# Exploring Various Deep Learning Techniques for DeepFake Video Classification

A.Nymisha Nandini Reddy[1] ,B.Hemalatha[2] ,Neha Saw[3] ,Vikram Kumar[4] ,Bhaskar Das[5]
[1,2,3,4]Student,[5]Assistant Professor
[1]Computer Science with specialization in Artificial Intelligence and Machine Learning,
[1]Hyderabad Institute of Technology and Management,Hyderabad,India

**Abstract** - The rapidly developing use of deepfake technology affects digital media credibility and public trust.Deepfakes are created using advanced machine learning techniques, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), to produce highly realistic synthetic images, videos, and audio by manipulating original media.The current study develops a powerful deepfake video detection system using advanced deep learning algorithms. We fine-tuned three CNN architectures—InceptionResNetV2, EfficientNetB5, and VGG16—using around 5GB of data from the Kaggle Deepfake Detection Challenge dataset. Transfer learning improved our models' efficiency and performance. This study emphasizes the necessity for a robust deepfake detection technique. Future research will center on full-body deepfake detection and real-time capabilities for enhanced digital media authentication.

**Index Terms** - Deepfake,Convolutional Neural Network(CNN),Deep Learning,Transfer Learning

## I.INTRODUCTION

Deepfake, a term derived from "deep learning" and "fake," refers to synthetic media where the likeness of an individual is replaced with someone else's. This technology leverages advanced machine learning techniques, particularly Generative Adversarial Networks (GANs), to create highly realistic fake videos, images, and audio. These deepfakes can be used for various malicious purposes, including spreading misinformation, identity theft, and creating non-consensual explicit content. The rise of deepfakes poses significant challenges to the authenticity of digital media and public trust.

Recent research has focused on developing detection methods to combat the proliferation of deepfakes.Despite these efforts, the rapid evolution of deepfake generation techniques necessitates continuous advancements in detection methodologies[1].The need for effective detection mechanisms is crucial, not only for safeguarding the integrity of digital content but also for protecting individuals' privacy and reputation.Enhancing the robustness of detection models to handle low-quality and compressed videos is increasingly important, given the prevalence of such formats on social media platforms. Additionally, interdisciplinary approaches combining technical solutions with legal and ethical frameworks are vital to address the broader societal implications of deepfake technology

The remainder of this paper is organized as follows: Section II delves into existing detection methodologies, discussing various techniques and their effectiveness in identifying deepfakes. In Section III, we present our proposed detection framework, detailing the underlying algorithms and implementation specifics. Section IV provides an evaluation of our framework, comparing its performance against other state-of-the-art methods. In Section V, we discuss the potential areas for future research.Finally, Section VI concludes the paper with a summary of our contributions and a discussion on the broader impact of deepfake detection technologies.

## II.LITERATURE SURVEY

In Deep Fake Video Detection Using Transfer Learning Approach, **Shraddha Suratkar & Faruk Kazi [2]** utilizes a sophisticated framework combining transfer learning in autoencoders and a hybrid model of convolutional neural networks (CNN) and recurrent neural networks (RNN) to detect deep fake videos. The approach leverages the strengths of CNNs in capturing spatial features and RNNs, particularly Long Short-Term

Memory (LSTM) networks, in handling temporal sequences, thus enhancing the detection of fake videos.However, the study acknowledges certain limitations, including the potential for overfitting due to the extensive fine-tuning process and the challenge of maintaining high accuracy when encountering novel, sophisticated deep fake techniques not represented in the training data.

