Enhancing Deepfake Security: A Transfer Learning Approach for Advanced Video Manipulation Detection

BY

Shahrin Islam
ID: 242-25-019

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Computer Science and Engineering

Supervised By

Shah Md Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
September 2025

APPROVAL

This Thesis titled "**Enhancing Deepfake Security: A Transfer Learning Approach for Advanced Video Manipulation Detection**", submitted by Shahrin Islam, ID No: 242-25-019 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-09-2025.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zahid Hasan — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

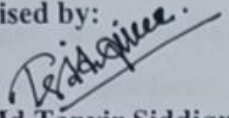
Dr. Naznin Sultana — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Nazibur Rahman — **External Examiner**
Head of IT Infrastructure
Networld Bangladesh PLC

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Shah Md. Tanvir Siddiquee, Assistant Professor, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



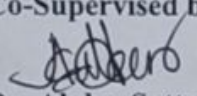
Shah Md Tanvir Siddiquee

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised by:



Dr. Abdus Sattar

Associate Professor

Department of CSE

Daffodil International University

Submitted by:

Shahrin Islam

Shahrin Islam

ID: 242-25-019

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible for us to complete the final year project/internship successfully.

I am grateful and wish to express my profound indebtedness to **Shah Md Tanvir Siddiquee**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka, deep knowledge & keen interest in the field of Deep Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Abdus Sattar,** Associate Professor**,** Department of CSE, for his kind assistance in completing our project, as well as to the other faculty members and staff of the CSE department at Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

The rapid advancement of deepfake technologies has emerged as a critical threat to the authenticity of digital media through the creation of highly realistic but fake visual content. The use of advanced video manipulation tools undermines public trust, poses security risks and challenges the integrity of information across digital platforms. Although significant progress has been made in deepfake image detection, research on identifying advanced video manipulation, particularly in low-resolution videos remains limited. To address this gap, this study propose a deepfake video detection framework by using transfer learning approach. Initially, our method analyzes each extracted frame using a CNN-based architecture to generate frame-level predictions, which are then aggregated by averaging. Based on a predefined threshold, the framework finally classifies the video as real or manipulated. For experimentation, we utilized a publicly available dataset named "FaceForensics++" containing a total of 2,000 real and manipulated videos of varying quality and resolution. We explored various CNN architectures including Xception, Densenet121, Inception ResNet V2, ResNet50, EfficientNet B3 along with rigorous hyperparameter tuning. Among these, the Xception architecture outperformed others by achieving a test accuracy of 94.5%. This research offers an effective solution to the growing challenge of deepfake detection that facilitates the development of robust and scalable tools that are vital for preserving information integrity in the industry 4.0.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

The rapid advancement of deepfake technologies, often driven by GANs and deep learning, generates highly realistic synthetic media which raises critical concerns for digital safety and public trust. Despite their technological sophistication, deepfakes pose a serious threat due to their potential for misuse [1,2]. Beyond spreading misinformation, deepfakes have been employed in malicious activities including political sabotage, identity impersonation, defamation, and cyberstalking [3,4]. Given the severity of these risks, to safeguard digital ecosystems it is imperative to develop robust and scalable deepfake detection systems [5].

There are many deepfake detection methods, if we classify them into four categories, they are deep learning-based [6], machine learning-based [7], media feature-based [8] and biometric-based [9]. Traditional deepfake detection methods often fall short against advanced generative techniques, as they rely on manually crafted features and forensic cues that are easily bypassed. In contrast, CNN-based deep learning model excels at deepfake detection by automatically learning discriminative patterns and identifying subtle manipulation artifacts. However, their accuracy declines on low-resolution or compressed videos, where fine details are lost and artifacts become harder to detect. This presents a key challenge in developing reliable detection systems for real-world applications.

In this paper, we propose a deepfake video detection framework using transfer learning, where frame-level predictions are generated using CNN-based architectures and aggregated to classify videos. Multiple architectures, including Xception, Densenet121, Inception ResNet V2, ResNet50, EfficientNet B3 are explored with extensive hyperparameter tuning. A comparative analysis and error evaluation are conducted to assess performance and identify sources of misclassifications.

1.2 Motivation

The swift progression of deepfake technologies has elicited much apprehension about the authenticity and integrity of digital media. Conventional detection approaches frequently depend on manually designed features and forensic indicators, which are often inadequate against novel and advanced generative techniques. Conventional methods often prove inadequate against minor modifications, rendering digital content susceptible to exploitation in domains like disinformation, identity theft, and cybercrime.

Conversely, deep learning methodologies, especially those employing Convolutional Neural Networks (CNNs), have exhibited an exceptional capacity to autonomously discern discriminative patterns from picture and video data. These models can discern intricate visual characteristics and nuanced artefacts frequently overlooked by conventional approaches, resulting in markedly enhanced detection precision. The Xception model, among different CNN architectures, has demonstrated exceptional performance in picture classification tasks owing to its capacity to effectively learn intricate spatial patterns. This renders it a promising contender for deepfake detection frameworks.

Nonetheless, difficulties persist in managing movies and photos of disparate quality. Although deepfake detection is typically proficient with high-resolution content, its efficacy frequently diminishes with low-resolution or heavily compressed movies, where small modification artefacts are more challenging to identify. This constraint drives the creation of a resilient and versatile detection framework that can sustain high accuracy across various video qualities and real-world situations.

By addressing these challenges, this study aims to leverage CNN-based deep learning and transfer learning to build a scalable, reliable, and effective deepfake detection system that can operate across a wide range of video qualities and manipulation techniques.

1.3 Research Objectives

This research proposes a deepfake video detection system utilising transfer learning, wherein frame-level predictions are produced through CNN-based architectures and consolidated to classify films. Various architectures, including as Xception, Densenet121, Inception ResNet V2,

ResNet50, and EfficientNet B3, are examined with comprehensive hyperparameter optimisation. A comparison analysis and error assessment are performed to evaluate performance and pinpoint sources of misclassifications.

The key contributions of this work are summarized as follows:

- Proposed Deepfake Detection Framework: We present a transfer learning-based model that evaluates each extracted video frame using CNN architecture to produce frame-level predictions. The predictions are consolidated through averaging, and a predetermined threshold is utilized to categorize the video as authentic or altered.
- Performance Evaluation: A thorough comparative analysis is performed between the proposed model and various existing architectures to build a solid baseline for deepfake video classification.
- Hyperparameter Optimization and Error Analysis: We conduct meticulous hyperparameter tuning to improve model performance and execute comprehensive error analysis to ascertain the principal causes of misclassifications.

1.4 Research Questions

In response to these challenges, our research focuses on exploring two key questions:

- RQ1: How can a robust and effective deep learning model be developed to accurately detect deepfakes and other advanced forms of video manipulation, even in low-resolution videos?
- RQ2: To what extent does the proposed model outperform existing approaches in terms of accuracy and reliability?
- RQ3: How well does the proposed detection framework generalize across diverse datasets and real-world scenarios with varying compression levels, resolutions, and manipulation techniques?

1.5 Expected Output

The principal objective of this study is to establish a resilient framework for deepfake video detection that can effectively differentiate between authentic and altered movies, even under

difficult conditions such as poor resolution or compression artefacts. The anticipated outcomes of this project comprise:

High-Accuracy Detection Model: A CNN-based deep learning model, focussing on the Xception architecture, trained and optimised to accurately detect altered video frames, achieving high precision, recall, and F1-score, hence providing balanced performance between actual and manipulated classes.

Frame-Level and Video-Level Predictions: The algorithm is anticipated to give precise frame-level predictions, which are consolidated to yield dependable video-level classifications, offering a thorough assessment of each video's authenticity.

Comparative Performance Analysis: A comprehensive comparison of various CNN designs, including DenseNet121, Inception-ResNet-V2, ResNet50, and EfficientNet-B3, emphasising the advantages and disadvantages of each model and determining the most efficacious method for deepfake detection.

Error Analysis and Insights: Identification and analysis of misclassifications, particularly in low-quality videos, to comprehend the limitations of existing algorithms and inform future enhancements in deepfake detection frameworks.

Scalable and Adaptable Framework: A transfer learning system that can be updated with fresh datasets and manipulation techniques, ensuring sustained flexibility and relevance in swiftly changing digital media landscapes.

Benchmarking Against Existing Methods: Comparative analysis of performance against cutting-edge deepfake detection techniques to validate the efficacy and dependability of the proposed framework in practical applications.

The anticipated outcome is a functional, dependable, and scalable deepfake detection system that can underpin further research and implementation in digital media verification and security applications.

1.6 Project Management and Finance

This study effort was conducted independently, without financial assistance from any individual, institution, or organisation. All facets of the study, encompassing data gathering, model construction, experimentation, and analysis, were conducted utilising existing resources and publically accessible datasets, including FaceForensics++. The project management employed a systematic methodology to guarantee punctual completion and superior results. Essential stages comprised:

- **Literature Review and Problem Identification:** Analysing current deepfake detection techniques and pinpointing research deficiencies to establish the study's objectives and scope.
- **Dataset Acquisition and Preprocessing:** Employing publicly accessible datasets and executing preprocessing, frame extraction, and augmentation for model training and assessment.
- **Model Development and Experimentation:** Executing and optimising several CNN architectures utilising open-source deep learning frameworks.
- **Analysis and Reporting:** Executing performance assessments, comparison evaluations, error investigations, and documenting of findings.

The study did not necessitate external financing and depended exclusively on accessible computing resources and software tools, including open-source libraries and frameworks. This self-directed strategy not only reduced expenses but also facilitated adaptable testing and iterative enhancement of the suggested deepfake detection system.

1.7 Report Layout

Chapter 1 delineates the introduction, aims, and principal research questions of the study. Chapter 2 presents succinct summaries of the literature review. Chapter 3 provides a comprehensive description of the suggested methodology. Chapter 4 delineates and analyses the experimental results of the paper. The fifth chapter addresses the sustainability plan, societal and environmental ramifications, and ethical considerations. The sixth chapter closes the current inquiry and delineates a strategy for future endeavours.

CHAPTER 2
BACKGROUND

2.1 Terminologies

In recent years, the rapid advancement of artificial intelligence (AI) and computer vision has enabled the development of sophisticated systems capable of detecting and classifying manipulated facial content in digital media. These systems play a critical role in identifying deepfakes and other forms of facial manipulation that threaten the authenticity of online information.

Many studies have focused on leveraging deep learning and image analysis techniques to accurately detect and categorize various types of manipulated facial content. Convolutional Neural Networks (CNNs) have proven highly effective due to their ability to extract rich, hierarchical features from images, such as fine-grained textures, facial structures, and subtle artifacts introduced during manipulation. By capturing these nuanced details, CNN-based models have significantly improved the accuracy and reliability of deepfake detection.

Recent research has also shown that combining transfer learning with advanced data augmentation techniques can enhance model robustness and generalization. This approach enables models to maintain high detection performance even under challenging conditions, such as low-resolution videos or heavily compressed content, where visual artifacts are less pronounced.

These technological advancements represent a major step forward in the field of digital media forensics. By enabling more accurate and reliable detection of manipulated content, they contribute to increasing the trustworthiness of online media and reinforcing public confidence in visual information'

2.2 Related Works

Deepfake detection has swiftly become an essential research domain, largely owing to the escalating complexity of synthetic media and its possible social, political, and ethical ramifications. As deepfake generation methods evolve, detection frameworks must also adapt to

ensure reliable identification of manipulated content. Recent years have therefore seen a surge of research efforts employing deep learning–based approaches to address this challenge.

Heidari et al. [10] presented a comprehensive review of state-of-the-art detection methods, identifying convolutional neural networks (CNNs) and region-based CNNs (RCNNs), particularly when combined with transfer learning and adversarial training strategies, as dominant techniques in the field. Their analysis revealed that such approaches consistently achieve more than 90% accuracy across standard benchmarks, reinforcing the effectiveness of deep feature extraction in manipulation detection.

Building on this foundation, Raza et al. [11] conducted an empirical evaluation of multiple deep learning architectures, including NASNet, Xception, MobileNet, and VGG16, along with their novel Deepfake Prediction (DFP) method. Their comparative study demonstrated that the DFP method surpassed existing models with a 94% detection accuracy, underscoring the value of tailored architectures specifically designed for deepfake detection. Similarly, Xu et al. [12] proposed NA-VGG, an enhanced variant of the traditional VGG16 network, which integrates Spatial Rich Model (SRM) filtering and image augmentation techniques. On the Celeb-DF dataset, NA-VGG achieved an AUC of 85.7%, outperforming baseline VGG16 and related methods, thereby highlighting the importance of noise residual analysis and data augmentation for robust detection.

To further enhance generalization across diverse forgeries, Hsu et al. [13] introduced the Contrastive Forgery Feature Network (CFFN), which incorporates a pairwise learning strategy. Their approach achieved 0.930 precision and 0.936 recall on the SA-GAN dataset, demonstrating significant improvements in feature discrimination and robustness against varied manipulation techniques. Likewise, Coccomini et al. [14] investigated the effectiveness of transformer-based models in comparison to CNN architectures. Their study revealed that EfficientNetV2-M attained higher accuracy (81.1%) on known forgeries, while ViT-Base exhibited superior generalization with 77.5% accuracy on unseen manipulations, suggesting that transformers may offer better resilience in cross-domain scenarios.

Expanding on this transformer-based line of work, Ghita et al. [15] trained a Vision Transformer (ViT) model on a large-scale Kaggle dataset consisting of 40,000 images. The model achieved 89.91% accuracy with strong convergence properties, positioning ViTs as competitive

alternatives to CNNs in terms of performance and training efficiency. Similarly, Joshi and Nivethitha [16] leveraged a transfer learning-based Xception architecture for both image and video deepfake detection, achieving an accuracy of 93.01%, which highlighted the adaptability of Xception to video-based manipulations.

Beyond deep learning networks, some researchers have explored alternative feature-based approaches. Younus et al. [17] proposed a method using Haar wavelet transforms to detect blur inconsistencies between facial regions and their surrounding backgrounds. Their model, tested on the UADFV dataset, achieved 90.5% accuracy, demonstrating that handcrafted feature extraction can still complement deep learning methods, particularly in detecting visual inconsistencies overlooked by CNNs.

