RESEARCH ARTICLE

DOI: 10.34256/ijcci2512

International Journal of Computer Communication and Informatics

Comparative Analysis on Different Deepfake Detection Techniques

Ashutosh Sharma [a,*], Aryan Dutt [a], Arpit Rastogi [a], Avinash Ratre [a]

[a] Department of Electronics & Communication Engineering, Delhi Technological University, India

* Corresponding Author: ashutoshsharma_ec20b12_32@dtu.ac.in

Received: 17-12-2024, Revised: 22-03-2025, Accepted: 31-03-2025, Published: 08-04-2025

Abstract: Advancements in deep learning have led to the emergence of highly realistic AI-generated videos known as deepfakes. These videos utilize generative models to expertly modify facial features, creating convincingly altered identities or expressions. Despite their complexity, deepfakes pose significant threats by potentially misleading or manipulating individuals, which can undermine trust and have repercussions on legal, political, and social frameworks. To address these challenges, researchers are actively developing strategies to detect deepfake content, essential for safeguarding privacy and combating the spread of manipulated media. This article explores current methods for generating deepfake images and videos, with a focus on facial features and expression alterations. It also provides an overview of publicly available deepfake datasets, crucial for developing and evaluating detection systems. Additionally, the research examines the challenges associated with identifying deepfake face swaps and expression changes, while proposing future research directions to overcome these hurdles. By offering guidance to researchers, the document aims to foster the development of robust solutions for deepfake detection, contributing to the preservation of the integrity and reliability of visual media.

Keywords: Deepak, CNN, Transformer, AUC, ViT, Dataset

1. Introduction

Lately, there has been rapid progress in the domain of Artificial Intelligence, particularly in deep learning. This advancement has empowered deep learning algorithms to handle vast datasets more effectively, propelling significant advancements in areas such as computer vision, NLP, and speech recognition. Despite these strides, deep learning is still in its early phases, prompting researchers to continuously refine its accuracy and efficiency. Concurrently, the proliferation of deepfake images, artificially created visuals using machine learning algorithms capable of altering the appearance and actions of individuals in videos and images—has become increasingly widespread. This surge has sparked considerable uncertainty surrounding the authenticity and trustworthiness of digital media, especially within political, journalistic, an

entertainment contexts. Identifying deepfakes is crucial to counter ongoing issues. Therefore, we'll explore a novel deep learning approach that distinguishes artificially generated hoax videos (deepfake videos) from authentic ones. Establishing a method to detect deepfakes is a significant challenge aimed at recognizing and curbing their dissemination online. To recognize a deepfake, understanding how the Generative Adversarial Network (GAN) creates it is essential. GAN takes an image ("target") and a video of a specific individual as input, generating another video ("source") that swaps the target person's face with a different face. The foundation of a deepfake involves a sophisticated adversarial neural network trained on facial images and target footage to seamlessly match the source face to the intended target. With appropriate post-processing, the resulting video can achieve a high level of authenticity. The GAN dissects the video into individual frames, substituting the initial image within each frame while also replicating the video. Typically, this process involves an autoencoder. Explore innovative deep learning approaches to accurately differentiate between deepfake (DF) videos and authentic ones. The complex process employed by GAN to generate a deepfake. The architecture is depicted in Figure 1.

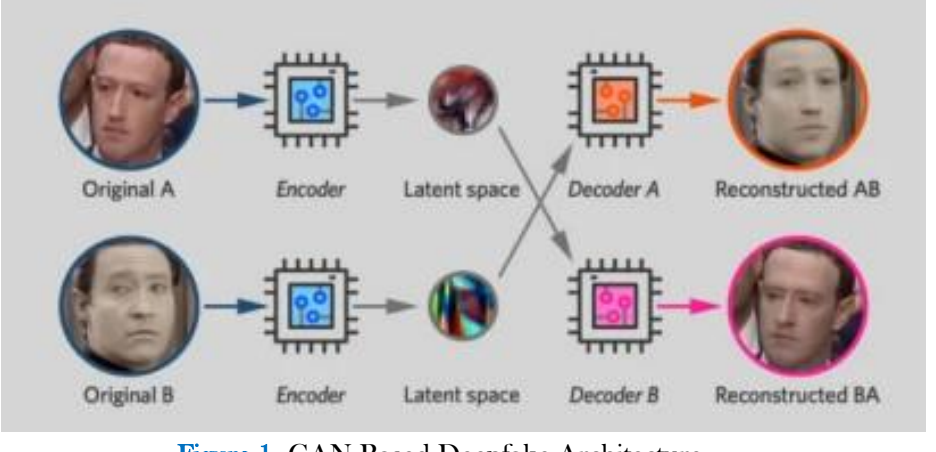

Figure 1. GAN Based Deepfake Architecture

The quest to detect these sophisticated manipulations in digital content is propelled by the pressing need to safeguard the integrity and credibility of media. Current detection systems, though valuable, fall short in their ability to discern intricate alterations, especially in audio and facial domains. This inadequacy highlights the urgency for more robust and comprehensive methodologies that can effectively scrutinize various modalities of content to root out falsified elements. We will compile a dataset containing both authentic and deepfake videos for a thorough examination of cutting-edge deepfake detection models. Our dataset will include voice-swapped, voice-cloned, and face-swapped deepfake videos to enhance their comprehensiveness and difficulty for detection. We will conduct a qualitative analysis by comparing the architecture, features, and detection techniques of chosen deepfake detection models. Additionally, we will quantitatively assess the performance of these models on a comprehensive dataset, using widely

recognized evaluation metrics such as accuracy, precision, recall, and F1 score. A schematic representation of deepfake detection system is demonstrated in Figure 2.

