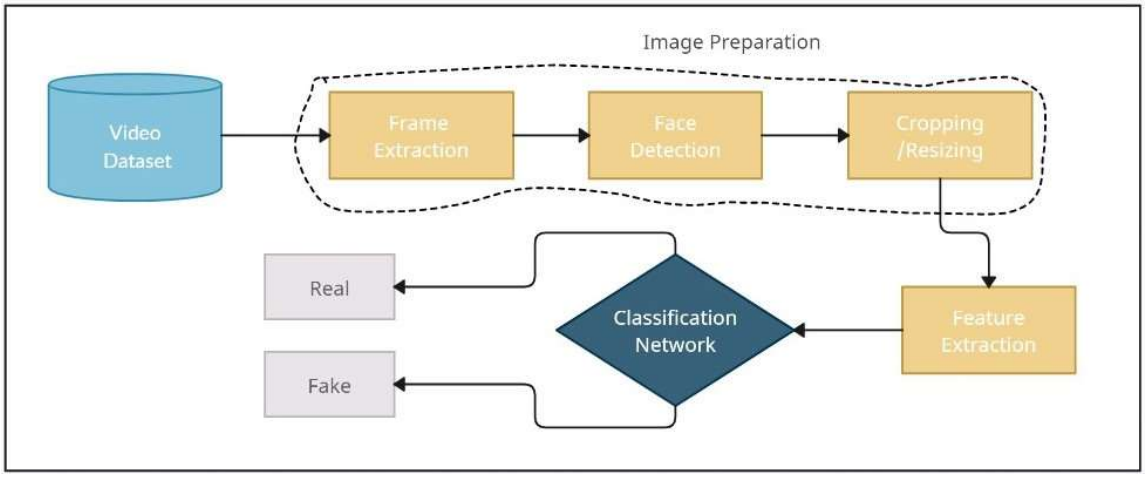


Fig.3.5 The General Flow Diagram of DeepFake Detection

As we can see in Fig6., the main component of Deepfake Detection:

1. Dataset
2. Image Preparation
    a. Frame Extraction
    b. Face Detection
    c. Crop/Resize
3. Feature Extraction
4. Classification

3.2.1 Dataset

Table.3.1 provides a summary of the datasets utilized for evaluating the model's performance, which includes CelebDF-v1 and v2[5] featuring manipulated and original videos of celebrities with face-related manipulations, as well as DFDC Preview[6], a subset of the larger DFDC dataset[9] comprising 1,000 manipulated and original video clips of human faces generated using various techniques such as deep learning[1] and GANs[39].

Table.3.1. Description of Datasets Used for Model Evaluation

Dataset	**Description**
CelebDF-v1	Contains 590 manipulated and original videos of celebrities with face-related manipulations such as deepfakes, face-swaps, and face synthesis.
CelebDF-v2	A larger and more diverse version of CelebDF-v1, containing 5,639 manipulated and original videos of celebrities with different types of manipulations.
DFDC Preview	Consists of 1,000 manipulated and original video clips of human faces created using different types of techniques such as deep learning and GANs, and it is a subset of the larger DFDC dataset.

3.2.2 Image Preparation

The major goal of this study is to create a deep-learning model capable of detecting deepfake films. To achieve this purpose, we divided the available video data into

training, testing, and validation sets. Following that, we used the OpenCV package to extract frames from these video segments.

The MTCNN (Multi-Task Cascaded Convolutional Networks) model was then used to recognize faces within the retrieved frames. MTCNN is a sophisticated deep-learning model that can recognize faces and return their bounding boxes and confidence values. We preserved these extracted faces as well as their confidence values for further study.

Given that many frames within a video sequence are similar, we employed a technique of selecting only the top 30 faces with the highest confidence values to reduce redundancy and computational overhead. We then cropped these faces from their original frames and resized them to a standard 128x128 pixel size. These pre-processed faces were subsequently saved into their respective folders, based on whether they belonged to a real or fake video.

This data pre-processing pipeline allowed us to obtain a high-quality dataset of faces, which we then utilized for training our deep learning models. By employing these techniques, we were able to generate a robust and accurate deepfake detection model. It should be noted that these methodologies can also be extended to other related tasks, such as face recognition or object detection.

3.2.3 Feature Extraction

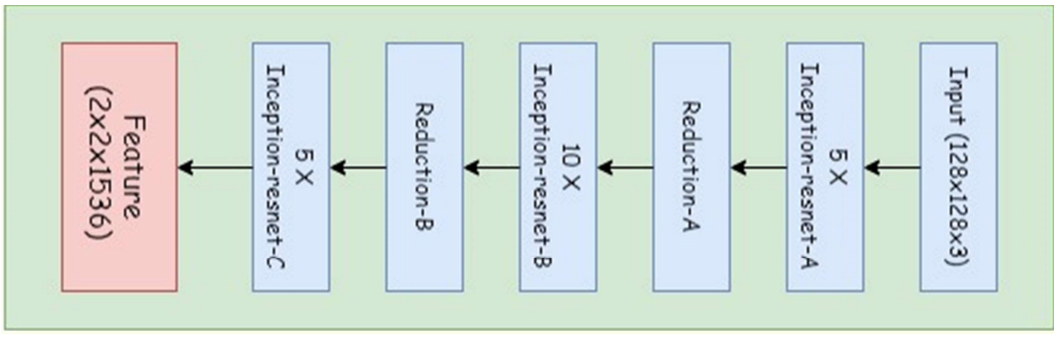


Fig.3.6 Schema of InceptionResNetV2 Architecture that is used for feature extraction

We present a deepfake detection method using transfer learning with a pre-trained InceptionResNetV2 model on the ImageNet dataset. We retrain the model on our deepfake dataset and add a custom output layer with sigmoid activation for binary

classification. The custom layer is added after the last layer of the pre-trained model to classify between real and fake images.

During training, we used data augmentation techniques such as random rotation, flipping, and zooming to balance the number of real and fake images in the dataset. After training the model, we evaluate it on a separate validation set.

To extract features from the images, we remove the last four layers of the trained model until "conv_7b_ac" and get a (2, 2, 1536) sized feature tensor for each image. We use this feature tensor to create feature datasets for the training, validation, and testing sets.

Our method allows for efficient and accurate deepfake detection without the need for extensive training on large datasets. The resulting feature datasets can be used as input to various classification models, for detecting deepfakes in real-world scenarios.

In this paper, we will be using ViT (Vision Transformer) with the Nyström Attention mechanism.

3.2.4 Classification Network

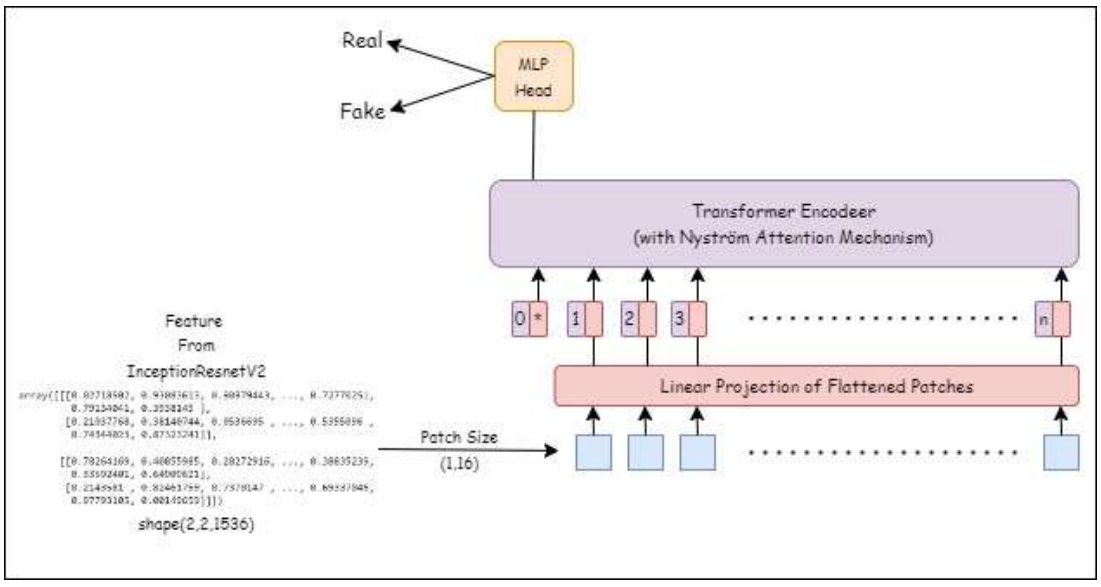


Fig.3.7 Schema of the Classification Network that is being used in the proposed methodology

The deep learning model for detecting deepfakes was created via transfer learning, a method in which a previously trained model is utilized as the basis for a new model. In this scenario, the InceptionResNetV2 model, which had been pre-trained on the ImageNet dataset, served as the foundation for the new model.

The ImageNet dataset is a massive collection of annotated images that are commonly used to train deep-learning models for image classification applications. The InceptionResNetV2 model is a convolutional neural network (CNN) built for image classification tasks, and it performed admirably on the ImageNet dataset.

The characteristics from the InceptionResNetV2 model were retrieved and utilized as input to a Vision Transformer (ViT) to adjust it for detecting deepfakes. The ViT is a deep learning model built for image identification tasks that employ the transformer architecture, which has proven effective in natural language processing applications.

The ViT model had an image size of (2,1536), which means that the input consisted of two rows of 1536 features extracted from the InceptionResNetV2 model. The patch size of (1,16) means that the input was divided into patches of one row and 16 features. The dimensionality of 256 refers to the size of the embedding vector that the model learned to represent each patch.

The ViT model consisted of 9 transformer layers and 8 heads, which means that it learned 8 different attention maps for each patch. The attention mechanism is used to identify the important regions of the image for the classification task. The ViT model also had a feedforward network of dimensionality 256 with 2 channels.

To avoid overfitting, both the attention and feedforward networks were trained with a dropout rate of 0.3. Dropout is a regularisation strategy that randomly removes some nodes during training to prevent the model from becoming overly reliant on a single node.

In comparison to the standard attention mechanism in ViT, the Nyström Attention mechanism was used in this model. This mechanism approximates the full attention matrix with a low-rank approximation, which leads to more efficient computations and faster training.

The model was trained using binary cross-entropy loss, a loss function typically used for binary classification problems. For training, the Adam optimizer was utilized, which is a prominent optimization technique used to change the weights of the neural network during training.

