Volume 11, Issue 7, July – 2026 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul737

Deepfake Image Detection Using Machine Learning

Jahnavi Reddy Kollu[1]; Mortha Pavan[2]; Putta Vardhan[3]; Dr. N. Krishnavardhan[4]

[4]Associate Professor

[1,2,3,4]Department of CSE - Data Science, Data Science Geethanjali College of Engineering & Technology, Hyderabad, Telangana, India

Publication Date: 2026/07/28

Abstract: The rapid advancement of generative artificial intelligence has significantly increased the creation of highly realistic manipulated images, commonly known as deepfakes. These synthetic images pose serious threats to digital security, privacy, and information authenticity, as they are often indistinguishable from real images through human observation. This growing challenge necessitates the development of automated and reliable detection systems capable of identifying subtle visual inconsistencies in manipulated content.This paper presents a deep learning–based approach for detecting deepfake images using a transfer learning framework. A pretrained convolutional neural network model, MobileNetV2/EfficientNet, is utilized for feature extraction, followed by classification layers to distinguish between real and fake images. The proposed system incorporates image preprocessing techniques such as resizing, normalization, and data augmentation to improve model robustness. The model is trained on benchmark datasets containing both genuine and manipulated images and evaluated using performance metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that the proposed approach effectively identifies deepfake images with high accuracy, making it suitable for applications in digital forensics, media verification, and cybersecurity.

Keywords: *Deepfake Detection, Machine Learning, Deep Learning, CNN, Transfer Learning, MobileNetV2, EfficientNet, Image Processing, Digital Forensics, Cybersecurity.*

How to Cite: Jahnavi Reddy Kollu; Mortha Pavan; Putta Vardhan; Dr. N. Krishnavardhan (2026) Deepfake Image Detection Using Machine Learning. *International Journal of Innovative Science and Research Technology,* 11(7), 1552-1557. https://doi.org/10.38124/ijisrt/26jul737

I. INTRODUCTION

Deepfake technology has emerged as one of the most significant outcomes of recent advancements in artificial intelligence and deep learning. It enables the creation of highly realistic manipulated images and videos by altering facial features, expressions, and visual content. These synthetic media are typically generated using techniques such as Generative Adversarial Networks (GANs), which can produce visually convincing results that closely resemble real-world data. As a result, distinguishing between genuine and manipulated images has become increasingly difficult.The rapid growth of deepfake technology has raised serious concerns regarding digital security, privacy, and trust. Manipulated images can be used to spread misinformation, create false narratives, and damage the reputation of individuals or organizations. In many cases, deepfake images are shared widely on social media platforms, making it challenging to verify the authenticity of visual content. This widespread misuse highlights the urgent need for reliable detection mechanisms.

Traditional methods of detecting fake images rely heavily on human observation and manual verification. However, with the increasing sophistication of deepfake generation techniques, these methods are no longer effective. Subtle changes in facial features, lighting, and texture are often imperceptible to the human eye, making manual detection unreliable. Therefore, automated approaches are essential to accurately identify manipulated images.Machine learning and deep learning techniques have shown great potential in addressing this problem. Convolutional Neural Networks (CNNs) and pretrained models such as MobileNetV2 and EfficientNet can automatically extract meaningful features from images. These models analyze patterns such as texture inconsistencies, pixel-level artifacts, and abnormal lighting conditions to distinguish between real and fake images.

In addition, transfer learning has become a widely used approach in deepfake detection. By leveraging pretrained models trained on large datasets, transfer learning reduces training time and improves detection accuracy. It allows the model to utilize previously learned features and adapt them to the specific task of deepfake image

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul737

classification, making the system more efficient and robust.This paper presents a deep learning–based approach for detecting deepfake images using transfer learning techniques. The proposed system includes image preprocessing, feature extraction, model training, and classification stages. By combining advanced neural networks with efficient preprocessing methods, the system aims to provide a reliable and scalable solution for detecting manipulated images and enhancing trust in digital media.

