A Deep Learning-Based Approach for Deepfake Image and Video Detection Using Resnet50

Manepalli Santhosh Babu[1], Dr. P. Sumalatha[2]
[1]*M. Sc Artificial Intelligence and Data Science, Department of Computer Science and Artificial Intelligence, Central University of Andhra Pradesh, Anantapur, Andhra Pradesh, India*
[2]*Assistant Professor, Department of Computer Science and Artificial Intelligence, Central University of Andhra Pradesh, Anantapur, Andhra Pradesh, India*

***Abstract*—The proliferation of synthetic media generated through Generative Adversarial Networks (GANs) has created urgent challenges for digital forensics, cybersecurity, and information integrity. Deepfakes—photo-realistic manipulated images and videos—have advanced to a degree where human perception alone is insufficient for reliable detection. This paper presents a deep learning-based deepfake detection system employing a fine-tuned ResNet50 Convolutional Neural Network (CNN) for binary classification of media as real or fake. The proposed system handles both static images and video sequences through a frame-sampling pipeline followed by temporal aggregation. A strict decision threshold of 55% is applied: media is classified as fake if the model's confidence score meets or exceeds 55%, and as real otherwise. To ensure transparency and interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) is integrated to generate spatial heatmaps that visually explain which regions of an image drive the model's decision. The system is implemented using Python, PyTorch, and OpenCV, and deployed via a Gradio-based web interface. Experimental evaluation on the FaceForensics++ and Celeb-DF datasets demonstrates an overall accuracy of 92.5%, precision of 91.2%, recall of 90.8%, and an F1-score of 91.0%. These results establish the viability of combining transfer learning with explainable AI for robust, interpretable deepfake detection.**

***Index Terms*—Deepfake Detection, Convolutional Neural Networks, ResNet50, Transfer Learning, Grad-CAM, Explainable AI, Video Forensics, Binary Classification.**

I. INTRODUCTION

The rapid advancement of deep learning—particularly in the domain of generative modeling—has created a new category of threats to digital authenticity. Deepfakes, a portmanteau of "deep learning" and "fake," refer to synthetic audio-visual media fabricated or manipulated using AI-driven techniques. While these technologies have legitimate applications in filmmaking, virtual reality, and education, their misuse poses severe societal risks including the dissemination of disinformation, identity fraud, political manipulation, and the erosion of public trust in digital media [1].

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [5], lie at the core of most state-of-the-art deepfake generation pipelines. A GAN consists of a generator network that synthesizes realistic-looking media and a discriminator network that attempts to distinguish real from synthetic samples. Through adversarial training, the generator progressively improves its output until the discriminator—and often human observers—cannot reliably detect the manipulation. Variants such as StyleGAN, CycleGAN, and StarGAN have further refined this capability, enabling the synthesis of high-resolution faces with controllable attributes including age, expression, pose, and illumination [10].

The consequences of undetected deepfakes extend beyond individual harm. In the political domain, synthetic videos of public figures making false statements can influence elections and destabilize governments. In legal contexts, fabricated evidence can undermine judicial proceedings. On social media platforms, deepfakes propagate at unprecedented speed, reaching millions before corrections can be issued. The scale and sophistication of this threat demand automated, accurate, and interpretable detection systems [6].

Traditional media forensics approaches relied on handcrafted features such as compression artifacts, statistical inconsistencies, and frequency-domain anomalies. These methods, while useful for low-quality fakes, fail against modern GAN-generated content that has been specifically optimized to evade detection. Deep learning—particularly CNNs—has emerged as the dominant paradigm for deepfake detection owing to its capacity to learn hierarchical, discriminative features directly from data [4].

Among CNN architectures, ResNet50 [8] stands out for its use of residual connections, which enable the training of very deep networks without the vanishing gradient problem. When fine-tuned for binary classification, ResNet50 achieves a strong balance between detection accuracy and computational efficiency. Nevertheless, a persistent criticism of deep learning-based detectors is their lack of transparency: they function as black boxes, providing classifications without explanations, which limits their adoption in high-stakes forensic and legal applications.