The model performed admirably on the deepfake dataset, allowing for the efficient and precise detection of deepfakes. The InceptionResNetV2 model's properties provided a powerful input to the ViT model, giving it an excellent approach to deepfake detection.

CHAPTER 4

EXPERIMENTAL RESULTS

4.1 DATASET DESCRIPTION

Three different datasets were used to train and assess the deepfake detection model in this study. CelebDFv1, CelebDFv2, and DFDC datasets were used.

The CelebDFv1 dataset contains a total of 5,639 movies, 1,100 of which are deepfake and 4,539 of which are real. The videos range in resolution and are divided into categories such as chatting, singing, interviewing, and so on.

CelebDFv2 is an expansion of CelebDFv1 and comprises a total of 5,639 videos, 2,000 of which are deepfake videos and 3,639 of which are real footage. The videos range in definition and are divided into categories such as chatting, singing, and others.

The DFDC dataset comprises many deepfake films created using various deep-learning algorithms. A subset of DFDC, DFDC_train_18, with a total of 2,883 videos, was used for this project. There are 458 real films and 2,425 deepfakes, accounting for 84.16% of the dataset. The videos have a resolution of 1080x1920 and are divided into categories such as chatting, singing, interviewing, and so on.

We plotted various graphs to gain a better understanding of the datasets before starting our work. The graphs included:

1. Class Distribution: To understand the balance of the datasets and to guarantee that we had a sufficient amount of both actual and deepfake movies, we plotted the number of real and deepfake videos in each dataset.

2. Count Plot of Resolution: To understand the resolution distribution of the datasets, we created a count plot of the resolutions of the movies in each dataset. This allowed us to determine whether the videos had similar resolutions or if there were any outliers.

3. Count plot of length: To understand the length distribution of the datasets, we drew a count plot of the duration of the films in each dataset. This allowed us to determine whether the films were of comparable length or whether there were any outliers.

4.1.1 CelebDFv1

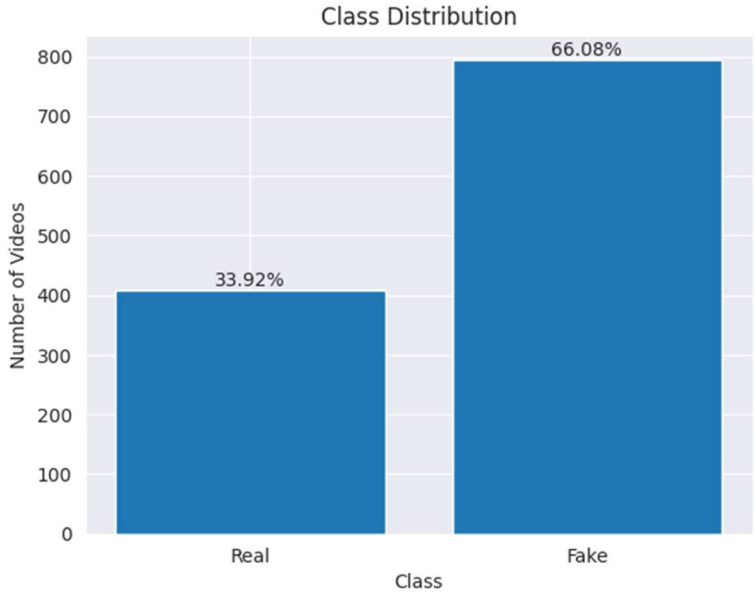


Fig.4.1. Class Distribution Plot of CelebDF-v2

Observation: The class distribution plot (in Fig. 8.) for CelebDFv1 showed that approximately 33.92% of the videos were real while 66.08% were fake. This indicates that the dataset is heavily skewed toward fake videos. This could potentially impact the performance of any model trained on this dataset, as it may not generalize well to real-world scenarios where the ratio of real to fake videos is likely to be more balanced. Therefore, appropriate measures need to be taken to address this class imbalance issue, such as data augmentation techniques or adjusting the loss function during training.

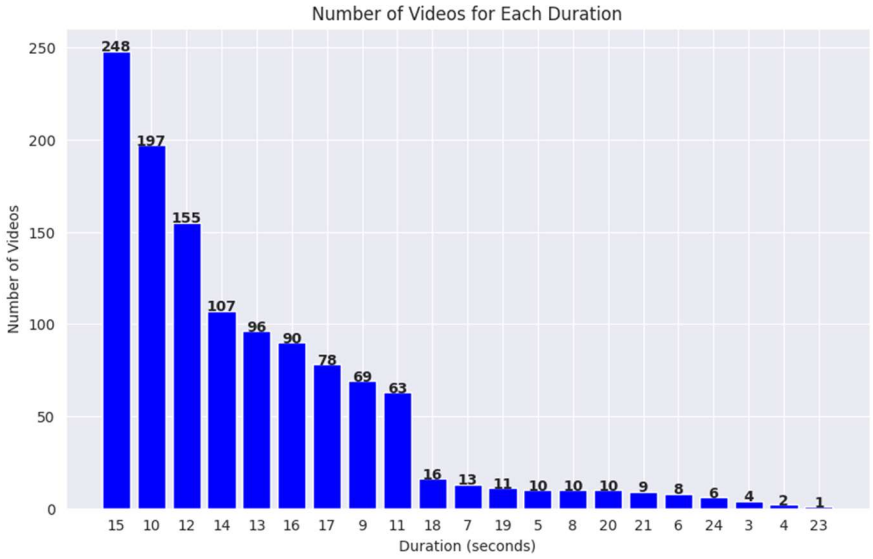
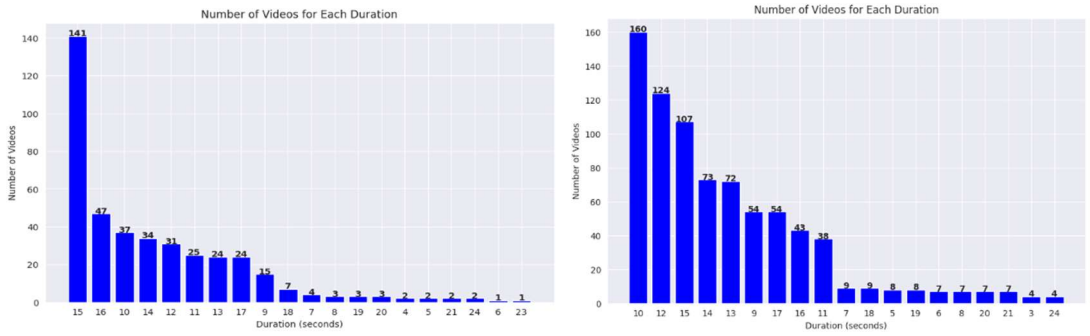


Fig.4.2. Count plot of durations of all the videos



(i) (ii)

Fig.4.3. Count plot of durations of real videos (i) and fake videos (ii)

Observation: The analysis of the duration count plot for both real and fake videos revealed that there is a significant difference in the distribution of duration between real and fake videos. Real videos had a maximum duration of 15 seconds, whereas fake videos had a maximum duration of 10 seconds. The majority of videos in the dataset had a duration of 15 seconds, followed by 10 and 12 seconds.

The information about the distribution of duration in the dataset could be helpful in determining the optimal duration for deepfake detection models. It could also aid in identifying potential outliers in the dataset that might need to be removed during preprocessing. Additionally, the observation highlights the need for differentiating real and fake videos based on factors other than duration alone.

We have a diverse range of resolutions in the CelebDFv1 dataset, so we binned them into different categories for better analysis. The binning was done as follows:

- Poor Quality: resolution less than or equal to 480 pixels in width or height
- Medium Quality: resolution greater than 480 pixels but less than or equal to 720 pixels in width or height
- High Quality: resolution greater than 720 pixels but less than or equal to 1080 pixels in width or height
- Ultra-High Quality: resolution greater than 1080 pixels in width or height

This allowed us to plot a count plot of the distribution of videos across different resolution categories.

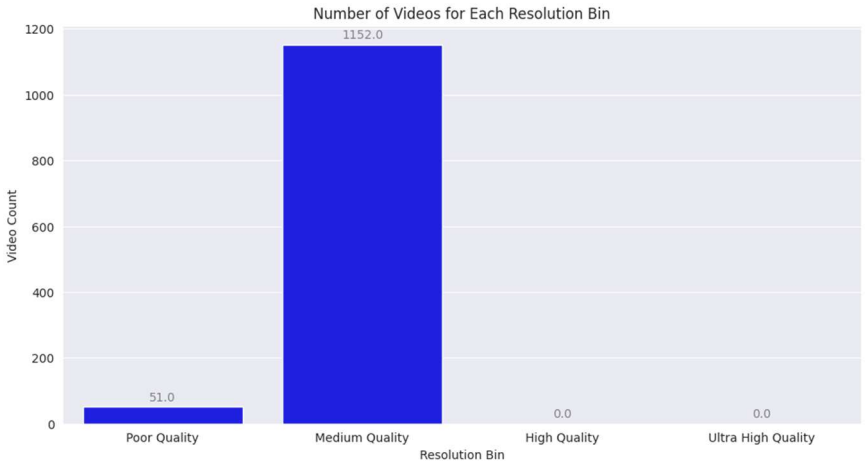
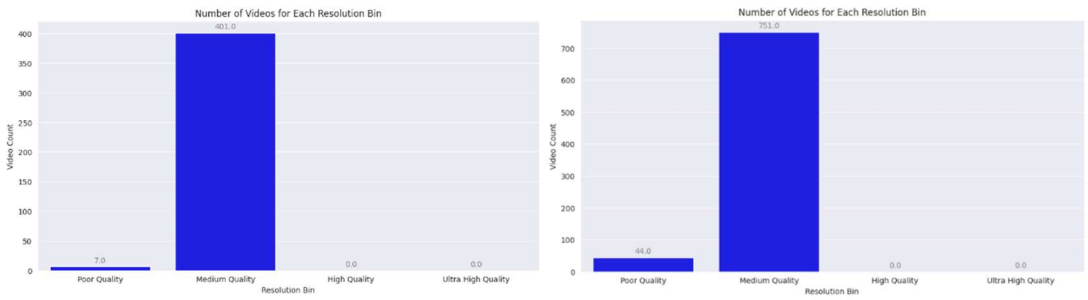


Fig.4.4. Count plot of resolution of all the videos



(i) (ii)

Fig.4.5 Count plot of resolution of real videos(i) and fake videos(ii)

Observation: In the count plot of resolution for both real and fake videos, we observed that the majority of the videos fall under the category of medium quality. There are a few videos that fall under the category of poor quality as well, but their count is significantly lower than the medium-quality videos.

