ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2035

Deepfake Detection Using Convolutional Neural Network

Dr. Anant N. Kaulage[1]; Pranjali Kolawale[2]; Anika Tuli[3]; Isha Liddad[4]; Tanvi Jadhav[5]

[1]School of Computing MIT-ADT University, Pune, India
[2]School of Computing MIT-ADT University, Pune, India
[3]School of Computing MIT-ADT University, Pune, India
[4]School of Computing MIT-ADT University, Pune, India
[5]School of Computing MIT-ADT University, Pune, India

Publication Date: 2026/05/06

Abstract: Deepfake technology, driven by modern AI models, has made video manipulation more convincing and harder to detect. This project aims to develop a deep learning based system that can automatically identify whether a video is real or artificially generated. The approach uses the EfficientNetB0 architecture as a feature extractor, trained on frames extracted from both authentic and manipulated video datasets. Frames were preprocessed, resized, and augmented to enhance the model's learning ability to learn the given pattern. The training process involved two stages: initial training of top layers and subsequent fine-tuning of the entire network for higher accuracy. Further refinement was done using real video samples to enhance prediction stability. A threshold calibration method was applied to decide the real or fake nature of videos based on average prediction scores across frames. The final model efficiently distinguishes between real and fake content, demonstrating strong performance and potential use in video authenticity verification.

Keywords: *Convolutional Neural Network, Deepfake, Fine Tuning.*

How to Cite: Dr. Anant N. Kaulage; Pranjali Kolawale; Anika Tuli; Isha Liddad; Tanvi Jadhav (2026) Deepfake Detection Using Convolutional Neural Network. *International Journal of Innovative Science and Research Technology*, 11(4), 3387-3393. https://doi.org/10.38124/ijisrt/26apr2035

I. INTRODUCTION

In the today's modern digital age, where artificial intelligence and machine learning serve as the foundation of innovation, the distinction between reality and fabrication has become increasingly blurred. Among the most transformative yet controversial advancements is the development of deepfake technology, which uses deep neural networks capable of manipulating or replacing human faces and voices with remarkable realism. While this technology offers positive applications in entertainment, virtual reality, and digital media, it simultaneously poses threats to personal privacy, social trust, and authenticity of information.

Traditional methods for detecting video forgeries often rely on handcrafted features or simple inconsistencies in pixel values and lighting. However, these techniques struggle to cope with the high realism of Deepfakes generated by modern AI models. To address these challenges, deep learning techniques have gained importance due to their ability to automatically learn complex spatial and temporal features that help differentiate real content from manipulated media. This project implements a deep learning model based on the EfficientNetB0 architecture to effectively detect Deepfake videos. EfficientNetB0 is widely known for its balanced scaling of network depth, width, and resolution, enabling efficient feature extraction while keeping computational efficiency.

The proposed system operates on a frame-level analysis approach, where videos are first converted into individual image frames using OpenCV. These frames are resized and preprocessed to maintain consistency and improve the model's ability to handle different samples. Data augmentation techniques such as rotation, flipping, and zooming, are applied to further enrich the training data and minimize overfitting. The model training phase is divided into two stages: in initial stage, only the top layers are trained while the base EfficientNet layers remain frozen, in the second stage, all layers are unfrozen for fine-tuning at a lower learning rate to optimize performance. The two phase process ensures that the model retains useful pretrained features while adapting effectively to the Deepfake dataset.

To further enhance model accuracy, an additional quick fine-tuning step is performed using real video samples. A threshold calibration technique is then introduced to determine the optimal boundary for classifying videos as real

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2035

or fake based on their mean prediction scores. This allows the system to provide a more stable and interpretable decision. The final model is capable of accurately differentiating authentic videos from manipulated ones, offering a dependable and efficient approach for detecting a deepfake.

Through this work, the project demonstrates the practical potential of transfer learning and deep neural networks in addressing one of the most pressing challenges in digital media integrity.

II. LITERATURE REVIEW

The area of deepfake detection has transitioned rapidly from handcrafted, artifact-based techniques to data-driven deep learning solutions. Early foundational work on deep convolutional architectures demonstrated the value of deep feature hierarchies for visual recognition, laying the groundwork for later forensic models [1]. At the same time, studies in legal and societal analyses highlighted the severe risks created by synthetic media motivating technical countermeasures that span computer vision and multimedia forensics [2].