In this pursuit, the ultimate goal is not just to identify and flag manipulated content but to devise proactive measures to preemptively thwart falsified information's spread and impact. By continuously refining detection systems, researchers aim to establish a robust defense mechanism that safeguards the authenticity and credibility of digital content in an increasingly sophisticated landscape of misinformation and deceit.

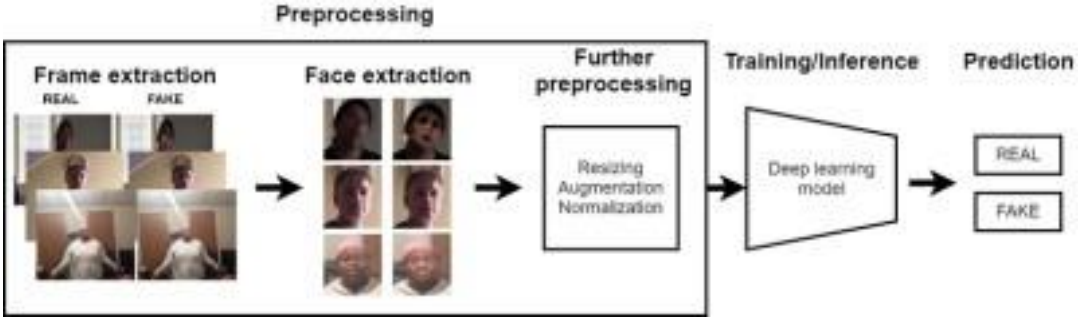

Figure 2. General Block Diagram of Deepfake Detection

2. Recent Work

The neologism "deepfake" signifies the synthesis of manipulated videos through the convergence of advanced computer vision algorithms and deep learning techniques. Characterized by their realism, these fabricated videos primarily encompass some categories: face-swapping, where a target individual's facial features are seamlessly replaced with another's, and face reenactment, where expressions and movements are emulated onto a different individual's visage.

Two primary generative approaches exist for obtaining realistic faces in videos: Generative Adversarial Networks (GANs) [1] and Variational AutoEncoders (VAEs) [2]. GANs function through a dichotomous network structure: a discriminator striving to discern reality from fabricated videos, and a generator adept at manipulating videos with convincing realism to outwit the discriminator. This framework has yielded highly authentic outcomes, prompting the development of diverse variations like StarGAN [3] and DiscoGAN [4].

Deepfake detection techniques can be broadly categorized into three groups based on their feature extraction methodologies which are:

2.1 Visual Feature Dependence Detection

Focuses on analyzing inconsistencies in readily observable facial features. This includes variations in blink patterns, head pose movements, and discrepancies in facial organ shapes [5].

Initial research by Yang et al within this category explored eye blink patterns as the primary feature for detection, hypothesizing a difference in blink patterns between the original video and the manipulated deepfake [6]. Subsequent work investigated head pose inconsistencies, examining misalignments not only between facial features but also between the head and other body parts like the neck and shoulders. This approach often leverages facial landmark detection techniques to quantify these inconsistencies (e.g., using 68 facial landmarks to estimate head pose variations).

Another category centers on Visual Artifact Detection, which aims to exploit imperfections arising from limitations in deepfake creation resources [7]. These imperfections, referred to as visual artifacts, can manifest as color inconsistencies between facial features (e.g., eye color differences), unnatural shadows (particularly around the nose), a lack of detailed light reflections, or a reduction in the geometric complexity of facial structures like teeth. Color and geometry extraction techniques applied to specific facial regions (eyes, nose, lips, etc.) can be employed to identify these visual artifacts as shown in Fig 3. As mentioned by Hady A. Khalil and Shady A, this approach offers initial effectiveness, its utility diminishes as deepfake generation techniques become more sophisticated and capable of producing increasingly realistic content.

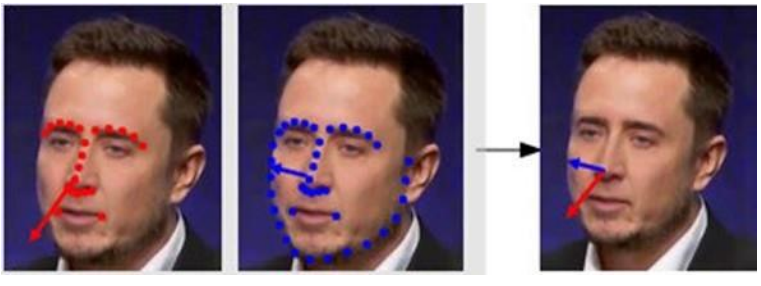

Figure 3. General Block Diagram of Deepfake Detection

2.2 Seni-Local Characterstics

It represents another prominent approach, extracting features directly from individual image pixels through techniques like pixel-based segmentation. This method offers advantages in reliability compared to analyzing readily observable visual features. Early research within this category combined features derived from image convolution with those obtained from steganalysis to pinpoint manipulated regions within facial images. These initial efforts paved the way for the development of more sophisticated local and deep feature-based deepfake detection methods. Additional feature extraction techniques employed in this domain include Pyramid of Histogram of Oriented Gradients (PHOG), Local Phase Quantization (LPQ), and Local Binary Pattern (LBP). While research suggests that methods like Image Quality Metric (IQM) can be effective in identifying deepfakes when compared to techniques like Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), the evolving nature of deepfake

generation algorithms necessitates the exploration of increasingly complex features for robust differentiation between original and manipulated content.

2.3 Deep Learning Based Frame and Temporal Features

Shares similarities with local feature extraction but leverages deep learning architectures with multiple layers to extract more intricate features from individual pixels. This enables the capture of complex relationships within the data compared to simpler methods . Convolutional Neural Networks (CNNs) like DenseNet, InceptionNet, and XceptionNet have been explored for deepfake image detection, demonstrating promising results [5]. Rossler et al. employed XceptionNet to identify deepfakes within the FF++ dataset [8]. Expanding upon this concept, Li et al. introduced a ResNet-50 architecture with a Spatial Pyramid Pooling layer, termed DeepFD, achieving good performance against Generative Adversarial Network (GAN)-generated images [9].