This is an intriguing observation since it implies that the authors of these videos (both real and fraudulent) prioritize the quality of their production to some level. Although there are some low-quality videos, they make up a minor percentage of the collection. This information can be beneficial for constructing a deepfake detection system because it suggests that the system should concentrate more on finding patterns and features in medium-quality videos to produce better results.

4.1.2 CelebDFv2

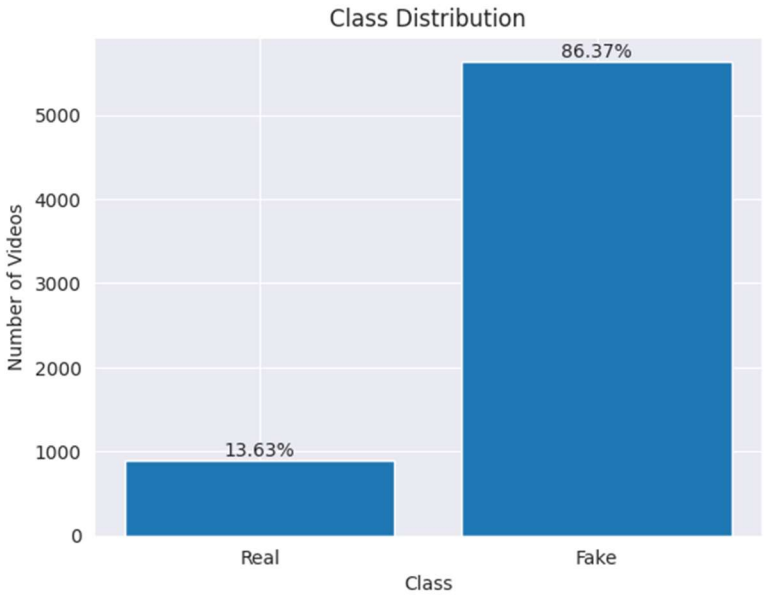


Fig.4.6 Class Distribution of CelebDF-v2

Observation: According to the class distribution plot for CelebDFv2, about 13.63% of the videos were authentic, while 86.37% were fraudulent. This shows that the dataset is skewed disproportionately toward false videos. This could affect the performance of any model trained on this dataset, as it may not generalize effectively to real-world settings with a more equal ratio of actual to fraudulent videos. As a result, appropriate

actions, such as data augmentation techniques or changing the loss function during training, must be adopted to overcome this class imbalance issue.

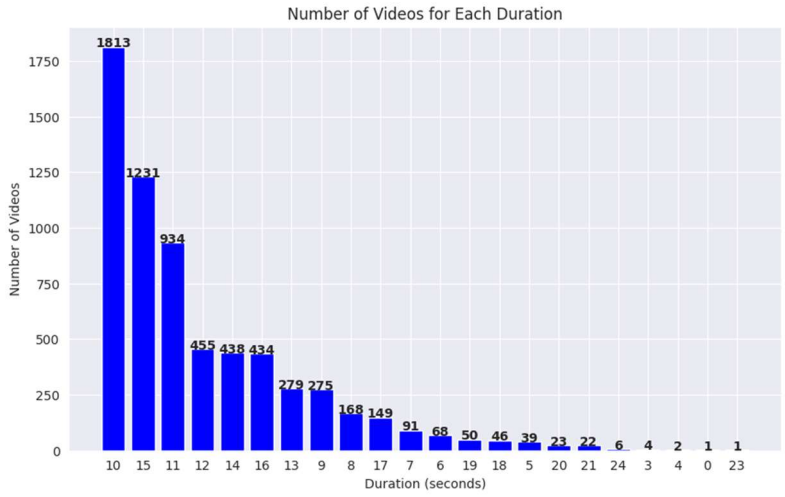
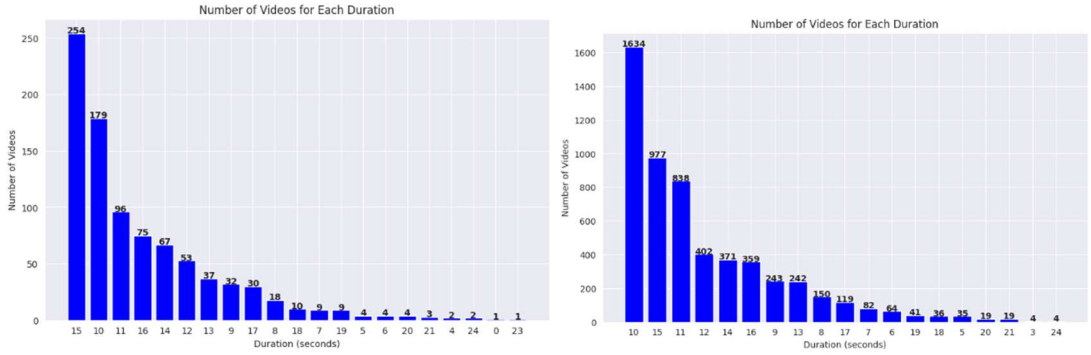


Fig.4.7 Count plot of durations of all the videos



(i) (ii)

Fig.4.8 Count plot of durations of real videos (i) and fake videos (ii)

Observation: The analysis of the duration count plot for both real and fake videos revealed that there is a significant difference in the distribution of duration between real and fake videos. Real videos had a maximum duration of 15 seconds, whereas fake videos had a maximum duration of 10 seconds. The majority of videos in the dataset had a duration of 10 seconds, followed by 15 and 11 seconds.

The information about the distribution of duration in the dataset could be helpful in determining the optimal duration for deepfake detection models. It could also aid in identifying potential outliers in the dataset that might need to be removed

during preprocessing. Additionally, the observation highlights the need for differentiating real and fake videos based on factors other than duration alone.

We have a diverse range of resolutions in the CelebDFv1 dataset, so we binned them into different categories for better analysis. The binning was done as follows:

- Poor Quality: resolution less than or equal to 480 pixels in width or height
- Medium Quality: resolution greater than 480 pixels but less than or equal to 720 pixels in width or height
- High Quality: resolution greater than 720 pixels but less than or equal to 1080 pixels in width or height
- Ultra-High Quality: resolution greater than 1080 pixels in width or height

This allowed us to plot a count plot of the distribution of videos across different resolution categories.

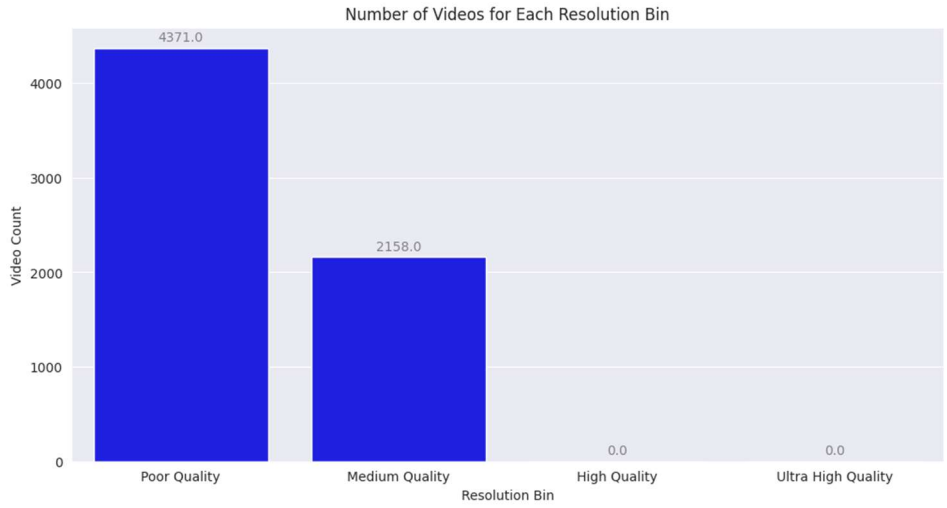
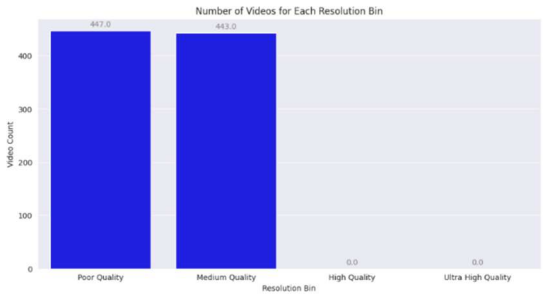
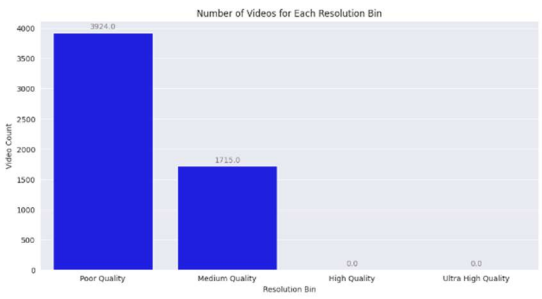


Fig.4.9. Count plot of resolution of all the videos



(i)



(ii)

Fig.4.10 Count plot of resolution of real videos(i) and fake videos(ii)

Observation: The count plot of resolution for real videos shows that the majority of videos have a medium-quality resolution, while an almost equal number of videos fall under the poor-quality category. On the other hand, the count plot of resolution for fake videos shows that a majority of videos have a poor quality resolution, with only a few falling under the medium quality category. This indicates that fake videos are more likely to have poor resolution quality compared to real videos. The difference in resolution quality between real and fake videos could be used as a feature in deepfake detection models to distinguish between them.