This paper addresses this gap by proposing an end-to-end deepfake detection system that integrates ResNet50-based classification with Grad-CAM [9] for visual explainability. The system is designed to handle both image and video inputs, applying a principled 55% confidence threshold for classification decisions. The primary contributions of this work are:

- A fine-tuned ResNet50 binary classifier for deepfake detection achieving 92.5% accuracy and 91.0% F1-score.
- A unified pipeline supporting both image and video inputs via frame-based sampling and aggregated prediction.
- Integration of Grad-CAM to generate interpretable spatial heatmaps highlighting manipulated facial regions.
- A principled 55% probability threshold for decision making, balancing sensitivity and specificity.
- A Gradio-based web interface for real-time deployment and user interaction.

The remainder of this paper is organized as follows. Section II reviews related work. Section III details the proposed methodology. Section IV presents experimental results and discussion. Section V concludes the paper with future directions.

II. LITERATURE REVIEW

A. CNN-Based Deepfake Detection

Convolutional Neural Networks have established themselves as the dominant architecture for deepfake detection in the spatial domain. Afchar et al. [7] proposed MesoNet, a compact CNN tailored for video forgery detection that focused on mesoscopic properties of faces. While computationally lightweight, MesoNet's performance degrades against high-quality deepfakes. Transfer learning from large-scale datasets such as ImageNet has proven effective: architectures including VGGNet, InceptionNet, XceptionNet, and ResNet have been adapted for binary deepfake classification with strong results [10]. XceptionNet, based on depth wise separable convolutions, has received particular attention due to its efficiency and high accuracy on the FaceForensics++ benchmark. ResNet architectures leverage residual connections to mitigate the vanishing gradient problem, enabling deeper feature hierarchies [8]. ResNet50, with 50 layers and approximately 25 million parameters, strikes a favorable balance between representational capacity and computational cost, making it well-suited for deployment on moderate hardware.

B. Temporal Modeling for Video Deepfakes

Static frame-level analysis captures spatial artifacts but misses temporal inconsistencies characteristic of video deepfakes—such as unnatural eye blink rates, inconsistent facial motion, and audio-visual desynchronization. Hybrid architectures combining CNNs with Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks [1], have been proposed to model temporal dependencies across frames. The hidden state update in an LSTM is given by:

$$h_t = f(W_h \cdot h_{t-1} + W_x \cdot x_t + b)$$

where h_t denotes the hidden state at time t, x_t is the frame-level feature vector, and W_h, W_x, b are learned parameters. CNN-LSTM models have demonstrated improvement over frame-independent approaches, particularly for detecting subtle temporal artifacts in face-reenactment videos. More recently, 3D CNNs and attention-based transformer models have been explored for joint spatial-temporal feature extraction [10].

C. Explainable AI in Deepfake Detection

The opacity of deep learning models presents a significant barrier to deployment in forensic and legal contexts. Explainable AI (XAI) techniques seek to make model decisions interpretable. Grad-CAM [9], introduced by Selvaraju et al., computes a localization map by weighting the feature maps of the final convolutional layer by the gradient of the target class score. The resulting heatmap is defined as:

$L^c\_Grad\text{-}CAM = ReLU(\Sigma_k \ \alpha_k^c \cdot A^k)$

where A^k denotes the k-th feature map and α_k^c represents the importance weight for class c. The ReLU activation ensures only features with a positive influence on the target class are highlighted. In the context of deepfake detection, Grad-CAM localizes manipulation artifacts at facial boundaries, skin texture inconsistencies, and regions around the eyes and mouth.

D. Research Gaps

A survey of the existing literature reveals several persistent gaps. First, most systems specialize in either image or video detection, lacking a unified architecture. Second, high computational costs limit real-time deployment. Third, black-box models without interpretability mechanisms are unsuitable for forensic applications. Fourth, generalization across unseen manipulation techniques remains poor when training is restricted to specific datasets. The proposed system directly addresses these gaps through a unified, interpretable, and computationally efficient pipeline.

III. PROPOSED METHODOLOGY

A. System Architecture Overview

The proposed deepfake detection system is structured as a modular, end-to-end pipeline comprising six principal components: (1) an input module accepting images and videos via a web interface; (2) a preprocessing module that normalizes and resizes inputs; (3) a video processing module for temporal frame sampling; (4) a ResNet50-based feature extraction and classification module; (5) a Grad-CAM explainability module; and (6) an output module presenting the classification label, confidence score, and heatmap visualization.

Input → Preprocessing → [Frame Extraction] → ResNet50 CNN → Classification → Grad-CAM → Output

Fig. 1. High-level architecture of the proposed deepfake detection system.