Collectively, these studies underscore significant progress in deepfake detection, with most methods achieving high accuracy on benchmark datasets. However, several challenges remain unaddressed. Most existing works focus primarily on image-level analysis, which may not effectively generalize to low-resolution or compressed video content commonly encountered in real-world applications. Furthermore, while CNNs excel at texture and spatial artifact detection, they often struggle with temporal inconsistencies present in video data, whereas transformer-based methods, though promising, are still underexplored in large-scale cross-dataset video scenarios.

To address these limitations, our research proposes a deepfake video detection framework that integrates transfer learning across multiple CNN and transformer architectures. By combining the spatial feature extraction strengths of CNNs with the contextual modeling capabilities of transformers, and adapting these models through transfer learning, we aim to improve detection performance on real-world, low-quality video data. This approach seeks to bridge the gap between benchmark-level success and practical, deployment-ready deepfake detection systems.

Table 2.2 Summary of Recent Deepfake Detection and Related Works

Ref	Year	Goal	Feature	Algorithm	Approach	Result	Limitation
[18]	2025	Review deepfake detection (image,	Image/video manipulation,	ML & DL, multimodal models	Visual/audio analysis, multimodal fusion,	Accuracy, precision,	Vulnerable to adversarial attacks

		video, audio, multimodal)	audio spoofing, multimodal synthesis		adversarial training, blockchain	recall, efficiency, robustness	
[19]	2025	Develop a generalizable deepfake detector using real face modeling.	Real Face Foundation Representation (RFFR), Masked Image Modeling (MIM)	Anomaly detection based on input vs reconstructed faces	Trained only on real face datasets, MIM for robust feature learning	Outperforms state-of-the-art in cross-manipulation tests	Requires large-scale real face data
[20]	2023	Provide a practical, high-quality, flexible face-swapping framework.	Tools for face extraction, training, merging	Convolutional autoencoders, encoder–decoder networks, face alignment & masking	Autoencoder-based face swap with loss optimization	Cinema-level, high-fidelity swaps; outperforms other tools	Manual setup is complex for beginners
[21]	2023	Review and compare image forgery detection methods.	Detect and localize image forgeries	Traditional: Block matching, keypoint analysis, compression artifacts; Deep learning: CNN classifiers, segmentation, attention	Traditional vs deep learning approaches	Deep learning achieves higher accuracy	Traditional methods lack generalization and fail under complex forgeries

				mechanisms			
[22]	2024	Address fairness and generalization in deepfake detection.	Demographic-agnostic and domain-agnostic forgery features	Xception backbone with disentanglement learning	Fair loss function, disentangled feature fusion, cross/intra-domain evaluation	+8.90 % FDP on DFDC; FFPR improved by 11.69 % (Celeb-DF), 7.94% (DFD); higher AUC consistently	Performance heavily dataset-dependent
[23]	2024	Detect social media deepfakes using a large-scale explainable framework.	300k images (AI-generated, tampered, authentic)	Large multimodal models (SIDA-7B, SIDA-13B)	SID-Set dataset, mask prediction for localization, textual explanations, fine-tuning	SIDA-13B: Accuracy 93.6%, F1-score 93.5%	Needs further evaluation on unseen real-world social media data

2.3 Research Gap

Despite the significant progress achieved in deepfake detection, several limitations remain unaddressed in existing literature. Many studies have demonstrated the effectiveness of convolutional neural networks (CNNs) and their variants in achieving high accuracy on benchmark datasets [10,11,16]. However, these models often struggle to generalize across diverse real-world scenarios, particularly when dealing with low-resolution or highly compressed videos [12,13]. Moreover, while transformer-based models such as ViT and

EfficientNetV2 have shown promise in enhancing cross-domain generalization [14,15], their application in large-scale video-based detection remains underexplored.

Another critical limitation lies in overreliance on image-level analysis. Most prior works, including CNN-based approaches [10,11,16], focus primarily on spatial artifacts, often overlooking temporal inconsistencies that are characteristic of manipulated video sequences. Although feature-based approaches such as wavelet transforms [17] provide complementary strengths, they are insufficient as standalone solutions for robust detection. Additionally, fairness and demographic generalization issues have been reported, as many models exhibit dataset dependency and performance biases across domains [22].

Consequently, the present study tackles these deficiencies by amalgamating transfer learning with various CNN and transformer-based architectures. This study seeks to augment robustness, boost cross-domain generalisation, and provide more reliable deepfake video identification in real-world, low-quality contexts by utilising the spatial feature extraction capabilities of CNNs alongside the contextual modelling strengths of transformers.

2.4 Challenges

Detecting deepfakes is a formidable challenge owing to the swift advancement of manipulation methods and the intricacy of differentiating authentic from fabricated media. A key problem is the heightened realism of contemporary deepfakes, wherein subtle artifacts are diminished, complicating the identification of altered content for both people and machines [24]. A further difficulty is the limited generalisation of detection models, those trained on a certain dataset or manipulation type frequently exhibit subpar performance when confronted with unfamiliar datasets or novel deepfake creation methods.

Moreover, real-world problems such as video compression, frame loss, low resolution, and noise significantly impair the efficacy of detection algorithms, as these elements hide the nuanced discrepancies upon which models depend. Dataset restrictions constitute a considerable impediment: numerous extant datasets are either imbalanced, featuring a predominance of genuine samples over fake ones, or lack diversity, hence constraining the robustness of trained models.

From a computational standpoint, video-level detection necessitates the analysis of numerous frames, rendering it resource-intensive and time-consuming. Moreover, the evolution of deepfake generation techniques outpaces that of detection systems, resulting in an ongoing competition between perpetrators and protectors. These issues underscore the pressing necessity for more resilient, adaptable, and effective deepfake detection methods capable of functioning reliably in varied and dynamic real-world environments.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Proposed Methodology

The subtle and varied manipulations present in video content make deepfake detection a difficult task. The methodology is a systematic and step-by-step framework that integrates advanced deep learning techniques with effective data preprocessing and aggregation strategies to address this issue. The objective is to maintain efficiency and generalization across a variety of video qualities and resolutions while accurately identifying manipulated video content. The provided diagram 3.1 outlines a structured methodology for deepfake detection using deep learning.

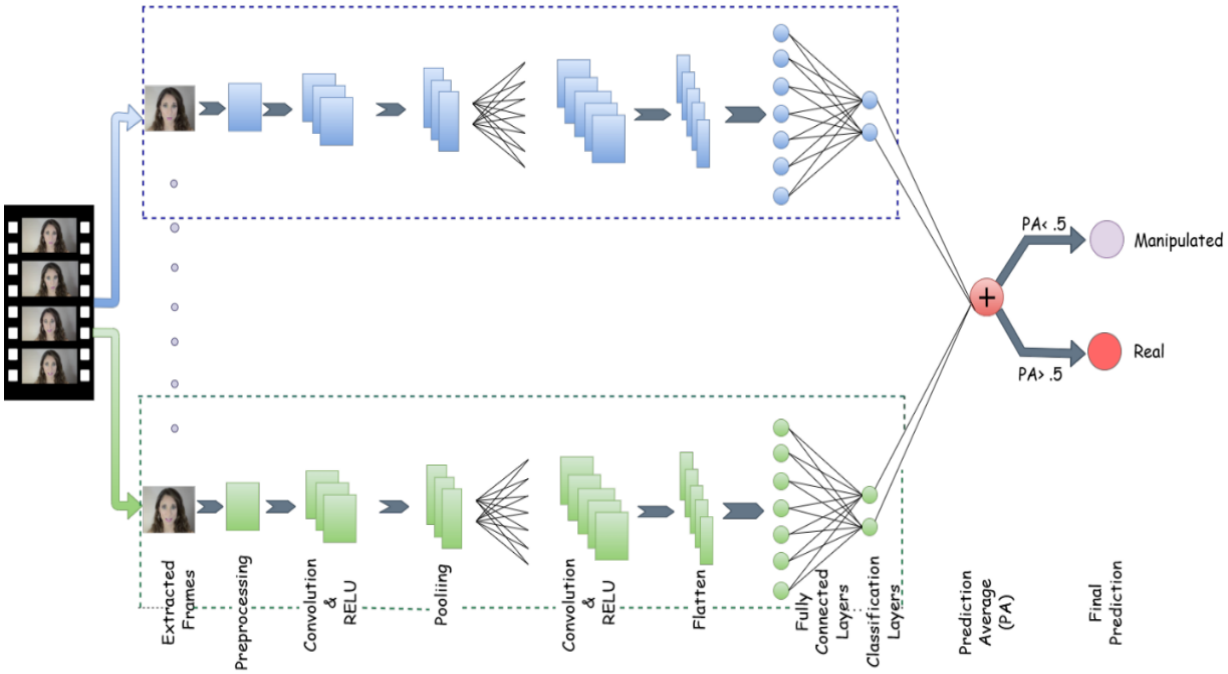


Figure 3.1: Workflow diagram of proposed methodology

In the study on deepfake detection, the methodology is designed as a structured and sequential process to analyse video content and accurately determine its authenticity. The process begins with the extraction of representative frames from input videos, where one frame per second is selected to provide a balanced dataset that captures essential visual information while minimising redundancy.

The extracted frames undergo a preprocessing pipeline involving resizing and normalisation, ensuring that each image is formatted consistently for compatibility with the input requirements of the model. These pre-processed frames are then fed into a Convolutional Neural Network (CNN) architecture composed of alternating convolutional and pooling layers. After each convolutional layer, the Rectified Linear Unit (ReLU) activation function is applied, introducing non-linearity and effectively mitigating the vanishing gradient problem, thereby enhancing the network's learning capabilities.

The convolutional layers provide feature maps, which are then flattened and sent through a fully connected layer. This generates frame-level classifications, designating each frame as either authentic or altered. The final video-level forecast is obtained by averaging the classification scores from all frames within each video. A threshold value of 0.5 is applied: if the average prediction score (PAP_APA) surpasses 0.5, the video is categorised as authentic; otherwise, it is deemed altered.

The model training process is conducted with meticulous attention to hyperparameter optimisation, as these parameters greatly affect classification results. The dataset distribution followed an 80:10:10 ratio, where the largest portion was used for training, while the smaller portions were used for validation and testing. The training set is utilised to learn model parameters, the validation set aids in hyperparameter tweaking and mitigating overfitting, and the testing set—consisting of unseen data—functions as the ultimate evaluation benchmark.

Training is performed across 50 epochs, with the model version that attains the highest validation accuracy retained for final evaluation. Through comprehensive experimentation, the ideal hyperparameter configurations are determined, guaranteeing that the model attains resilient and dependable performance in deepfake detection. The final system combines frame-level and video-level analysis, providing a thorough method for detecting modified video content.

3.2 Data Collection Procedure/Dataset Utilized

The dataset employed in this study for deepfake detection was sourced from the FaceForensics+ benchmark dataset, a widely recognized resource in the domain of multimedia forensics and deepfake research [25,26]. This dataset consists of a total of 2,000 video samples, equally divided into 1,000 authentic videos and 1,000 manipulated counterparts. The original, unaltered videos

were collected from YouTube, featuring front-facing individuals exhibiting a range of facial expressions and natural body movements. The authentic videos were thereafter altered through diverse deepfake creation methods to create equivalent synthetic copies.

To construct a frame-based dataset suitable for training and evaluation, one frame per second was extracted from all videos, resulting in approximately 38,000 frames in total. Each frame is assigned a binary label: 1 for real and 0 for manipulated. These labelled frames form the foundation for the model's training, validation, and testing phases. Figure 3.2 illustrates a sample deepfake manipulation from the FaceForensics++ dataset, showcasing the extent and quality of synthetic alterations applied to facial features.

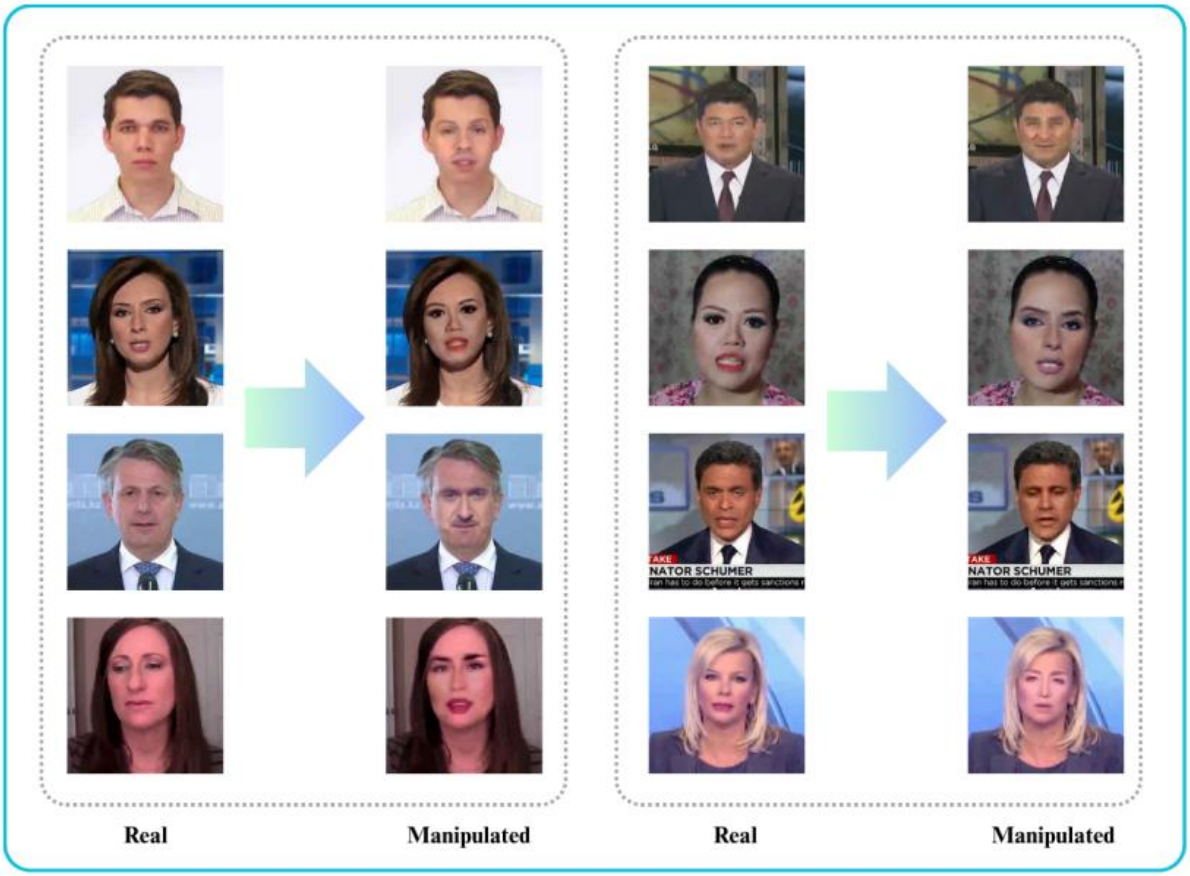


Figure 3.2: Sample deepfake manipulation in FaceForensics++ dataset

Prior to model training, the extracted frames undergo a preprocessing stage designed to ensure uniformity in input dimensions and enhance model robustness. Frames are scaled to 128×128×64 pixels to maintain consistent input shapes and lessen computational demand. In the training set,

a series of data enhancement techniques are applied, including horizontal flipping, rotation, and colour jittering, in addition to resizing. These augmentations introduce controlled variability into the dataset, improving the model's generalisation capabilities. In contrast, the validation set undergoes only resizing and normalisation to maintain consistency and prevent data distribution shifts during performance evaluation.