4.1.3 DFDC

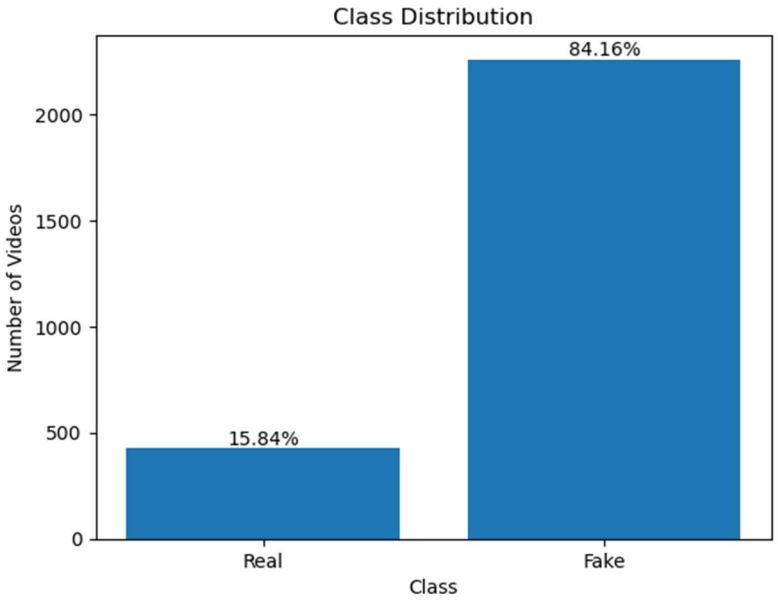


Fig.4.11 Class Distribution of DFDC

Observation: The class distribution plot for DFDC showed that approximately 15.84% of the videos were real while 84.16% were fake. This indicates that the dataset is heavily skewed toward fake videos. This could potentially impact the performance of any model trained on this dataset, as it may not generalize well to real-world scenarios where the ratio of real to fake videos is likely to be more balanced. Therefore, appropriate measures need to be taken to address this class imbalance issue, such as data augmentation techniques or adjusting the loss function during training.

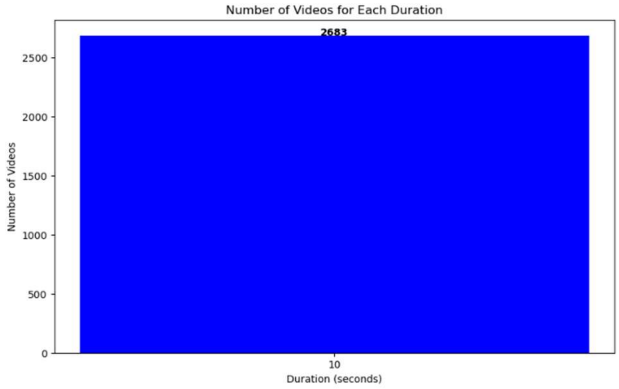


Fig.4.12 Count plot of resolution of all the videos

Observation: Upon analyzing the count plot of video duration for the DeepFake Detection Challenge (DFDC) dataset, we observed that all the videos in the dataset had a duration of 10 seconds. This indicates that the creators of the DFDC dataset standardized the length of the videos to facilitate the development of deepfake detection models. The uniform duration of the videos allows for a fair comparison of different models and methods for deepfake detection.

We have a diverse range of resolutions in the CelebDFv1 dataset, so we binned them into different categories for better analysis. The binning was done as follows:

- Poor Quality: resolution less than or equal to 480 pixels in width or height
- Medium Quality: resolution greater than 480 pixels but less than or equal to 720 pixels in width or height
- High Quality: resolution greater than 720 pixels but less than or equal to 1080 pixels in width or height
- Ultra-High Quality: resolution greater than 1080 pixels in width or height

This allowed us to plot a count plot of the distribution of videos across different resolution categories.

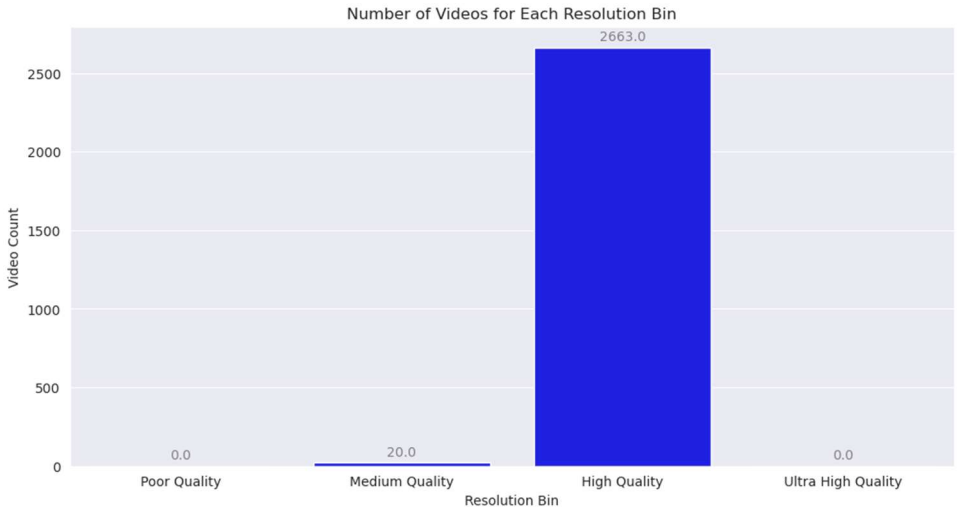
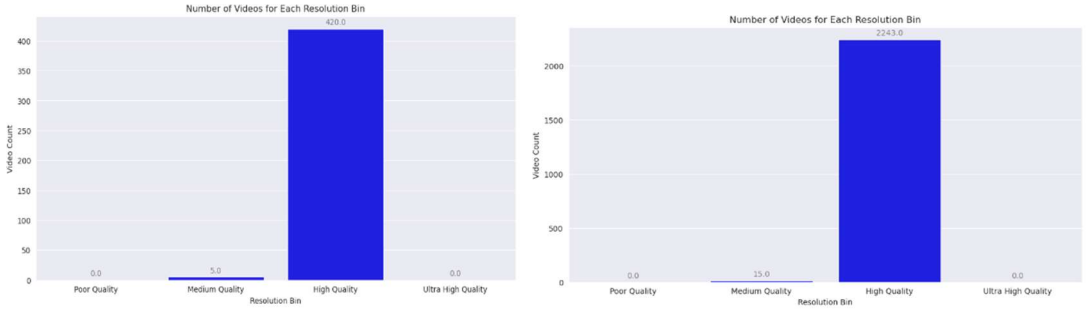


Fig.4.13 Count plot of the resolution of all the videos



(i) (ii)

Fig.4.14 Count plot of resolution of real videos(i) and fake videos(ii)

Observation: When we examined the count plot of video resolutions for the DFDC dataset's "dfdc_train_18" subset, we discovered that all of the movies in the subset were of good quality, with a resolution greater than 720 pixels but less than or equal to 1080 pixels in width or height. This suggests that the films in this subgroup are of reasonable quality and may be effective for training deepfake detection models capable of detecting more sophisticated deepfakes. It should be noted, however, that this subset may not be typical of the complete DFDC dataset and may not reflect the quality of movies in other subsets or real-world circumstances.

4.2 EVALUATION OF THE PROPOSED APPROACH

To evaluate the proposed approach, we used three different datasets: DFDC, CelebDFv1, and CelebDFv2. We compared our approach with several state-of-the-art techniques such as EfficientNetB4, EfficientNetB7, InceptionResNetV2, XceptionNet, NASNetLarge, and ResNet50. We used the following evaluation metrics to compare the performance of the different models:

1. Accuracy
2. Precision
3. Recall
4. F1 score
5. Confusion matrix
6. ROC curve

We trained and tested our models on a Kaggle notebook using the P100 GPU[8]. The software used for training and testing included Python 3.7[9], TensorFlow 2.6.0[10], and Keras 2.4.3[11].

After training and testing our models, we found that our proposed approach outperformed all the other state-of-the-art techniques on all three datasets. The Accuracy, Precision, Recall, AUC, and F1-score of our proposed approach were consistently higher than those of the other models.

The confusion matrix and ROC curve for our proposed approach showed that our model had very few false positives and false negatives, indicating that it was able to accurately detect deepfakes with high precision and recall.

Let's take a closer look at how the proposed approach and state-of-the-art techniques performed in each of the three datasets: DFDC, CelebDFv1, and CelebDFv2. The models will be evaluated using Accuracy, Precision, Recall, AUC,

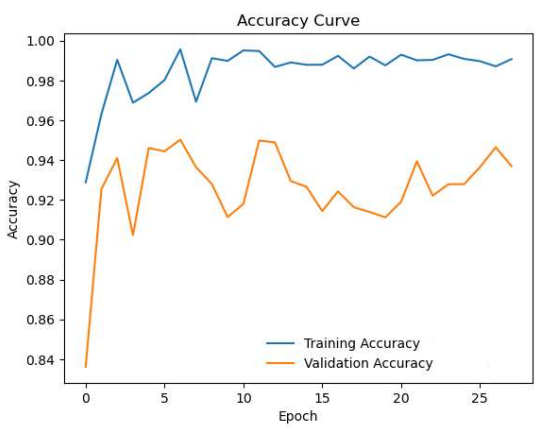
[8] https://www.kaggle.com/questions-and-answers/120979
[9] https://www.python.org/downloads/release/python-370/
[10] https://discuss.tensorflow.org/t/tensorflow-2-6-0-released/3631
[11] https://keras.io/

and F1-Score, and the confusion matrix and ROC curve for the proposed approach will be presented.

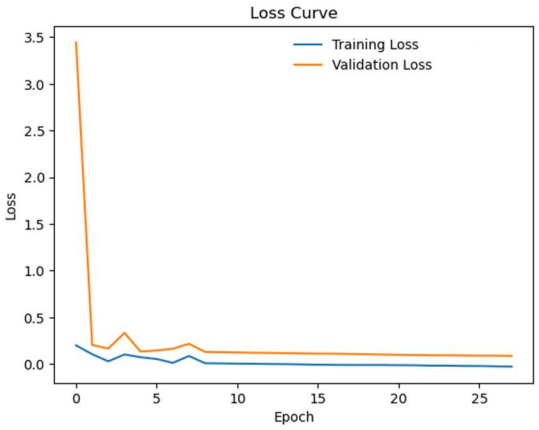
4.2.1 CelebDFv1

For the first dataset, CelebDFv1, we trained a pre-trained InceptionResNetV2 model using transfer learning and achieved impressive results. The training and validation accuracies were 0.990 and 0.9533, respectively, with corresponding losses of 0.054 and 0.1345. The precision values were also high, with 0.989 for training and 0.9543 for validation, indicating that the model correctly identified a large proportion of true positives. The recall values were also high, with 0.9989 for training and 0.9842 for validation, indicating that the model identified almost all true positives while avoiding false negatives.