Several detection strategies exploit physiological and geometric inconsistencies introduced by synthesis. Li et al. showed that temporal cues such as abnormal eye-blinking patterns are reliable indicators of manipulation and proposed temporal models to leverage this behavior [3]. Face-warping artifacts arising from imperfect alignment or blending were identified as a distinct forensic cue in subsequent studies, and algorithms that detect these spatial artifacts achieved notable accuracy on manipulated datasets [4]. Complementing these approaches, Afchar et al. introduced MesoNet, a compact mesoscopic CNN that targets mid-level facial anomalies and offers a computationally efficient detection option suitable for constrained environments [5].

Geometric analysis of head pose and motion dynamics has also proven effective: inconsistent head orientation between source and synthesized frames can reveal tampering, particularly in reenactment scenarios [6]. To capture both spatial and temporal discrepancies, hybrid architectures combining CNNs with sequential models such as RNNs and LSTMs were introduced; these yield improved robustness to subtle temporal artifacts [13], [17]. Capsule networks were explored as an alternative that can model part-whole relationships and resist certain adversarial perturbations [7], providing another path toward reliable forgery detection.

Researchers have also focused on exploiting diverse visual artifacts and network-specific weaknesses. Matern et al. and Marra et al. analyzed lighting, blending, and distributional artifacts that persist after synthesis, demonstrating that artifact-aware detectors generalize better across manipulation methods and compression levels [8], [19]. Benchmarking initiatives involving models and datasets, including the DFDC preview dataset, have played an important role in evaluating system performance and revealing real-world challenges such as compression artifacts, occlusions, and low-resolution inputs [14].

Systematic reviews and vulnerability assessments emphasize the arms-race nature of this domain: as generative models and training data improve, previously useful artifacts become less reliable, and detection must evolve accordingly [12], [16]. Practical detection systems therefore require not only strong spatial feature extractors (e.g., VGG-style or Inception modules for local patterns [1], [11]) but also architectures that balance accuracy with computational efficiency (e.g., EfficientNet scaling strategies) and resilience to compression [12].

Other contributions address complementary problems that affect detection pipelines: robust face tracking and alignment are critical preprocessing steps for stable feature extraction [18], while domain diversity and large annotated corpora (FaceForensics++, DFDC) improve generalization during training [2], [14]. Legal and ethical scholarship further underscores the need for detection methods that are explainable and legally admissible [2], [9], and real-world deployment demands lightweight models that can operate under constrained resources [5], [20].

Table 1 Literature Review

Title & Authors	Summary	Limitations
"In-Depth Fake Face Detection using Eye Blinking" – Li et al., 2018	Proposed detection of deepfakes by analyzing unnatural eye blinking patterns, effective for early deepfakes.	Fails on high-quality GAN-generated videos, limited robustness.
"FaceForensics++: Learning to Detect Manipulated Facial Images" – Rossler et al., 2019	Introduced large-scale dataset (FaceForensics++) and CNN-based benchmark models for image/video deepfake detection.	Requires large datasets, computationally intensive, limited explainability.
"Exposing DeepFake Videos by Detecting Face Warping Artifacts" – Nguyen et al., 2019	Used CNN + LSTM to capture spatial and temporal inconsistencies across video frames for detection.	Limited performance on unseen GAN manipulation techniques, high model complexity.
"GAN-Generated Video Detection" – Korshunov & Marcel, 2020	Evaluated detection methods against modern GAN-generated videos, emphasized need for robustness.	Existing methods often fail on new GANs, lacks real-time applicability.
"Multi-modal Deepfake Detection using AudioVisual Cues" – Chugh et al., 2020	Detected inconsistencies between lip movements and audio to identify fake videos.	Only applicable to videos with audio, less effective on silent deepfakes.