The structured integration of a balanced dataset, systematic frame extraction, and targeted preprocessing ensures that the proposed deepfake detection framework is trained on high-quality, representative data, thereby enhancing its ability to accurately distinguish between authentic and manipulated video content.

3.3 Image Pre-processing

Image pre-processing is a vital step in the proposed deepfake video detection pipeline, as it standardizes the visual input and ensures that critical features are preserved for accurate classification. Since deep learning models are sensitive to variations in image scale, illumination, and orientation, appropriate pre-processing improves both the stability and generalization of the model.

3.3.1 Frame Extraction

From each video in the dataset, one frame per second is extracted to create a representative set of samples. This sampling rate strikes a balance between computational efficiency and maintaining sufficient temporal diversity. Each frame inherits the label of its source video 1 for real and 0 for manipulated. In total, this process generated 38,396 frames that were subsequently used for training, validation, and testing.

3.3.2 Resizing and Normalization

To ensure consistency throughout the dataset, all extracted frames were scaled to 128×128 pixels with three RGB color channels. This resolution preserves essential facial details while reducing computational costs. Following resizing, pixel intensity values are normalized to a range of [0, 1] by dividing by 255, which accelerates model convergence and mitigates gradient instability during training.

3.3.3 Data Augmentation

To improve model robustness and reduce overfitting, several augmentation techniques are applied to the training data:

- Horizontal flipping introduces left–right orientation variations.
- Rotation (±15 degrees) simulates different camera perspectives.
- Brightness, contrast, and saturation are altered through color jittering to replicate different illumination conditions.

No augmentation is applied to the validation and test sets, which undergo only resizing and normalization to ensure fair and unbiased evaluation.

3.3.4 Batch Preparation

After pre-processing, the dataset is organized into batches to facilitate efficient training of the deep learning model. To optimize both GPU utilization and training stability, frames were divided into fixed-size batches containing 32 samples each. Each batch contains a mix of real and manipulated frames to ensure balanced learning, and the training data is shuffled at the start of every epoch to prevent the model from memorizing sample sequences and to improve generalization. Batch processing offers several advantages: it reduces memory consumption by loading only a subset of data per iteration, stabilizes optimization by averaging gradients over multiple samples, and accelerates training through parallel computation on GPUs. By combining batch preparation with pre-processing, the model receives consistent, well-structured input while benefiting from efficient, stable, and scalable training, which is essential for handling large video datasets such as FaceForensics++.

3.3.5 Importance of Pre-processing

The chosen pre-processing strategy balances efficiency with detail preservation, which is crucial for detecting subtle deepfake artifacts. Standardized frame dimensions and normalization ensure compatibility with CNN architectures, while controlled augmentation enhances generalization to unseen data.

3.4 Deep Learning Models

Convolutional Neural Networks (CNNs) and other deep learning models have proven highly effective in identifying deepfake videos, owing to their capacity to automatically extract distinguishing features from images [27,28]. In this research, multiple CNN architectures were explored to evaluate their capability in identifying subtle manipulations in video frames, including Xception, DenseNet121, Inception ResNet V2, ResNet50, and EfficientNet B3. These models differ in their depth, convolutional operations, and parameter efficiency, which impacts their performance on both high- and low-resolution video data.

The Xception architecture leverages depthwise separable convolutions to reduce computational cost while preserving spatial features, making it particularly effective for detecting fine-grained manipulations. DenseNet121 utilizes dense connections between layers to promote feature reuse and mitigate the vanishing gradient problem, while Inception ResNet V2 combines Inception modules with residual connections to capture multi-scale information efficiently. ResNet50 employs residual blocks to allow training of deeper networks without degradation, and EfficientNet B3 balances network depth, width, and resolution to optimize performance relative to computational resources.

Each model receives pre-processed frames organized in batches and outputs frame-level predictions, which are later aggregated to classify entire videos. Weights pre-trained on large image datasets were leveraged and fine-tuned on the FaceForensics++ dataset using a transfer learning approach. A comprehensive hyperparameter tuning process was carried out to determine the most suitable learning rates, batch sizes, dropout rates, and optimizers, ensuring stable training and improved accuracy.

Among the evaluated architectures, Xception achieved the highest test accuracy of 94.5%, demonstrating superior capability in detecting subtle manipulations in low-resolution videos. This highlights the importance of selecting architectures that efficiently capture spatial details while maintaining computational efficiency, which is crucial for real-world deepfake detection applications.

3.4.1 Xception

As an extension of the Inception model, the Xception architecture replaces conventional Inception modules with depthwise separable convolutions to improve feature extraction. This design allows the network to decouple spatial and cross-channel correlations, significantly reducing computational complexity while preserving critical spatial features necessary for accurate image classification[29,30].

In the context of deepfake detection, Xception is particularly effective because it can capture subtle manipulation artifacts in video frames, such as irregular textures, unnatural facial boundaries, and minor inconsistencies introduced by generative models. The network consists of a series of convolutional layers, followed by depth wise separable convolutions, stride convolutions for down sampling, and global average pooling (GAP) layers, which help aggregate feature maps into a robust representation suitable for classification. A final SoftMax layer outputs the probability of a frame being real or manipulated.

For this research, the pre-trained Xception model was fine-tuned on the FaceForensics++ dataset, leveraging transfer learning to adapt the network to deepfake detection. Frames extracted from videos were passed through the Xception architecture in batches, with hyperparameters such as learning rate, batch size, and dropout rate optimized for maximum performance. The model's ability to learn discriminative patterns and generalize across low- and high-resolution videos contributed to achieving a test accuracy of 94.5%, outperforming other evaluated CNN architectures.

The Xception model's efficiency and high accuracy make it a robust choice for frame-level deepfake detection, and its modular design allows further adaptation for more complex video manipulation tasks in future research.

Furthermore, visualization approaches like Class Activation Maps (CAM) or Grad-CAM might improve the interpretability of Xception-based models by highlighting portions of the frame that have the most influence on the model's choice. This not only improves transparency, but also aids in identifying the specific visual cues used by the detector. Furthermore, merging Xception with ensemble approaches or temporal sequence models like LSTMs and Transformers has the potential to improve video-level predictions by mixing spatial and temporal information, opening the way for more advanced and trustworthy deepfake detection frameworks.

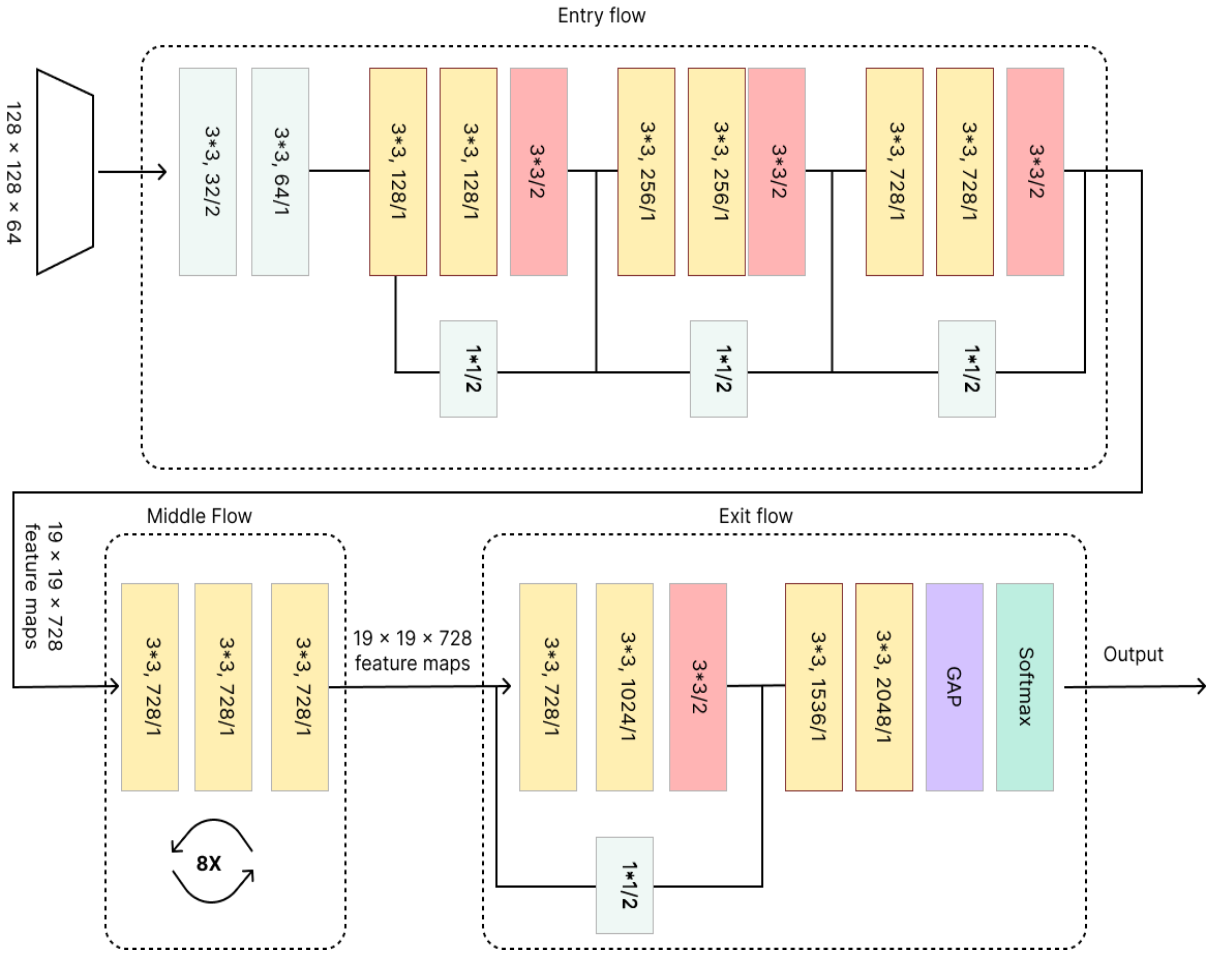


Figure 3.4.1: Architecture of Xception

3.4.2 DenseNet121

DenseNet121 is a deep convolutional neural network characterised by its dense connection architecture, wherein each layer takes inputs from all preceding levels and transmits its feature maps to succeeding layers. This connectedness facilitates feature reuse, alleviates the vanishing gradient issue, and allows for effective learning with a reduced number of parameters relative to conventional CNNs.

In deepfake detection, DenseNet121 can identify intricate spatial information from video frames, including subtle facial artefacts and inconsistencies created by deceptive generative models. The network consists of several dense blocks linked by transition layers, incorporating convolutional and pooling techniques to progressively derive hierarchical feature representations. A global

average pooling (GAP) layer, succeeded by a fully linked layer and SoftMax activation, generates frame-level classification probabilities.

This research involved fine-tuning the pre-trained DenseNet121 model on the FaceForensics++ dataset through a transfer learning methodology. Pre-processed video frames were input into the model in batches, while hyperparameters including learning rate, batch size, and dropout rate were optimised to ensure steady training and enhanced performance. DenseNet121 exhibited robust feature extraction abilities, especially in identifying small alterations, attaining a test accuracy of 93.5%, thereby establishing itself as a dependable alternative to existing deep learning architectures.

The model's intricate connection and effective feature propagation render it highly suitable for deepfake detection tasks, particularly when the aim is to discern subtle distinctions between authentic and altered frames.