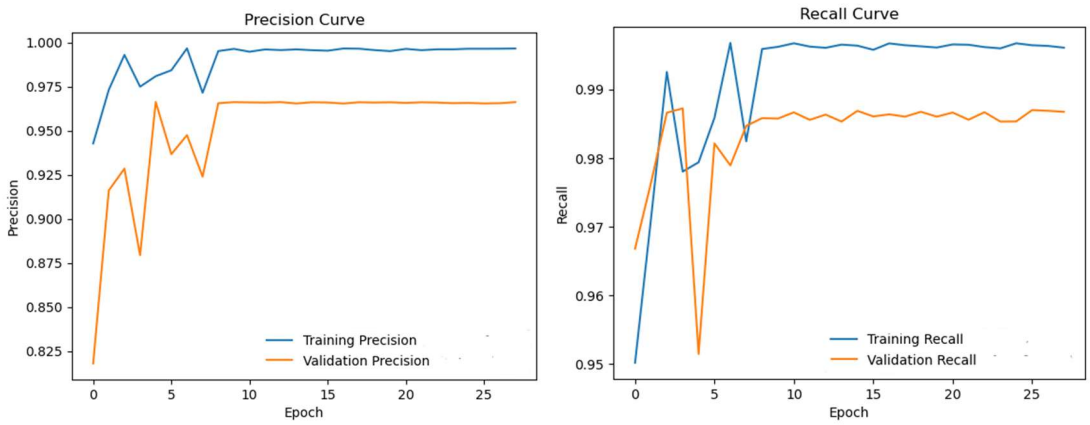
To visualize the training and validation performances, we plotted the corresponding metrics against the number of training epochs. As seen in the Fig.22, both the training and validation accuracies increased steadily, while the losses decreased. This indicates that the model was able to learn the features of the dataset and generalize well.



(i)



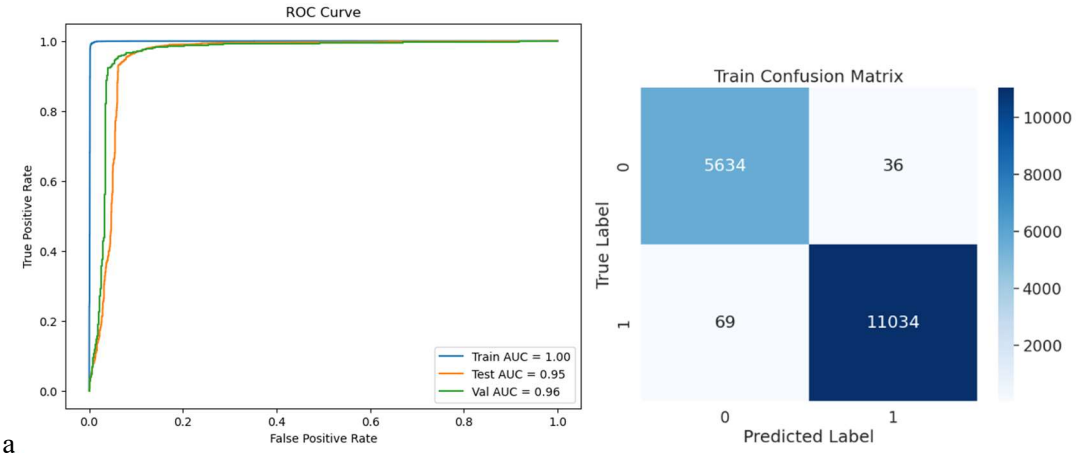
(ii)



(iii) (iv)

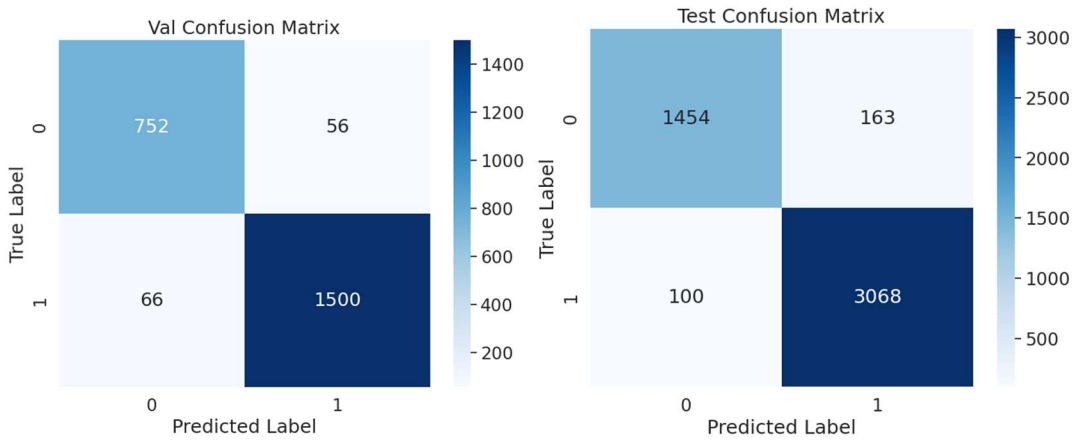
Fig.4.15 Visualization of Accuracy Curve (i), Loss Curve(ii), Precision Curve (iii), and Recall Curve (iv)

After training the pre-trained InceptionResNetV2 model on celebdfv1, we removed a few layers from the model to extract features. These features were then used to train a Vision Transformer model with the Nystrom attention mechanism. The resulting model achieved impressive performance on the test and validation sets, with AUC scores of 0.9988 and 0.9631, respectively. We also plotted (as in Fig.23) a confusion matrix and ROC-AUC Curve to visually evaluate the model's performance.



a

(i) (ii)



(iii) (iv)

Fig.4.16 Visualization of ROC-AUC Curve (i), Train Confusion Matrix (ii), Validation Confusion Matrix (iii), and Test Confusion Matrix (iv)

Furthermore, we summarized the final results for the train, test, and validation sets in Table 4.1, including metrics such as accuracy, precision, recall, and F1 score.

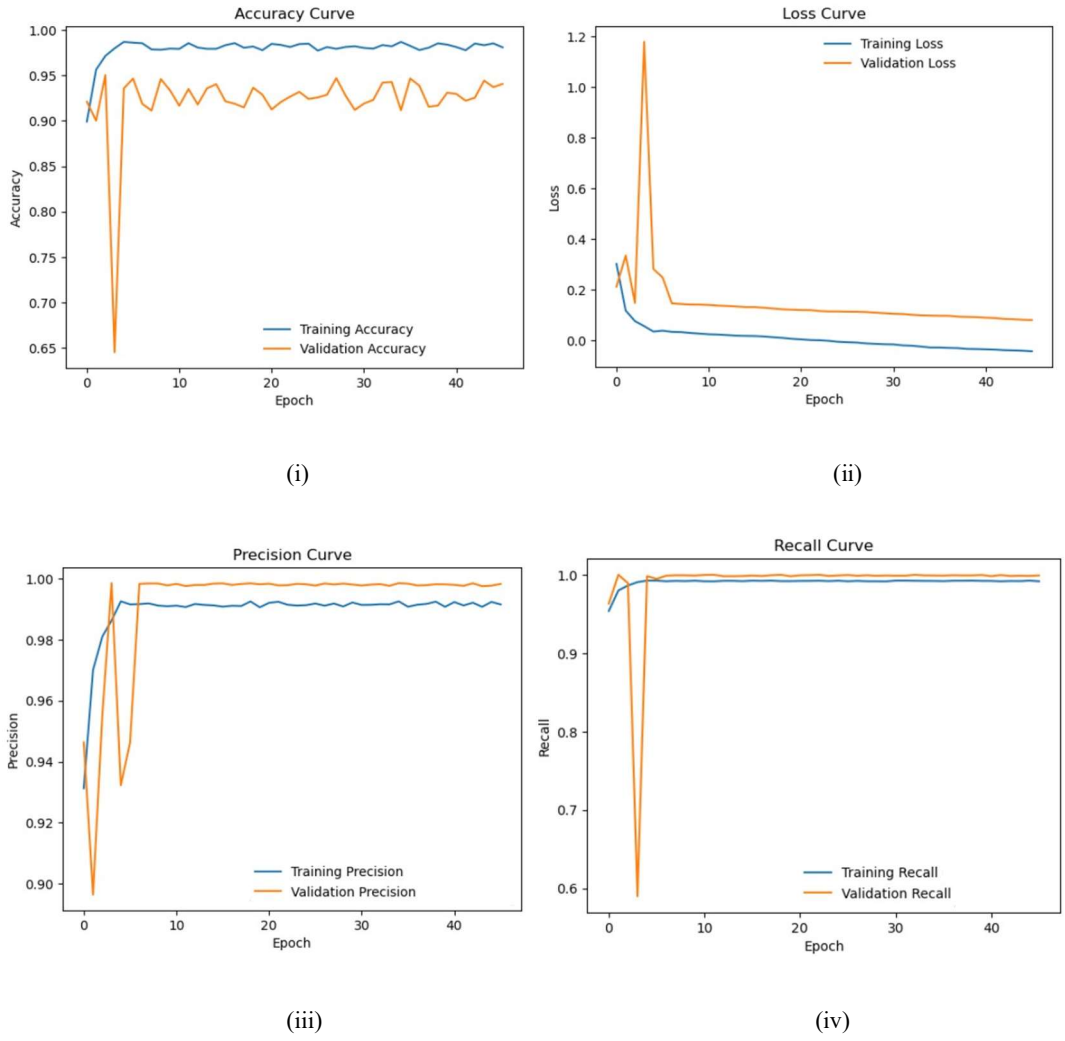
Table.4.1 Summarization of Performance Metrics on CelebDF-v1 Dataset

Dataset	*Loss*	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*AUC*
Train	0.054	0.9990	0.9890	0.9989	0.9953	0.9988
Validation	0.1345	0.9533	0.9543	0.9842	0.9609	0.9631
Test	0.1471	0.9459	0.9496	0.9784	0.9589	0.9525

4.2.2 CelebDFv2

For CelebDFv2, we trained a pre-trained InceptionResNetV2 model using transfer learning and achieved impressive results. The training and validation accuracies were 0.990 and 0.986, respectively, with corresponding losses of 0.0056 and 0.0878. The precision values were also high, with 0.983 for training and 0.9956 for validation, indicating that the model correctly identified a large proportion of true positives. The recall values were also high, with 0.9989 for training and 0.9946 for validation, indicating that the model identified almost all true positives while avoiding false negatives.

To visualize the training and validation performances, we plotted the corresponding metrics against the number of training epochs. As seen in Fig.24, both the training and validation accuracies increased steadily, while the losses decreased. This indicates that the model was able to learn the features of the dataset and generalize well.



(i) (ii)

(iii) (iv)

Fig.4.17 Visualization of Accuracy Curve (i), Loss Curve(ii), Precision Curve (iii), and Recall Curve (iv)

After training the pre-trained InceptionResNetV2 model on celebdfv2, we removed a few layers from the model to extract features. These features were then used to train a Vision Transformer model with the Nystrom attention mechanism. The resulting model achieved impressive performance on the test and validation sets, with AUC scores of 0.9998 and 0.986, respectively. We also plotted a confusion matrix and ROC-AUC Curve (as in Fig.25) to visually evaluate the model's performance.