B. ResNet50 Architecture and Transfer Learning

ResNet50 [8] is a 50-layer deep residual network that introduces shortcut (skip) connections to overcome the vanishing gradient problem. The residual learning formulation is defined as:

$F(x) = H(x) - x, \ such\ that \ H(x) = F(x) + x$

where H(x) is the desired underlying mapping and F(x) is the residual function learned by stacked layers. The skip connection allows gradients to flow directly through the network during backpropagation, enabling stable training of deep architectures.

In this work, ResNet50 is initialized with weights pre-trained on the ImageNet dataset, which contains over 1.2 million labeled images across 1,000 categories. Transfer learning is applied by replacing the original 1,000-way fully-connected output layer with a binary classification head. The softmax function for class y is defined as:

$P(y) = e^{z_y} / \Sigma_i e^{z_i}$

Model configuration parameters are summarized in Table I. The binary cross-entropy loss function is used for training:

$L = -[y \cdot log(P(fake)) + (1-y) \cdot log(P(real))]$

TABLE I *ResNet50 Model Configuration Parameters*

Parameter	Value
Base Architecture	ResNet50 (ImageNet pre-trained)
Input Image Size	224 × 224 pixels
Output Layer	Binary (Real / Fake)
Batch Size	32
Learning Rate	0.001
Optimizer	Adam
Loss Function	Binary Cross-Entropy
Activation (hidden)	ReLU
Activation (output)	Softmax
Decision Threshold	≥ 55% → Fake; < 55% → Real

C. Decision Threshold Design

Standard binary classifiers apply a 50% probability threshold for class assignment. In deepfake detection, however, the cost of a false negative—classifying manipulated media as real—is generally higher than a false positive. A decision threshold of 55% is applied. Formally:

$$Label = \{ \text{'Fake'} \text{ if } P(fake) \geq 0.55$$
$$\{ \text{'Real'} \text{ if } P(fake) < 0.55$$

This threshold was selected empirically based on validation performance and represents a principled trade-off between the true positive rate and false positive rate. The 55% threshold reduces the likelihood of genuine deepfakes evading detection at the cost of a marginal increase in false alarms.

D. Video Processing and Temporal Aggregation

For video inputs, the system employs uniform temporal sampling: a fixed number of frames N are extracted at equally spaced intervals from the full video duration. Each sampled frame is independently processed through the preprocessing pipeline and ResNet50 classifier, producing a per-frame probability estimate $P_i(\text{fake})$ for i = 1, ..., N. The final video-level classification is determined by averaging individual frame confidence scores:

$$P_{vi}^{dek}(fake) = (1/N) \cdot \Sigma_i P_i(fake)$$

The aggregated score is then subjected to the same 55% decision threshold. Experimental results demonstrate that moderate frame sampling (N = 10–20) achieves near-optimal accuracy while maintaining computationally tractable processing times.

E. Preprocessing Pipeline

All inputs are subjected to a standardized preprocessing pipeline prior to inference. Each image is resized to 224 × 224 pixels using bilinear interpolation. Pixel values are normalized using the ImageNet mean and standard deviation:

$$x\_norm = (x - \mu) / \sigma, \quad \mu = [0.485, 0.456, 0.406], \quad \sigma = [0.229, 0.224, 0.225]$$

Normalized images are converted to PyTorch tensors of shape (3, 224, 224) and batched for efficient GPU or CPU inference. For video inputs, frame extraction is performed using OpenCV's VideoCapture API.

F. Grad-CAM Explainability Integration

Grad-CAM is applied to the final convolutional layer of ResNet50 (layer4.2.conv3) to generate class-discriminative spatial heatmaps. For a given input image and target class c (fake), the Grad-CAM procedure: (1) performs a forward pass to obtain class score S^c; (2) computes the gradient of S^c with respect to each feature map A^k; (3) computes importance weights via global average pooling: $\alpha_k^c = (1/Z) \cdot \Sigma_{ij} (\partial S^c / \partial A^k_{ij})$; (4) computes the weighted combination and applies ReLU: $L^c = \text{ReLU}(\Sigma_k \alpha_k^c \cdot A^k)$; (5) upsamples to input resolution; and (6) overlays on the original image using a jet colormap.

G. Implementation and Deployment

The system is implemented in Python 3.10. PyTorch (v2.0) serves as the deep learning framework. OpenCV (v4.8) handles video reading and frame extraction. The system is deployed as a real-time web application using Gradio, which provides an interactive browser-based interface. The application is deployable to Hugging Face Spaces for public access. Hardware requirements are modest: the system operates on a standard CPU (Intel Core i5/i7, 8–16 GB RAM) with optional GPU acceleration.