III. PROPOSED SYSTEM

The proposed system introduces an end-to-end Deepfake Detection Framework designed to automatically detect manipulated or synthetically generated videos through detailed frame-level analysis and adaptive decision calibration. The framework focuses on extracting spatial features from facial regions using a deep CNN and aggregating the predictions across frames to make a final decision at the video level. By employing the EfficientNetB0 architecture as the backbone model and integrating adaptive thresholding for improved decision boundaries, the system gains a balance between computational efficiency and detection accuracy.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2035

The framework follows a sequential pipeline that includes video preprocessing, feature extraction, model training, fine-tuning, threshold calibration, and final prediction generation. Each component of this pipeline contributes significantly to ensuring that the system performs reliably across various datasets, including videos generated by different manipulation approaches such as face replacement, lip-synchronization, and reenactment.

Fig 1 Frame Samples

➢ *System Overview*

The system operates by decomposing input videos into static frames and evaluating them individually using a deep learning model trained on both authentic and manipulated video samples. By aggregating frame-wise predictions, the system estimates a mean confidence score that determines whether a given video is real or fake. This approach enables robust detection performance, even when manipulations occur in specific temporal segments rather than across entire videos.

➢ *Data Preprocessing*

Each input video is first processed using OpenCV to extract uniformly spaced frames. Frames are resized to 128×128 pixels to maintain standardized input size compatible with the EfficientNet architecture. The dataset is structured into labeled folders for real and fake samples. To enhance generalization, data augmentation methods such as horizontal flipping, image rotation, and zoom transformation are applied using the ImageDataGenerator module. This ensures that the model remains resilient to variations in pose, lighting conditions, and facial expressions.

➢ *Model Architecture*

The detection model employs EfficientNetB0, a CNN architecture pre-trained on ImageNet, known for its high accuracy-to-computation efficiency ratio. The pre-trained base is integrated with a custom classification head comprising:

GlobalAveragePooling2D → Dropout(0.4) → Dense(1, activation='sigmoid')

This configuration enables the network to map high-dimensional features to a binary classification output, distinguishing real frames from fake ones. The model is optimized using the Adam optimization algorithm, while binary cross-entropy loss function and the performance evaluated using accuracy as the primary metric.

➢ *Training and Fine-Tuning Strategy*

The training process is divided into two phases:

- *Initial Training Phase:*

During the first phase of the, EfficientNetB0 layers are kept frozen, and only the newly added Dense layers are trained. This enables the model to learn features specific to the target task while preserving the knowledge already learnt from the pre-trained network.

- *Fine-Tuning Phase:*

In the next stage all layers of the network are unfrozen and trained with a smaller learning rate. This helps the model refine its understanding and capture subtle differences between real and manipulated frames. Techniques such as Advanced EarlyStopping and ReduceLROnPlateau are applied to control overfitting and automatically adjust the learning rate during training.

Following the main training, a secondary fine-tuning stage is conducted using additional real samples to recalibrate

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2035

the model's distribution and improve detection reliability in authentic video contexts.

➢ *Threshold Calibration*

After model training, an adaptive threshold calibration process is applied to enhance classification precision. The mean prediction scores of real and fake datasets are computed, and the optimal decision threshold is determined as the midpoint between these two means:

$$T = \frac{(\text{mean}_{real} + \text{mean}_{fake})}{2}$$

This dynamic threshold adjusts the decision boundary based on the dataset's inherent distribution, resulting in improved classification stability and reduced bias toward either class.

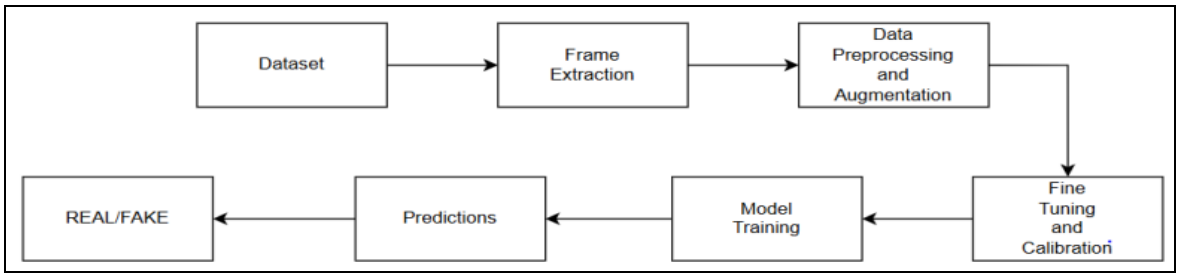

Fig 2 Threshold Calibration

➢ *Predictions*

For prediction, each test video undergoes frame extraction and pre-processing. The model generates individual frame predictions, and their mean value represents the confidence score of the video being real. A score higher than the calibrated threshold indicates a real video, while a lower score signifies a fake one. This aggregated prediction mechanism ensures robust video-level inference while minimizing frame-level noise or transient inconsistencies.

