Advances in Consumer Research
ISSN (Print): 0098-9258
ISSN(Online): 3079-1766

Vol. 2, Issue 4 (2025) https://acr-journal.com/

Optimizing Deepfake Image Classification with Transfer Learning: Insights from Models Inception V3 and Inception V4

Salina Adinarayana[1], Bhuvan Unhelkar[2], Siva Shankar S[3]

[1]Post Doc Researcher, Muma College of Business, University of South Florida 8350 N. Tamiami Trail, Sarasota, Florida. USA.

[1]Professor, Dept. of CSE (Data Science), Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India,

Email ID: sadinrayana.pdf@gmail.com

[2]Professor, Muma College of Business, University of South Florida 8350 N. Tamiami Trail, Sarasota, Florida. USA.

Email ID: bunhelkar@usf.edu

[3]Professor & IPR Head, Department of Computer Science and Engineering, KG Reddy College of Engineering and Technology, Hyderabad, Telangana, India,

Email ID: drsivashankars@gmail.com

Cite this paper as: Salina Adinarayana, Bhuvan Unhelkar, Siva Shankar S, (2025) Optimizing Deepfake Image Classification with Transfer Learning: Insights from Models Inception V3 and Inception V4 *Advances in Consumer Research*, 2 (4), 1586-1600

KEYWORDS

Deepfake classification, Deepfake image classification, Transfer Learning in Deepfake classification, Inception models in Computer vision.

ABSTRACT

In the recent past facial forgery had been widespread and for this there is no any requirement of possessing technical skills. Nowadays due to the development of Generative Adversarial Networks & diffusion models (DMs), generation of quality of deepfake has been intensified. Audio or video data if exists as it doesn't create any problems. In this digital world manipulating although easy, such data creates a lot of problems. They often lead to theft of identity, misinformation and cybercrime in various formats. The resultant technology of manipulating audio or video data using advanced technologies in other words known as Deep fake technology and its rise created a provision to new frontiers in current digital world. Although the advancement offers potentiality on the positive edge, also there are significant risks to the security as well integrity of the data & information. In one perspective, Deepfakes due to their malicious applications in order to attain political, economic and social reputation goals, the technology is becoming quite dishonourable. In this paper, deep fake image classification has been performed on a state-of-the-art dataset by employing transfer learning and ensemble models. Initially, image classification has been performed using Inception V3 model with variants of batch sizes 16, 32 and 64; subsequently Inception V4 is employed due to which the performance has been improved significantly...

1. INTRODUCTION

The term "deepfake" is a merger of the words "deep" & "fake" which refers to the creation of forged content of photos, videos, or audio, using advanced and unique DL techniques. The spread of deepfakes, a manipulated media is creating a substantial threat to society and deprives the reliability on the information available on the internet. The two sides to the coin of deepfake technology in the world of digital technology has the disadvantages that include it could also be used to alter videos, which would pose serious threats to society and national security besides having advantages.

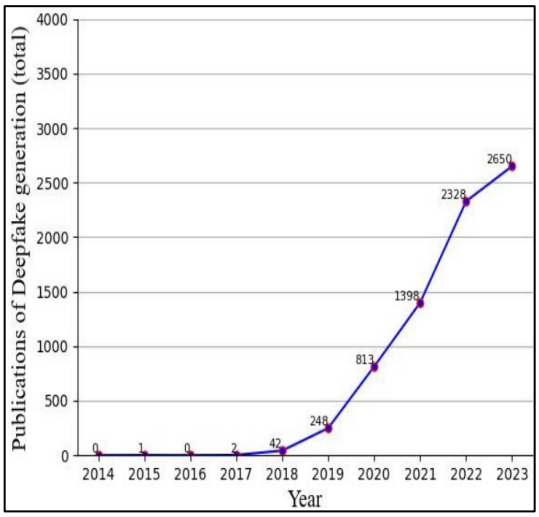
Due to the enhanced interest in deepfake technology, there has been a surge in related research. Over the past couple of years, there has been notable advancement in the development of novel detection techniques. At the same time, despite the progress, crucial challenges remain unresolved in current deepfake detection approaches. Due to the advancement in deepfake creation techniques being momentous, the quality of created fake videos is proportionately increasing. Conventional approaches may not be adequate for identifying the changes done by advanced deepfake algorithms. Moreover, focus on visual quality analysis or inconsistency in detection is lacking .

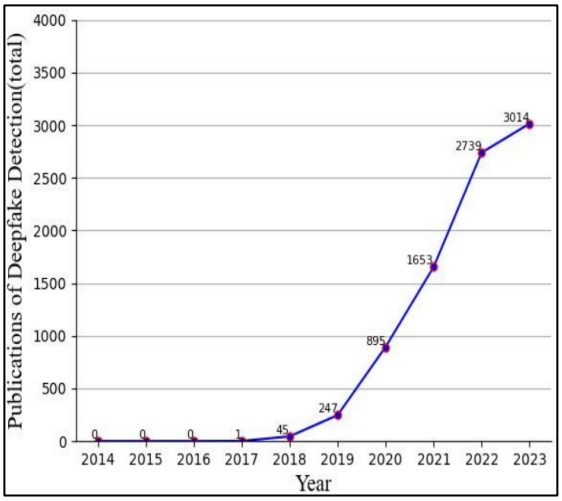
and hence there is an urgent need to get updated with the trend and also to be on par with the ever-changing technologies associated with.

In numbers, the prevalence of deepfake data is growing exponentially, with an annual growth rate of around 300% [16]. During recent years, various researchers have made a deliberate and focused attempt to devise techniques for detecting Deep Fakes. Initially the research efforts concentrated on detecting visual discrepancies within individual frames, employing various techniques that either utilised biological signals or selected features and Convolutional Neural Networks (CNNs). Although recent algorithms for Deep Fake detection have achieved excellent accuracy, they still require a comprehensive solution that can effectively handle different video alterations, especially those related to voice biometrics. Therefore, they must exhibit greater reliability in detecting Deep Fakes in current real-world scenarios.

Basing on the offensive and defensive approach, Research on Deep Fake as mentioned is rapidly increasing. The graphs in figures 1 and 2 represent the number of publications of deepfake generation and detection algorithms respectively. During the past decade the contributions are less in number in the first half. But in the remaining half decade there is a tremendous momentum in the contributions of both generation and detection.

Fig





1: Number of Deepfake generation Research publications [2] Gong et.al.

Fig 2: Number of Deepfake detection Research publications [2] Gong et.al.

Understanding the phenomenal growth, it is quiet important to investigate and predict the progress of deepfake-related research and accordingly the protective mechanisms to be implemented.

This section gives a brief introduction regarding the research being done on Deepfake technology and in the succeeding section the reviewed literature is presented. In the third section the observations and other findings are cited. The final section concludes the paper.

2. LITERATURE REVIEW

Deepfake fraud Evolution over the last five years:

After the emergence of deepfake technology happened in 2019, the proliferation since then is quite disruptive. The deepfake creation tools are very user friendly and they are easily accessible due to which the technology is widespread. Later in 2021 the technological advancements made the deepfake scam gain the momentum and also mainstream attention. Since 2022 there is a spike in the deepfake fraud incidents in various segments. Although advanced techniques are being adopted, detection of the frauds are challenging. Also, in this year the first step was made towards inter-disciplinary endeavour so as to combat the deepfake deception. In the current scenario the race between the security teams and fraudsters is continuous and seems to be never ending. Currently, the risk of Deepfake is widely recognised all over the world. The deepfake detection across all sectors worldwide has been increased by tenfold from the year 2022 to 2023.