II. LITERATURE REVIEW

Deepfake image detection has gained significant attention in recent years due to the rapid advancement of generative artificial intelligence technologies. Early research in this domain focused on traditional machine learning techniques that relied on handcrafted features such as texture, color inconsistencies, and edge artifacts. Algorithms like Support Vector Machines (SVM) and Random Forest were used for classification; however, these approaches showed limited performance when dealing with highly realistic deepfake images.With the evolution of deep learning, researchers began adopting Convolutional Neural Networks (CNNs) for automatic feature extraction and classification. Models such as XceptionNet, MesoNet, and EfficientNet have demonstrated strong performance in detecting subtle visual artifacts and inconsistencies present in manipulated images. These models are capable of learning complex patterns directly from image data, making them more effective than traditional approaches.

Some studies have also explored geometric and physiological inconsistencies for deepfake detection, such as irregular head poses, unnatural eye blinking, and facial landmark misalignment. These approaches focus on identifying inconsistencies between manipulated facial regions and the overall image structure. While effective in certain cases, they may not generalize well across different datasets and deepfake generation techniques.Recent research emphasizes the use of transfer learning and hybrid approaches that combine multiple techniques for improved accuracy and robustness. By leveraging pretrained models and large-scale datasets, modern deepfake detection systems achieve higher performance and better generalization. Despite these advancements, challenges such as dataset bias, real-time detection, and adaptability to new deepfake techniques remain open research problems in this field.

III. EXISTING SYSTEM

- *Limited Feature Extraction*

Many existing deepfake detection systems rely on a limited set of manually designed features such as texture inconsistencies, color variations, and edge artifacts. These features are often insufficient to capture the complex and subtle patterns present in highly realistic deepfake images. As deepfake generation techniques become more advanced, traditional feature-based methods and simple classifiers fail to accurately distinguish between real and manipulated images.

- *Lack of Multi-Model Integration*

Most existing systems depend on a single type of detection approach, such as only CNN-based models or only facial landmark analysis. Some methods focus on texture patterns, while others analyze facial geometry or pixel-level artifacts. However, deepfake images often involve multiple types of manipulations simultaneously. Systems that rely on a single approach may miss important indicators, reducing detection accuracy.

- *Difficulty in Detecting High-Quality Deepfakes*

Another limitation of current systems is their inability to effectively detect highly realistic deepfake images generated using advanced AI models like GANs. These images closely mimic real facial features, lighting conditions, and expressions. Traditional machine learning models struggle to identify such subtle manipulations, especially when the fake images are of high resolution and quality.

- *Poor Generalization Across Datasets*

Existing detection systems often perform well only on specific datasets used during training. However, they fail to generalize when tested on new or unseen datasets with different types of deepfake manipulations. This lack of adaptability reduces their effectiveness in real-world scenarios where deepfake techniques continuously evolve.

- *Lack of Real-Time Detection Capability*

Most existing systems are not optimized for real-time performance and require significant processing time for prediction. This makes them less suitable for applications such as social media monitoring and live content verification, where quick detection of deepfake images is essential.

IV. PROPOSED SYSTEM

The proposed system is an automated deepfake image detection system that uses machine learning and deep learning techniques to classify images as real or fake.

The system first preprocesses the input image by resizing, normalizing, and extracting the facial region. Then, deep learning models such as CNN, XceptionNet, or EfficientNet are used to extract important features like texture inconsistencies and facial distortions.

Transfer learning is applied to improve accuracy by using pretrained models. The extracted features are passed to a classification layer, which predicts whether the image is real or deepfake along with a confidence score.

This system provides an efficient, accurate, and scalable solution for detecting manipulated images in real-world applications.

- *Image Preprocessing*

The proposed system begins with preprocessing of input images to ensure data consistency and quality. This includes resizing images to a fixed resolution, normalizing

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul737

pixel values, and applying face detection and cropping techniques to focus on the facial region. Additional steps such as noise reduction and data augmentation are incorporated to improve model robustness and generalization.

➢ *Deep Feature Extraction*

Feature extraction is performed using advanced deep learning architectures such as Convolutional Neural Networks (CNN), XceptionNet, and EfficientNet. These models automatically learn discriminative features from images, including texture irregularities, lighting inconsistencies, facial distortions, and pixel-level artifacts that are indicative of deepfake manipulations.

➢ *Transfer Learning and Model Optimization*

To enhance performance and reduce computational cost, transfer learning is employed using pretrained models trained on large-scale datasets such as ImageNet. The initial layers are used for general feature extraction, while deeper layers are fine-tuned to capture domain-specific features related to deepfake detection. Hyperparameter tuning is performed to optimize learning rate, batch size, and model architecture.