In An attention-based DeepFake detection (ADD) approach, **Aminollah Khormali & Jiann-Shiun Yuan [3]** focuses on the fine-grained and spatial locality attributes of artificially synthesized videos for enhanced detection. The ADD framework consists of face close-up and face shut-off data augmentation methods and is applicable to any classifier based on convolutional neural network architecture. The method first locates potentially manipulated areas of the input image to extract representative features and then forces the detection model to pay more attention to these forgery regions in the decision-making process through a particular focus on interpreting the sample in the learning phase. The ADD's performance is evaluated against two challenging datasets of DeepFake forensics, Celeb-DF (V2) and WildDeepFake, and demonstrates significant improvement in detection performance. However, the study acknowledges certain limitations, including limitation of the research that the datasets used in the study, Celeb-DF (V2) and WildDeepFake, were collected from public Internet and YouTube videos, and no consents were obtained. This may raise ethical concerns regarding the use of data without proper consent and the potential impact on privacy rights. Additionally, the study is limited by the availability and quality of the data.

In DeepVision: Deepfakes detection using human eye blinking pattern, **Tackhyun Jung & Sangwon Kim [4]** utilizes the DeepVision methodology leverages significant changes in human eye blinking patterns to detect deepfakes generated by GANs. By analyzing eye blinking influenced by factors such as gender, age, activity, and time of day, DeepVision tracks and verifies the integrity of these patterns against a pre-configured database. This process combines heuristic methods with machine learning, integrating insights from medicine, biology, brain engineering, and statistical algorithms to differentiate between real and fake videos. The method demonstrated an accuracy rate of 87.5% in detecting deepfakes. However, the study identifies limitations, particularly regarding individuals with mental illnesses or nerve conduction pathway issues that may alter natural blinking patterns, potentially affecting the accuracy of the detection algorithm.

In "Exposing DeepFake Videos By Detecting Face Warping Artifacts",**Yuezun Li & Siwei Lyu [5]** exploits a key limitation of DeepFake algorithms: their inability to generate high-resolution face images, necessitating warping to fit synthesized faces into target videos, which introduces distinctive artifacts. This method employs Convolutional Neural Networks (CNNs) to detect these artifacts, eliminating the need for DeepFake-generated images as negative training examples by simulating resolution inconsistencies in affine face warpings. The approach involves training CNN models (VGG16, ResNet50, ResNet101, and ResNet152) on real face images and simulated negatives, showcasing high effectiveness in distinguishing real from fake videos across multiple datasets. However, the method's limitations include potential challenges in maintaining accuracy with videos subjected to multiple compressions and the need for further refinement of network structures to enhance detection efficiency and robustness against evolving deepfake techniques.

In Deepfake Video Detection Using Recurrent Neural Networks, **David Güera & Edward J. Delp [6]** leverages a temporal-aware pipeline that integrates convolutional neural networks (CNN) to extract frame-level features and recurrent neural networks (RNN), specifically a convolutional LSTM structure, to classify manipulated videos by analyzing temporal sequences. This approach addresses the inherent weaknesses in deepfake video generation, such as scene inconsistency and the lack of temporal awareness, which can lead to anomalies like flickering in manipulated videos. The model demonstrates high accuracy, achieving a 99% detection rate. However, the limitations include potential challenges in generalizing across different types of deepfake videos, particularly those with more sophisticated manipulation techniques that may not exhibit the same temporal inconsistencies or flickering artifacts. Additionally, the reliance on the quality and variability of the training data

means that the model's performance could vary with different datasets, potentially affecting its robustness and generalizability in real-world applications.

Major challenges in deepfake detection include difficulties in generalizing to new data, the exclusion of audio data, and issues with real-time scalability. Furthermore, reliance on specific features like eye blink patterns and the rapidly evolving nature of deepfake technology necessitates the development of more robust and comprehensive detection approaches. Ongoing research is essential to create adaptable and effective deepfake detection systems.

## III.METHODOLOGY

The input to the system consists of videos which are processed to extract individual frames. The frame extraction is performed at a rate of one frame per second to balance the computational load and ensure sufficient data for training.The frames are then processed using a face detection algorithm to isolate and crop face regions. The cropped face images are resized to a standardised dimension of 128x128 pixels for consistency in model training.After extracting frames, the images are augmented using various techniques to increase the robustness and diversity of the training data.