Gayar et.al. [1] introduce a new way to detect deep fake videos using a GNN, which is abbreviated as graph neural network. A mini-batch graph CNN and a quad-block CNN Network with Convolutions in addition to the Batch Normalisation and an activation function, make up the architecture of the detection process. Convolutional layers are connected to the dense layer in the last stage, after which flattening is done. In order to combine these two steps, three separate fusion networks, Additive fusion (FuNet-A), Element-wise multiplicative fusion (FuNet-M) and Concatenation fusion (FuNet-C) have been used. The evaluation results in terms of training and validation accuracies of the proposed research defend that the model applied on many datasets is quite impressive, with accuracies of 99.3% post 30 epochs. The proposed mechanism overcomes the

limitations of other models, such as minimization of processing demands, excellent detection accuracy, overfitting problems, and so on. The model can adapt to new deepfake techniques due to its adaptability and scalability to GNN. The goal of fusion techniques is to improve deep fake video detection accuracy by creating a trainable network that combines features from mini–Graph Neural Networks (miniGNNs) and Convolutional Neural Networks (CNNs). Multi-fusion between GNN and CNN produces mini GNNs, which, when compared to Convolutional Neural Networks (CNNs) train the network so as to detect deep fake videos.

DeepFaceLive & Roop are a few amongst the majority of online tools available open source. The primary technique for generating deepfake datasets is the Basic DeepFakemaker. In **[2] Gong et.al.** discussed datasets in detail and also the detection methods, which are described below. The methods are traditional CNN-based detection methods, Transformer-based detection methods, and biological signal detection methods.

Traditional CNN-Based Detection Methods

Sabour et.al. [19] proposed the Capsule Network, which is capable of analysing the relationships between 3D spatial information, overcoming the classic CNN model in this aspect. It achieved the capability with a reduced number of parameters. Nevertheless, the limitation of Capsule Network lies in its limited ability to accurately forecast unfamiliar input due to its weak generalisation capacity.

Semi-Supervised Learning in CNN Backbone

To determine how similar the representations of two data augmentations as well as feature extraction, are, the popular method known as Consistency Representation Learning of Forgery Detection (CORE) is used. This is done because we assume that different data augmentations should not affect the consistency of the other types of data. The main benefit of this approach is that it uses a new loss function that considers different data augmentations and combines cross-entropy loss with representation similarity. Data augmentation is crucial to the efficacy of the mentioned approach due to the reason that different augmentations will have different effects on the assessment metrics used.

T-Face the "Dual-Contrastive Learning Detection," mechanism uses a number of data augmentation approaches to reliably identify real faces. DCL is a method that combines two forms of contrastive learning: intra-instance and inter-instance contrastive learning. The objective of DCL is to make the distances between embeddings of distinct classes larger while decreasing the distances between embeddings of same classes.

It has been demonstrated that learning temporal artifacts can enhance detection resilience. Spatio- Temporal Inconsistency Learning STIL focuses on identifying temporal & spatial inconsistencies in fake videos through the process of learning. STIL comprises three modules TIM (Temporal Inconsistency Module), SIM (Spatial Inconsistency Module), and ISM (Information Supplement Module). The purpose of these modules is to build a complete representation by capturing information that is both spatial and temporal in nature.

Transformer-Based Detection

The adoption of the transformer framework for classification although less prominent; is progressively replacing standard CNN classification. End-to-End Transformer Detection is a deep fake detection system that utilises visual transformers. It addresses the limitation of typical CNNs in capturing the connection between spatial patches and the loss of information loss due to the receptive field. The transformer exhibits superior generalisation capabilities and high efficacy, overcoming the backbone-based CNN methods.

ISTVT: A Video Transformer based Deepfake Detection model is made up of two different modules: a self-subtract module and a spatial-temporal self-attention module. These components have been developed with the purpose of successfully capturing spatial characteristics and temporal irregularities in films.

Biological Signal Detection

The key benefit of this approach is that it improves the identification of deep fakes by including biological signals into the classification of the residuals of generative models, which in turn makes it possible to identify the source of the fake. The experiments demonstrate the increase of accuracy by 47%.

The cited literature states that the models must exhibit robust generalisation capabilities and high efficiency. Rather than traditional CNN models, models with CNN backbone with semi supervised learning are performing well. Also, it is to understand that spatial and temporal features are highly important for the models to perform well. Transformer based detection methods outperform the remaining ones besides biological signal detection methods are effectively used in deep fake detection.

Usually, softwares such as Adobe Photoshop are used for image editing and to do so, the humans need expertise, significant amount of time and effort. Machines do not require specialised skills, rather if the algorithm is efficient, within less time the deep fakes could be generated. **Kaur et.al. [3]** cited the noticeable findings such as there is no predominant focus on image and video-based detection and the real-world testing is insufficient. Also, there are gaps in the quality and relevance of

datasets. To address these issues, their contributions include consolidation of existing knowledge, the detection challenges are classified and represented comprehensively. Deep insights are given regarding the deepfake datasets.

Gambín et al. [4] highlights the need of collaboration among researchers, governments, and corporate groups in developing and executing effective strategies for detecting and preventing deepfakes. The researchers had an in-depth discussion on block chain technology and the capacity of distributed ledgers to augment cybersecurity measures against deepfakes.

The purpose of this study done by **Bray et.al. [5]** is to evaluate the human capability to recognise altered images of human faces, known as deepfakes that are generated by the StyleGAN2 algorithm which are a part of FFHQ dataset. Also, there are non-deepfake images in the dataset which are randomly selected and used for study. Additionally, the study aims to determine the efficiency of manual methods and enhancing the accuracy of deepfake detection. Each participant has been presented 20 photographs which are in a random sequence in a collection out of 50 deepfake and 50 genuine human face images. Participants were instructed to determine if each image was generated by artificial intelligence (AI) or not, provide their level of confidence, and rationalize. The accuracy of participants was only 62%. Besides the accuracy in identifying the images varied between 30% to 85%. Additionally, one in every five images had an accuracy below 50%. This raised an alarm and indicated for an immediate action to address this threat.

The study of **Rafique et.al. [6]** presents an automated approach for categorising deep fake images using advanced techniques in Deep Learning and Machine Learning. An Error Level Analysis of the image is carried out initially by the suggested framework in order to determine whether the image is altered or not. The Convolutional Neural Networks are given this image in order to carry out the process of deep feature extraction. Following the completion of hyper-parameter optimisation, the feature vectors that were produced are classified by employing Support Vector Machines and K-Nearest Neighbours classification methods. By utilising Residual Network (ResNet18) and K-Nearest Neighbour, the proposed method was able to achieve an accuracy of 89.5%. Both the effectiveness and the resiliency of the proposed methodology are demonstrated by the observations. As a result, the proposed model could be promoted for the purpose of identifying deep false images and mitigating the potential threats. Both ML & DL techniques were used in the proposed model. The framework begins by resizing the image so that it is compatible with the input layer of the CNN. Subsequently the ELA image is given to the underlying GoogleNet, ResNet18, and SqueezeNet, so as to extract deep features from the input image.

Saxena et.al. [7] introduced a new framework for detecting deepfake videos. Deep learning is utilised in this approach, and it is based on a pre-trained model XceptionNet that is constructed using deep convolutional neural networks (CNNs). The data retrieved from movies relevant to a variety of face characteristics is done through the use of facial landmark recognition. Then this data is further used to provide the DL model the ability to differentiate between genuine and deepfake videos. The Xception Neural Network model which is capable of taking multiple input operates is based on CNN and is used by the deepfake detection method. The model is trained with the Dataset 'Dessa' and partial Deepfake Detection Challenge Dataset. Their model achieved a remarkable classification accuracy of 96% and AUC value of 0.97. To further improve the performance of model, it must be trained in accurately identifying a wider range of persons and also people from diverse racial backgrounds and tune the performance by optimizing the parameters. Also, it is mentioned that the geographical and temporal information need to be integrated with the facial data so that the further evaluation becomes quite impressive.

In **[8] Janutenas et.al.** selected LRNet approach as the foundation for subsequent research due to its exceptional performance. Dual-stream recurrent neural networks (RNNs) are utilised in this model. The concept used in the method is analyzing the image changes over time and analyzing the face area. In other words the calibration module is said to improve the precision of geometric feature identification as time progresses. Altering the parameters effectively the model exhibited with high accuracy and the results demonstrated the performance of model. Their insight is that parameter tuning plays an important role to improve the model performance.