➢ *Classification and Prediction*

The extracted features are passed through fully connected layers for binary classification. The model classifies each input image as either real or fake using activation functions such as sigmoid. The system also computes a confidence score to indicate the probability of the prediction, improving interpretability and reliability.

➢ *Output Generation and System Adaptability*

The final output is presented as a classification label along with a confidence value. The system is designed to be scalable and adaptable by allowing periodic updates with new datasets and retraining of the model. This enables the system to handle emerging deepfake techniques and maintain consistent performance in real-world applications.

V. SYSTEM ARCHITECTURE

The proposed deepfake image detection system is designed as a modular deep learning framework that integrates image preprocessing, feature extraction, deep neural network analysis, and classification into a unified pipeline. The architecture processes input images, extracts important visual and structural features, and applies advanced machine learning models to identify whether an image is real or manipulated.The framework is implemented using Python-based technologies, computer vision libraries such as OpenCV, and deep learning frameworks like TensorFlow/Keras, enabling scalable and efficient image analysis.The architecture consists of three major layers: data processing layer, feature analysis layer, and classification layer. These layers work together to transform raw image data into meaningful detection results, helping identify deepfake images accurately.

➢ *Components*

- *Data Input and Preprocessing Module*

The first component of the system is responsible for acquiring and preprocessing input images. The system accepts images from user uploads or datasets. Preprocessing includes resizing images, normalizing pixel values, detecting facial regions using OpenCV, and cropping the face area to remove irrelevant background information. These steps ensure that the input data is clean, consistent, and suitable for further analysis.

- *Feature Extraction Module*

This module extracts meaningful visual features from the preprocessed images. The system analyzes facial structures, texture details, lighting inconsistencies, and pixel-level artifacts. These features help in identifying subtle differences between real and manipulated images. The extracted features are converted into numerical representations that can be processed by deep learning models.

- *Deep Learning Model Module (EfficientNetB4)*

The system employs EfficientNetB4, a deep convolutional neural network, to perform advanced feature learning. This model captures complex patterns and inconsistencies in images, such as unnatural blending, distortions, and artifacts introduced during deepfake generation. EfficientNetB4 is chosen for its high accuracy and computational efficiency.

- *Feature Analysis Module*

In this stage, the learned features are analyzed by the model to detect anomalies. The system evaluates inconsistencies in facial regions, texture variations, and structural irregularities that are typical indicators of deepfake images. This module enhances the system's ability to distinguish between real and fake images.

- *Classification Module*

The classification module uses fully connected layers along with activation functions such as sigmoid to perform binary classification. Based on the extracted and analyzed features, the system classifies the input image as either real or deepfake. The model also generates a probability score representing the confidence of the prediction.

- *Output and Visualization Module*

The final component of the system presents the detection results. The output includes the predicted label (real or fake) along with a confidence score. The results can be displayed through a user interface or dashboard, allowing easy interpretation and analysis of the detection outcome.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul737