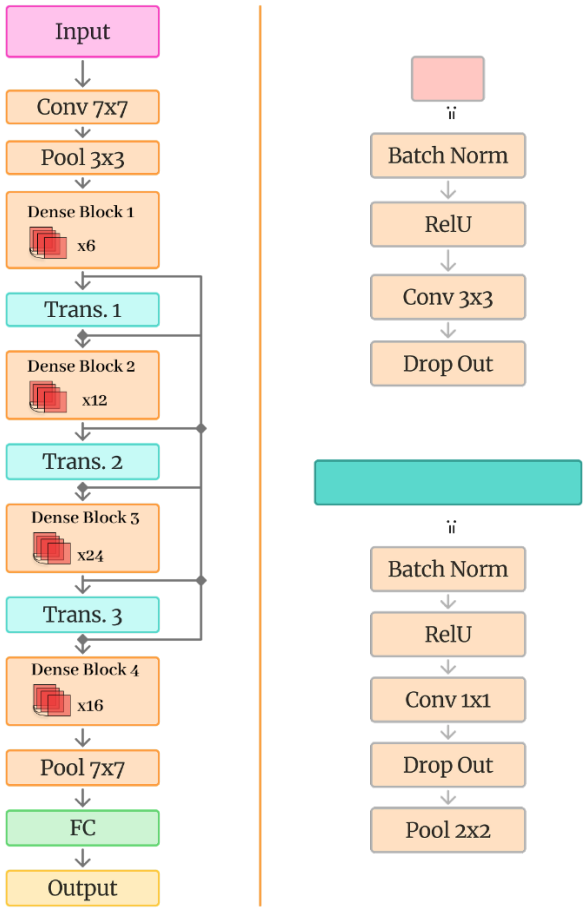


Figure 3.4.2: Architecture of DenseNet121

3.4.3 Inception ResNet V2

Inception ResNet V2 is a hybrid convolutional neural network architecture that enhances gradient propagation and guarantees robust training in extremely deep networks by integrating the multi-scale feature extraction capabilities of the Inception module with residual connections. The network can capture features across a variety of spatial scales, from global facial structures to fine local textures, by applying multiple convolutional filters of varying sizes in parallel on the Inception modules. In the interim, residual connections assist in the mitigation of the vanishing gradient issue, thereby facilitating the training of more complex models while preserving critical information throughout the network.

For deepfake detection, Inception ResNet V2 is particularly effective in learning both fine-grained and holistic facial features, making it capable of identifying subtle inconsistencies like unnatural facial boundaries, irregular lighting, or artifacts introduced during video manipulation. The architecture is composed of repeated Inception-ResNet blocks, followed by global average pooling (GAP) to aggregate spatial features and a SoftMax layer for frame-level classification.

In this research, the pre-trained Inception ResNet V2 model was fine-tuned on the FaceForensics++ dataset, using pre-processed frames arranged in batches. Extensive hyperparameter tuning, including optimization of learning rate, batch size, and dropout rate, was performed to improve generalization. Although it effectively captures multi-scale features, the model achieved a test accuracy of 93.0%, slightly lower than Xception and DenseNet121, indicating that while it is strong in feature extraction, it may struggle with the most subtle manipulations in low-resolution videos.

Moreover, Inception ResNet V2 shows promise when combined with temporal sequence learning models such as Long Short-Term Memory (LSTM) or 3D CNNs, which could help capture frame-to-frame inconsistencies in videos. Visualization techniques like Grad-CAM can also be applied to better understand which facial regions influence the model's decision-making, increasing interpretability. Future work may also explore ensemble strategies that integrate Inception ResNet V2 with lighter architectures to balance accuracy and computational cost, making it more suitable for large-scale or real-time deepfake detection scenarios.

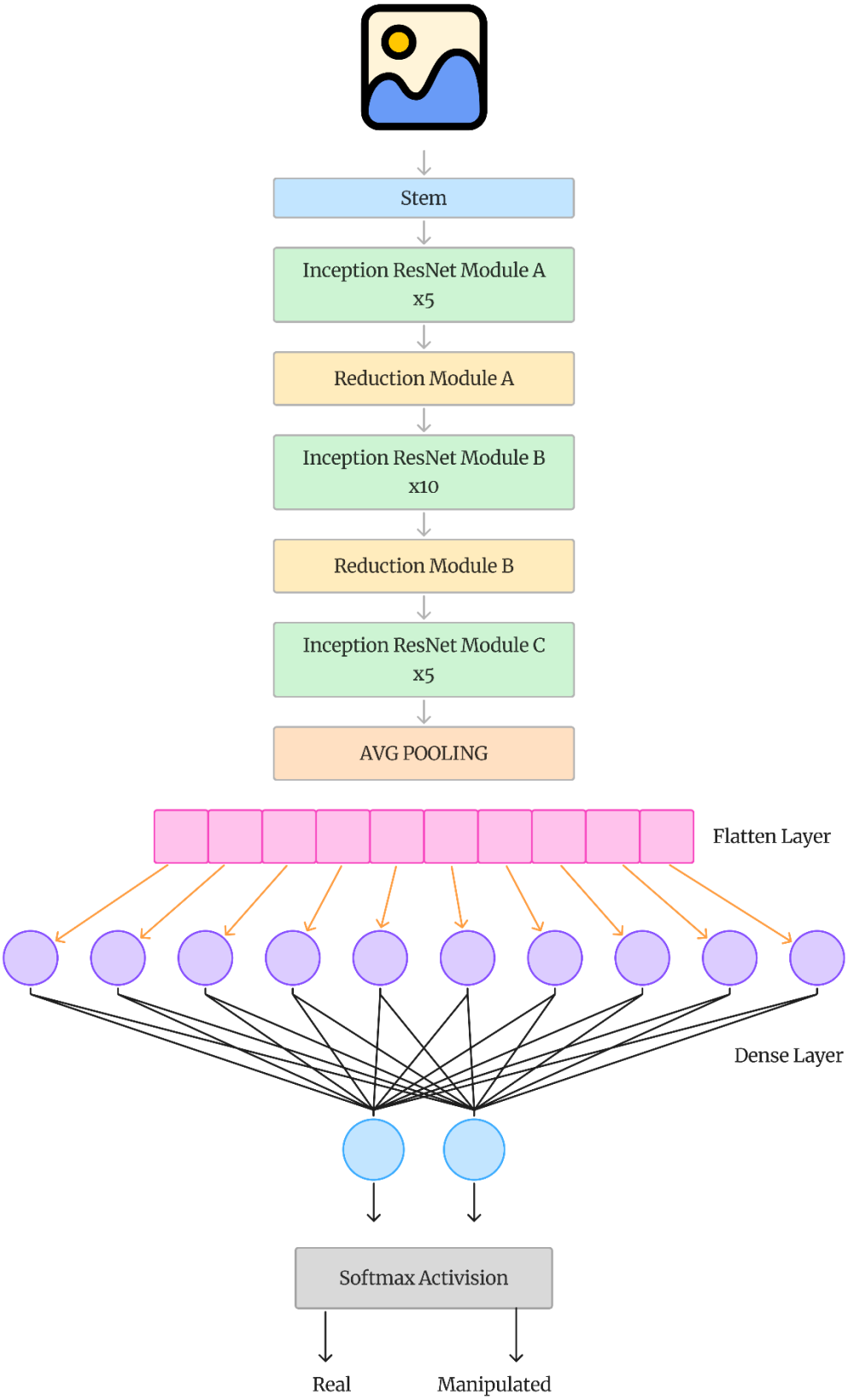


Figure 3.4.3: Architecture of Inception ResNet V2

3.4.4 ResNet50

ResNet50 is a deep convolutional neural network that utilises the residual learning framework to introduce shortcut connections that bypass one or more convolutional layers. This design resolves the degradation issue, which occurs when extremely deep networks experience performance saturation or even a decrease in accuracy, by enabling direct gradient flow through the shortcuts.

ResNet50 is highly capable of learning hierarchical feature representations across multiple layers in the context of deepfake detection, beginning with basic edges and textures in the early layers and progressing to intricate facial features in the deeper layers. Composed of 50 layers, the architecture is organised into identity and convolutional units. After this, a SoftMax layer is used for frame-level classification, which is followed by global average pooling.

ResNet50 was fine-tuned on the FaceForensics++ dataset for this investigation by utilising batch-wise pre-processed frames. The model was able to adapt to the specific task of deepfake detection while leveraging knowledge from large-scale image datasets using transfer learning. The model achieved a test accuracy of 91.5%, demonstrating stable and reliable performance. In comparison to more specialised architectures such as Xception, ResNet50 demonstrated slightly lower sensitivity in detecting extremely subtle manipulations, particularly in low-resolution or compressed videos, despite its effectiveness in capturing prominent facial anomalies.

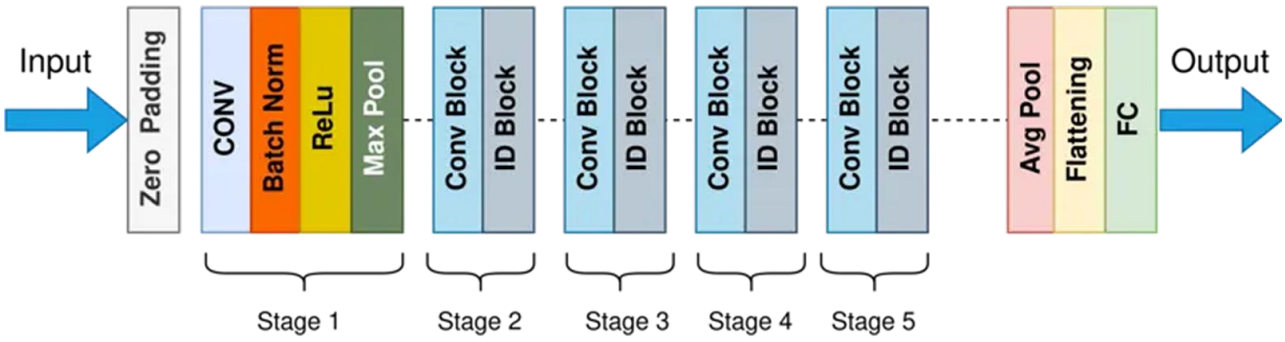


Figure 3.4.4: Architecture of ResNet50

3.4.5 EfficientNet B3

EfficientNet B3 is part of the EfficientNet family, which uses a compound scaling method to optimize network depth, width, and input resolution simultaneously. This systematic scaling enables EfficientNet B3 to achieve high accuracy while maintaining computational efficiency, which is crucial for processing large-scale video datasets.

The network is composed of Mobile Inverted Bottleneck Convolution (MBConv) blocks, which enhance feature representation and efficacy by combining depthwise separable convolutions with squeeze-and-excitation modules. The design of EfficientNet B3 enables the capture of multi-level spatial features, such as subtle texture irregularities, fine facial details, and inconsistencies that are introduced during deepfake generation. The architecture culminates with a Softmax classifier for frame-level prediction and a global average pooling layer.

In this study, EfficientNet B3 was fine-tuned on the FaceForensics++ dataset, utilizing batch-wise input and optimized hyperparameters for learning rate, batch size, and dropout. Although the architecture is highly efficient and performs well in capturing general features, it achieved a test accuracy of 87.5%, slightly lower than other evaluated CNN models. This indicates that while EfficientNet B3 is computationally attractive, it may not be as sensitive to subtle manipulations in low-resolution frames as architectures like Xception or DenseNet121.

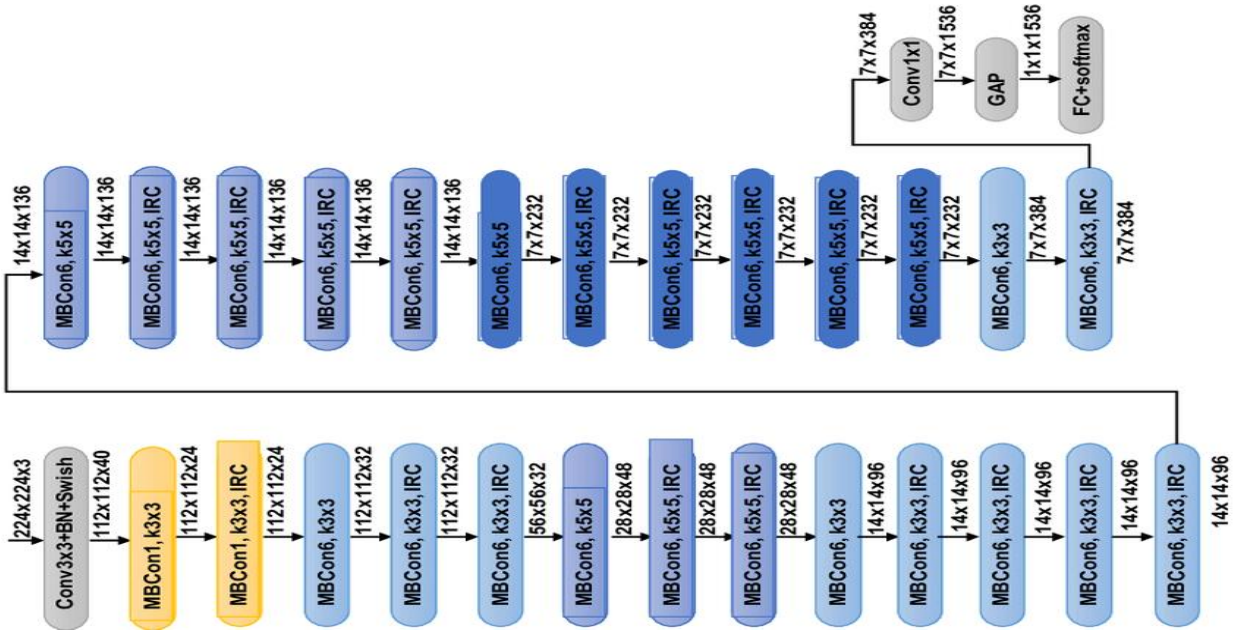


Figure 3.4.5: Architecture of EfficientNet B3

3.4.6 Hyperparameter Settings

The performance of deep learning models is highly dependent on the selection of appropriate hyperparameters, which control various aspects of the training process. Critical hyperparameters including learning rate, optimizer, batch size, number of epochs, and dropout rate were meticulously adjusted to attain optimal model performance. An empirical approach was adopted, where multiple candidate values were tested within predefined ranges, and the combination yielding the highest validation accuracy was selected.

The learning rate has a substantial impact on the stability of the training process and the speed at which the model converges. Optimizers were explored to identify the one that balances convergence efficiency with generalization ability. Similarly, different batch sizes were evaluated to optimize the trade-off between computational efficiency and stability of gradient updates. The number of epochs was chosen to allow sufficient training without overfitting, while dropout rates were fine-tuned to introduce regularization and reduce the risk of overfitting.

Through systematic experimentation, the final hyperparameter configuration was established, as summarized in Table 3.4.5. These settings were consistently applied across the models during training to ensure comparability of results.