Beyond deep learning approaches, researchers have explored exploiting inherent differences between real and fake videos through preprocessing techniques. For instance, Agarwal et al. investigated "micro-expression" features, focusing on the inconsistencies in facial action units between different individuals. While offering robustness against compression and noise, this method is limited to specific individuals.

Building upon previous efforts, recent studies have explored Vision Transformers for deepfake detection. Notably, this work achieves promising results by fusing a convolutional network, responsible for extracting facial patches from videos, with a Transformer architecture. This combined approach demonstrates its effectiveness in identifying manipulated content [10].

The state-of-the-art was further elevated by through a knowledge distillation approach. By extracting knowledge from a pre-trained EfficientNet B7 network fine-tuned on the DFDC dataset, they transferred it to a Vision Transformer model. Specifically, patch features from both the EfficientNet B7 and Vision Transformer were combined via global pooling and fed into the Transformer encoder. Additionally, a "distillation token" was injected into the Transformer network, facilitating the transfer of learned knowledge from the EfficientNet B7. This innovative approach resulted in further advancements in deepfake detection performance.

2.3.1 Datasets

In any deep learning application, it's crucial to utilize a sizable, comprehensive, and top-notch dataset to, among other purposes, prevent overfitting. This means the application functions effectively beyond just the training data.

There were several rationales behind this selection. Firstly, its accessibility in the public domain is noteworthy. It offers comprehensive documentation regarding both the dataset itself and its application. Secondly, a predominant characteristic of the extracted images is the presence

of a solitary, unobstructed frontal face, facilitating straightforward tracking. Thirdly, it encompasses videos generated through an extensive array of prevalent deepfake video production techniques. This implies that the eventual deepfake detection solution should be capable of accommodating outputs from a diverse spectrum of deepfake generation methodologies, rather than being tailored to a singular source. Fourthly, Datasets like Face Forensics++ offers a variety of video quality options for data download, catering to constraints related to both time and bandwidth. The presence of multiple quality levels is crucial for deepfake detection, as distinguishing a high-resolution deepfake from a lower-resolution counterfeit video is comparatively simpler. Some used datasets are

- FaceForensics++ offers a dataset comprising 1000 original video sequences, subject to manipulation through four distinct automated face alteration methods: Face2Face, FaceSwap, NeuralTextures, and Deepfakes. The 5000 videos produced were created with the University of Melbourne's SPARTAN High Performance Computing System. Employing a 23x compression rate facilitated efficient utilization of time and storage resources during the downloading process for all h264 videos. The dataset encompasses five distinct sets of videos, delineated in Table 1, encompassing both the original videos and their corresponding deepfake iterations generated via Deepfakes, Face2Face, FaceSwap, and NeuralTextures methodologies.
- The Google DFD dataset was generated using undisclosed deepfake generation technologies, potentially contributing to its omission from consideration in the majority of previous research endeavours. Despite certain prior studies utilizing CNN architecture and achieving AUC performance surpassing a specified threshold, this dataset's exclusion from extensive examination suggests a need for further investigation and scrutiny.
- The DFDC Preview dataset, an earlier iteration and subset of the comprehensive DFDC dataset, has posed challenges for researchers attempting to consistently achieve high detection rates. During the Kaggle DFDC Challenge, participants primarily endeavored to devise ensembles of detection models based on deep CNN architectures within the confines of a 9-hour testing time limit. However, a comprehensive and systematic analysis of CNN models specifically tailored to the DFDC dataset was largely unexplored. There exists a pressing need for further research to meticulously develop and evaluate the effectiveness of individual models, as opposed to relying solely on ensemble methods.

3. Convolutional Neural Networks

Current research in the fields of computer vision, image processing, and NLP underscores the remarkable efficacy of Convolutional Neural Networks (CNNs) due to their

potent learning capabilities. These networks are characterized by their multi-stage feature extraction architecture, enabling them to autonomously learn data representations and effectively capture spatiotemporal dependencies within signals. Ongoing research efforts primarily concentrate on exploring novel activation and loss functions, parameter optimization techniques, regularization methods, and most crucially, architectural innovations for CNNs. Notably, recent advancements in CNN architecture have yielded significant enhancements in representational capacity.

LeNet, introduced in 1989, marked the pioneering CNN architecture, employing backpropagation for handwritten zip code recognition [11]. Subsequent years witnessed significant advancements, exemplified by AlexNet, championing both classification and localization tasks at the 2012 Large Scale Visual Recognition Challenge. This achievement was credited to its deeper architecture and enhanced channel consideration. InceptionNet further evolved the field in 2014 by introducing multi-scale feature extraction and increasing model width through parallel varying kernel sizes, deviating from the solely depth-focused approach. The same year, VGGNet (Visual Geometry Group at University of Oxford) demonstrated the efficacy of extremely small (3x3) convolution filters, achieving depths of 16-19 weight layers, surpassing prior works. These milestones collectively illustrate the continuous progress in CNN architecture design.

Addressing the vanishing gradient problem hindering deep CNNs, ResNet emerged in 2016 [12]. Its innovative residual connections offered alternate gradient paths bypassing intermediate layers, enabling the training of exceptionally deep models with superior performance. The following year, XceptionNet [13] drew inspiration from InceptionNet, replacing its modules with depthwise separable convolutions. Evaluations on ImageNet demonstrated XceptionNet's outperformance compared to InceptionNet-V3. Continuing the trend, EfficientNet, proposed in 2019, focused on balancing network depth, width, and resolution. Notably, it introduced a compound coefficient method for uniform scaling across all three dimensions.