➢ *Implementation*

The proposed framework was implemented in Python using:

- TensorFlow / Keras for model design and training.
- OpenCV for video frame extraction and pre-processing.
- NumPy for numerical computation.
- Flask for web-based representation.

The pseudocode representation of the working pipeline is summarized below:

Start
Load model "deepfake_detector_final.keras"
Define extract_frames(video):
Open video using cv2
Sample 80 frames (resize 224x224)
Return frame list
Define predict_video(video_path):
frames = extract_frames(video_path)
preds = model.predict(frames)
mean_score = average(preds)
If mean_score > threshold → "REAL"
Else → "FAKE"
Return result
Run Flask app for real-time prediction
End

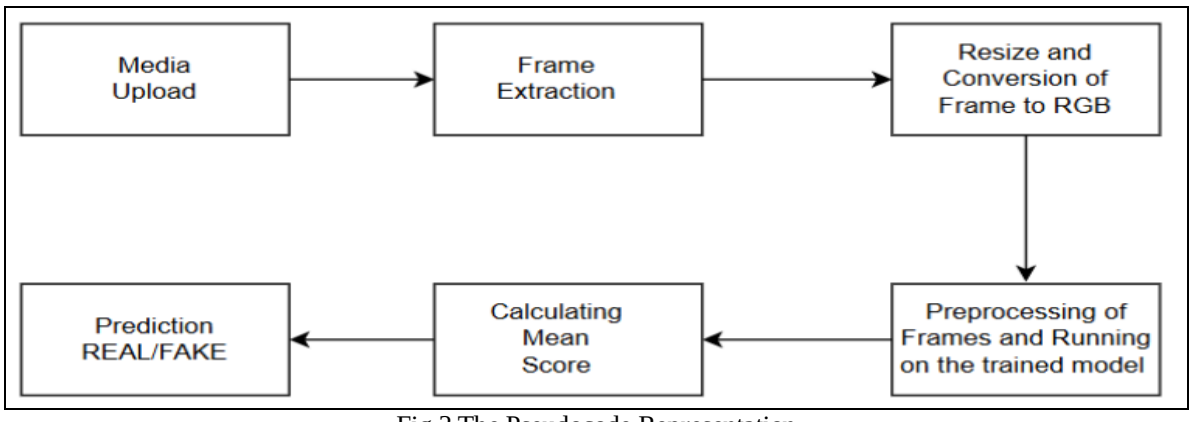

Fig 3 The Pseudocode Representation

The uploaded video is first processed for frame extraction, where key frames are captured from the media. Each frame is then resized and transformed into RGB format to ensure consistency across all images. The converted frames undergo pre-processing and are passed through the trained deepfake detection model. The model outputs prediction scores for each frame, which are then averaged to compute the mean score. Based on this mean score, the system predicts whether the uploaded video is REAL or FAKE.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2035

This framework enables real-time analysis of uploaded videos through a user-friendly web interface, producing instantaneous authenticity reports.

➢ *Summary*

The methodology integrates transfer learning, data augmentation, and temporal analysis to develop a reliable deepfake detection system. Through EfficientNetB0's efficient scaling and frame-level aggregation, the model effectively captures deep generative inconsistencies, providing robust detection performance on challenging datasets.

IV. RESULTS ANALYSIS

The proposed Deepfake detection approach was trained using the EfficientNetB0 model integrated with transfer learning to leverage its pre-trained ImageNet features. Throughout the training phase, the model exhibited a steady improvement in both training and validation accuracy, along with a noticeable decline in loss values. This consistent learning pattern indicates that the network successfully captured the essential discriminative features required to differentiate between authentic and manipulated frames. The implementation of data augmentation techniques reduced the risk of overfitting and enhanced the model's capability to generalize new samples.