Online platforms have extensively exploited deepfake videos, particularly focussing politicians and celebrities. Various solutions have been outlined in the research to address such challenges. **Mukta et.al. [9]** in their work conducted a comprehensive analysis by assessing and comparing two types of deepfake concepts one being major research contributions in the field of deepfake technology and the other being tools related to deepfake technology that are extensively used. They relied on the concept of Multitask learning that involves the integration of forgery location and deepfake detection. The opine that accurate placement of the forgery plays a vital role in the effectiveness of deepfake detection. They claim that there are flaws in present forensics innovation and there is a need in novel anti forensic technique. Also, they mentioned that there is a need of large qualitative data.

Narayan et.al. [10] in their study replicated the real-life situation of creating deepfake videos and introduced the DFPlatter dataset. The dataset comprises both low and high resolution deepfakes produced with the help of various techniques. It also consists of deepfakes featuring single and multiple subject scenarios, using images of individuals of Indian ethnicity. Past results indicated a notable decrease in performance while detecting deepfakes with images or videos of low resolution. In addition, current methods result in reduced accuracy when detecting deepfakes involving numerous subjects. The proposed dataset shall be of enhanced standard and improves the performance of the detection algorithms in coordination to the real-life situations.

Prezja et.al. [13] as a part of their research introduced Generative Adversarial Neural networks (GAN Networks) that are capable to generate authentic knee joint X-ray pictures with different levels of osteoarthritis severity. They provided a dataset consisting of 320,000 synthetic (DeepFake) X-ray pictures, which were generated using 5,556 genuine images for training purpose. The medical correctness of the proposed model was authenticated by consulting 15 medical professionals. Additionally, the augmentation effects are evaluated by performing a task involving the evaluation of osteoarthritis severity. While consulting medical professionals 30 authentic photos and 30 deep fake images of X-ray pictures were given for evaluation. The utilisation of DeepFakes resulted in enhanced accuracy in classifying the severity of osteoarthritis, even with a limited amount of genuine data available. This improvement was achieved through the implementation of transfer learning techniques. Furthermore, in the identical categorization test, the authentic training data had been substituted with DeepFakes towards which 3.79% decrease in accuracy was attained when compared to the initial baseline performance in identifying genuine osteoarthritis X-rays.

This amazing study indicates the usage of deep fakes in solving critical issues easily and the positive side of the deep fakes. Also, the study indicates the employment of deep fakes in various sectors.

Narayan et.al. [14] presented an innovative dataset DeePhy which is also known as Deepfake Phylogeny. Usually, the datasets don't provide the model that is employed as target labels. Moreover, by providing information related to the generative model that was employed; model attribution helps to enhance the explainability of the detection results. This study addresses these questions easily. A total of 5040 deep-fake videos were generated with the help of three different generating algorithms. Among these videos 840 videos which were generated by swapping the faces once, 2520 videos were generated by swap ping the faces twice and 1680 videos in which faces have been swapped three times. The database is of size of 30 GB plus which was created over a period of 1100 hours' time using a total of 18 GPUs with a combined memory capacity of 1,352 GB. Additionally, the benchmarks are presented using six deepfake detection algorithms. The findings emphasise the necessity to advance the study of model attribution for deepfakes and generalize the process over a range of DF generation strategies. The intent is to create robust and generalized models across all origins and ethnicities.

Hussain et.al. [15] made efforts in analyzing the vulnerabilities of different DNN based deepfake systems. The deepfake detection algorithms are of per-frame as well per-sequence based. Modern Deepfake detection techniques utilise Deep Neural Networks (DNNs) to differentiate between artificially generated fake videos and authentic videos. Their study is intended at demonstrating the feasibility of bypassing the detection algorithms through adversarial alteration of fake videos created by standard techniques. Additionally, they proved that the adversarial perturbations exhibit resilience against image and video compression codecs, due to which they cause a threat in real-world scenarios. Ultimately, they developed the attack scenario which led to an alarming situation. Also, the mitigation procedures are discussed

Suratkar et.al. [16] presented a novel method for detecting fake videos using a hybrid model that combines CNN and RNN by employing transfer learning in autoencoders. Unknown test input data is given to determine the model's generalizability. Further, the effect of residual image input on the model's accuracy is investigated. Results in both cases with and without transfer learning are presented to validate the efficacy of the model. The study demonstrated how fine-tuning and transfer learning can enhance the accuracy of deep fake detection models. The future research indicates the focus on detection of deep fakes involving single fake face or multiple fake faces in a video. By combining the suggested video deep fake detection with audio deep fake detection, a multimodal deep fake detection system may be built.

Xu et al. [17] evaluated the contributions done on deepfake video content creation, detection and how the techniques misled the users with respect to the content. They illustrated the two groups namely the offenders (Deepfake creators) and defenders (Deepfake detectors). They made an exhaustive survey citing 300 plus references. The landscape of their research is highly useful to the researchers to understand the nature and devise novel mechanisms.

Debasish Samal et.al. [18] employed Xception model for deepfake detection by considering a deepfake image dataset containing 190K images and their model achieved performance in terms of 88%, 0.95 precision, 0.79 recall and 0.86 F1-score. To further improve the performance, ensemble approach could be used.

In 2019, **Nguyen et.al.** presented an overview on the deepfake strategies whereas in 2020, **Tolesana et.al.** discussed regarding the facial image alternation methods. **Verdoliva et.al.** done their work on visual media integrity verification using traditional and deep learning methods. Agarwal et.al. contributed on the detection methods analyzing the image changes over time, **Ciftci et.al.** discussions are on the detection methods pertaining to face area while the focus of **Durali et.al.** is on frequency domain analysis.

In 2021, **Mirsky et.al.** contributed with reenactment approaches for DNN based generating architecture. **Yu et.al.** an approach for detecting deepfake videos and model evaluation benchmarks was proposed. **Khormali et.al.** proposed on attention-based video authentication mechanism. Both **Lee et.al.** and **Sun et.al.** worked out on facial characteristic changes with respect to temporal aspects. **Tanaka et.al.** proposed a mechanism on image comparison based on hash values. **Yavuzkilic et.al.** contributed the model which is trained on three different types of images.

In 2022, **Rana et.al.** contributed on the categorisation of detection methods into four groups. Gu et.al. proposed a method that compares 3D face over time. Ganguly et.al. worked out an attention-based model while Wolter et.al. made an analysis of GAN spatial and frequency features.

In 2023, **Patil et.al.** proposed a novel method based on biological classifiers. Masood et.al. made their contribution on ML-based audio and video manipulations. Akhtar et.al. reviewed various contributions and provided insights on the detection methodologies. Epetalgy et.al. as well Ilyas et.al., proposed methods on video and audio comparison over time working on the dataset FakeAVClub, in which they attained an accuracy of 98.51% and 90.94% respectively. Ganguly et.al. again made their contribution attention-based mechanism on three datasets, FaceForensics++, Celeb-DF & DeepFakes and achieved around 99% in all three contributions.

A brief note on Datasets available

Incredible datasets are compiled by various researchers in the past five years. The dataset UADF is contributed in the year 2018 which contains 49 real and 49 fake images which are collected from youtube and generated using FakeApp. But the limitation with this dataset is the data is very less. Similarly in 2019, the datasets FaceForensics++, DeepFakeDetection and Celeb-DF data are provided which are collected/generated from/using youtube/FaceSwap, Face2Face, NeuralTextures, FaceShifter, DeepFakes, Improved DeepFakes and Generative Adversarial Network (GAN) mechanisms. The advantages are limited as well the limitations include less data and visible manipulated facts. However, due to GAN, there is a rise in the datasets introduced in 2020 both in the quality as well quantity. The datasets are DFDC, DeeperForensics-1.0 and iFakeFaceDB and DFFD. The data is in numbers of multiples of thousands which had the advantage comprising multi scenarios and multi techniques. The limitations with these datasets are also less mentioned. DFDC dataset is generated through the challenge launched and continued by the IT giants Facebook, Amazon, Microsoft and other research institutions successively from 2020 onwards. Celeb-DF V2 is the improved dataset of its earlier version Celeb-DF. Other datasets include Flickr-Faces-HQ, 100K-Faces, WildDeepFake and VGGFace2. Besides these, there are several other datasets in Kaggle and other web sources.