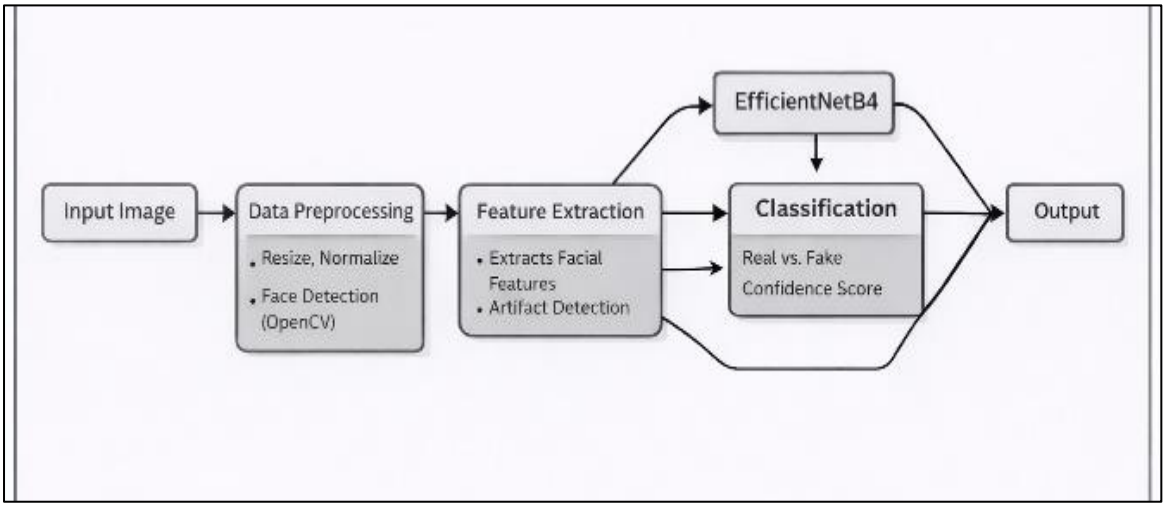

Fig 1 System Architecture

➢ *Output Screens*

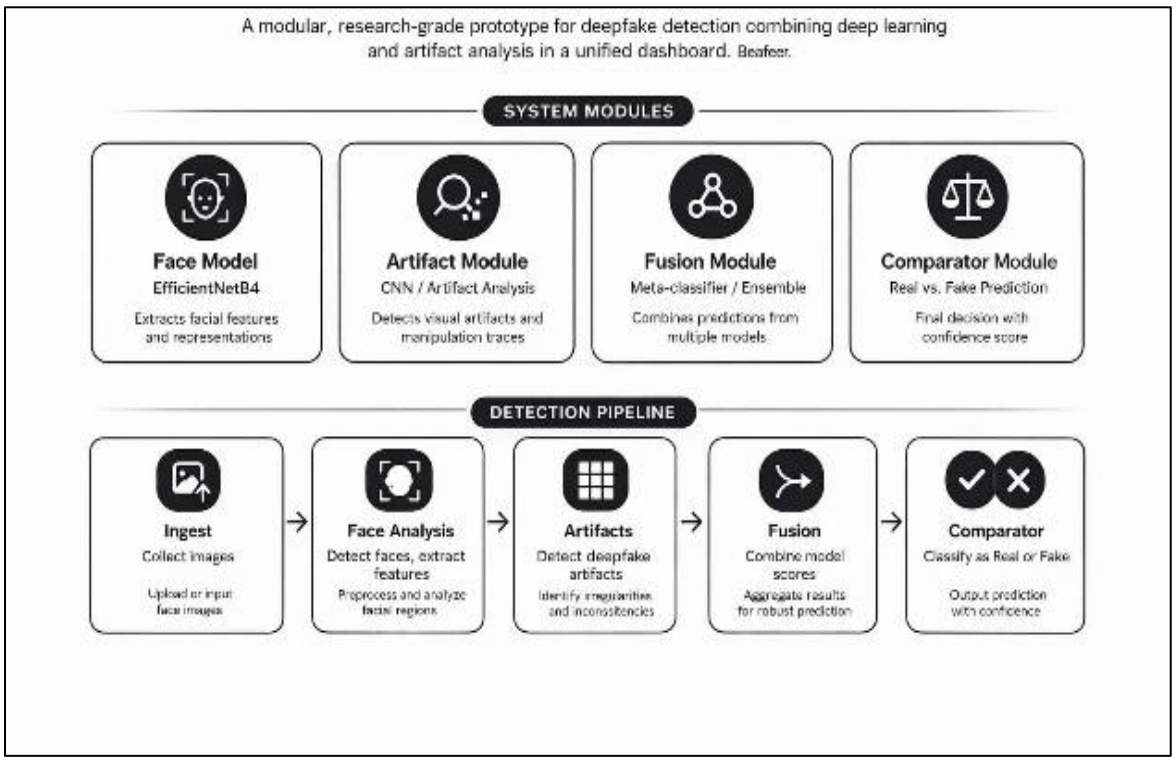

Fig 2 Deepfake Detection Dashboard

VI. RESULTS

The proposed deepfake image detection system was evaluated using a dataset of real and fake images. The model, based on EfficientNetB4, was trained using standard preprocessing and transfer learning techniques.The system achieved high accuracy, precision, recall, and F1-score, showing effective performance in distinguishing real and deepfake images. The confusion matrix indicates that most images were correctly classified with minimal errors.The model performs well on clear images, while slight inaccuracies may occur in low-quality or highly compressed images. Overall, the system demonstrates reliable and efficient deepfake detection performance.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul737