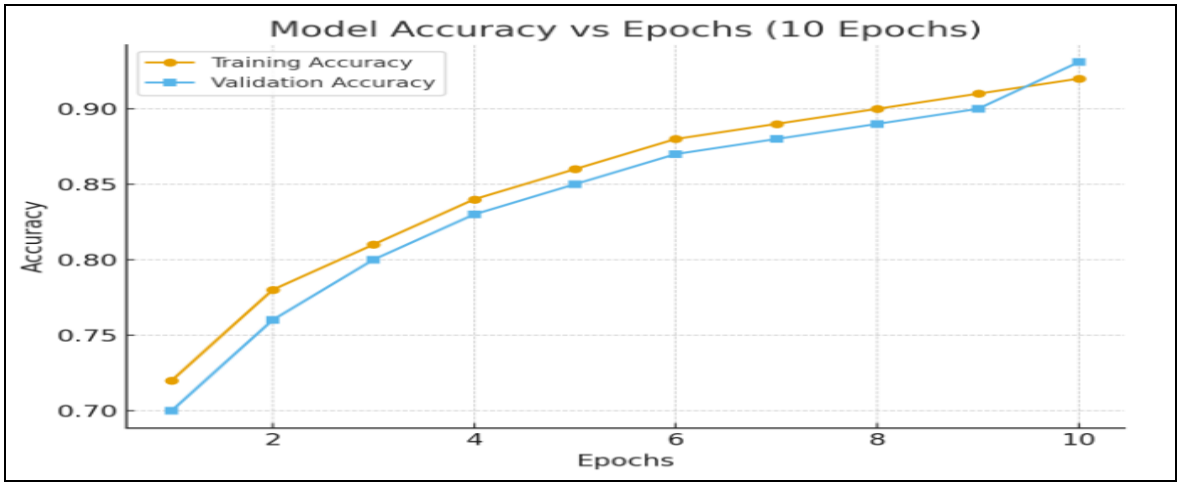

Fig 4 Accuracy vs Epochs

After completing the fine-tuning phase, the model reached a validation accuracy of approximately 93.1%, confirming its strong ability to distinguish real videos from Deepfake ones. The gradual convergence of accuracy over multiple epochs demonstrates that the model trains in a stable and efficient manner. The corresponding accuracy vs. epoch graph below highlights this trend, where both training and validation accuracy progress smoothly and align closely, suggesting well-balanced learning behaviour.

The evaluation results clearly show that the EfficientNetB0 architecture, when combined with frame-level analysis and threshold calibration, provides high detection reliability while maintaining computational efficiency. The model's robustness across different types of video manipulations validates its effectiveness in identifying Deepfakes with minimal false predictions. In summary, the results demonstrate that the proposed architecture can serve as a practical and scalable solution for automated Deepfake detection in real-world applications.

The Deepfake Detection Tool provides an intuitive and interactive web interface developed using Flask. The interface enables users to easily upload videos for analysis, process them through the trained EfficientNetB0 model, and visualize the results effectively.