Observations from the literature reviewed

Having gone through the literature following are the observations made in general.

i) Initially the image editing tools such as Adobe Photoshop are limited and are done manually.

ii) Now due to the technological advancement and rapid growth of Artificial Intelligence, the image, audio and video editing tools are implemented using Deep learning algorithms and the editing job is being done by the machine itself.

iii) Various types of Deep learning Neural Networks such as DNN, CNN, GAN are being used in Deep Fake generation as well detection.

iv) Initially the datasets are of poor quality but gradually the quantity and quality of deepfake data had become high.

v) Earlier works are on datasets of less size while now the scalability is being increased.

vi) Parameter tuning is one important aspect with which effective results can be yielded.

vii) Deepfake datasets or Deepfake multimedia data if utilized correctly, it is quite beneficial in all sectors. But at the same time, due to high expectations in terms of money, reputation, and often entertainment, this data is becoming detrimental. Hence always the Défense mechanisms must always be proactive.

viii) Transformers play a vital role, and utilizing them in frame-based detection approaches shall make the detection a dominant strategy. Even the feature extraction may use the transformer architecture so that the combination yields the best results.

ix) Biological signal-based detectors are one more contributor to impressive Deepfake detection mechanisms. Biological signals such as PPG are fused by them, due to which multimodality data could be generated. This supplements the diversity of counterfeit information and enhances the learning ability of the model

x) Various detection methods involving spatial and temporal aspects could be developed at large. Video transformers could play a substantial role in performing the spatial and temporal feature extraction.

3. MATERIALS & METHODOLOGY

This section explains the dataset used and the mechanism implemented in detecting the deepfakes.

3.1 Overview of the Dataset

Deepfake Detection Challenge (DFDC) is the dataset used in this research that contains labelled real and deepfake images. The dataset is available at [20]. It contains 1,90,335 images comprising real and fake ones, which are divided into 140002, 10905, and 39428 for training, validation, and testing, respectively.

3.2 Research Objectives

The following are the objectives of this research:

Adopting Transfer Learning mechanism in Deepfake Image Classification which can improve the accuracy.

Employing an Ensemble model in addition to the transfer learning mechanism to further improve the accuracy in Deep Fake Image Classification.

Designing a high-profile deep learning framework in classifying DeepFake data in terms of computational complexity.

3.3 Steps involved in the complete process

3.3.1 Preprocessing

The preprocessing step is critical to the success of the classification task. Image rescaling has been done, after which various preprocessing parameters, rotation range (RR), width shift range (WSR), height shift range(HSR), shear range(SR), and zoom range(ZR) have been applied. Flipping is also performed. Then the image normalization is done. Finally, images are resized to 299 x 299 in RGB color mode.

3.3.2 Model Selection (Transfer Learning)

Multiple deep learning models have been investigated which are mentioned in this sub section. All the models are thoroughly analyzed.

EfficientNet: This model tends to balance computational efficiency with accuracy, making it ideal for tasks involving high-resolution face images. Its scaling capabilities allow for more efficient learning across different dataset sizes.

XceptionNet: Known for its performance in deepfake detection tasks, Xception employs depthwise separable convolutions that help capture subtle differences between real and deepfake images.

ResNet50: ResNet50 introduces residual learning, which allows the model to go deeper without facing the vanishing gradient problem. Its deep architecture can learn complex features necessary for classification.

InceptionNet: A hybrid model that combines the strengths of EfficientNet and Xception could be employed to take advantage of both models' strengths. This could help capture fine-grained features while maintaining computational efficiency.

Why Inception Net has been chosen for our research work?

Having studied all the transfer learning models, it could be identified that Inception model is a light weight and an efficient model. Inception Net is a Google's Inception CNN based architecture comprising 48 Layers. It is an excellent Transfer Learning model that could classify images up to 1000 classes. The architectural design separates the process in three steps – Stem (Data Ingestion), Body (Data Processing) and Head (Prediction). The popular GoogleNet which is based on Inception model won the ILVRC 2014 competition. Inception Net has different variants V1 to V4. Inception Net V1 was introduced in 2014 while, V2 and V3 were introduced in 2015. In 2016, V4 was presented. The variants improved one after another with certain features. The basic version is V1 to which batch normalization is the prime component. V3 Network got redesigned by Convolutional Kernel refactorization and in V4 is a simplified version of V3 with high performance. It offers higher performance than the ensemble model Inception-ResNet comprising Inception V3 and ResNet 50.

3.3.3 Experimental Setup & Model Training

Loss Function: "Categorical Cross entropy (CCE)" is the loss function used, given the binary classification task.

Optimizer: "Adam" is the optimizer with an adaptive learning rate is used to enhance the convergence speed.

Activation Function: "Relu" is the activation function used in the intermediate layers and "Softmax" is the activation function used in the fully connected layer for classification.

Evaluation Metrics: Evaluation of the models will be done basing on accuracy (A), precision (P), recall (R), F1-score (F1), and AUC. A separate test set will be used for validation to ensure the generalizability of the results.

Cross-Validation: For the model to be robust and perform excellent, 5-fold cross-validation will be employed. Also the model is expected to perform extremely well across various subsets of data.

Hyperparameter Tuning: Grid search will be used to fine-tune hyperparameters viz. learning rate (lr), batch size, and the number of epochs.

Inception V3 & V4 Architectures

Inception V3 & V4 architectures constitute a stem block followed by Inception-A, B & C blocks subsequently reduction A & B blocks in an alternative manner in the I-R-I-R-I format. However, number of blocks in both the network architectures vary and are as given in table 1.

	No. of Inception A Blocks	No. of Inception B Blocks	No. of Inception C Blocks	Stem, Reduction A & B Blocks
Inception V3	3	4	2	1 per each
Inception V4	4	7	3	1 per each

Table 1: Various blocks in Inception V3 and V4

In addition to the number of blocks in which the two architectures differ, the internal architectures of the Stem block differ. However, the internal architectures of Inception – A, B & C blocks and reduction blocks remain the same. The Inception Net architectures of the two variants, as well as the internal block structures, are as shown in Figures 3 and 4.