Building upon the EfficientNet design principles, Tan and Le introduced a series of eight scaled architectures, termed EfficientNet B0 to B7 [14]. Evaluations on the ImageNet dataset revealed that EfficientNet B7 surpassed established models like InceptionNet-V4 (80.0%), XceptionNet (79.0%), ResNet152 (77.8%), and ResNet50 (76.0%), achieving a remarkable top-1 accuracy of 84.3%. This feat underscores the эффективность of the balanced scaling approach employed in EfficientNet. Subsequently, in 2020, Sun et al. proposed HRNet, a novel architecture prioritizing the preservation of high-resolution representations. This design choice aimed to provide a more robust backbone for position-sensitive tasks in computer vision, including human pose estimation, object detection, and semantic segmentation.

4. Transformer

The Transformer architecture (Figure 4) has emerged as a pivotal model for neural sequence transduction tasks, marked by its encoder-decoder structure. Originally introduced for natural language processing, Transformers have since widened their applicability to various domains, including computer vision. Unlike traditional convolutional neural networks (CNNs), Transformers operate on sequences of symbol representations, facilitating a more comprehensive understanding of input data. In the context of deepfake detection, this architecture proves promising due to its ability to discern intricate patterns and nuances within sequences of images or video frames, enabling effective discrimination between authentic and manipulated.

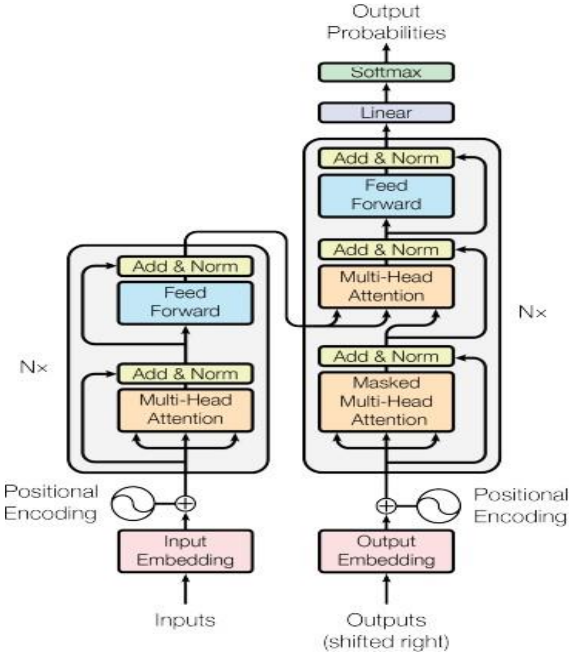

Figure 4. Transformer-Architecture

The evolution of Transformer models in the field of computer vision has catalyzed significant advancements in deepfake detection methodologies. Vision Transformer (VIT), introduced in 2020, demonstrated the efficacy of applying Transformer architectures directly to sequences of image patches for image classification tasks. This breakthrough marked a departure from conventional CNN-based approaches, showcasing the Transformer's potential in visual data analysis. Subsequent innovations, such as the Bidirectional Encoder representation from image Transformers (BEiT) and the Swin Transformer, further enhanced the capabilities of Transformer-based models in image understanding tasks, achieving remarkable performance on benchmarks like ImageNet. These developments underscore the adaptability and versatility of

Transformer architectures in tackling complex visual data manipulation scenarios, including deepfake generation and detection.

Recent research efforts have extended the Transformer paradigm to address specific challenges inherent in deepfake detection. Models like Class-attention in image Transformers (CaiT) have been designed with specialized attention mechanisms tailored for image classification tasks, optimizing the processing of image patches while enhancing feature extraction and classification accuracy. By integrating learnable diagonal matrices and separating transformer layers for self-attention and class attention, CaiT demonstrates a sophisticated approach to guiding attention processes and extracting meaningful features from manipulated image data. Such advancements signify a paradigm shift in the application of Transformer architectures, offering new avenues for robust deepfake detection systems capable of mitigating the proliferation of synthetic media for malicious purposes.

5. Methodology

In this study, we outline our methodology for developing and evaluating deep learning models for the detection of deepfake content. Initially, we conducted a systematic partitioning of each of the five datasets into distinct sets for training, validation, and testing, ensuring non-overlapping data subsets. For the DFDC dataset, which already provided predefined splits, we adhered to these partitions. However, for the remaining datasets, we ensured a split where approximately 15% of the data was allocated for validation and 15% for testing, while the remaining 70% was designated for training. Furthermore, we meticulously maintained the integrity of source videos used in deepfake creation within the same split, thereby preserving consistency across real and fake distributions while adhering to the specified minimum percentages.

Subsequently, we conducted frame extraction from both authentic and deepfake videos, followed by facial detection and extraction processes to establish balanced distributions of real and fake instances within the training and validation datasets. The testing dataset retained its original video format, and during the detection phase, we extracted video frames and facial regions for analysis to obtain test results. The results obtained from the analyzed frames were averaged for each video.

Convolutional Neural Networks (CNNs) exhibit a capability to discern spatial intricacies within data, rendering them well-suited for the analysis of visual media such as photographs and videos. Within the domain of deepfake detection, CNNs prove instrumental in scrutinizing video frame data to discern distinctive features distinguishing deepfake content from authentic footage. For instance, CNNs can adeptly identify subtle variations in skin texture or facial appearance indicative of deepfake manipulation. The utilization of CNNs involves processing preprocessed images to ascertain the authenticity of video content. This process entails training the model on

a dataset comprising both genuine and synthetic videos, followed by employing a CNN architecture to classify individual frames as fake or real.

The algorithmic procedure encompasses the following steps:

Model Training: The CNN model is trained utilizing a dataset comprising genuine and fabricated videos to facilitate learning discriminative features.

Frame Extraction and Preprocessing: Frames are extracted from the videos and subjected to preprocessing techniques aimed at noise reduction and enhancing clarity.