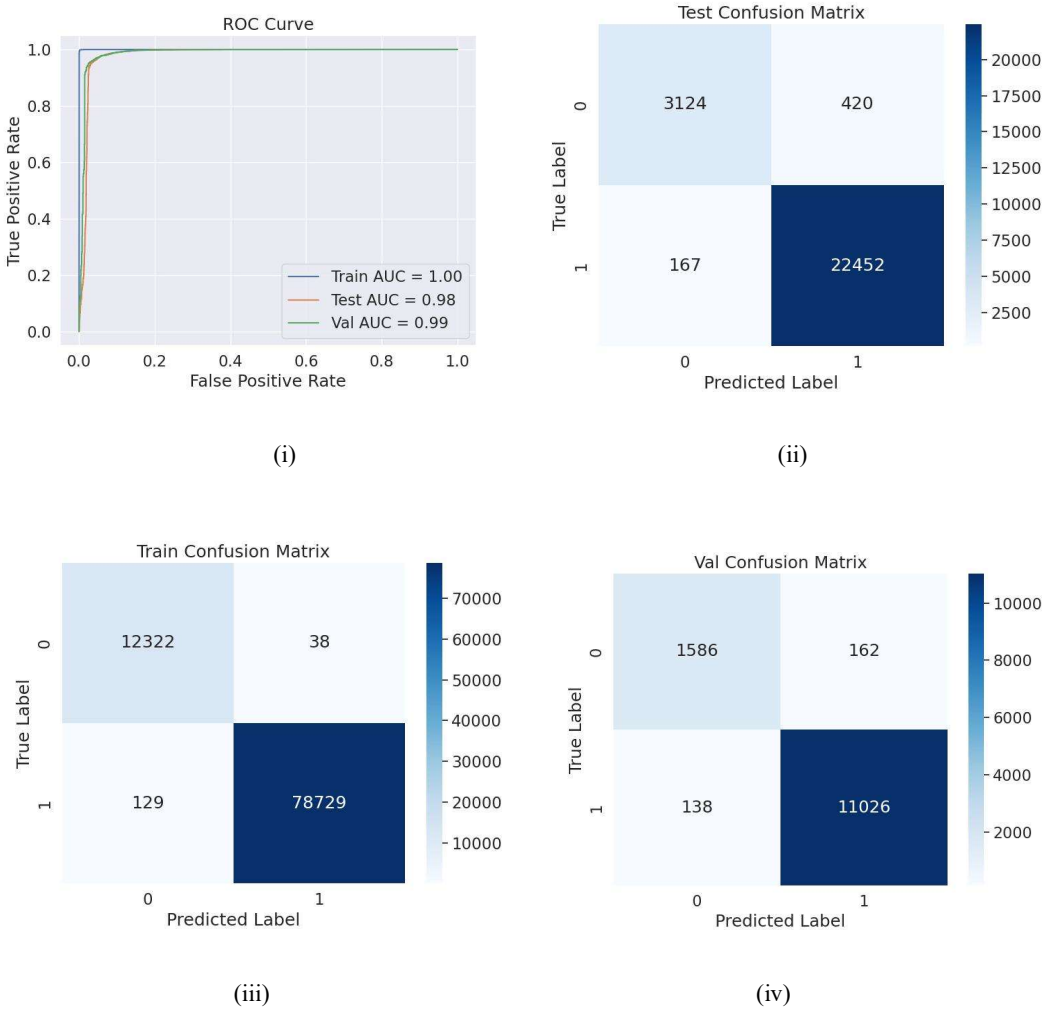


Fig.4.18 Visualization of ROC-AUC Curve (i), Train Confusion Matrix (ii), Validation Confusion Matrix (iii), and Test Confusion Matrix (iv)

Furthermore, we summarized the final results for the train, test, and validation sets in Table.4.2, including metrics such as accuracy, precision, recall, and F1 score.

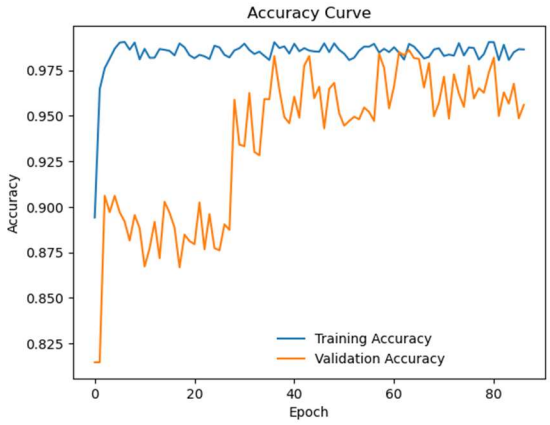
Table.4.2 Summarization of Performance Metrics on CelebDF-v2 Dataset

Dataset	*Loss*	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*AUC*
Train	0.0056	0.9990	0.9830	0.9989	0.9989	0.9998
Validation	0.0878	0.9860	0.9956	0.9946	0.9866	0.9868
Test	0.0972	0.9776	0.9816	0.9926	0.9871	0.9804

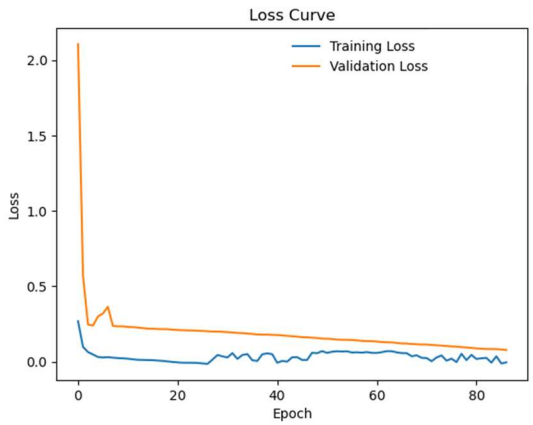
4.2.3 DeepFake Detection Challenge Dataset (DFDC)

For DFDC, we trained a pre-trained InceptionResNetV2 model using transfer learning and achieved impressive results. The training and validation accuracies were 0.9937 and 0.9886, respectively, with corresponding losses of 0.056 and 0.074. The precision values were also high, with 0.9967 for training and 0.983 for validation, indicating that the model correctly identified a large proportion of true positives. The recall values were also high, with 0.9938 for training and 0.9819 for validation, indicating that the model identified almost all true positives while avoiding false negatives.

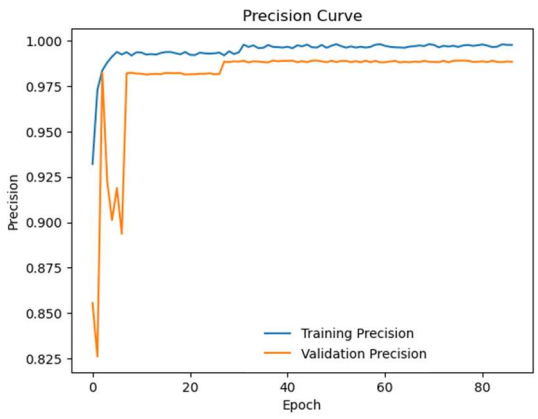
To visualize the training and validation performances, we plotted the corresponding metrics against the number of training epochs. As seen in Fig.26, both the training and validation accuracies increased steadily, while the losses decreased. This indicates that the model was able to learn the features of the dataset and generalize well.



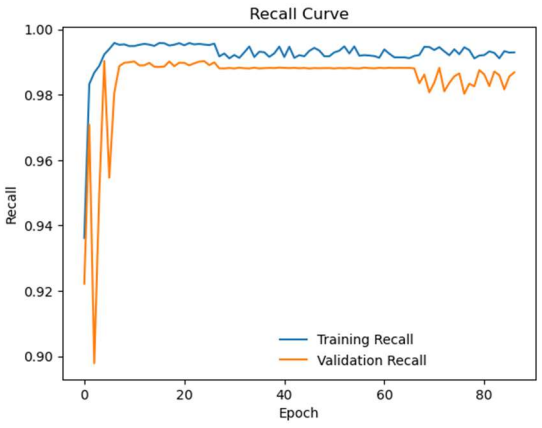
(i)



(ii)



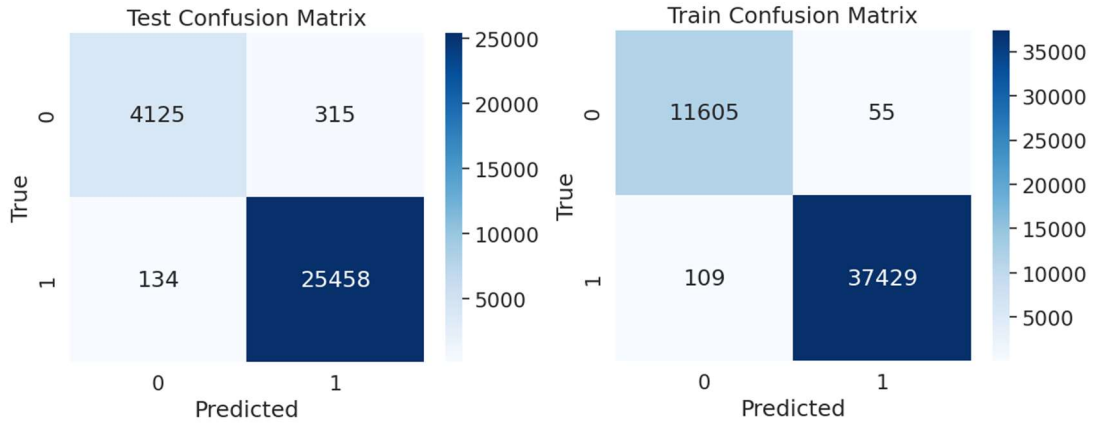
(iii)



(iv)

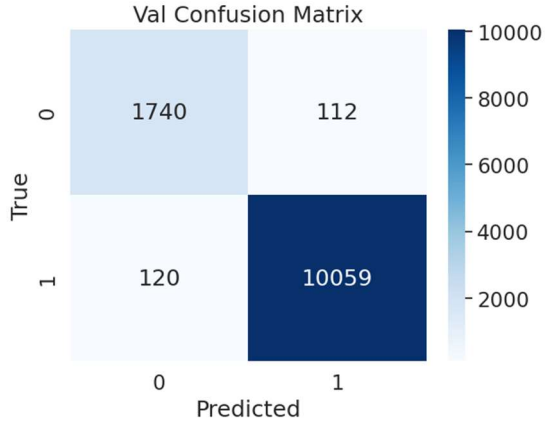
Fig.4.19 Visualization of Accuracy Curve (i), Loss Curve(ii), Precision Curve (iii), and Recall Curve (iv)

After training the pre-trained InceptionResNetV2 model on celebdfv2, we removed a few layers from the model to extract features. These features were then used to train a Vision Transformer model with the Nystrom attention mechanism. The resulting model achieved impressive performance on the test and validation sets, with AUC scores of 0.9988 and 0.9831, respectively. We also plotted a confusion matrix (as in Fig.27) to visually evaluate the model's performance.