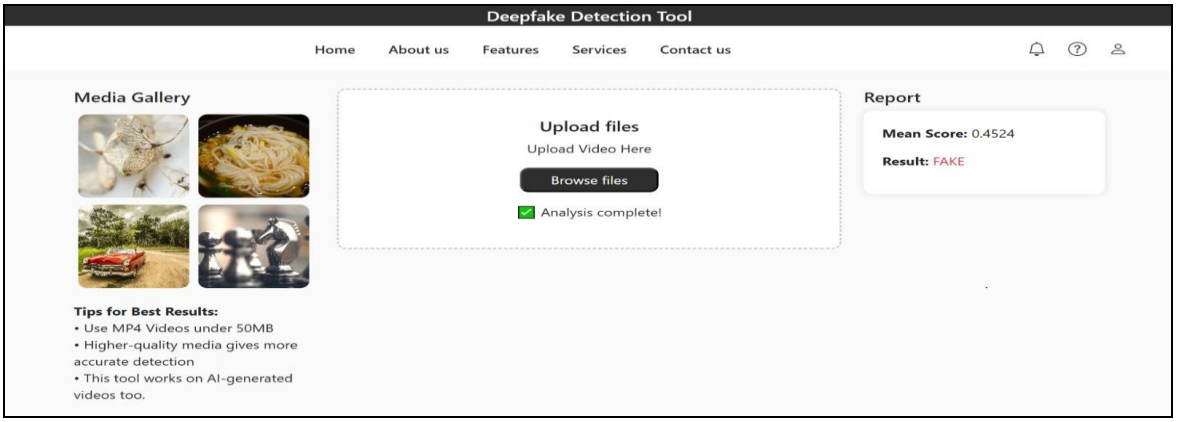

Fig 5 Output

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2035

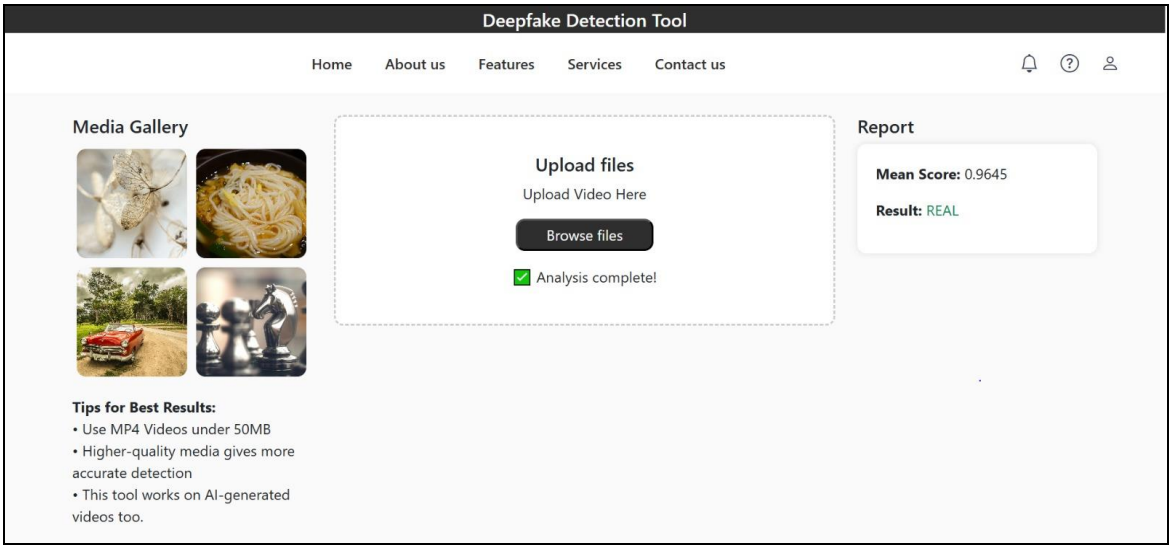

Fig 6 Output

As shown in the screenshots, the interface includes options for uploading media, displaying progress during analysis, and presenting the final prediction results. The output section displays the Mean Score (indicating prediction confidence) and the final classification result as either *Real* or *Fake*.

Overall, the system interface is designed for simplicity and efficiency, ensuring smooth interaction for users while delivering accurate detection outcomes with clear visual feedback.

V. CONCLUSION

The proposed Deepfake detection system demonstrates an effective approach for identifying manipulated videos using deep learning and transfer learning techniques. By leveraging the EfficientNetB0 architecture and fine-tuning it with frame-level analysis, the model achieves reliable differentiation between real and synthetic video content. The integration of data preprocessing, augmentation, and threshold calibration enhances model generalization and reduces false predictions.

This study highlights the potential of convolutional neural networks in safeguarding digital media authenticity. Through extensive training and validation on a balanced dataset, the system exhibits strong predictive accuracy and robustness against varying video qualities. Moreover, the threshold-based decision mechanism allows for interpretable results, making it suitable for real-world deployment.

In future work, the system can be extended to include multimodal indicators such as audio inconsistencies and facial motion irregularities for more comprehensive detection. Integration of real-time video analysis and edge-based deployment can further enhance its scalability and efficiency. Overall, the proposed framework contributes to the growing need for automated Deepfake detection systems capable of ensuring trust and transparency in digital communication.

REFERENCES

[1]. Karen Simonyan and Andrew Zisserman, “Very Deep Convolutional Networks for Large–Scale Image Recognition”, ICLR 2015, arXiv:1409.1556v6 [cs.CV], 10 Apr 2015.