CNN Processing: Preprocessed frames are fed into the CNN model, which extracts pertinent features crucial for discerning between authentic and manipulated content.

Feature Extraction: Utilizing a pre-trained CNN model, hidden features are extracted from the input frames, leveraging established architectures such as MesoNet or other suitable alternatives.

Classification: Post-feature extraction, the model classifies videos based on the authenticity of their content, thereby discerning between genuine and falsified footage.

This systematic approach underscores the efficacy of CNNs in discerning deepfake content, thereby contributing to the advancement of robust detection methodologies

5.1 ResNet-50

A convolutional neural network architecture known in Figure 5 as ResNet-50 has been applied in the identification of deepfake videos. Initially, it utilizes multiple convolutional layers to learn and extract features from input images, followed by fully connected layers for classification purposes. ResNet-50 is adept at analyzing video frames and discerning characteristics that distinguish deepfake content from authentic ones within the realm of deepfake video detection. To develop a broad understanding of features, the network is typically pre-trained on extensive datasets comprising genuine images. These generalized features are further refined through fine-tuning on a smaller dataset that includes both authentic and deepfake videos, specifically tailored for deepfake identification tasks. One of the principal advantages of ResNet-50 lies in its capability to grasp features at diverse scales, facilitating the capture of nuanced details within the data, which is crucial for accurate deepfake identification given the subtle discrepancies between genuine and fake videos. The utilization of ResNet-50 for deepfake detection generally involves several steps: (i) assembling a dataset consisting of both genuine and fraudulent videos along with appropriate labelling, (ii) preprocessing the data by extracting frames from the videos and normalizing pixel values, (iii) by processing each video frame through the ResNet-50 model, features are extracted, resulting in a high-dimensional feature vector that captures the fundamental elements of the input frame, (iv) aggregating temporal information by combining feature vectors of all frames using techniques like max or average pooling to create a

fixed-length representation of the video, (v) employing a classifier to ascertain the authenticity of the fixed-length video representation through processing, and (vi) training the model involves utilizing labeled data for backpropagation and gradient descent, followed by evaluation on a separate test dataset using metrics such as accuracy, precision, and recall.

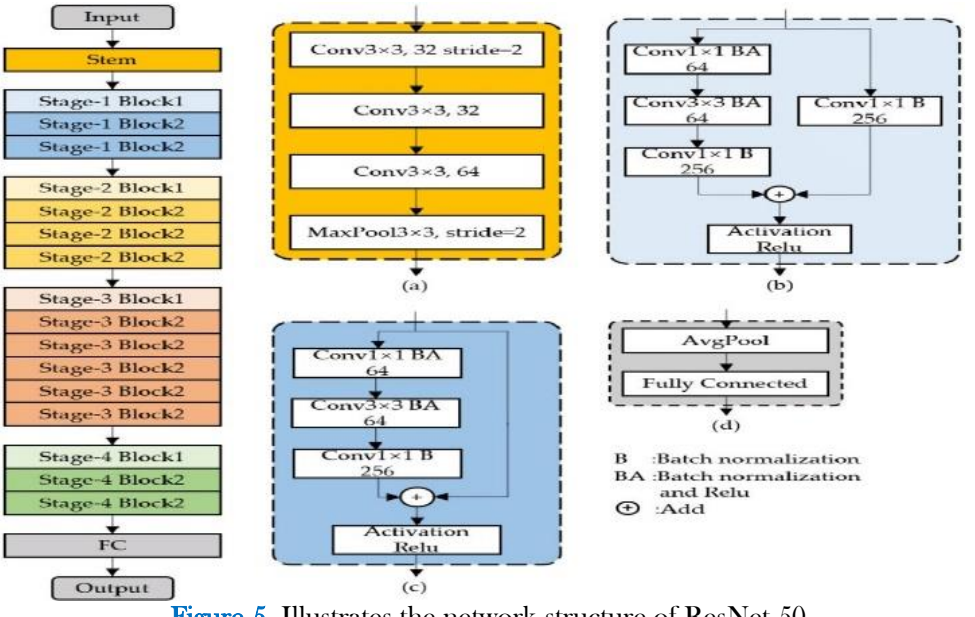

Figure 5. Illustrates the network structure of ResNet-50

5.2 Meso-Net 4

Our experimentation commenced with intricate architectures, progressively simplifying them until arriving at a streamlined model that delivers equivalent outcomes with enhanced efficiency. The devised network initiates with a series of 4 layers pooling operations and successive convolutions, succeeded by a dense network featuring a solitary layer which is hidden. To bolster generalization capabilities, ReLU activation functions are applied within the convolutional layers to introduce non-linearities, while Batch Normalization techniques are employed to regularize their outputs and mitigate the vanishing gradient phenomenon. Additionally, Dropout mechanisms[15] are incorporated within the fully connected layers to bolster regularization efforts and fortify their resilience against overfitting.

5.2.1 Mesoinception-4

A viable modification involves substituting the initial pair of convolutional layers within Meso4 with a variation of the inception module pioneered by Szegedy *et al* [16]. The core concept of this module entails amalgamating the outputs of multiple convolutional layers

featuring diverse kernel shapes, thereby expanding the functional space within which the model operates.