Table 3.4.5 Hyperparameter Settings for Deep Learning Models

Hyperparameter	Hyperparameter Space	Selected Hyperparameter
Learning rate	[1e-5, 1e-4, 3e-4, 1e-3]	3e-4
Optimizer	[Adam, Nadam, AdamW]	AdamW
Batch size	[8, 16, 32, 64]	32
Number of epochs	[5, 10, 20, 25, 50]	50
Dropout rate	[0.1, 0.15, 0.25, 0.35, 0.5, 0.55]	0.1

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Results of Segmentation

In our deepfake detection framework, segmentation refers to the frame-level classification of video frames into real or manipulated categories. The model processes each extracted frame independently and generates a prediction score indicating the likelihood of manipulation. These results provide insights into how effectively the model identifies subtle artifacts present in deepfake videos.

4.1.1 Frame Extraction and Prediction

To effectively analyze deepfake videos, we first extract representative frames from each video in the dataset. Specifically, one frame per second is selected, ensuring a diverse yet manageable set of samples that captures variations in facial expressions, lighting, and movement. This approach provides a comprehensive view of each video while reducing computational complexity.

Once extracted, each frame undergoes a preprocessing pipeline designed to standardize inputs for the CNN model. This encompasses the application of normalisation techniques to consistently scale pixel values and the resizing of frames to a uniform dimension. These steps are crucial to ensure that the model focuses on subtle manipulations rather than variations caused by frame size or intensity differences.

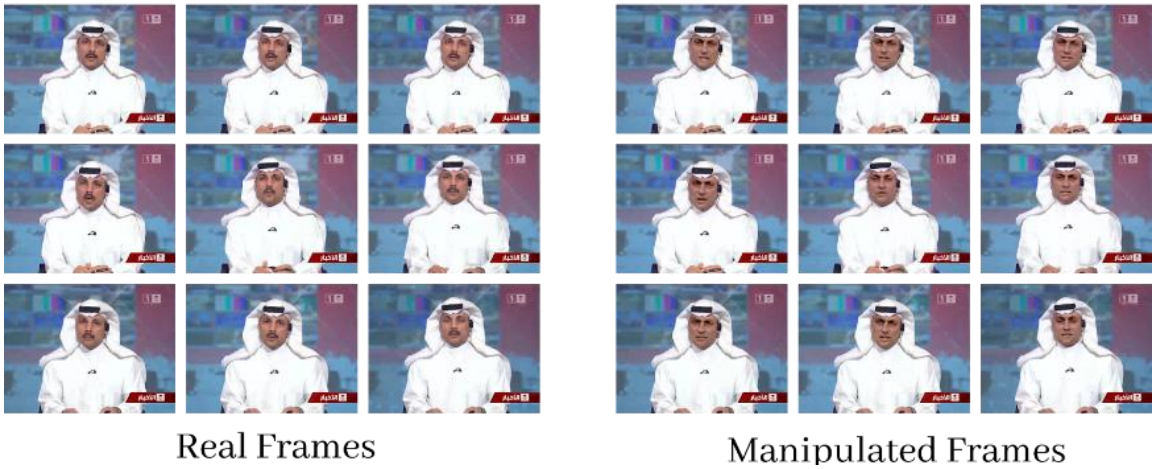


Figure 4.1.1.1: Sample real and manipulated frames

The original frames, as shown in Figure 4.1.1.1, depict both real and manipulated content from the FaceForensics++ dataset. These frames serve as the input to the model and provide a visual reference for the type of manipulations the system is expected to detect.

After preprocessing, frames are passed individually through the CNN-based deepfake detection model, which generates a prediction score for each frame. Based on this score, each frame is classified as either real or manipulated. The resulting predicted frames, illustrated in Figure 4.1.1.2, demonstrate the model's ability to distinguish authentic content from forgeries at the frame level. This step forms the foundation for subsequent video-level aggregation and overall classification.

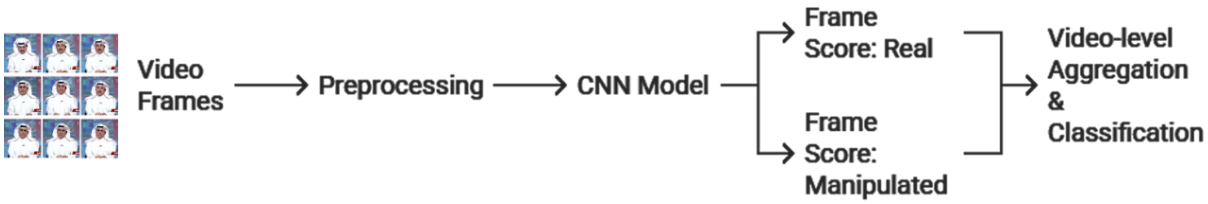


Figure 4.1.1.2: Frame-level classification of real and manipulated content

4.1.2 Video-Level Aggregation

A video-level classification is generated by aggregating frame-level predictions, which offers a more comprehensive assessment of the model's performance. The aggregation is performed by averaging the prediction scores across all frames in a video. If the average exceeds a predefined threshold, the video is classified as manipulated; otherwise, it is labeled as real. This approach ensures that transient misclassifications in individual frames do not disproportionately affect the overall video-level result.

The video-level aggregation allows the model to account for both the number and distribution of manipulated frames, providing a robust measure of video authenticity.

4.1.3 Performance Metrics

The effectiveness of the applied deepfake detection models is quantitatively evaluated using standard performance metrics at the video level, including accuracy, confusion matrix, and ROC

curve with AUC. These metrics provide a comprehensive assessment of each model's ability to distinguish real from manipulated videos.

Accuracy: The percentage of videos that are correctly classified among all test videos is referred to as accuracy. Comparing accuracy across the five models—Xception, DenseNet121, Inception ResNet V2, ResNet50, and EfficientNet B3—allows us to identify the best-performing architecture. Models such as Xception achieved the highest accuracy, followed by DenseNet121, highlighting their strong feature extraction capabilities. Other architectures, such as Inception ResNet V2, ResNet50, and EfficientNet B3, showed slightly lower performance, particularly in more complex scenarios.

Confusion Matrix: The confusion matrix allows for a comprehensive evaluation of each model's classification performance, highlighting both correct and incorrect predictions. By presenting the matrices side by side for all five models, it is possible to observe differences in how each model handles misclassifications. Figures 4.1.3.1–4.1.3.5 display the confusion matrices for Xception, DenseNet121, Inception ResNet V2, ResNet50, and EfficientNet B3, respectively, illustrating the distribution of correct and incorrect predictions.

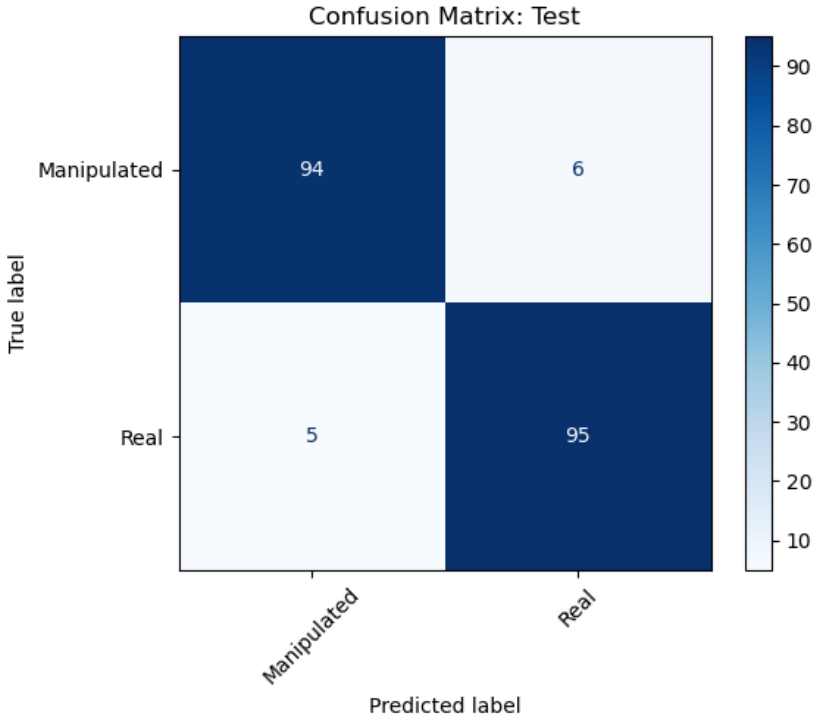


Figure 4.1.3.1: Confusion Matrix of Xception model for deepfake detection

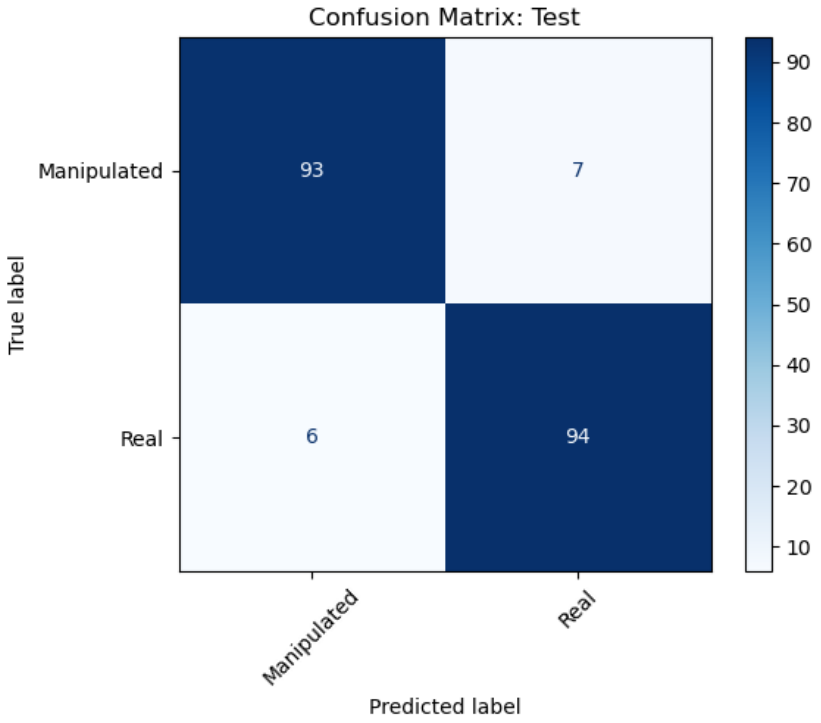


Figure 4.1.3.2: Confusion Matrix of DenseNet121 model for deepfake detection

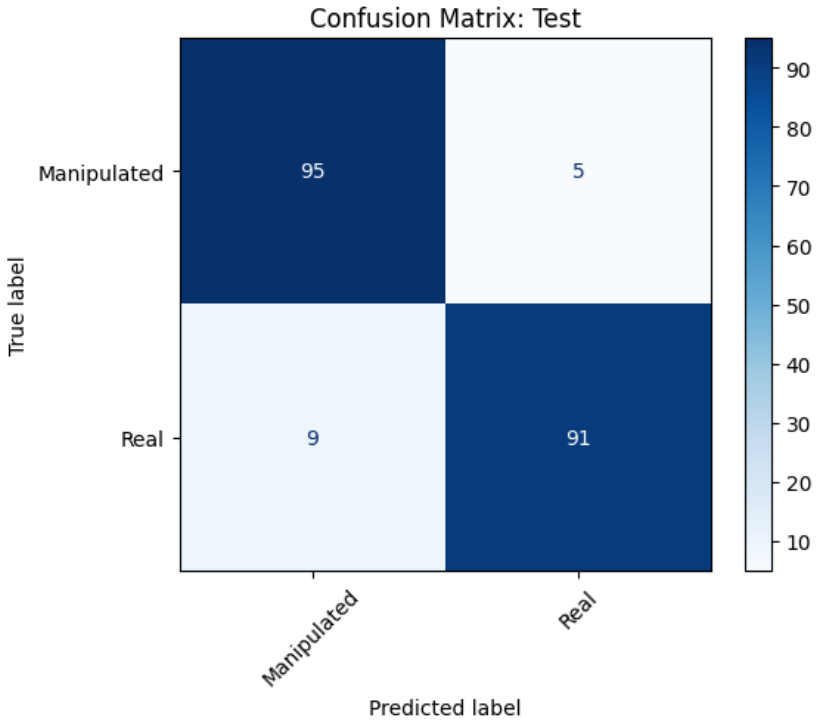


Figure 4.1.3.3: Confusion Matrix of Inception ResNet V2 model for deepfake detection