The dataset is then split to facilitate effective model training and evaluation.The model has been trained on a dataset consisting of approximately 800 videos sourced from the Kaggle Deepfake Detection Challenge. These videos, generated using advanced machine learning techniques, serve as a vital resource for developing deepfake detection systems. With a size of around 5GB, the dataset ensures access to diverse deepfake scenarios, enhancing model robustness.Three advanced deep learning models were employed for feature extraction and classification: InceptionResNetV2, EfficientNetB5, and VGG16.

InceptionResNetV2 combines the strengths of Inception and ResNet modules, excelling in capturing intricate features and dependencies within images. Transfer learning is leveraged using a pre-trained InceptionResNetV2 model initially trained on ImageNet.A `Sequential` model was constructed, starting with the InceptionResNetV2[8] base model followed by a `GlobalAveragePooling2D` layer to reduce dimensionality. A `Dropout` layer with a rate of 0.5 was added to mitigate overfitting, followed by a final `Dense` layer with 2 units and a `softmax` activation function for binary classification.The model was compiled using the Adam optimizer with a learning rate of 1e-5, $\beta1=0.9$, $\beta2=0.999$, and $\varepsilon=1e-7$. The binary cross-entropy loss function was chosen to optimize model weights, with accuracy as the evaluation metric.

The EfficientNetB4 model[7], a state-of-the-art convolutional neural network architecture, was selected for its strong performance in image classification tasks. A `Sequential` model was constructed,with the EfficientNetB4 base model.The model is trained with a binary cross-entropy loss function and fine-tuned using transfer learning.The model was compiled using the Adam optimizer with a learning rate of 1e-5, $\beta1=0.9$, $\beta2=0.999$, and $\varepsilon=1e-7$. The binary cross-entropy loss function was chosen to optimize model weights, with accuracy as the evaluation metric.

The VGG16 architecture, known for its simplicity and effectiveness, was chosen as a comparative model. A `Sequential` model was constructed with the VGG16 base model followed by a stack of fully connected `Dense` layers, employing `relu` activation to extract high-level features. The final layer consisted of 2 units with a `softmax` activation for binary classification.The SGD optimizer was employed for VGG16 with a learning rate of 0.001, a momentum of 0.9, and Nesterov acceleration disabled. Binary cross-entropy loss was used for training, with accuracy as the evaluation metric.

The design involves a comprehensive pipeline starting with data import and preparation, followed by data splitting for training and validation as mentioned in Fig.1. It incorporates data augmentation techniques such as random cropping and flipping. Frames are extracted from videos, normalized, resized, and subjected to face detection. The combined dataset is split into training and test sets. Three models (InceptionResNetV2, EfficientNetB5, VGG16) are trained, featuring dense layers and softmax outputs. Model performance is assessed through accuracy and loss calculations, along with confusion matrices, and models are saved and tested for validation.