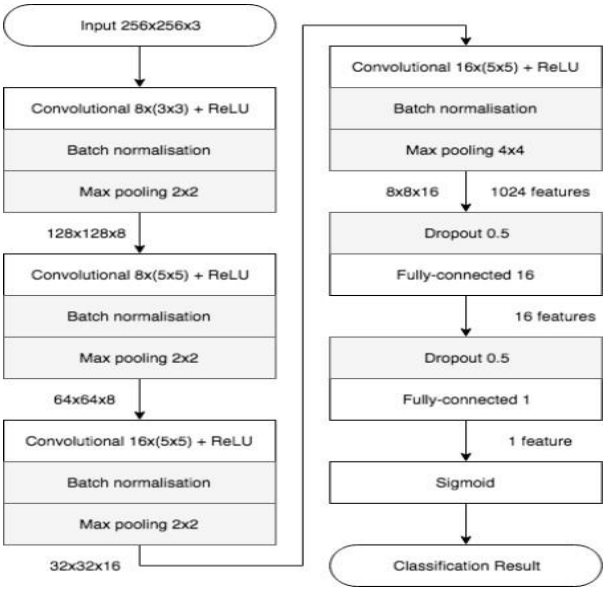

Figure 6. The image illustrates the layered network architecture of Meso-4, showing the layers, parameters, and output sizes through boxes and annotated arrows.

Departing from the original module's utilization of 5×5 convolutions, we advocate for the adoption of 3×3 dilated convolutions to circumvent excessive semantic impact. While the notion of employing dilated convolutions within the inception module has been explored in prior works as a strategy to address multi-scale information, we augment this approach by integrating 1×1 convolutions preceding the dilated convolutions for dimensionality reduction. Additionally, an extra 1×1 convolution is introduced in parallel to serve as a skip-connection linking successive modules. Detailed schematics elucidating these modifications are depicted in Figure 7.

Replacing more than two layers with inception modules did not yield superior classification results. The parameters (ai, bi, ci, di) chosen for each layer are as follows: a value of 1 for layers 1 and 2, a value of 4 for b in layers 1 and 2, a value of 4 for c in layers 1 and 2, and a value of 1 for d in layer 1 and 2 in layer 2. These hyperparameters resulted in a total of 28,615 trainable parameters for the network.

5.3 Xception-Net

Xception leverages a novel deep convolutional neural network architecture centred on Depthwise Separable Convolutions (DSCs). Introduced by researchers at Google, this design incorporates inception modules that bridge the gap between standard convolutions and their depthwise separable counterparts. This innovative approach builds upon the foundation laid by

Inception modules, ultimately leading to the development of Xception – an architecture aptly named for its emphasis on "Extreme Inception."

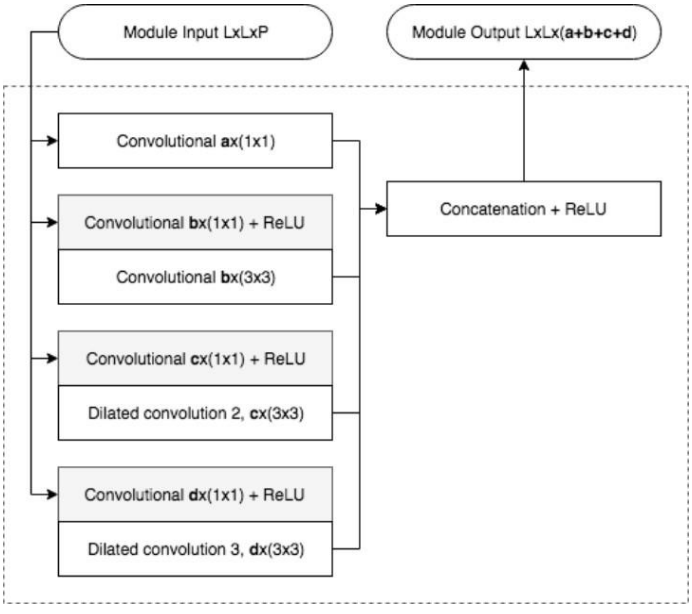

Figure 7. Structure of the inception modules utilized in MesoInception-4

As illustrated in Figure 8, the Xception architecture employs a total of 36 convolutional layers meticulously organized into 14 distinct modules. These modules collaboratively perform feature extraction throughout the network. Notably, all layers except for the initial and final ones (as depicted in Figure 9) incorporate linear residual connections that circumvent the modules. It's important to acknowledge that the Xception network is designed to handle input images with dimensions varying between 71 x 71 and 299 x 299 pixels.

As detailed in Figure 9, depthwise separable convolution constitutes a two-step process encompassing a depthwise convolution followed by a pointwise convolution. During depth wise convolution, a separate spatial convolution is applied to each input channel. Subsequently, the outputs from these individual convolutions are combined through a pointwise convolution. This approach enables depthwise separable convolution layers to extract richer feature representations while simultaneously incurring lower computational costs and requiring a reduced number of parameters. Remarkably, this technique achieves comparable or even superior performance and scalability. As highlighted in [6], combining multiple, simpler convolutions into a block fosters the overall depth of the neural network, ultimately facilitating the extraction of more intricate and nuanced features.

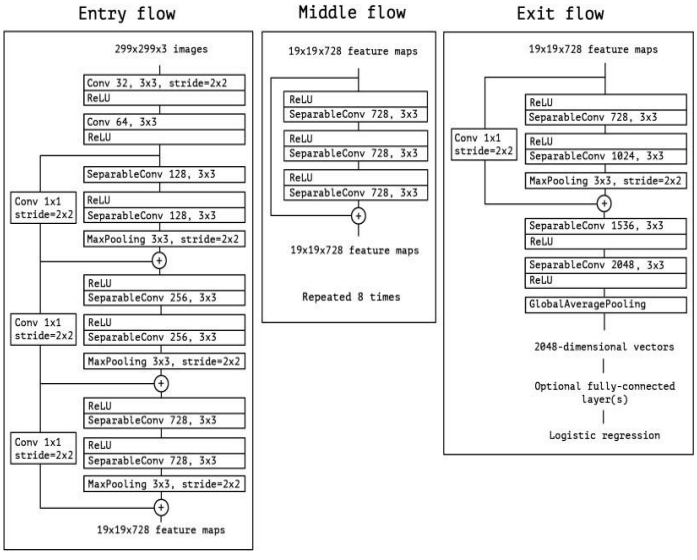

Figure 8. Xception net Architecture [7]