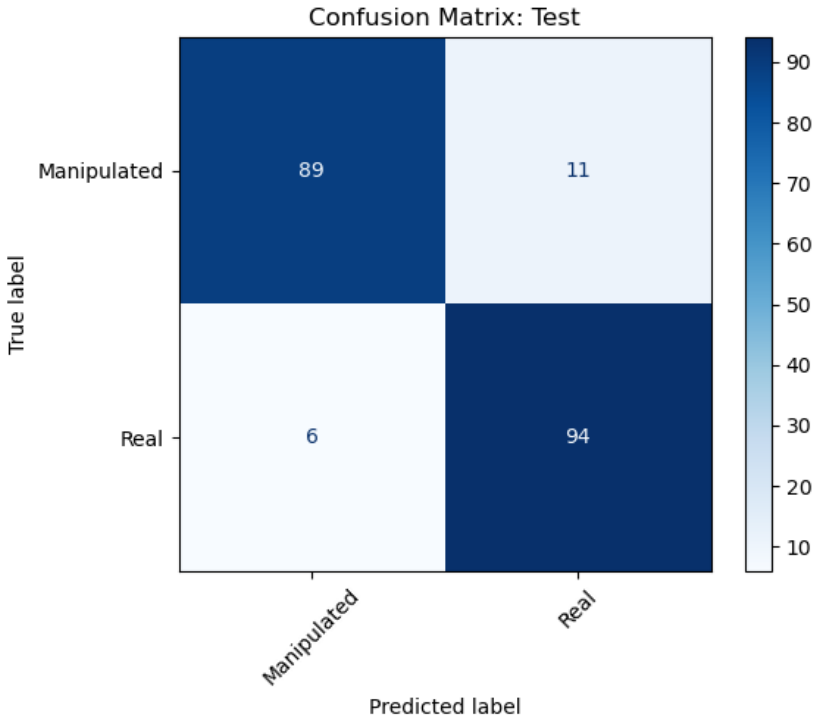


Figure 4.1.3.4: Confusion Matrix of ResNet50 model for deepfake detection

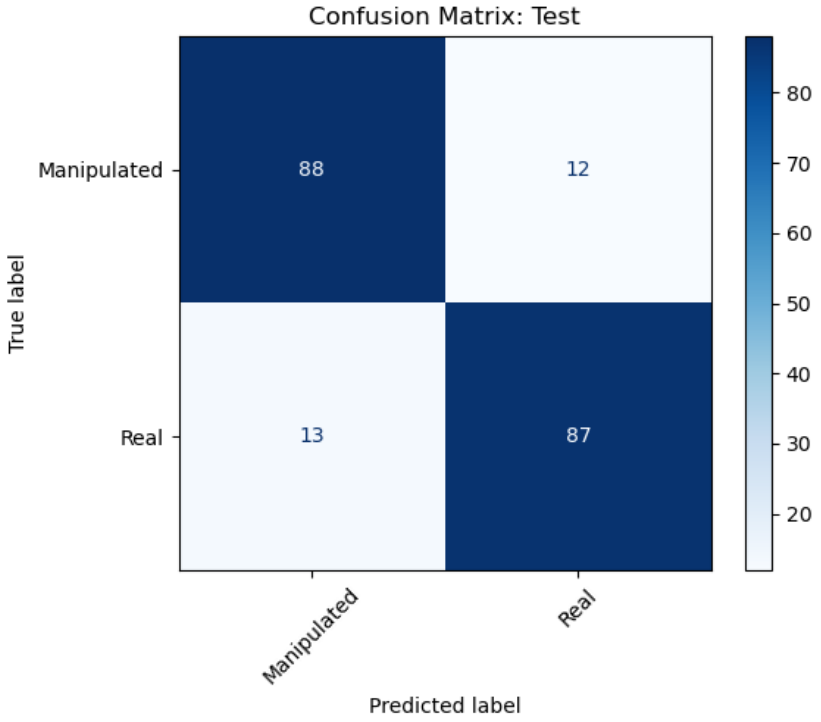


Figure 4.1.3.5: Confusion Matrix of EfficientNet B3 model for deepfake detection

ROC Curve and AUC: The Receiver Operating Characteristic (ROC) curve illustrates the true positive rate versus the false positive rate across different categorization thresholds, while the Area Under the Curve (AUC) offers a threshold-independent assessment of performance. A comparative view of model performance, based on ROC analysis, is provided in Figures 4.1.3.6 to 4.1.3.10. Models exhibiting curves nearer to the top-left corner and elevated AUC values, such as Xception and EfficientNet B3, exhibit superior discriminative capability, adeptly distinguishing between authentic and altered films.

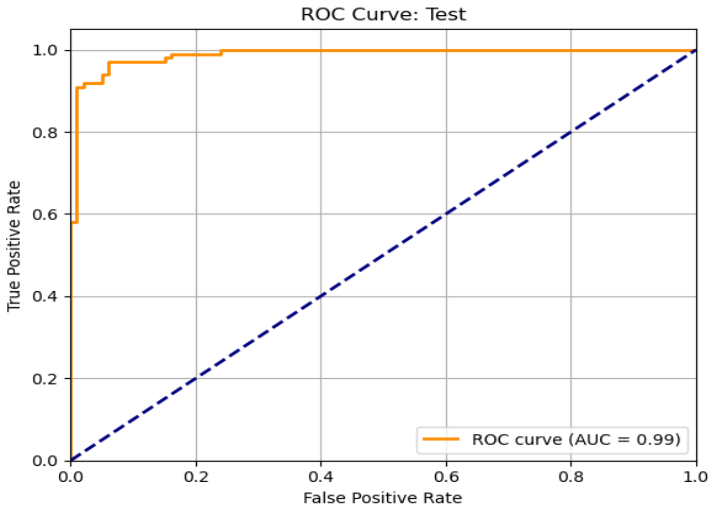


Figure 4.1.3.6: ROC curve of Xception model for deepfake detection

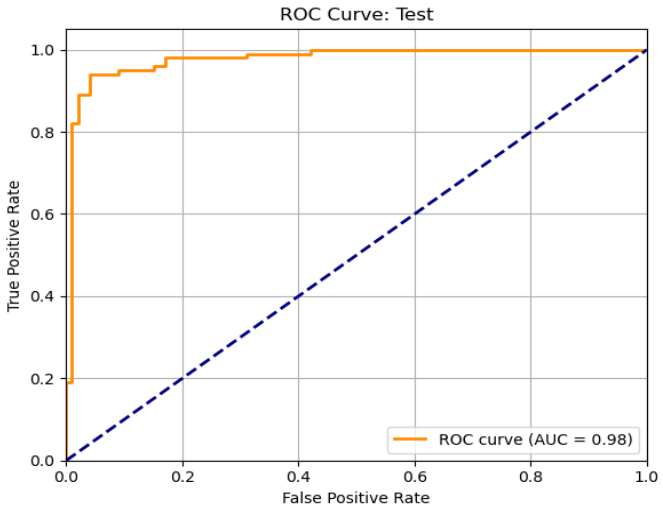


Figure 4.1.3.7: ROC curve of DenseNet121 model for deepfake detection

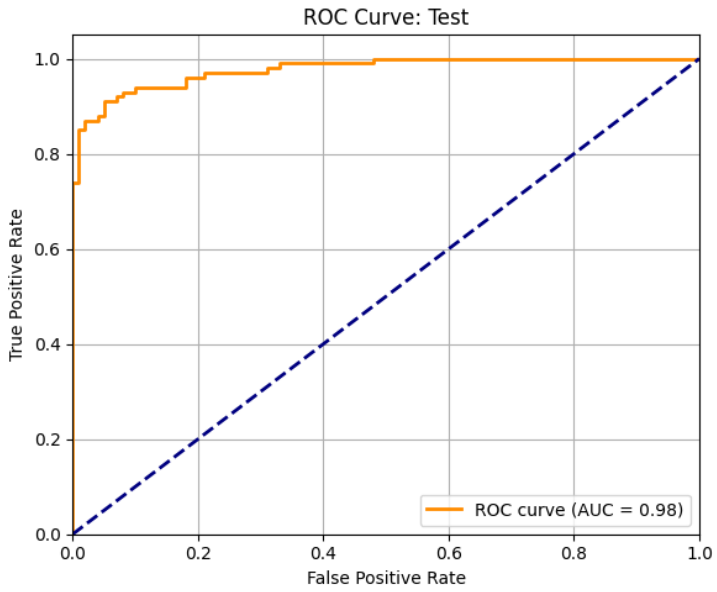


Figure 4.1.3.8: ROC curve of Inception ResNet V2 model for deepfake detection

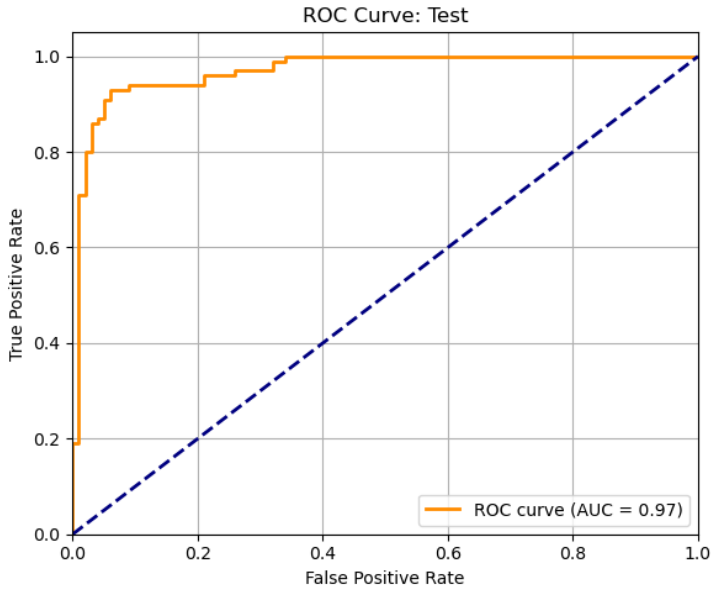


Figure 4.1.3.9: ROC curve of ResNet50 model for deepfake detection

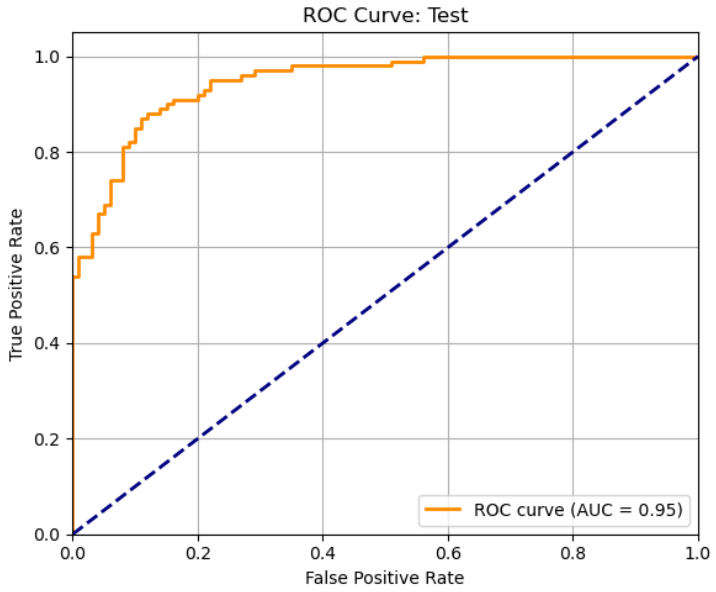


Figure 4.1.3.10: ROC curve of EfficientNet B3 model for deepfake detection

By evaluating these metrics across all five models, we can rank the architecture based on overall performance, highlight specific strengths and weaknesses of each model, and justify the selection of the best-performing architecture for deepfake detection.

4.1.4 Loss and Accuracy Trends

Observing the training process via loss and accuracy trends yields essential insights into the models' learning and generalisation efficacy. The loss curves capture the cross-entropy error over epochs, reflecting the degree to which each model minimizes misclassification, while the accuracy curves show the progression of frame-level prediction performance. Together, these curves allow for a detailed assessment of model convergence, stability, and the balance between training and validation performance. Across the five applied models—Xception, DenseNet121, Inception ResNet V2, ResNet50, and EfficientNet B3—the training and validation loss curves consistently decrease over time, confirming effective convergence. Similarly, accuracy curves steadily rise, ultimately stabilizing at high values, which demonstrates strong learning capacity and effective generalization to unseen data. For Xception, the loss steadily declines while accuracy rapidly improves, stabilizing at strong performance levels. DenseNet121 shows smooth optimization behavior, with tightly aligned training and validation curves, indicating robust learning without overfitting. Inception ResNet V2 converges gradually yet consistently,

achieving high accuracy due to its hybrid architecture that combines inception blocks with residual connections.ResNet50 leverages residual learning, which helps align training and validation trends closely, promoting stable optimization. Finally, EfficientNet B3 demonstrates rapid convergence and balanced performance, with validation accuracy closely tracking training accuracy, reflecting the strength of its compound scaling approach. Figures 4.1.4.1 to 4.1.4.5 display the loss and accuracy trends for both training and validation across the five models. Collectively, these visualizations confirm that each architecture effectively captures discriminative features, avoids severe overfitting, and achieves strong generalization on manipulated video frame classification tasks.