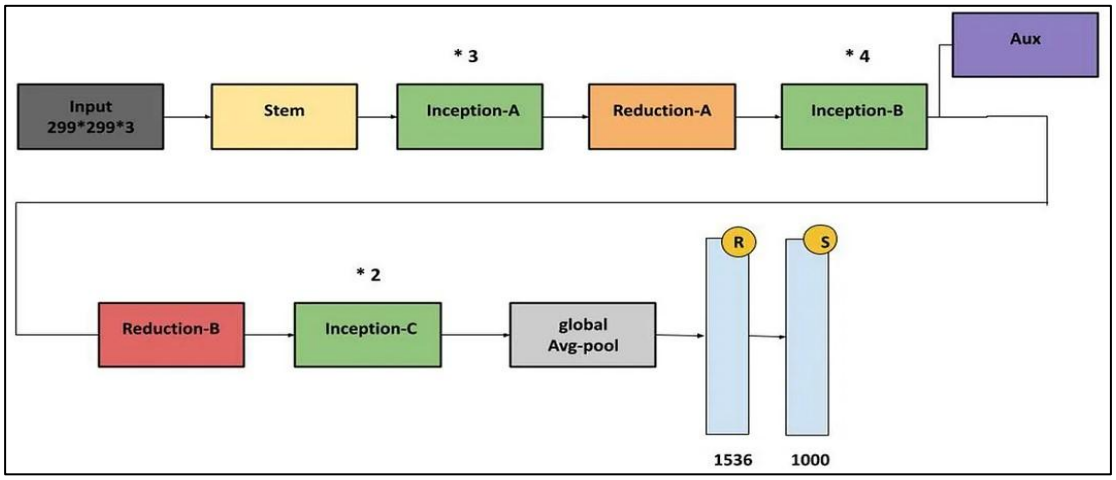


Fig 3: Inception V3 architecture

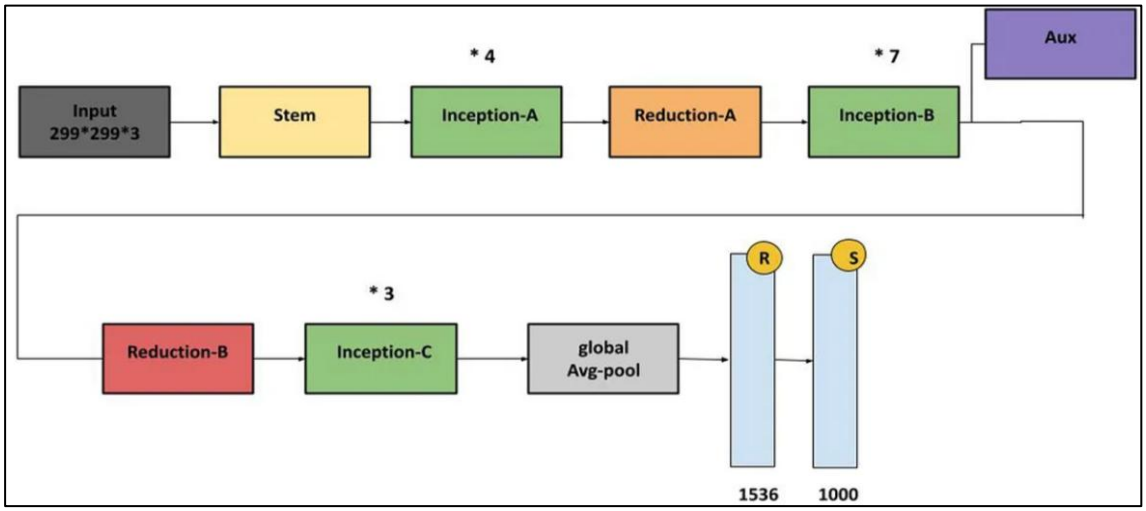


Fig 4: Inception V4 architecture

Once received the input from input layer, the stem block performs feature extraction using a series of convolution layers. The stem blocks of both the architectures are as shown in figures 5 and 6 respectively.

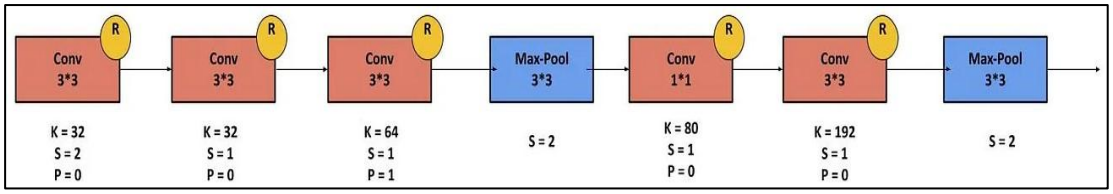


Fig 5: Stem block of V3 architecture

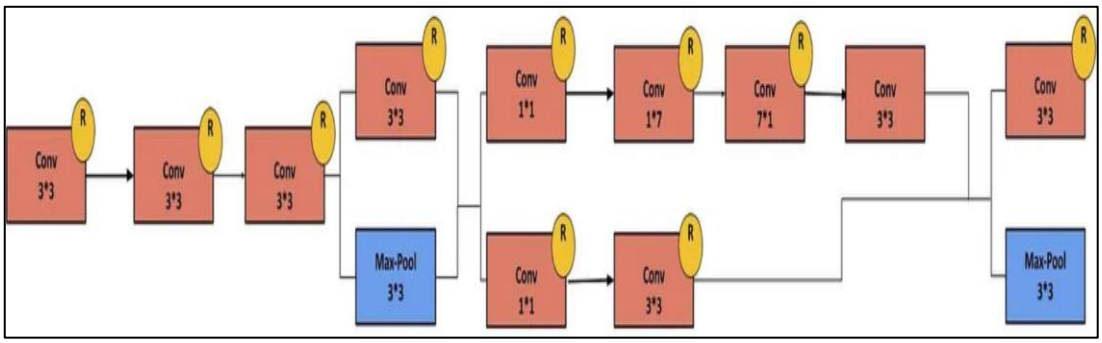


Fig 6: Stem block of V4 architecture

The inception blocks are meant for combining parallel convolutions with different kernel sizes that extracts features at multiple scales. The internal architecture of Inception blocks A, B & C are as shown in figures 7 to 9.