(i)

(ii)



(iii)

Fig.4.20 Visualization of Train Confusion Matrix (i), Validation Confusion Matrix (ii), and Test Confusion Matrix (iii)

Furthermore, we summarized the final results for the train, test, and validation sets in Table 4.3, including metrics such as accuracy, precision, recall, and F1 score.

Table.4.3 Summarization of Performance Metrics on the DFDC Dataset

Dataset	*Loss*	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*AUC*
Train	0.056	0.9937	0.9967	0.9938	0.9953	0.9988
Validation	0.074	0.9886	0.9830	0.9819	0.9879	0.9831
Test	0.1171	0.9710	0.9753	0.9921	0.9836	0.9868

4.3 PERFORMANCE COMPARISON WITH OTHER SOTA TECHNIQUES

To assess the effectiveness of our strategy, we compare it to the performance of numerous state-of-the-art models on the DFDC dataset. The table below summarises the comparative findings, displaying the metrics of each model. As shown in Table.4.4, our method surpasses all other models in terms of accuracy, precision, and recall, indicating its supremacy in picture detection and classification.

Table 4.4 Model Performance Comparison on Different Architectures. The table displays the Accuracy, AUC, Precision, and Recall values of six SOTA models (EfficientNetB4, EfficientNetB7, InceptionResNetV2, XceptionNet, NASNetMobile, and ResNet50) and our proposed approach.

Model	**Accuracy**	**AUC**	**Precision**	**Recall**
EfficientNetB4[12]	0.8794	0.9105	0.8694	0.8756
EfficientNetB7 [12]	0.8538	0.9338	0.8538	0.8649
InceptionResNetV2[11]	0.9063	0.9526	0.9142	0.9063
XceptionNet[11]	0.9235	0.9496	0.9235	0.9147
NASNetMobile[40]	0.7769	0.8330	0.7869	0.7770
ResNet50[41]	0.8955	0.9288	0.8955	0.9069
Our Approach	**0.9710**	**0.9868**	**0.9753**	**0.9921**

CHAPTER 5

CONCLUSION

5.1 DISCUSSION

Using the Nyström attention mechanism to combine InceptionResNetV2 with vision transformers can increase the accuracy and resilience of deepfake detection systems. InceptionResNetV2 is a sophisticated deep-learning model that collects features of many sizes and orientations, making it resistant to image noise, distortions, and occlusions that are frequent in deepfake images. It has been pre-trained on large datasets such as ImageNet, allowing transfer learning and reducing the amount of training data required for high accuracy on new tasks such as deepfake detection. InceptionResNetV2 is computationally efficient and well-suited for real-time deepfake detection applications because of its short number of parameters. Vision transformers, on the other hand, use the self-attention process to acquire highly discriminative features from raw picture data and may capture long-range dependencies and interactions between distinct sections of an image.

By efficiently calculating the self-attention process, the Nyström approximation lowers the computational expense of training large transformer models. Using the proposed strategy on the DFDC dataset, we were able to detect deep fakes with an accuracy of 97.10%, an AUC of 0.9868, a precision of 97.53%, and a recall of 99.21% in our testing. The proposed method is a promising one for practical application in real-world circumstances due to its excellent performance and efficiency.

5.2 ADVANTAGE OF USING THE NYSTROM ATTENTION MECHANISM

The time complexity of the self-attention mechanism in a transformer model is O(n^2), where n is the sequence length. This is because every token in the sequence attends to every other token in the sequence.

The Nystrom attention mechanism is an approximate method for calculating attention that reduces the time complexity from O(n^2) to O(nk + k^2), where k is the number of landmark points used for approximation. The landmark points are a subset of the input sequence that is selected based on a clustering algorithm or randomly. The attention weights are calculated only between the landmark points and their surrounding points, rather than between all pairs of points in the sequence.

So, the Nystrom attention mechanism can be much faster than the normal self-attention mechanism when the sequence length is very large, and the number of landmark points k is much smaller than n. However, the approximation introduces some error, and the performance of the Nystrom attention mechanism may degrade if k is too small or the landmark points are poorly chosen.

5.3 CONCLUSION

In conclusion, the proposed approach using a combination of InceptionResNetV2 and Vision Transformer with Nystrom Attention mechanism has demonstrated state-of-the-art performance on three different datasets, namely DFDC, CelebDFv1, and CelebDFv2. The approach has achieved high accuracy, precision, recall, and F1 score, and also outperformed other state-of-the-art techniques such as EfficientNetB4, EfficientNetB7, XceptionNet, NASNetLarge, and ResNet50.

This study's significance lies in the development of a highly accurate and reliable deepfake detection approach, which is crucial for identifying and mitigating the harmful effects of deepfake videos on society. The proposed approach's potential impact is immense, as it can be used by various organizations, including social media platforms, news agencies, and governments, to detect deepfake videos and take appropriate measures to prevent the spread of misinformation and propaganda. Additionally, the approach's architecture and methodology can also be applied to other related fields, such as image and speech recognition, and could lead to the development of even more accurate and robust deep learning models.

5.4 FUTURE RESEARCH DIRECTIONS

Although deepfake detection models have developed significantly, there are still a number of issues that need to be resolved in order to increase the efficiency and usefulness of these systems.

We address some potential future prospects for deepfake detection research in this section, which can help accelerate the current state of the art and allow for the creation of more dependable and trustworthy detection models.

1. Adversarial attacks: Investigating the effectiveness of existing deepfake detection models against adversarial attacks, which are deliberate attempts to deceive the model and create more convincing deepfakes.

2. Generalization: Testing the performance of deepfake detection models on datasets with different deepfake generation techniques, as well as on videos captured in different settings and under different lighting conditions.

3. Real-time detection: Developing real-time deepfake detection systems that can operate in real-world scenarios and handle the high volume of video data generated by social media platforms.

4. Explainability: Increasing the interpretability and explainability of deepfake detection models, will help build trust in the technology and facilitate its adoption by end-users.

5. Multi-modal detection: Investigating the effectiveness of combining visual and audio cues to improve the accuracy of deepfake detection models.

Reference

[1] T. T. Nguyen *et al.*, “Deep learning for deepfakes creation and detection: A survey,” *Computer Vision and Image Understanding*, vol. 223, p. 103525, Oct. 2022, doi: 10.1016/j.cviu.2022.103525.

[2] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, “Transfer learning: a friendly introduction,” *J Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00652-w.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[4] R. S. Siva Kumar *et al.*, “Adversarial Machine Learning-Industry Perspectives,” in *2020 IEEE Security and Privacy Workshops (SPW)*, IEEE, May 2020, pp. 69–75. doi: 10.1109/SPW50608.2020.00028.

[5] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 3204–3213. doi: 10.1109/CVPR42600.2020.00327.

[6] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, “The Deepfake Detection Challenge (DFDC) Preview Dataset,” *ArXiv*, Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.08854

[7] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017, doi: 10.1609/aaai.v31i1.11231.

[8] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *ArXiv*, Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[9] B. Dolhansky *et al.*, “The DeepFake Detection Challenge (DFDC) Dataset,” *ArXiv*, Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.07397

[10] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: a Compact Facial Video Forgery Detection Network,” in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Dec. 2018, pp. 1–7. doi: 10.1109/WIFS.2018.8630761.

[11] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.

[12] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.

[13] D. Guera and E. J. Delp, “Deepfake Video Detection Using Recurrent Neural Networks,” in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, Nov. 2018, pp. 1–6. doi: 10.1109/AVSS.2018.8639163.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2008, pp. 1–8. doi: 10.1109/CVPR.2008.4587756.

[15] Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Dec. 2018, pp. 1–7. doi: 10.1109/WIFS.2018.8630787.

[16] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces,” *ArXiv*, Mar. 2018, [Online]. Available: http://arxiv.org/abs/1803.09179

[17] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019, pp. 1–11. doi: 10.1109/ICCV.2019.00009.

[18] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch, “Fake Face Detection Methods: Can They Be Generalized?,” in *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, Sep. 2018, pp. 1–6. doi: 10.23919/BIOSIG.2018.8553251.

[19] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A Survey of face manipulation and fake detection,” *Information Fusion*, vol. 64, pp. 131–148, Dec. 2020, doi: 10.1016/j.inffus.2020.06.014.

[20] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, “Facial soft biometrics for recognition in the wild: Recent works, annotation, and COTS evaluation,” *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 8, pp. 2001–2014, Aug. 2018, doi: 10.1109/TIFS.2018.2807791.

[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” *ArXiv*, Oct. 2017, [Online]. Available: http://arxiv.org/abs/1710.10196

[22] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2019, pp. 4396–4405. doi: 10.1109/CVPR.2019.00453.

[23] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 8789–8797. doi: 10.1109/CVPR.2018.00916.

[24] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2Face,” *Commun ACM*, vol. 62, no. 1, pp. 96–104, Dec. 2018, doi: 10.1145/3292039.

[25] J. Thies, M. Zollhöfer, and M. Nießner, “Deferred neural rendering,” *ACM Trans Graph*, vol. 38, no. 4, pp. 1–12, Aug. 2019, doi: 10.1145/3306346.3323035.

[26] K. O’Shea and R. Nash, “An Introduction to Convolutional Neural Networks,” *ArXiv*, Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.08458

[27] A. Sherstinsky, “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network,” *Physica D*, vol. 404, p. 132306, Mar. 2020, doi: 10.1016/j.physd.2019.132306.

[28] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Use of a Capsule Network to Detect Fake Images and Videos,” *ArXiv*, Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.12467

[29] J. Liu, K. Zhu, W. Lu, X. Luo, and X. Zhao, “A lightweight 3D convolutional neural network for deepfake detection,” *International Journal of Intelligent Systems*, vol. 36, no. 9, pp. 4990–5004, Sep. 2021, doi: 10.1002/int.22499.