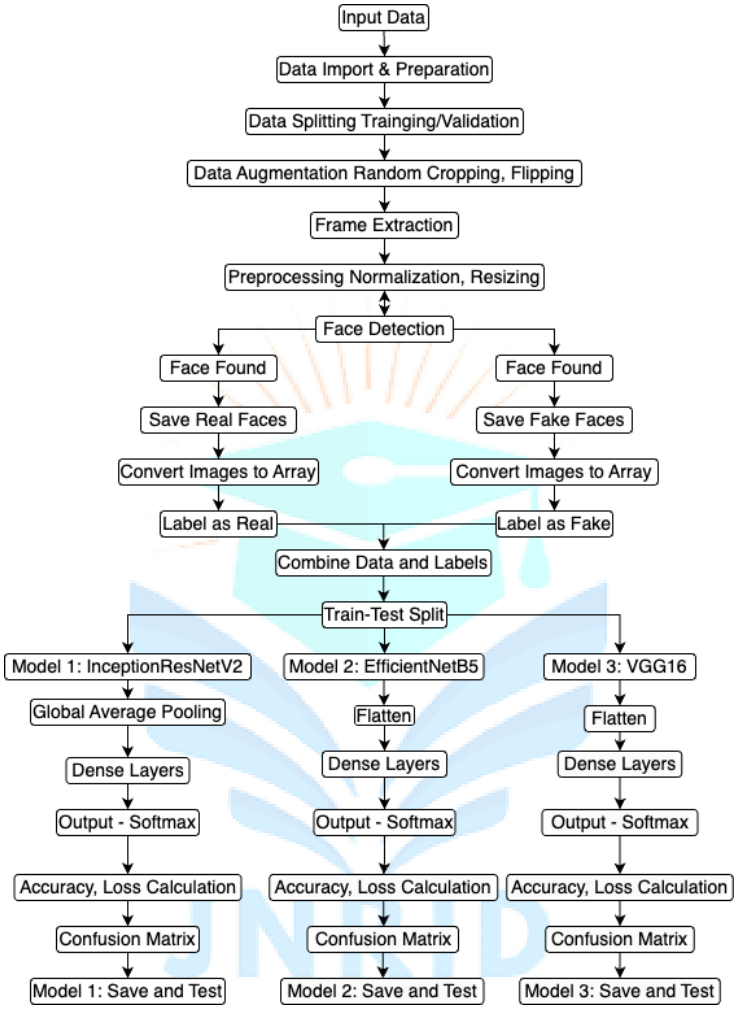


Fig.1 Design of the Model

## IV. Performance Evaluation

Each model is trained on the training dataset with early stopping and learning rate reduction callbacks to prevent overfitting and optimize the learning process.The training history is saved and visualized to monitor the model's performance over epochs.

The models are evaluated on the validation set using accuracy and loss metrics. The best-performing model which is InceptionResNet V2 is selected based on its accuracy and computational efficiency.

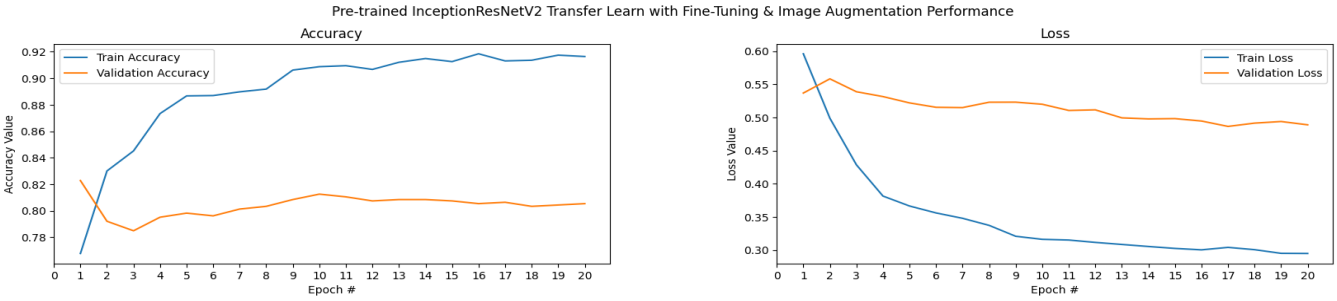


Fig.2 Design of the Model

The model exhibited a steady increase in training accuracy as shown in Fig.2 , starting from 0.78 and reaching approximately 0.92 by epoch 20.Validation accuracy showed initial improvement but plateaued around 0.80 to 0.82 after epoch 4.Training loss decreased significantly from around 0.60 to 0.30, indicating effective minimization of error on the training data.Validation loss initially decreased but stabilized between 0.45 and 0.50, mirroring the trend observed in validation accuracy.

## V.FUTURE AREAS OF RESEARCH

While our study has demonstrated the effectiveness of fine-tuning pre-trained models such as InceptionResNetV2, there remain several avenues for enhancing the robustness and generalization capabilities of deepfake detection systems.