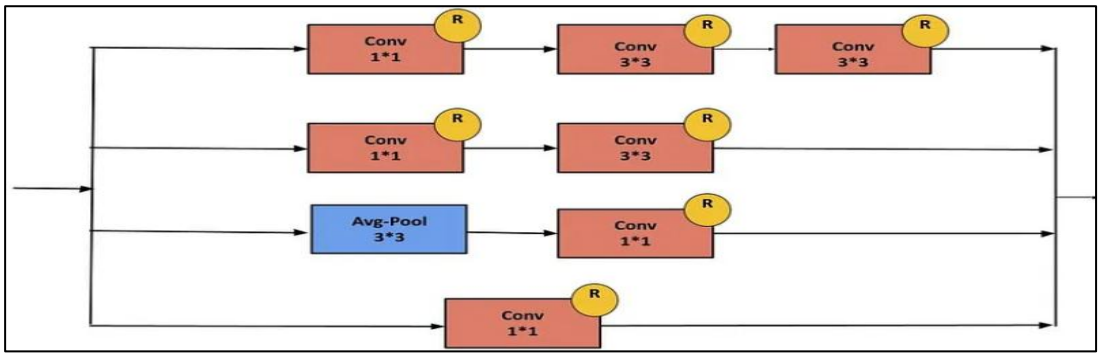


Fig 7: Inception A block

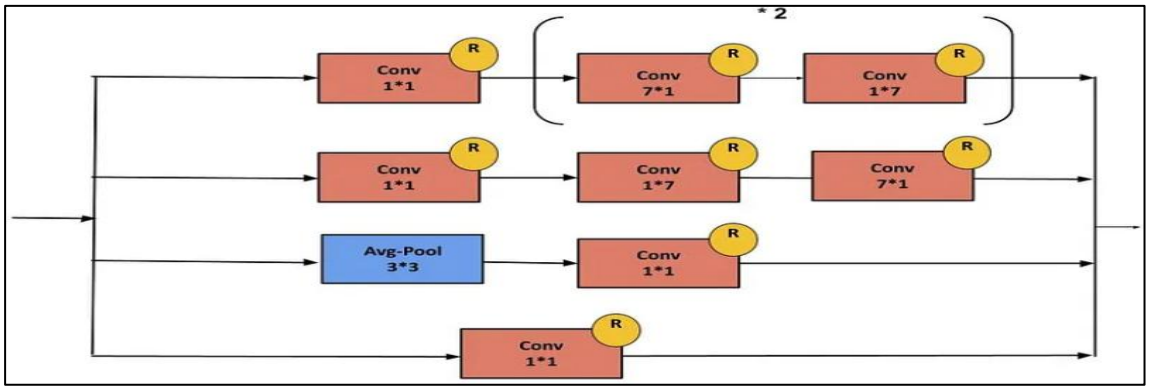


Fig 8: Inception B block

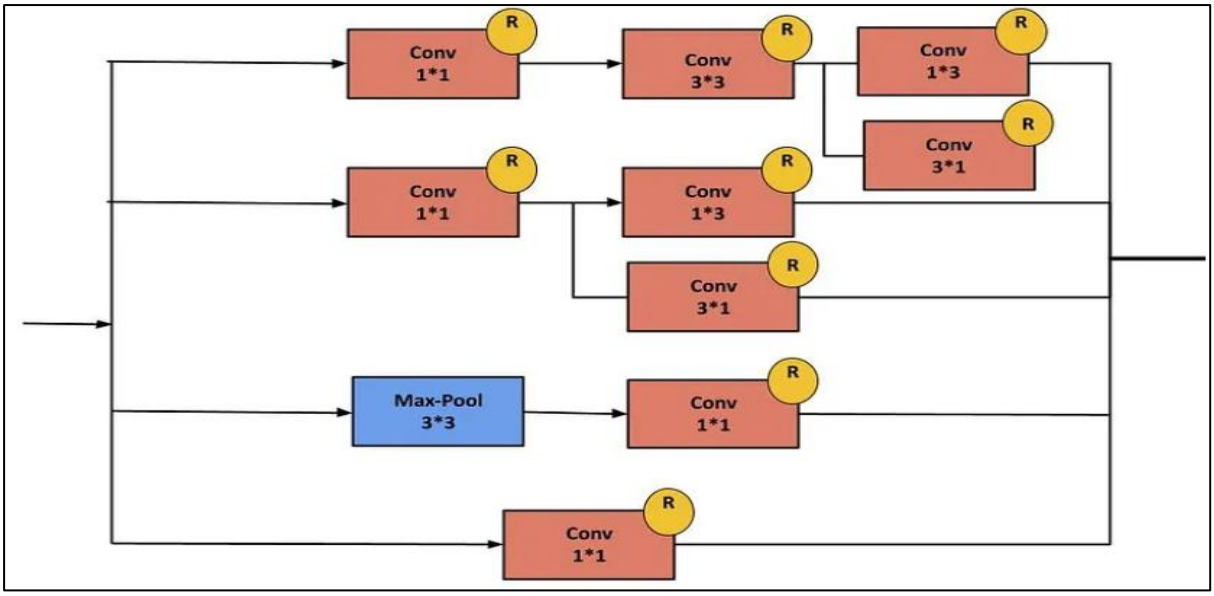


Fig 9: Inception C block

It being the highlight feature of Inception Net, the reduction block is a specialized block that is designed to decrease the spatial dimensions in the feature maps during the transition to deeper layers. This property effectively reduces the computational requirements. The reduction blocks are as shown in Figures 10 and 11.

Fig 10: Reduction A block

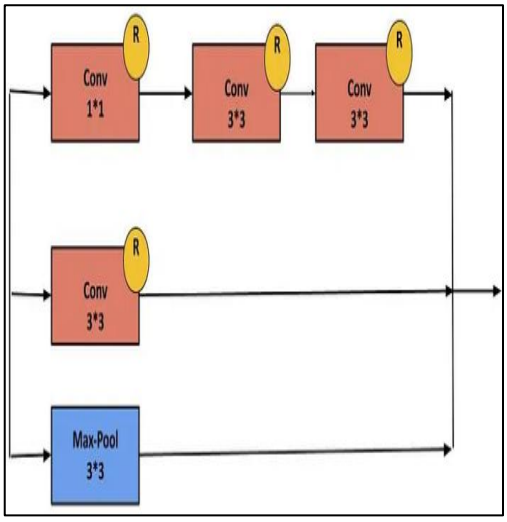
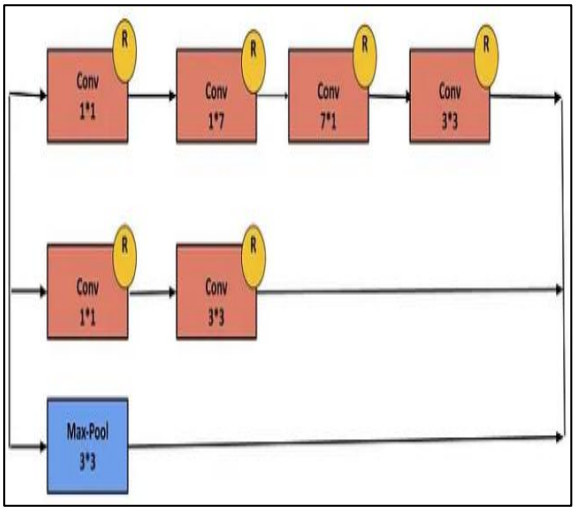


Fig 11: Reduction B block



An auxiliary block is also part of the architecture. An auxiliary block acts as a regularization technique and intermediate learning due to which the features are learnt at different levels of the network. At the end, a global average pooling layer is placed which computes the average value across the height & width of each feature map. Output from this layer is fed to the final classification layer which is a fully connected (FC) layer comprising a dropout layer that outputs 1536 classes and output layer that outputs 1000 classes. The classification block is depicted in figure 12.

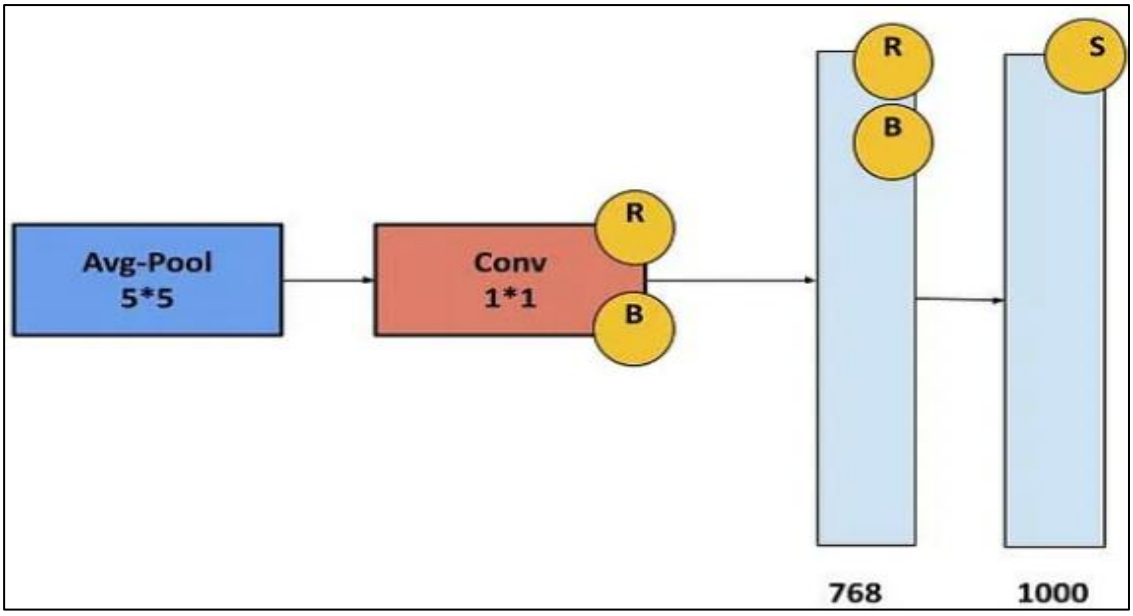


Fig 12: Classification block

Process Flow

The process in classifying the images using the pretrained network involves the preprocessing of the input image as mentioned in earlier section and the converted image of size 299 x 299 x 3 shall be fed as input into the Deep Learning network. The process involves training, validation and testing with 75%, 5% and 20% of images respectively. As Inception V3 is highlighted for Batch normalization the classification is performed with batch sizes 16, 32 & 64 in using Inception V3. Similarly, the classification is performed with batch sizes 32 & 64 using Inception V4. The process flow is as depicted in figure 13. The figure shows the implementation using V3. However, V4 architecture replaces V3 architecture when the other variant is used.

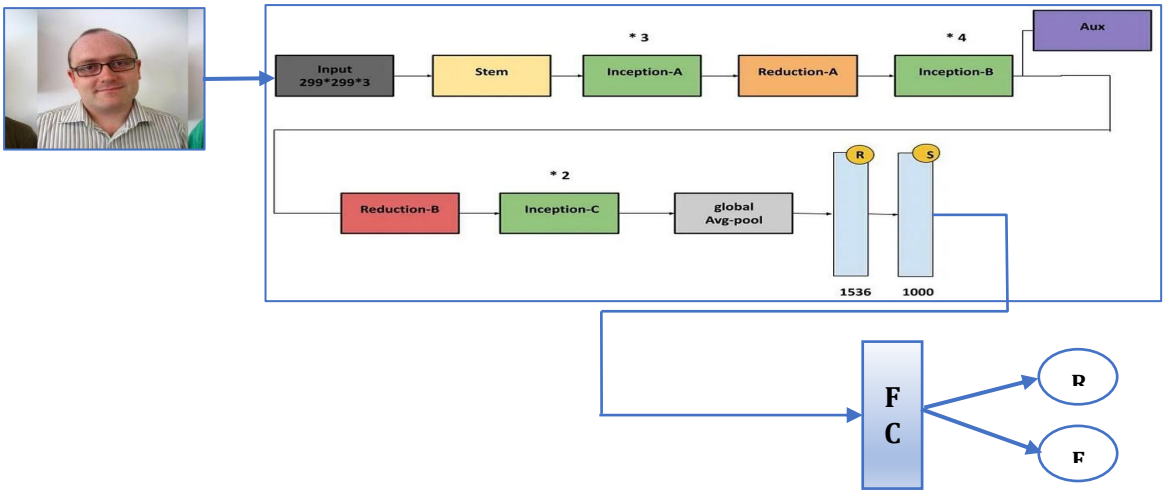


FIGURE 13: PROCESS FLOW

4. RESULTS & DISCUSSION

In this section, the results and comparison of various implementations shall be discussed.