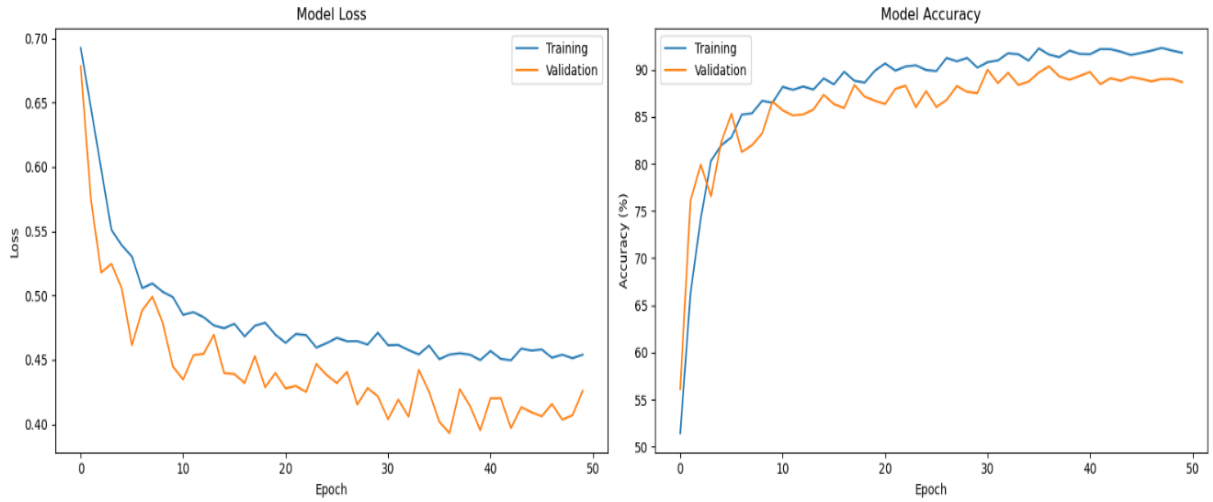


Figure 4.1.4.1: Loss and Accuracy curve of Xception model

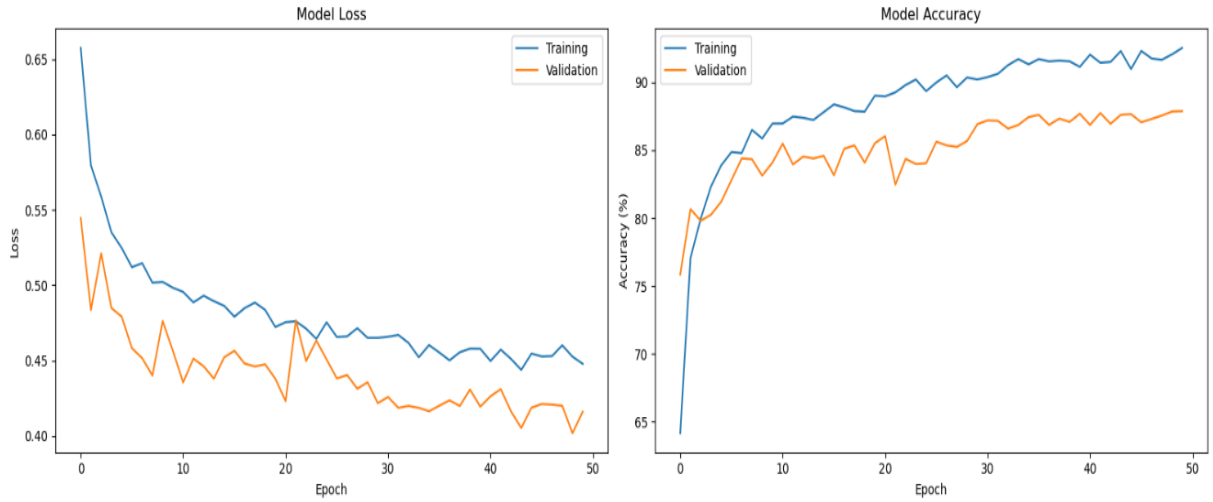


Figure 4.1.4.2: Loss and Accuracy curve of DenseNet121 model

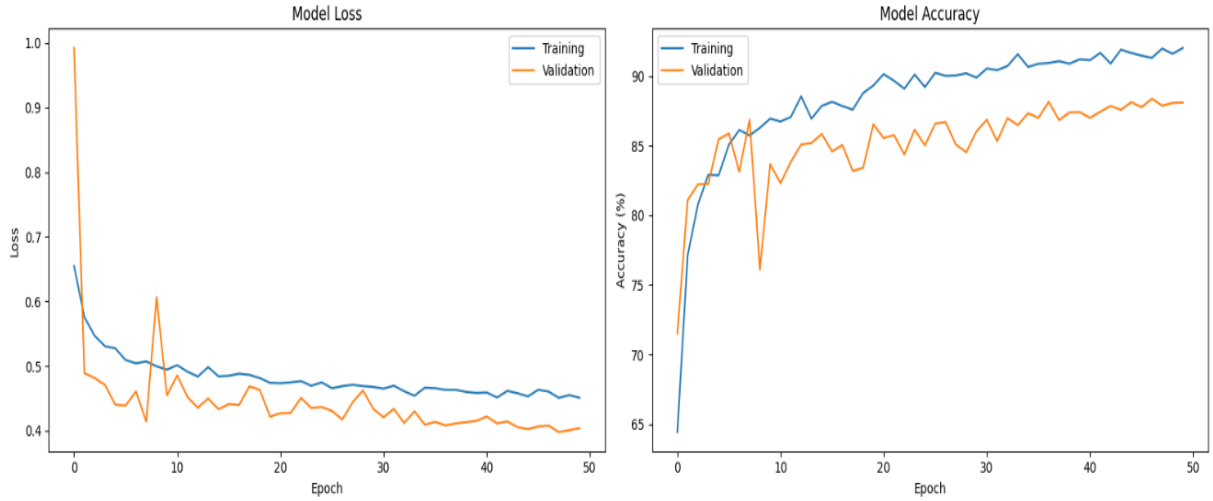


Figure 4.1.4.3: Loss and Accuracy curve of Inception ResNet V2 model

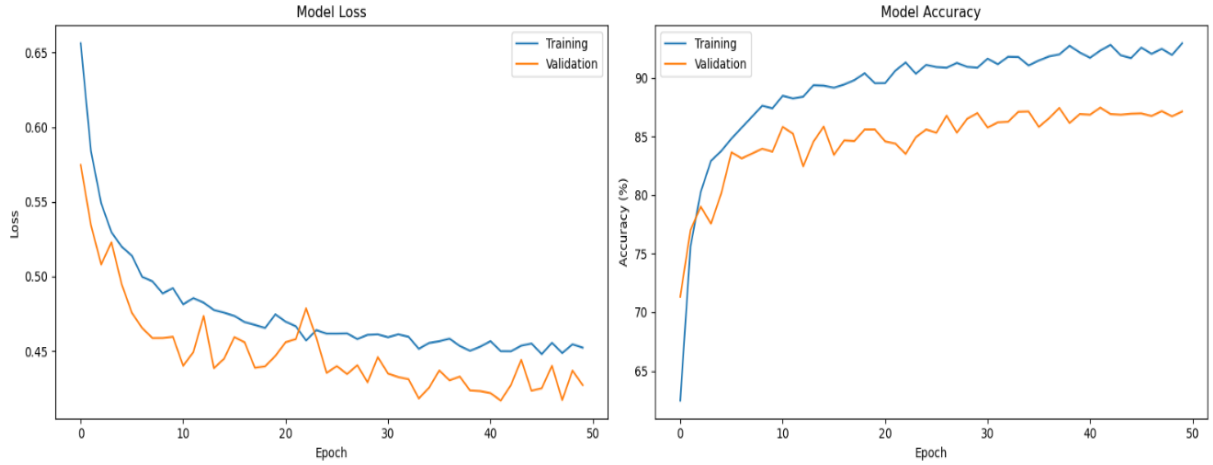


Figure 4.1.4.4: Loss and Accuracy curve of ResNet50 model

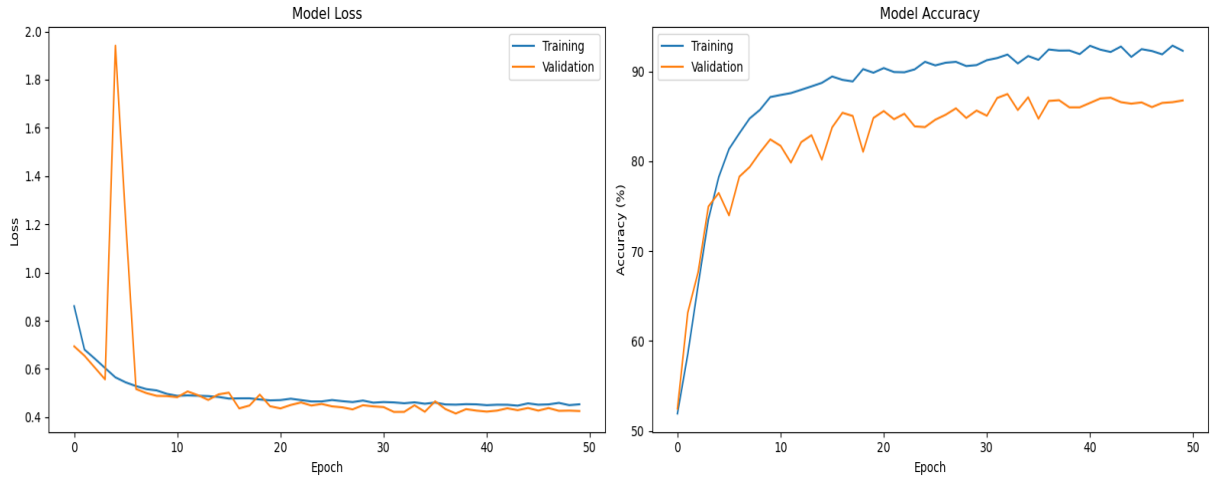


Figure 4.1.4.5: Loss and Accuracy curve of EfficientNet B3 model

4.2 Evolution Methods

After segmentation, the categorized images are analyzed via transfer learning models. The efficiency of these models is assessed using a confusion matrix, which classifies prediction outputs into four categories. True Positives (TP) denote the quantity of manipulated frames accurately recognized as manipulated, whereas False Positives (FP) signify genuine frames erroneously categorized as manipulated. FN represents altered frames mistakenly classified as real, while TN indicates genuine frames accurately recognized as authentic.

Several performance measures are derived from these values. The initial metric, Accuracy, assesses the overall correctness of the categorisation and is computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Although accuracy offers a broad perspective on performance, it may not consistently represent the model's efficacy in differentiating between genuine and altered frames, particularly in imbalanced datasets.

To address this, Precision is used, which determines the proportion of correctly predicted manipulated frames out of all frames predicted as manipulated:

$$Precision = \frac{TP}{FP + FN} \tag{2}$$

A higher precision indicates that the model makes fewer false positive errors, meaning it rarely mislabels authentic frames as manipulated.

Another important metric is Recall, which measures the ability of the model to identify actual manipulated frames. It is calculated as:

$$Recall = \frac{TP}{FP + FN} \tag{3}$$

High recall indicates that very few manipulated frames are missed, which is critical in deepfake detection where overlooking a fake instance can have serious consequences. he F1-Score provides an integrated measure of precision and recall, calculated using their harmonic mean, defined as:

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

4.3 Experimental Results & Analysis

This section will discuss the paper's findings. To evaluate the effectiveness of the proposed deepfake detection framework, we conducted extensive experiments using multiple CNN architectures, including Xception, DenseNet121, Inception-ResNet-V2, ResNet50, and EfficientNet-B3. The evaluation was performed on the FaceForensics++ test set, comprising 200 videos equally distributed between Real and Manipulated classes. Key performance metrics used for assessment include precision, recall, F1-score, and accuracy.

4.3.1 Classification Performance

The overall classification results indicate that the Xception model achieved the best performance, with an accuracy of 94.50%, demonstrating its capability to effectively distinguish between manipulated and real videos. DenseNet121 and Inception-ResNet-V2 also performed well, achieving accuracies of 93.50% and 93.00%, respectively. ResNet50 and EfficientNet-B3 achieved comparatively lower accuracies of 91.50% and 87.50%. Table 4.3.1 below summarizes the detailed classification metrics for each model:

Table 4.3.1 Classification Metrics of Deepfake Detection Models

Model	**Class**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Xception	Manipulated	0.9495	0.9400	0.9447	0.9450
	Real	0.9406	0.9500	0.9453	
DenseNet121	Manipulated	0.9394	0.9300	0.9347	0.9350
	Real	0.9307	0.9400	0.9353	

Inception-ResNet-V2	Manipulated	0.9135	0.9500	0.9314	0.9300
	Real	0.9479	0.9100	0.9286	
ResNet50	Manipulated	0.9368	0.8900	0.9128	0.9150
	Real	0.8952	0.9400	0.9171	
EfficientNet-B3	Manipulated	0.8713	0.8800	0.8756	0.8750
	Real	0.8788	0.8700	0.8744	

From the table, it is evident that Xception provides the most balanced performance across both classes, achieving high precision and recall. DenseNet121 also demonstrates strong performance, slightly lower than Xception, while EfficientNet-B3, despite having balanced macro metrics, shows the lowest overall accuracy.

4.3.2 Error Analysis

An in-depth error analysis of the Xception model was performed using the confusion matrix to gain insights into its classification behavior on the test set of 200 videos. The model correctly identified 94 manipulated videos as manipulated (true positives) and 95 real videos as real (true negatives), demonstrating strong overall performance. Nonetheless, some misclassifications occurred: five authentic videos were wrongly identified as manipulated (false positives), while six manipulated videos were incorrectly labeled as real (false negatives). The slightly higher number of false negatives indicates that the model occasionally struggles to detect subtle manipulations, particularly in low-resolution or highly compressed videos where fine facial details and manipulation artifacts are less pronounced. These observations highlight the challenges inherent in deepfake detection for low-quality video content and suggest potential avenues for further improvement, such as enhanced preprocessing or more sophisticated feature extraction techniques.

4.3.3 Comparison with Existing Methods

To validate the effectiveness of the proposed framework, we compared the fine-tuned Xception model against existing methods reported in the literature. Table 4.3.3 summarizes this comparison:

Table 4.3.3 Comparison of Deepfake Detection Methods

Paper	Methodology	Accuracy
Raza et al. [11]	DFP	94.00%
Joshi et al. [16]	Xception	93.01%
Younus et al. [17]	Haar Wavelet Transform-based	90.50%
Ghita et al. [7]	Vision Transformer (ViT)	89.91%
Proposed Method	Fine-Tuned Xception Model	**94.50%**

The comparison clearly demonstrates that the proposed Xception-based framework not only outperforms traditional methods but also surpasses recent deep learning-based approaches in terms of accuracy and balanced performance across classes.

4.3.4 Discussion

The experimental results highlight several important findings regarding the performance of the proposed deepfake detection framework. One of the most significant observations is the effectiveness of depthwise separable convolutions, which form the foundation of the Xception model. These specialized layers not only reduce computational complexity but also enable the network to capture subtle manipulation artifacts in video frames with high precision, thereby contributing to the superior performance of Xception compared to other models.

Another key insight is the importance of transfer learning. By leveraging pre-trained CNN architectures, the models were able to achieve strong generalization even when trained on a

relatively small dataset. This approach helped maintain both accuracy and robustness, underscoring the value of transfer learning in deepfake detection tasks.

The results also demonstrate that models such as Xception and DenseNet121 achieved a balanced performance across precision, recall, and F1-scores. This balance indicates their reliability in detecting both real and manipulated videos, reducing the likelihood of bias toward one class. However, despite these strengths, the study also reveals certain limitations. In particular, the models occasionally misclassified low-resolution or highly compressed videos, suggesting that there is room for improvement through the inclusion of advanced preprocessing techniques or the exploration of more sophisticated architecture.

Overall, the findings validate the robustness, reliability, and superiority of the proposed deepfake detection framework compared to existing approaches. Specifically, the results address both research questions: first, a robust CNN-based model was successfully developed to detect deepfakes, even under the challenging condition of low-resolution videos (RQ1); and second, the proposed model consistently outperformed existing approaches, thereby confirming its effectiveness and applicability in real-world scenarios (RQ2).