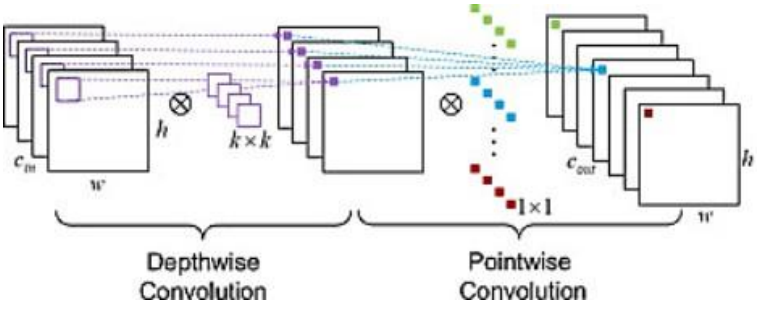

Figure 9. Depthwise separable convolution Architecture [8].

5.4 ViT

The proposed shallow Vision Transformer for deepfake detection introduces a transformer-based network optimized to perform efficiently within constrained computational and memory resources. By streamlining the model architecture to include only 12 transformer layers and 6 heads, we significantly reduce the number of parameters to 5,217,642, approximately 16.48 times less than conventional ViT variants. This reduction in complexity allows the model to achieve optimal performance while requiring fewer training images to learn its parameters. Furthermore, our approach leverages the attention mechanism to emphasize crucial regions within input images, enabling accurate discrimination between real and fake content. This targeted attention to specific image patches is vital, as differences between authentic and manipulated images often manifest in localized areas. Therefore, by analyzing the distribution of the attention vector in the shallow ViT, we can effectively identify and prioritize significant image features for deepfake detection.

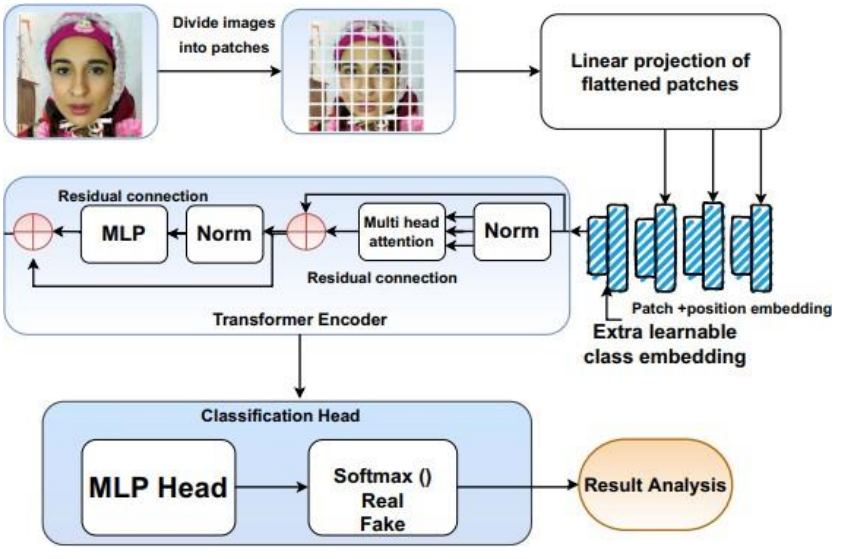

Figure 10. Architecture ViT model for deepfake detection

6. Results

In our study, along with the CNN and Transformer models, we introduced another model called Cross-ViT, which is a separate architecture in its own right. Initially, we segmented each of the five datasets into distinct portions for training, validation, and testing purposes. While the DFDC dataset already had predefined divisions, we partitioned the videos into other datasets to allocate roughly 15% for validation, 15% for testing, and the remaining data for training. This ensured a balanced distribution while maintaining consistency by keeping specific source videos used in deepfake creation within the same segment. Subsequently, we conducted frame extraction and facial detection to create balanced datasets. The testing dataset retained its original video format, and during detection, we extracted frames and facial regions to analyze the results.

For model construction, we utilised ResNet152, MesoNet-4, MesoInception-4, and Xception as CNN architectures, alongside ViT as the Transformer architecture, and Cross-ViT which is a convolution of CNN and Transformer architecture. Each model underwent the same training process and was evaluated based on the highest validation accuracy. Finally, cross-dataset evaluation tests were conducted, with results summarized in Table 1 and Table 2. To optimize the model's performance, the cross-entropy loss function was employed. This function effectively measures the disparity between the predicted probability distribution and the actual distribution. Additionally, the Stochastic Gradient Descent (SGD) solver was utilized for optimization, facilitating the iterative refinement of model parameters to minimize the chosen loss function.

Table 1. Accuracies of different models

Train Data	Architect ure	Test Data (Balanced Accuracy Results in %)		
		FF+ 2020	Google DFD	DFDC
FF+ 2020	ResNet	88.01	58.01	55.00
	MesoNet- 4	93.53	76.0	63.77
	MesoInce ption-4	95.94	78.3	68.41
	Xception	87.32	67.11	52.81
	VIT	79.66	50.88	68.11
	Cross-Vit (Conv)	73.33	69.69	78.55
Google DFD	ResNet	52.80	96.01	65.34
	MesoNet- 4	64.3	98.54	78.93
	MesoInce ption-4	65.8	99.33	80.22
	Xception	43.31	93.84	72.24
	VIT	58.80	86.72	66.10
	Cross-Vit (Conv)	63.35	88.57	74.21
DFDC	ResNet	52.10	91.0	78.21
	MesoNet- 4	60.42	93.71	89.33
	MesoInce ption-4	64.81	94.88	91.55
	Xception	58.94	87.31	89.01.
	VIT	57.21	82.66	90.03
	Cross-Vit (Conv)	61.11	90.67	79.05