4.1 Inception V3 implementation:

Firstly, the inception v3 architecture has been implemented with batch sizes of 16, 32, and 64, and the respective classification accuracies are mentioned in Table 2 and Figure 14.

Model	Accuracy (%)
Inception V3(16)	85.67
Inception V3(32)	84
Inception V3(64)	85.67

Table 2: Accuracy comparison of Inception V3 variants

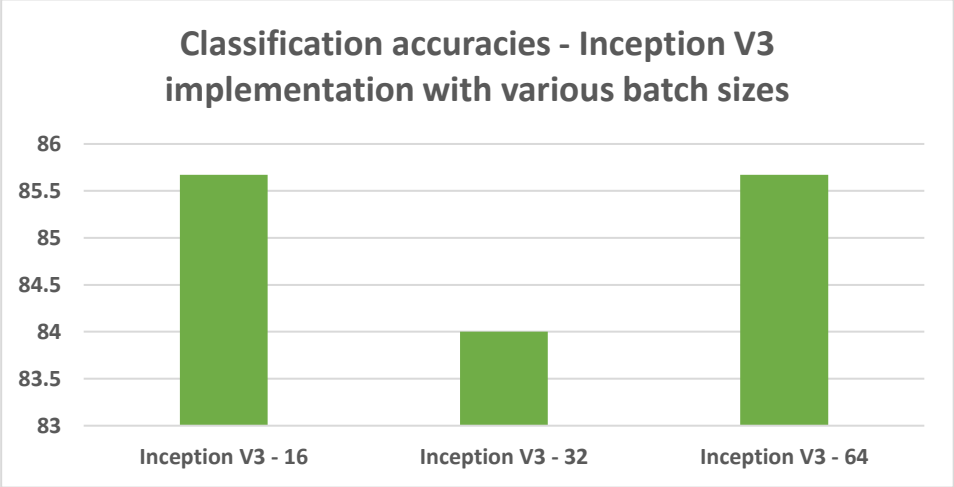


Figure 14: Accuracy comparison of Inception V3 variants

4.2 Inception V3 vs V4 implementation

Inception v4 architecture has been implemented with batch sizes of 32 and 64, and the classification accuracies are found to be the same with both variants. However, the number of parameters as well required memory significantly vary in both implementations. Parameter and memory comparison of the implementation with batch sizes 32 and 64 is as mentioned in Table 3.

	Inception V4 (Batch size 32)	Inception V4 (Batch size 64)
Trainable parameters	54,542,280	5,306,722
Memory required	208.06 MB	20.24 MB
Non-trainable parameters	63,232	12,288
Memory required	247 KB	48 KB
Total parameters	54,605,512	5,319,010
Memory required	208.30 MB	20.29 MB

Table 3: Comparison of parameters and memory requirements between variants of Inception V4 implementation

It could be observed that there is a drastic decrease in parameters and required memory in Inception V4 when compared to V3. Both the inception variants are compared in terms of classification accuracies and are as mentioned in Table 4 and Figure 15.

Model	Accuracy (%)
Inception V3(16)	85.67
Inception V3(32)	84

Inception V3(64)	85.67
Inception V4	94.02

Table 4: Accuracies of Inception V3 and V4 implementations

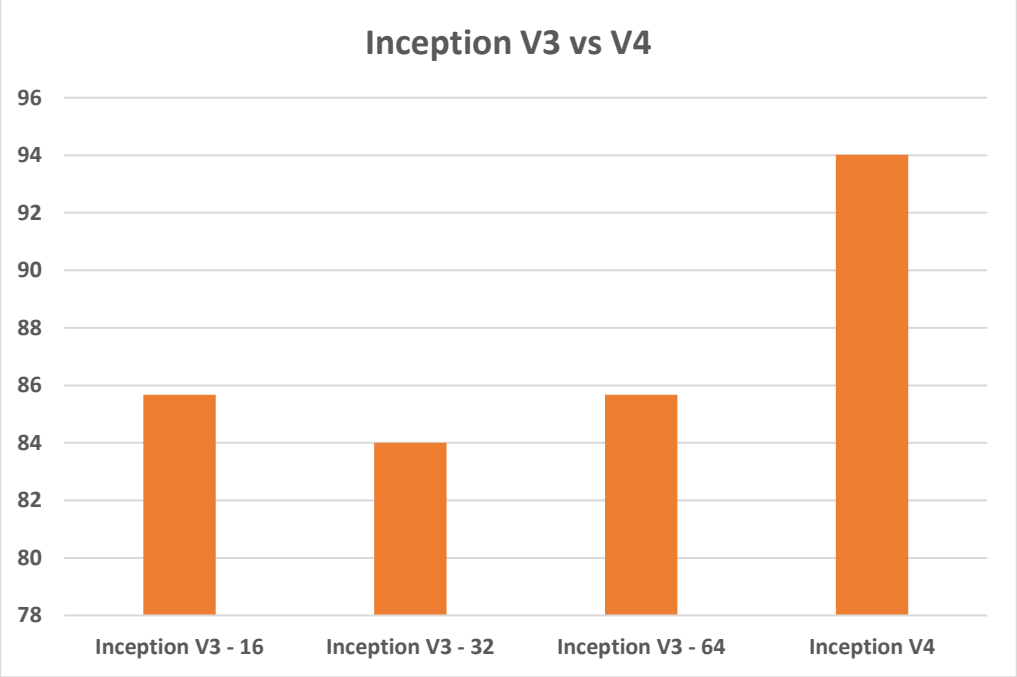


Figure 15: Accuracies of Inception V3 and V4 implementations

4.3 Inception V4 performance

In this subsection, the comparisons of classification accuracy of the Inception V4 implementation is done with the accuracies of earlier contributions made by various researchers. It could be understood that the Inception V4 model outperformed. The results are as mentioned in Table 5 and Figure 16.

Model	**Accuracy (%)**
MobileNet [11]	82.78
ResNet50 [11]	83.33
Xception [11]	84.07
InceptionV3 [11]	85
DenseNet201 [11]	86.58
XceptionNet [12]	88
Inception V4 (Current implementation)	94.02

Table 5: Comparison of Accuracies of Inception V4 with other research contributions