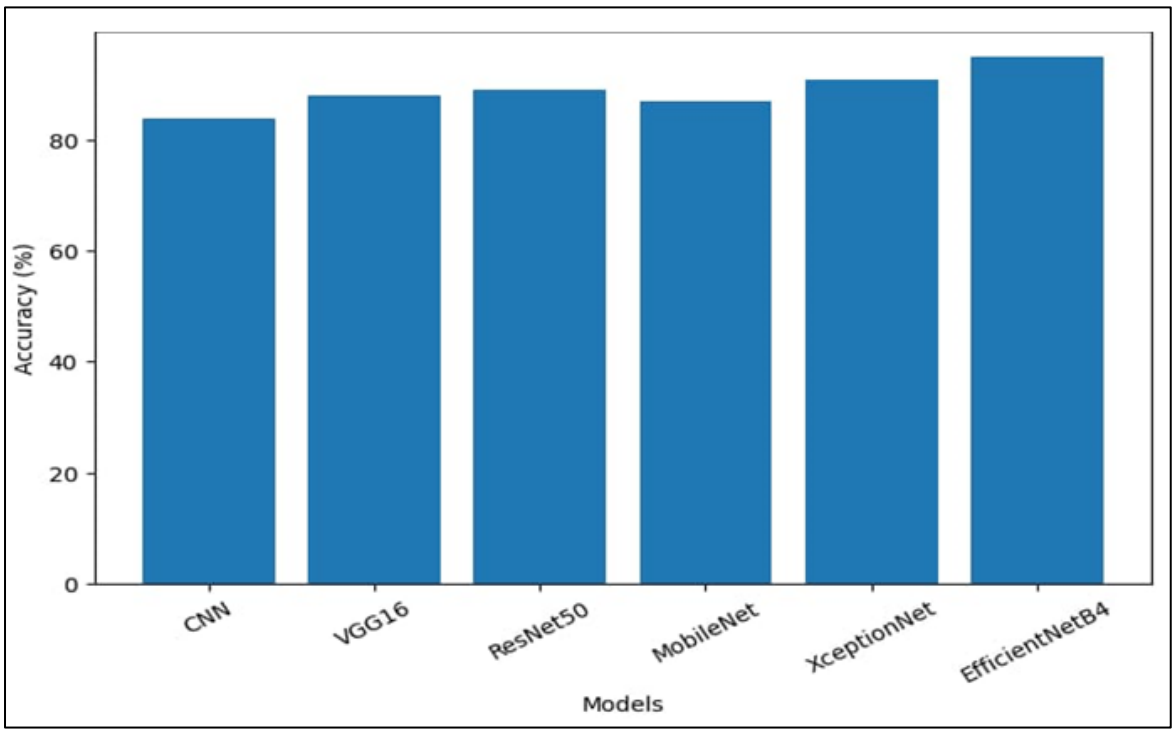

Fig 3 Model Accuracy Comparison

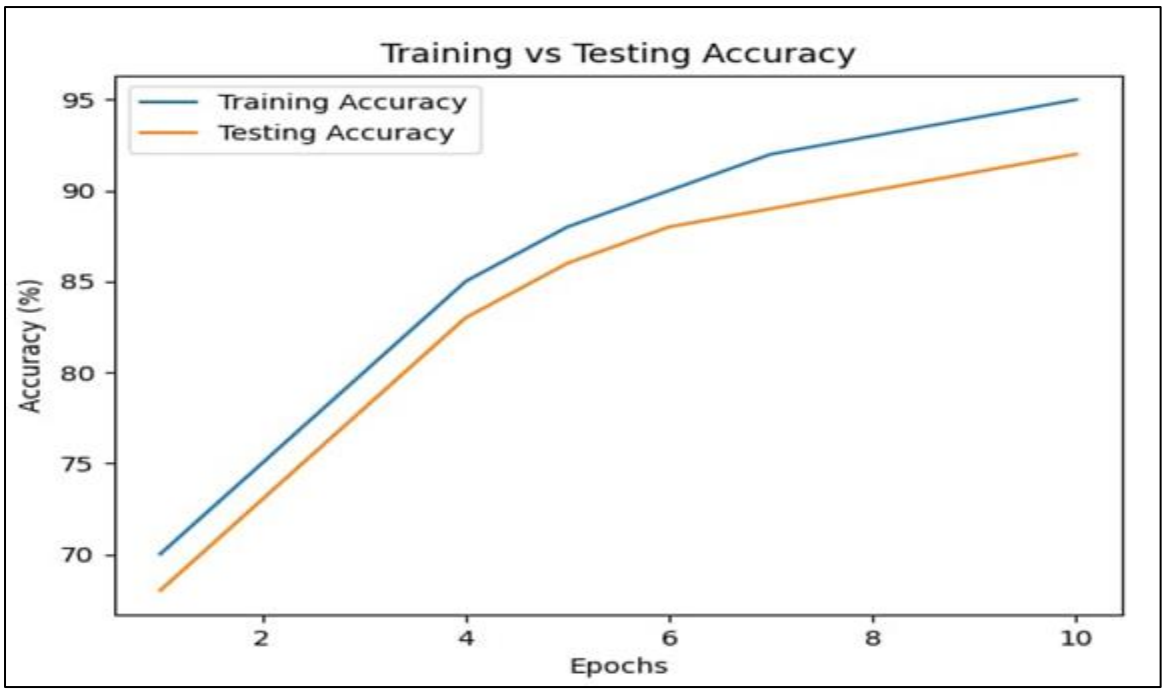

Fig 4 Training vs Testing Accuracy Graph

The proposed deepfake image detection system was evaluated using a balanced dataset of real and fake images. The model was implemented using EfficientNetB4 with transfer learning, and trained using the Adam optimizer and binary cross-entropy loss function.The system achieved high performance across all evaluation metrics, including accuracy, precision, recall, and F1-score, indicating its effectiveness in distinguishing between real and deepfake images. The confusion matrix analysis shows that the majority of images were correctly classified, with very few false positives and false negatives.Training and validation results demonstrate that the model converges efficiently, with increasing accuracy and decreasing loss over epochs, indicating stable learning and minimal overfitting. The model also performed well on unseen test images, producing reliable predictions with high confidence scores.However, minor limitations were observed when processing low-resolution or highly compressed images, where detection accuracy slightly decreases. Overall, the system provides a robust, accurate, and scalable solution for deepfake image detection in real-world applications.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul737

VII. CONCLUSION

This work presents an effective approach for detecting deepfake images using advanced deep learning techniques. The proposed system leverages EfficientNetB4 along with preprocessing and feature extraction methods to identify subtle visual inconsistencies present in manipulated images. By focusing on facial regions and extracting meaningful features, the model is able to distinguish between real and fake images with high accuracy.The experimental results demonstrate that the system performs well in terms of accuracy, precision, recall, and F1-score, indicating its