Table 2. AUC scores of different models

Train Data	Architec ture	Test Data (AUC Results in %)		
		FF+ 2020	Google DFD	DFDC
FF+ 2020	ResNet	93.26	87.61	68.44
	MesoNe t-4	78.12	60.52	58.44
	MesoInc eption-4	82.00	66.44	64.33
	Xception	99.65	91.18	65.07
	VIT	92.3	84.86	76.34
	Cross-Vi t(Conv)	63.32	55.00	40.12
Google DFD	ResNet	60.52	90.23	70.00
	MesoNe t-4	58.03	94.14	68.32
	MesoInc eption-4	60.01	97.21	68.55

	Xception	57.43	99.89	78.54
	VIT	62.63	98.07	72.59
	Cross-Vi t(Conv)	42.31	88.01	44.31
DFDC	ResNet	73.37	46.94	94.83
	MesoNe t-4	76.41	50.08	96.44
	MesoInc eption-4	77.22	56.44	97.21
	Xception	71.16	75.71	95.98
	VIT	69.40	62.53	95.97
	Cross-Vi t(Conv)	56.32	51.10	91.13

6.1 Optimal Detection Outcomes

We noticed that the top accuracies attained in identifying deepfakes from FF++ 2020, Google DFD, and DFDC test datasets were 95.94%, 99.33%, and 91.55%, respectively. Similarly, the highest AUC scores reached for detecting deepfakes from FF++ 2020, Google DFD, and DFDC test datasets were 99.65%, 99.89%, and 97.21%, respectively.

6.2 Equivalent Training and Testing Datasets

Both CNNs and Transformer models demonstrated commendable performance, with accuracies exceeding 85% and AUC surpassing 98% when tested on FF++ 2020 and Google DFD datasets. However, VIT exhibited comparatively lower performance, with accuracies ranging from around 50.88% to 84.86% and AUC from around 58.11% to 76.34%. Performance declined as the models were tested on larger and newer datasets. CNNs generally matched or outperformed Transformer models, with ResNet, MesoNet-4, Xception, and VIT showing notable detection capabilities on their respective test datasets.

6.3 Assessments across Different Datasets

Models trained using the FF++ 2020 dataset demonstrated effectiveness when applied to the DFDC test dataset. Specifically, the Xception, VIT, and CrossVit (Conv) models trained on FF++ 2020 achieved accuracies of 52.81%, 68.11%, and 78.55%, respectively, when tested on the DFDC dataset. Additionally, their respective AUC scores on the DFDC dataset were 65.07%, 76.34%, and 40.12%.

7. Conclusion

In our research, we explored the capabilities of various deep learning models in detecting deepfakes across different publicly available datasets. We evaluated the performance of four convolutional neural networks (CNNs) and four transformer models on the same train-to-test

and cross-dataset scenarios. Through comprehensive cross-dataset evaluations and overall model performance analysis, we uncovered the relationships between the FF++ 2020, Google DFD, and Celeb-DF datasets. Additionally, we identified the unique strengths and characteristics of the Deeper Forensics, DFDC, and FF++ 2020 datasets. Our findings suggest that these datasets will continue to play a crucial role in future research, especially as new and more advanced deepfake techniques emerge, necessitating further investigations.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets. *Advances in neural information processing systems*, (2014) 27.

[2] D.P. Kingma, M. Welling, (2013). Auto-encoding variational bayes. *arXiv preprint*,

[3] Y. Choi, M. Choi, M. Kim, J.W. Ha, S. Kim, J. Choo (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *In Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE*, USA. https://doi.org/10.1109/CVPR.2018.00916

[4] T Kim, M Cha, H Kim, JK Lee, J Kim, Learning to discover cross-domain relations with generative adversarial networks. *International Conference on Machine Learning*, PMLR, (2017) 1857–1865.

[5] J. Chai, H. Zeng, A. Li, E.W. Ngai, (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6, 100134. https://doi.org/10.1016/j.mlwa.2021.100134

[6] X. Yang, Y. Li, S. Lyu, (2019). Exposing deep fakes using inconsistent head poses. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, UK. https://doi.org/10.1109/ICASSP.2019.8683164

[7] P. Zhou, X. Han, V.I. Morariu, L.S. Davis, (2017). Two-stream neural networks for tampered face detection. In 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW), IEEE. USA. https://doi.org/10.1109/CVPRW.2017.229

[8] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, IEEE, Korea (South). https://doi.org/10.1109/ICCV.2019.00009

[9] Z Li, F Liu, W Yang, S Peng, J Zhou, A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Transactions on Neural Networks and Learning Systems, 33(12), (2021) 6999-7019. https://doi.org/10.1109/TNNLS.2021.3084827

[10] D Wodajo, S Atnafu, (2021) Deepfake video detection using convolutional vision transformer. arXiv preprint.

[11] P. Korshunov, S. Marcel, Deepfakes: a new threat to ace recognition? Assessment and detection. arXiv preprint.

[12] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, (2020). The deepfake detection challenge (DFDC) dataset. arXiv preprint.

[13] L. Jiang, R. Li, W. Wu, C. Qian, C.C. Loy, (2020). Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, USA. https://doi.org/10.1109/CVPR42600.2020.00296

[14] M. Tan, Q. Le, (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning. PMLR.

[15] Z. Chen, P.H. Ho, (2019). Global-connected network with generalized ReLU activation. *Pattern Recognition,* 96, 106961. https://doi.org/10.1016/j.patcog.2019.07.006

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, (2015). Going deeper with convolutions. *In Proceedings of the IEEE conference on computer vision and pattern recognition.*

Funding

No funding was received for conducting this study.

Conflict of interest

The Author's have no conflicts of interest to declare that they are relevant to the content of this article.

About The License