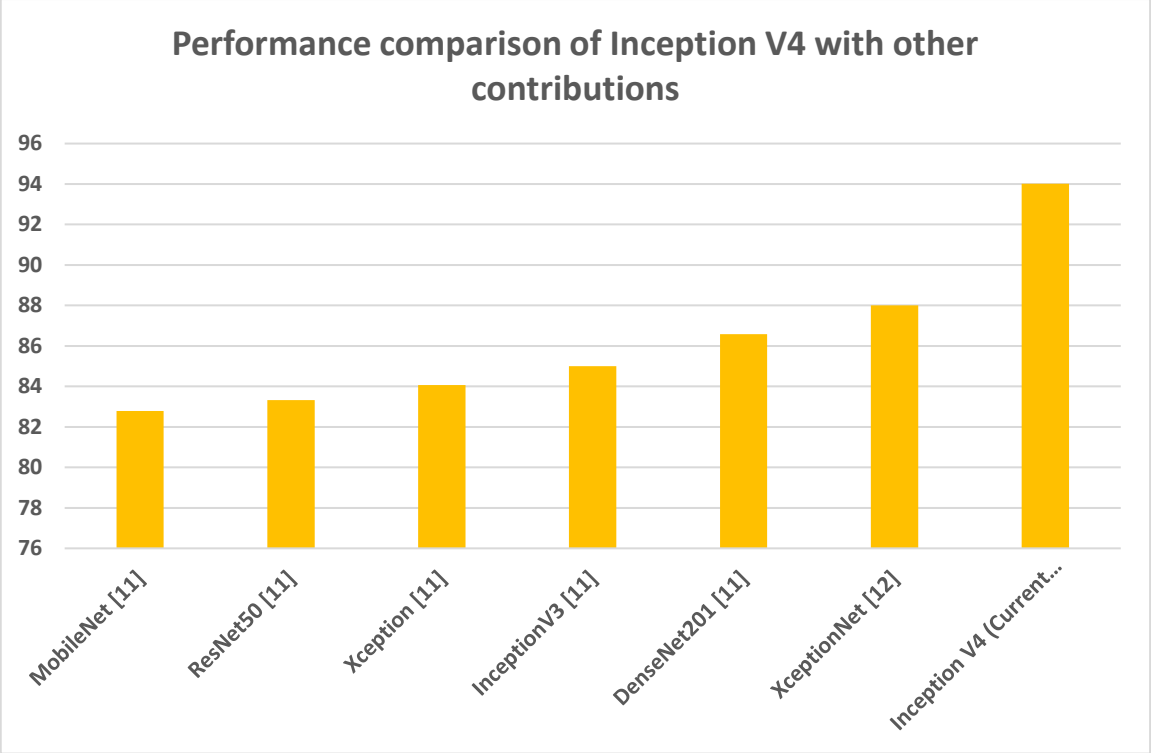


Figure 16: Accuracy comparison of Inception V4 with other research contributions

Moreover, the other metrics, Precision, Recall, F1-score, and Support, are also mentioned as part of the performance of the Inception V4 model. The macro and weighted averages of the above parameters are as mentioned in Table 6. The performance metrics also seem to be commendable.

Metric	Macro average	Weighted average
Precision(P)	0.9426	0.9426
Recall(R)	0.9406	0.9402
F1-score(F1)	0.9416	0.9414
Support(S)	3942	3942

Table 6: Performance metrics of Inception V4 implementation

5. CONCLUSION & FUTURE SCOPE

The access to digital image editing tools and advancements in technology enabled individuals to create and manipulate fake images that seem to be real, leading to a few advantages. On the other hand, Deep-fake videos, which are generated using deep learning techniques, are becoming a major concern, causing harm to individuals as well as society. Therefore, the development of accurate and efficient deepfake detection methods is needed of the hour. In this research, the Inception V3 transfer learning architecture has been adopted, and image classification is done in three variants of Inception V3: Inception V3–16, Inception V3–32, and Inception V3–64, which means the variants are in terms of batch sizes. The classification accuracies with Inception V3 obtained are 85.67, 84, and 85.67 for IV3-16, IV3-32, and IV3-64, respectively. The model is augmented with ResNet, employing an ensemble model to further improve the model accuracy. Further, Classification is carried out with Inception V4, and the accuracy obtained is 94.02. The performance is remarkable when compared with those of other contributions. Finally, we recommend Inception models to be used in computer vision due to their lightweight nature and efficiency. The future insights are that Inception V4 could be augmented with other models and implemented as an ensemble so that the performance still improves.

REFERENCES

1. El-Gayar, M.M., Abouhawwash, M., Askar, S.S. et al. A novel approach for detecting deep fake videos using graph neural network. J Big Data 11, 22 (2024). https://doi.org/10.1186/s40537-024-00884-y.

2. Gong, Liang Yu, and Xue Jun Li. 2024. "A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges" *Electronics* 13, no. 3: 585. https://doi.org/10.3390/electronics13030585.
3. Kaur, A., Noori Hoshyar, A., Saikrishna, V. et al. Deepfake video detection: challenges and opportunities. Artif Intell Rev 57, 159 (2024). https://doi.org/10.1007/s10462-024-10810-6.
4. Gambín ÁF, Yazidi A, Vasilakos A et al (2024) Deepfakes: current and future trends. Artif Intell Rev 57(3):64
5. Sergi D Bray, Shane D Johnson, Bennett Kleinberg, testing human ability to detect 'deepfake' images of human faces, *Journal of Cybersecurity*, Volume 9, Issue 1, 2023, tyad011, https://doi.org/10.1093/cybsec/tyad011.
6. Rafique, R., Gantassi, R., Amin, R. et al. Deep fake detection and classification using error-level analysis and deep learning. Sci Rep 13, 7422 (2023). https://doi.org/10.1038/s41598-023-34629-3.
7. 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, Vol. 40, No. 3, pp. 835-846. https://doi.org/10.18280/ts.400301.
8. Janutenas, L.; ˙ Janutenait ˙ e-Bogdanien ˙ e, J.; Šešok, D. ˙ Deep Learning Methods to Detect Image Falsification. Appl. Sci. 2023, 13, 7694. https://doi.org/10.3390/ app13137694.
9. Mukta, M. S. H., Ahmad, J., Raiaan, M. A. K., Islam, S., Azam, S., Ali, M. E., & Jonkman, M. (2023). An Investigation of the Effectiveness of Deepfake Models and Tools. Journal of Sensor and Actuator Networks, 12(4), 1-43. Article 61. https://doi.org/10.3390/jsan12040061.
10. K. Narayan, H. Agarwal, K. Thakral, S. Mittal, M. Vatsa and R. Singh, "DF-Platter: Multi-Face Heterogeneous Deepfake Dataset," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 9739-9748, doi: 10.1109/CVPR52729.2023.00939. keywords: {Deepfakes;Computer vision;Image coding;Databases;Face recognition;Benchmark testing;Skin;Datasets and evaluation},
11. Atwan J, Wedyan M, Albashish D, Aljaafrah E, Alturki R, Alshawi B. Using Deep Learning to Recognize Fake Faces. Int J Adv Comput Sci Appl. 2024; 15(1):Article 113.Available from: http://dx.doi.org/10.14569/IJACSA.2024.0150113
12. Debasish Samal, ,Prateek Agrawal, VishuMadaan(2024 IMPROVED FAKE IMAGE DETECTION AND CLASSIFICATION USING XCEPTION MODEL. Library Progress International, 44(3), 24541-24549
13. Prezja, F., Paloneva, J., Pölönen, I. *et al.* DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. *Sci Rep* **12**, 18573 (2022). https://doi.org/10.1038/s41598-022-23081-4
14. Narayan, Kartik, Harsh Agarwal, Kartik Thakral, S. Mittal, Mayank Vatsa and Richa Singh. "DeePhy: On Deepfake Phylogeny." *2022 IEEE International Joint Conference on Biometrics (IJCB)* (2022): 1-10.
15. Shehzeen Hussain, Paarth Neekhara, Brian Dolhansky, Joanna Bitton, Cristian Canton Ferrer, Julian McAuley, and Farinaz Koushanfar. 2022. Exposing Vulnerabilities of Deepfake Detection Systems with Robust Attacks. Digit. Threat.: Res. Pract. 3, 3, Article 30 (September 2022), 23 pages. https://doi.org/10.1145/3464307
16. Suratkar, S., Kazi, F. Deep Fake Video Detection Using Transfer Learning Approach. *Arab J Sci Eng* **48**, 9727–9737 (2023). https://doi.org/10.1007/s13369-022-07321-3
17. Xu FJ, Wang R, Huang Y et al (2022) Countering malicious deepfakes: survey, battleground, and horizon. Int J Comput Vis. https://doi.org/10.1007/s11263-022-01606-8
18. Zhang, L.; Lu, T. Overview of Facial Deepfake Video Detection Methods. J. Front. Comput. Sci. Technol. 2022, 17, 1–26.
19. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic Routing between Capsules. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3859–3869.
20. Karki M. Deepfake and Real Images [Data set]. Kaggle. 2023. Available from https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images/data

fffff