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Recently, deepfake technology has been rapidly growing and is becoming more common. Besides, it has serious effects on society. It damages people's confidence, privacy, and digital security. Deepfakes can help spread false information, manipulate politics, defame people, and commit fraud by making fake audio and video that look and sound very authentic. This technology makes it very hard for people to tell the difference between true and fake news by making the media less trustworthy. This feature not only makes it harder for police and cybersecurity to do their jobs, but it also makes it harder for democracy and society to work properly. To deal with these hazards, we need powerful detection methods, public awareness, and strong policies to make sure that synthetic media is used safely and ethically.

5.2 Impact on the environment

Deepfake detection models, like the Xception-based architecture used in this research, require intensive computational resources for training on large-scale manipulated video datasets. This high demand for GPU and CPU power results in significant energy consumption, contributing to carbon emissions. We can optimize the model architecture if we use efficient training strategies such as advanced data augmentation, transfer learning, and minimizing unnecessary retraining. This research aims to reduce the overall computational burden. On the other hand, deploying lightweight and efficient detection models helps lower energy use during real-time inference. As a result, this work not only addresses the growing threat of deepfakes but also supports more sustainable and environmentally conscious AI development.

5.3 Ethical Aspects

The ethical concern for deepfake detection research is related to privacy, data security, and the reliability of detection algorithms. It is very important to ensure that any personal data used in

training or evaluation is handled with strict confidentiality and protected against misuse. For building public trust and clarifying how manipulated content is identified, its very necessary to maintain transparency in algorithm design and decision-making processes. also, there is a moral responsibility to address potential biases in datasets, which could affect detection accuracy across different demographic groups and lead to unfair outcomes. Upholding fairness, interpretability, and accountability is vital to ensure that these systems are not only effective but also ethically sound and socially responsible.

5.4 Sustainability Plan

The sustainability of the proposed deepfake detection framework is crucial to ensure its long-term applicability and relevance, particularly given the rapid evolution of deepfake generation techniques. To maintain and enhance the system's effectiveness over time, several strategies have been incorporated into the sustainability plan.

First, the framework is built upon transfer learning and modular CNN architectures, which allows for easy updating and retraining of models as new datasets and manipulation techniques emerge. By leveraging pre-trained networks, future adaptations can be implemented with minimal computational overhead, ensuring the model remains relevant without requiring full-scale retraining from scratch.

Second, the system incorporates a scalable data pipeline capable of continuously ingesting new videos for testing and evaluation. This ensures that the model can adapt to real-world scenarios by gradually learning from newly encountered manipulations and video qualities. Periodic retraining with updated datasets, including emerging deepfake generation methods, is recommended to prevent model performance degradation.

Third, to promote long-term usability in diverse environments, the framework is designed to be platform-independent and resource-efficient. The use of depthwise separable convolutions in the Xception architecture reduces computational cost, enabling deployment on systems with limited hardware resources, such as edge devices or cloud-based platforms.

Finally, the sustainability plan emphasizes continuous monitoring and evaluation. Regular assessment using performance metrics such as accuracy, precision, recall, and F1-score ensures

that the model maintains its reliability and robustness. Any observed decline in performance can trigger model fine-tuning or the integration of additional architecture or features, thus safeguarding the framework's long-term effectiveness.

In summary, the sustainability plan ensures that the proposed deepfake detection framework remains adaptive, scalable, and robust against evolving threats, making it a reliable tool for preserving the integrity of digital media over time.

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1 Summary of the Study

This study focused on developing a robust and effective deepfake video detection framework using transfer learning and Convolutional Neural Network (CNN) architectures. With the growing sophistication of deepfake generation techniques, the authenticity of digital media has become increasingly vulnerable, posing threats to information integrity, security, and public trust. The primary goal of this research was to design a model capable of accurately identifying manipulated videos, including those with low resolution or compression artifacts, which are commonly encountered in real-world scenarios.

The proposed methodology involved extracting frames from input videos and processing them through a preprocessing pipeline, followed by classification using CNN-based architectures. Frame-level predictions were aggregated to generate video-level decisions. Multiple pre-trained architectures, including Xception, DenseNet121, Inception-ResNet-V2, ResNet50, and EfficientNet-B3, were evaluated to determine the most effective model. Extensive hyperparameter tuning and performance evaluation were conducted to optimize model accuracy and reliability.

Experimental results demonstrated that the Xception model achieved the highest performance, with an overall test accuracy of 94.5%. It also showed balanced precision, recall, and F1-score for both real and manipulated classes. DenseNet121 and Inception-ResNet-V2 also provided strong performance, while ResNet50 and EfficientNet-B3 delivered moderate results. Error analysis highlighted that most misclassifications occurred in low-quality or highly compressed videos, suggesting areas for future improvement.

Furthermore, the proposed framework outperformed existing methods reported in the literature, confirming its effectiveness in real-world applications. The study provides a scalable, adaptive, and reliable solution to the challenge of deepfake detection, laying a foundation for future research and practical deployment in digital media security.

6.2 Conclusions

The rapid evolution of deepfake technologies has created significant challenges for maintaining the authenticity and trustworthiness of digital media. This study addressed these challenges by developing a robust deepfake video detection framework leveraging transfer learning and CNN-based architectures. Through systematic experimentation and comparative analysis, the research demonstrates that deep learning models can effectively identify manipulated content, even in low-resolution videos where traditional detection methods often fail.

Among the architectures tested, the Xception model emerged as the most effective, achieving an overall test accuracy of 94.5% with balanced precision, recall, and F1-scores across both real and manipulated classes. The model's superior performance is largely attributed to its depthwise separable convolutional layers, which efficiently capture subtle spatial features and manipulation artifacts while reducing computational complexity. Other architectures, including DenseNet121 and Inception-ResNet-V2, also delivered strong results, confirming the effectiveness of transfer learning and CNN-based approaches in deepfake detection.

Error analysis revealed that misclassifications were primarily associated with low-quality or highly compressed video frames, indicating areas for potential improvement in preprocessing and feature extraction techniques. Comparative evaluation with existing methods further validated the superiority of the proposed framework, highlighting its ability to outperform recent state-of-the-art approaches in accuracy and robustness.

In conclusion, this study establishes that transfer learning-based CNN architectures provide a reliable and scalable solution for deepfake video detection. The proposed framework not only addresses current challenges in identifying manipulated content but also offers a foundation for future advancements in digital media security, making it a valuable contribution to the field.

6.3 Implication for Further Study

While the proposed deepfake detection framework has demonstrated strong performance, several avenues for future research could further enhance its effectiveness and applicability. One key direction is multi-dataset evaluation. Future studies should extend the framework to incorporate diverse datasets that include a variety of manipulation techniques, resolutions, and

real-world video conditions. By evaluating the model across multiple sources, researchers can improve its generalization and robustness, ensuring consistent performance under different scenarios.

Another important area for advancement lies in enhanced preprocessing and feature extraction. Low-resolution and highly compressed videos pose significant challenges for detection. Incorporating advanced preprocessing techniques such as super-resolution, noise reduction, and temporal feature analysis could improve the model's ability to detect subtle manipulations that are otherwise difficult to identify.

Given the rapid evolution of deepfake generation methods, adaptive and continuous learning represents another promising direction. Future models could incorporate online or incremental learning strategies, allowing them to dynamically adapt to new types of manipulations without requiring full retraining. This would ensure that detection systems remain up to date as new deepfake techniques emerge.

Multi-modal approaches could also enhance detection performance. By integrating audio cues, metadata, and behavioral signals alongside visual information, models may be better equipped to identify sophisticated deepfakes that exploit non-visual channels.

Additionally, there is potential in developing lightweight and real-time detection architectures. Optimizing models for deployment on edge devices or mobile platforms could expand the practical applicability of deepfake detection, enabling real-time monitoring and media verification in various environments.

Finally, future work should focus on robust evaluation metrics. This includes cross-dataset testing and stress-testing under extreme video conditions, which can ensure that detection models remain reliable in real-world scenarios. By addressing these areas, future research can strengthen deepfake detection systems, making them more adaptive, accurate, and applicable across diverse contexts. Incorporating multiple datasets will enhance the framework's generalization capabilities, ensuring its effectiveness against a wide spectrum of manipulation techniques.

Reference

[1] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1-11).

[2] Kim, E., & Cho, S. (2021). Exposing fake faces through deep neural networks combining content and trace feature extractors. IEEe Access, 9, 123493-123503.

[3] Basit, N., Khalid, F., Ain, Q. U., & Andleeb, M. (2025, February). Faceswap Finder: A Fusion-Based Deepfake Detection Technique. In 2025 6th International Conference on Advancements in Computational Sciences (ICACS) (pp. 1-6). IEEE.

[4] Alrawahneh, A. A. M., Abdullah, S. N. A. S., Abdullah, S. N. H. S., Kamarudin, N. H., & Taylor, S. K. (2025). Video authentication detection using deep learning: a systematic literature review. Applied Intelligence, 55(4), 239.

[5] Ying, L. X., Aman, M., & Hafizah, A. (2023). Malaysia Cyber Fraud Prevention Application: Features and Functions. Asia-Pacific Journal of Information Technology & Multimedia, 12(2).

[6] Haliassos, A., Mira, R., Petridis, S., & Pantic, M. (2022). Leveraging real talking faces via self-supervision for robust forgery detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14950-14962).

[7] Lanzino, R., Fontana, F., Diko, A., Marini, M. R., & Cinque, L. (2024). Faster than lies: Real-time deepfake detection using binary neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3771-3780).

[8] Liu, Z., Qi, X., & Torr, P. H. (2020). Global texture enhancement for fake face detection in the wild. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8060-8069).

[9] Li, Y., Chang, M. C., & Lyu, S. (2018, December). In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In 2018 IEEE International workshop on information forensics and security (WIFS) (pp. 1-7). Ieee.

[10] Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(2), e1520.

[11] Raza, A., Munir, K., & Almutairi, M. (2022). A novel deep learning approach for deepfake image detection. Applied Sciences, 12(19), 9820.

[12] Chang, X., Wu, J., Yang, T., & Feng, G. (2020, July). Deepfake face image detection based on improved VGG convolutional neural network. In 2020 39th chinese control conference (CCC) (pp. 7252-7256). IEEE.

[13] Hsu, C. C., Zhuang, Y. X., & Lee, C. Y. (2020). Deep fake image detection based on pairwise learning. Applied Sciences, 10(1), 370.

[14] Coccomini, Davide Alessandro, et al. "Cross-forgery analysis of vision transformers and cnns for deepfake image detection." Proceedings of the 1st International Workshop on Multimedia AI against Disinformation. 2022.

[15] Ghita, B., Kuzminykh, I., Usama, A., Bakhshi, T., & Marchang, J. (2024, June). Deepfake image detection using vision transformer models. In 2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom) (pp. 332-335). IEEE.

[16] Joshi, P., & Nivethitha, V. (2024, April). Deep fake image detection using Xception architecture. In 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST) (pp. 533-537). IEEE.

[17] Younus, M. A., & Hasan, T. M. (2020, April). Effective and fast deepfake detection method based on haar wavelet transform. In 2020 International Conference on Computer Science and Software Engineering (CSASE) (pp. 186-190). IEEE.

[18] Khan, A. A., Laghari, A. A., Inam, S. A., Ullah, S., Shahzad, M., & Syed, D. (2025). A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. Discover Computing, 28(1), 48.

[19] Shi, L., Zhang, J., Ji, Z., Bai, J., & Shan, S. (2025). Real face foundation representation learning for generalized deepfake detection. Pattern Recognition, 161, 111299.
[20] Liu, K., Perov, I., Gao, D., Chervoniy, N., Zhou, W., & Zhang, W. (2023). Deepfacelab: Integrated, flexible and extensible face-swapping framework. Pattern Recognition, 141, 109628.
[21] Mehrjardi, F. Z., Latif, A. M., Zarchi, M. S., & Sheikhpour, R. (2023). A survey on deep learning-based image forgery detection. Pattern Recognition, 144, 109778.
[22] Lin, L., He, X., Ju, Y., Wang, X., Ding, F., & Hu, S. (2024). Preserving fairness generalization in deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16815-16825).
[23] Huang, Z., Hu, J., Li, X., He, Y., Zhao, X., Peng, B., ... & Cheng, G. (2025). Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 28831-28841).
[24] Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019, May). Self-attention generative adversarial networks. In International conference on machine learning (pp. 7354-7363). PMLR.
[25] Ilyas, H., Irtaza, A., Javed, A., & Malik, K. M. (2022, December). Deepfakes examiner: An end-to-end deep learning model for deepfakes videos detection. In *2022 16th international conference on open source systems and technologies (ICOSST)* (pp. 1-6). IEEE.
[26] Lin, Y. H., & Xu, Y. S. (2024, June). Training Deepfake Detection Model from Photos with Face Mask. In *2024 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-6). IEEE.
[27] Park, J., Cho, D., Ahn, W., & Lee, H. K. (2018). Double JPEG detection in mixed JPEG quality factors using deep convolutional neural network. In Proceedings of the European conference on computer vision (ECCV) (pp. 636-652).
[28] Wang, Q., & Zhang, R. (2016). Double JPEG compression forensics based on a convolutional neural network. EURASIP Journal on Information Security, 2016(1), 23.
[29] Ganguly, S., Ganguly, A., Mohiuddin, S., Malakar, S., & Sarkar, R. (2022). ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection. *Expert Systems with Applications*, *210*, 118423.
[30] Saxena, A., Yadav, D., Gupta, M., Phulre, S., Arjariya, T., Jaiswal, V., & Bhujade, R. K. (2023). Detecting deepfakes: A novel framework employing XceptionNet-based convolutional neural networks. *Traitement du Signal*, *40*(3).

710003712

ORIGINALITY REPORT

11%	7%	6%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	1%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	1%
4	www.mdpi.com Internet Source	1%
5	Shankar Babu, Mahesh Babu Kota. "Synergies in Smart and Virtual Systems using computational intelligence", CRC Press, 2025 Publication	1%
6	fastercapital.com Internet Source	<1%
7	Ahsan, Sevinj Aliyeva. "Prediction of Covid-19 Using Procedures of Transfer Learning.", Khazar University (Azerbaijan), 2024 Publication	<1%
8	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1%
9	www.nature.com Internet Source	<1%