One promising area for future research is full-body deepfake analysis. Current deepfake detection methods predominantly focus on facial features, leaving other manipulated aspects of the body less scrutinized. Expanding the scope to include full-body detection can significantly improve the accuracy and reliability of these systems. DeepFakes can manipulate body movements, gestures, and other non-facial features that, when detected, provide additional evidence of tampering. By analyzing the entire body, deepfake detection models can capture inconsistencies and anomalies that are missed when focusing solely on the face, leading to more comprehensive and effective detection strategies.

Another crucial area is exploring advanced data augmentation techniques. Utilizing sophisticated methods, such as generative adversarial networks (GANs) for creating realistic synthetic data, could provide a richer and more diverse dataset for training deepfake detection models. GANs can generate high-quality, varied training samples that enhance the model's ability to generalize to unseen data. This approach helps to overcome the limitations of existing datasets, ensuring that the models are robust against a wider range of deepfake techniques and variations. By augmenting the training data with synthetic yet realistic examples, we can significantly improve the model's performance and resilience to new and evolving deepfake methods.

## VI.CONCLUSIONS

Among the trained models, we found that the InceptionResNetV2 model performed the best. Therefore, we have provided the performance metrics of this model, which demonstrates its effectiveness in detecting deepfakes while also highlighting areas where it can be further improved. The InceptionResNetV2 model achieved high accuracy on the training data, but the validation performance metrics plateaued, indicating a risk of overfitting.

Our contributions underscore the importance of ongoing research and development in this field, highlighting the need for innovative approaches to tackle the evolving challenges posed by deepfake technology. By focusing on full-body deepfake analysis and exploring advanced data augmentation techniques, future research can build upon our findings to create more robust and reliable deepfake detection systems. Additionally, integrating temporal

analysis of video frames and leveraging ensemble learning methods may further enhance detection accuracy and generalization capabilities.

The broader impact of deepfake detection technologies is profound, as they play a critical role in maintaining the integrity of digital media. Effective detection systems can help mitigate the spread of misinformation and protect individuals and organizations from malicious activities. However, it is essential to continue advancing these technologies while addressing ethical and privacy concerns to ensure their responsible use. Through ongoing innovation and collaboration, the field of deepfake detection can continue to evolve, providing critical tools for safeguarding digital content and fostering trust in digital communications.

## VII. REFERENCES

[1] Masood, M., Nawaz, M., Malik, K.M. *et al.* Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Appl Intell* **53**, 3974–4026 (2023). https://doi.org/10.1007/s10489-022-03766-z.

**[**2] 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.

[3] Aminollah, and Jiann-Shiun Yuan. 2021. "ADD: Attention-Based DeepFake Detection Approach" *Big Data and Cognitive Computing* 5, no. 4: 49. https://doi.org/10.3390/bdcc5040049.

[4] T. Jung, S. Kim and K. Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern," in *IEEE Access*, vol. 8, pp. 83144-83154, 2020, doi: 10.1109/ACCESS.2020.2988660.

[5] Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts. arXiv 2018. arXiv preprint arXiv:1811.00656.

[6] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Auckland, New Zealand, 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639163.

[7]A. A. Pokroy and A. D. Egorov, "EfficientNets for DeepFake Detection: Comparison of Pretrained Models," 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), St. Petersburg, Moscow, Russia, 2021, pp. 598-600, doi: 10.1109/ElConRus51938.2021.9396092. keywords: {Social networking (online);Training data;Media;Task analysis;Faces;Videos;Information integrity;deepfake videos;deep learning;digital media forensics;detection techniques}

[8] Shuai Peng, Hongbo Huang, and Weijun Chen. More trainable inception-resnet for face recognition. In Neurocomputing, pages 9–19. sciencedirect, 2020. https:// www.sciencedirect.com/science/article/abs/pii/S0925231220308572