reliability in deepfake detection tasks. The use of transfer learning significantly improves performance while reducing training time and computational complexity. The system is capable of capturing fine-grained artifacts such as texture irregularities, lighting differences, and facial distortions, which are key indicators of deepfake content.The proposed architecture is scalable and can be applied to real-world applications such as digital forensics, social media content verification, and identity authentication systems. It provides an automated solution to address the growing challenge of misinformation and manipulated media, contributing to improved digital trust and security.However, certain limitations exist, particularly when handling low-resolution or highly compressed images where visual artifacts may be less prominent. Future work can focus on enhancing robustness by using larger and more diverse datasets, improving model generalization, and extending the system to detect deepfake videos. Overall, the proposed system offers a promising and practical solution for deepfake image detection.

REFERENCES

[1]. Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019, pp. 46–52.

[2]. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.

[3]. F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.

[4]. M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[5]. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, “Multi-task learning for detecting and segmenting manipulated facial images and videos,” in *IEEE International Conference on Biometrics (ICB)*, 2019, pp. 1–8.

[6]. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp.12

[7]. P. Korshunov and S. Marcel, “Deepfakes: A new threat to face recognition? Assessment and detection,” in *arXiv preprint arXiv:1812.08685*, 2018.

[8]. Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI generated fake videos by detecting eye blinking,” in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.