[2]. Chesney, Robert and Citron, Danielle Keats, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (July 14, 2018). 107 California Law Review (2019, Forthcoming); U of Texas Law, Public Law Research Paper No. 692; U of Maryland Legal Studies Research Paper No. 2018-21.

[3]. Yuezun Li, Ming-Ching Chang and Siwei Lyu, In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking, arXiv:1806.02877v2 [cs.CV], 11 Jun 2018.

[4]. Yuezun Li and Siwei Lyu, “Exposing DeepFake Videos By Detecting Face Warping Artifacts”, arXiv:1811.00656v3 [cs.CV], 22 May 2019. https://doi.org/10.1016/S0969-4765(19)30137-7

[5]. Darius Afchar, Vincent Nozick, Junichi Yamagishi and Isao Echizen, “MesoNet: A Compact Facial Video Forgery Detection Network”, arXiv:1809.00888v1 [cs.CV], 4 Sep 2018. https://doi.org/10.1109/WIFS.2018.8630761

[6]. Xin Yang, Yuezun Li and Siwei Lyu, “Exposing Deep Fakes Using Inconsistent Head Poses”, ICASSP 2019 - 2019 IEEE ICASSP, 17 May 2019.

[7]. Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen, “Use of a Capsule Network to Detect Fake Images and Videos”, arXiv:1910.12467v2 [cs.CV], 29 Oct 2019.

[8]. Falko Matern, Christian Riess and Marc Stamminger, “Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations”, 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). https://doi.org/10.1109/WACVW.2019.00020

[9]. Jessica and Silbey Woodrow Hartzog, “The Upside of Deep Fakes”, Maryland Law Review, Volume 78, Issue 4, 2019.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2035

[10]. Schwartz, Oscar (12 November 2018). "You thought fake news was bad? Deep fakes are where the truth goes to die". The Guardian.
[11]. Sik-Ho Tsang, “Review: Inception-v3 — 1st Runner Up (Image Classification) in ILSVRC 2015”, https://medium.com/@sh.tsang/review-inception-v3-1st-runner-up-image-classification-in-ilsvrc-2015-17915421f77c
[12]. Pavel Korshunov, Sebastien Marcel, “DeepFakes: A New Threat to Face Recognition? Assessment and Detection”, citing arXiv:1812.08685 [cs.CV], 20 Dec 2018.
[13]. David Güera, Edward J. Delp, “Deepfake Video Detection Using Recurrent Neural Networks”, 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). https://doi.org/10.1109/AVSS.2018.8639163
[14]. Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, Cristian Canton Ferrer, “The Deepfake Detection Challenge (DFDC) Preview Dataset”, arXiv:1910.08854 [cs.CV], 19 Oct 2019.
[15]. Pavel Korshunov and Sebastien Marcel, “Vulnerability Assessment and Detection of Deepfake Videos”, IAPR International Conference, 2019.
[16]. Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Saeid Nahavandi, “Deep Learning for Deepfakes Creation and Detection”, arXiv:1909.11573 [cs.CV], 25 Sep 2019.
[17]. Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, Prem Natarajan, “Recurrent Convolutional Strategies for Face Manipulation Detection in Videos”, arXiv:1905.00582 [cs.CV], 2 May 2019.
[18]. Shuo Yuan, Xinguo Yu, Abdul Majid, “Robust Face Tracking Using Siamese-VGG with Pre-training and Fine-tuning”, 4th International Conference on Control and Robotics Engineering (ICCRE), 20–23 April 2019. https://doi.org/10.1109/ICCRE.2019.8724212
[19]. Francesco Marra, Diego Gragnaniello, Davide Cozzolino, Luisa Verdoliva, “Detection of GAN-Generated Fake Images over Social Networks”, IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 10–12 April 2018.
[20]. Shubhangi Tirpude, Naman Vidyabhanu, Hashir Sheikh, Shoeb Pathan, Zeeshan Ali Syed, Shivam Singh, “Abnormal X-Ray Detection System using Convolution Neural Network”, International Journal of Advanced Trends in Computer Science and Engineering, ISSN 2278-3091, Volume 9, No. 1, January–February 2020.