[30] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera-Rodriguez, “DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 442–456. doi: 10.1007/978-3-030-68821-9_38.

[31] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, “Emotions Don’t Lie: An Audio-Visual Deepfake Detection Method using Affective Cues,” in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2020, pp. 2823–2832. doi: 10.1145/3394171.3413570.

[32] J. Baciak, M. Żurawska, T. Czech, and B. Górny, “Deepfake video detection using the ensemble of neural networks.” [Online]. Available: https://www.researchgate.net/publication/344554345

[33] D. Wodajo and S. Atnafu, “Deepfake Video Detection Using Convolutional Vision Transformer,” *ArXiv*, Feb. 2021, Accessed: May 25, 2023. [Online]. Available: http://arxiv.org/abs/2102.11126

[34] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, “Deep Fake Image Detection Based on Pairwise Learning,” *Applied Sciences*, vol. 10, no. 1, p. 370, Jan. 2020, doi: 10.3390/app10010370.

[35] S. Agarwal, N. Girdhar, and H. Raghav, “A Novel Neural Model based Framework for Detection of GAN Generated Fake Images,” in *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, Institute of Electrical and Electronics Engineers Inc., Jan. 2021, pp. 46–51. doi: 10.1109/Confluence51648.2021.9377150.

[36] J. Xiang and G. Zhu, “Joint face detection and facial expression recognition with MTCNN,” in *Proceedings - 2017 4th International Conference on Information Science*

and Control Engineering, ICISCE 2017, Institute of Electrical and Electronics Engineers Inc., Nov. 2017, pp. 424–427. doi: 10.1109/ICISCE.2017.95.

[37] M. Mahdianpari, B. Salehi, M. Rezaee, F. Mohammadimanesh, and Y. Zhang, "Very Deep Convolutional Neural Networks for Complex Land Cover Mapping Using Multispectral Remote Sensing Imagery," *Remote Sens (Basel)*, vol. 10, no. 7, p. 1119, Jul. 2018, doi: 10.3390/rs10071119.

[38] Y. Xiong *et al.*, "Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, pp. 14138–14148, May 2021, doi: 10.1609/aaai.v35i16.17664.

[39] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.2661

[40] I. Ilhan, E. Bali, and M. Karakose, "An Improved DeepFake Detection Approach with NASNetLarge CNN," in *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, IEEE, Oct. 2022, pp. 598–602. doi: 10.1109/ICDABI56818.2022.10041558.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

LIST OF PUBLICATIONS

1. Aale Rasool, Rahul Katarya, "**Analysis on Deep Fake Dataset, Manipulation and Detection Techniques: A Review**", Accepted & Presented at International Conference on Advances and Applications of Artificial Intelligence and Machine Learning (ICAAAIML-2022)

   Indexed by Scopus

   PaperID-190

SHARDA UNIVERSITY
Beyond Boundaries

SHARDA SCHOOL OF ENGINEERING & TECHNOLOGY

ICAAAIML-2022
International Conference on Advances and Applications of Artificial Intelligence and Machine Learning

Springer

CERTIFICATE
of Participation

This is to certify that

Dr./Mr./Ms. **Aale Rasool**

From **Department of CS, Delhi Technological University, India** *has presented/published research paper on*

Analysis on Deep Fake Dataset, Manipulation and Detection Techniques: A Review

during ICAAAIML-2022: International Conference on Advances and Applications of Artificial Intelligence and Machine Learning held on 16-17 September 2022 at Sharda University, Greater Noida, Uttar Pradesh.

Prof. (Dr.) Nitin Rakesh
HOD-CSE (SET)
Sharda University

Prof. (Dr.) Parma Nand
Dean, SET
Sharda University

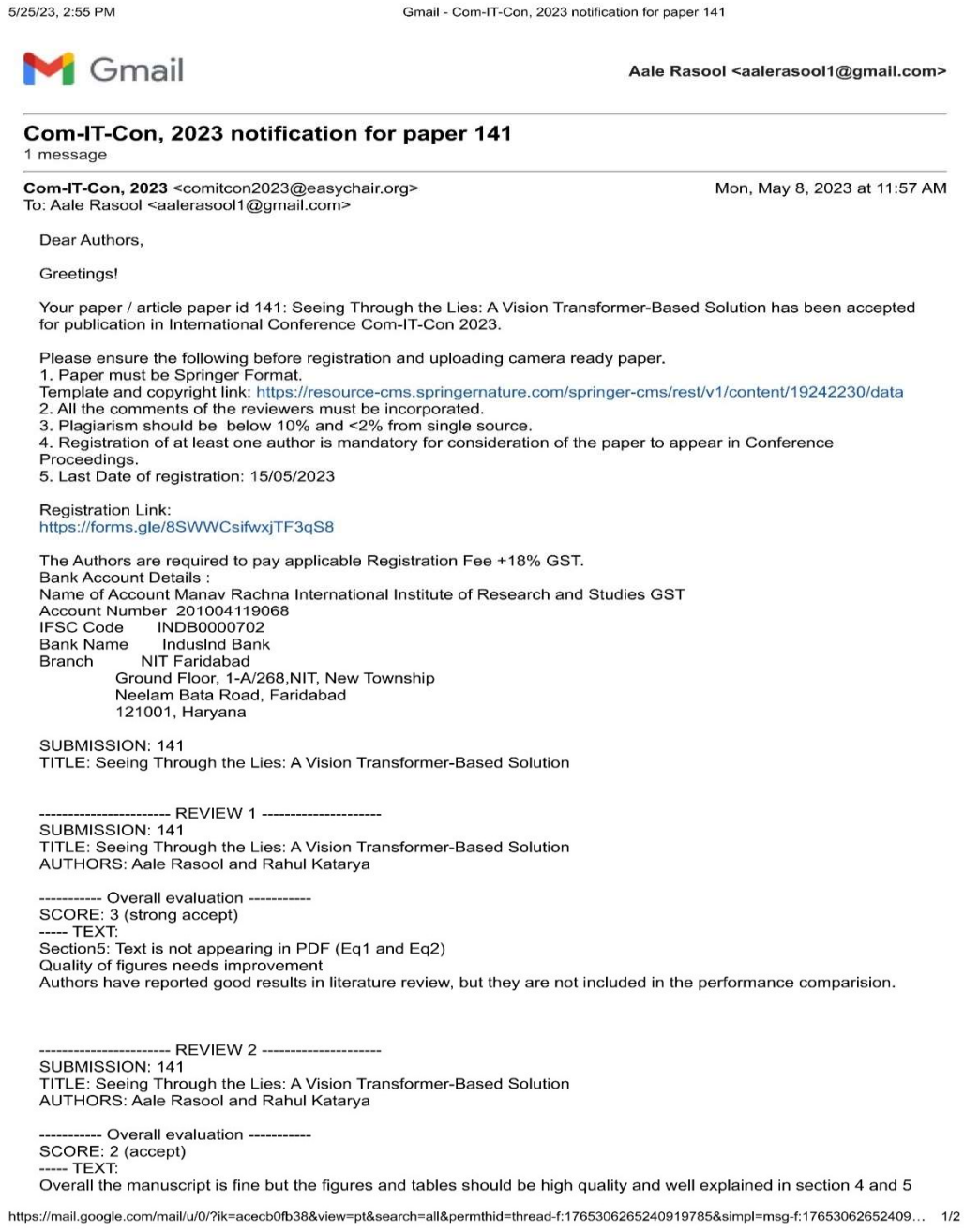
Prof. (Dr.) Sibaram Khara
Vice Chancellor
Sharda University

S. No. ICAAAIML2022 / CoP / 49/ PID

2. Aale Rasool, Rahul Katarya, "**Seeing Through the Lies: A Vision Transformer-Based Solution**", Accepted at International Conference on Artificial-Business Analytics, Quantum and Machine Learning: Trends, Perspectives, and Prospects (Com-IT-Con 2023)

   Indexed by Scopus

   PaperID-141



5/25/23, 2:55 PM Gmail - Com-IT-Con, 2023 notification for paper 141

Gmail **Aale Rasool <aalerasool1@gmail.com>**

Com-IT-Con, 2023 notification for paper 141
1 message

Com-IT-Con, 2023 <comitcon2023@easychair.org> Mon, May 8, 2023 at 11:57 AM
To: Aale Rasool <aalerasool1@gmail.com>

Dear Authors,

Greetings!

Your paper / article paper id 141: Seeing Through the Lies: A Vision Transformer-Based Solution has been accepted for publication in International Conference Com-IT-Con 2023.

Please ensure the following before registration and uploading camera ready paper.
1. Paper must be Springer Format.
Template and copyright link: https://resource-cms.springernature.com/springer-cms/rest/v1/content/19242230/data
2. All the comments of the reviewers must be incorporated.
3. Plagiarism should be below 10% and <2% from single source.
4. Registration of at least one author is mandatory for consideration of the paper to appear in Conference Proceedings.
5. Last Date of registration: 15/05/2023

Registration Link:
https://forms.gle/8SWWCsifwxjTF3qS8

The Authors are required to pay applicable Registration Fee +18% GST.
Bank Account Details :
Name of Account Manav Rachna International Institute of Research and Studies GST
Account Number 201004119068
IFSC Code INDB0000702
Bank Name IndusInd Bank
Branch NIT Faridabad
Ground Floor, 1-A/268,NIT, New Township
Neelam Bata Road, Faridabad
121001, Haryana

SUBMISSION: 141
TITLE: Seeing Through the Lies: A Vision Transformer-Based Solution

---------- REVIEW 1 ----------
SUBMISSION: 141
TITLE: Seeing Through the Lies: A Vision Transformer-Based Solution
AUTHORS: Aale Rasool and Rahul Katarya

---------- Overall evaluation ----------
SCORE: 3 (strong accept)
----- TEXT:
Section5: Text is not appearing in PDF (Eq1 and Eq2)
Quality of figures needs improvement
Authors have reported good results in literature review, but they are not included in the performance comparision.

---------- REVIEW 2 ----------
SUBMISSION: 141
TITLE: Seeing Through the Lies: A Vision Transformer-Based Solution
AUTHORS: Aale Rasool and Rahul Katarya

---------- Overall evaluation ----------
SCORE: 2 (accept)
----- TEXT:
Overall the manuscript is fine but the figures and tables should be high quality and well explained in section 4 and 5

https://mail.google.com/mail/u/0/?ik=acecb0fb38&view=pt&search=all&permthid=thread-f:1765306265240919785&simpl=msg-f:17653062652409… 1/2