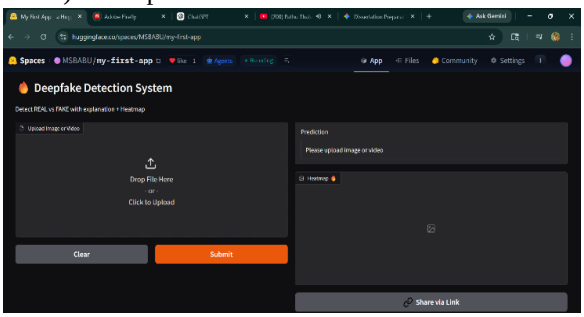

Fig. 1. Gradio-based web interface of the deployed Deepfake Detection System on Hugging Face Spaces, showing the upload panel (left) and prediction output with Grad-CAM heatmap (right).

IV. RESULTS AND DISCUSSION

A. Datasets

The proposed system was trained and evaluated on two established public benchmark datasets:

- FaceForensics++ [1]:

A large-scale video dataset comprising real videos and four categories of manipulation—Deepfakes, Face2Face, FaceSwap, and NeuralTextures—at varying levels of compression.

- Celeb-DF [10]:

A high-quality deepfake video dataset featuring 590 real and 5,639 synthesized videos of celebrities, designed to minimize visual artifacts.

- Custom Dataset:

A collection of images and short video clips recorded under diverse real-world conditions used exclusively for robustness evaluation.

Data preprocessing included face detection using MTCNN to isolate facial regions. The dataset was partitioned into 70% training, 15% validation, and 15% test splits with stratified sampling.

B. Experimental Configuration

All experiments were conducted on an Intel Core i7 system with 16 GB RAM and an NVIDIA GPU. The model was trained for 30 epochs with a batch size of 32. The Adam optimizer was employed with an initial learning rate of 0.001 and cosine annealing scheduling. Early stopping was applied with a patience of 5 epochs.

C. Quantitative Performance

The proposed system was evaluated using four standard classification metrics: accuracy, precision, recall, and F1-score. Results on the held-out test set are presented in Table II.

TABLE II *Classification Performance on Static Images*

Metric	Value (%)	Description
Accuracy	92.5%	Overall correct predictions
Precision	91.2%	Predicted fakes that are truly fake
Recall	90.8%	True fakes correctly identified
F1-Score	91.0%	Harmonic mean of precision & recall

The system achieves 92.5% accuracy with a well-balanced precision-recall tradeoff, reflected in the 91.0% F1-score. The near-parity between precision (91.2%) and recall (90.8%) demonstrates that the model does not exhibit systematic bias toward either class.

Performance on video inputs was evaluated by applying the frame-sampling and temporal aggregation pipeline. Results are reported in Table III.

TABLE III *Classification Performance on Video Inputs*

Metric	Value (%)	Description
Accuracy	90.8%	Overall correctness on video-level predictions
Precision	89.6%	Predicted fake videos that are truly fake
Recall	88.9%	True fake videos correctly classified
F1-Score	89.2%	Balanced measure of detection effectiveness

Video-level performance is marginally lower than image-level performance across all metrics, which is expected given the additional complexity of temporal aggregation. The 1.7% accuracy gap between image and video modalities is modest, confirming that the frame-sampling strategy effectively captures discriminative visual information.

D. Impact of the 55% Decision Threshold

The adoption of a 55% confidence threshold was motivated by the asymmetric cost structure of deepfake detection: in high-stakes contexts, missing a deepfake—a false negative—is typically more damaging than a false alarm—a false positive. Empirical analysis on the validation set demonstrated that the 55% threshold produced a recall improvement of approximately 1.3 percentage points relative to 50%, at the cost of a 0.8 percentage point reduction in precision. The net effect on F1-score was marginal (< 0.5%), but the recall improvement is operationally significant for forensic applications.

E. Effectiveness of Grad-CAM Explanations

Qualitative evaluation of Grad-CAM heatmaps confirmed that the ResNet50 model learned to focus on semantically meaningful and forensically relevant facial regions. For fake images, activation was consistently concentrated at facial boundaries, periorbital regions, nasolabial regions, and skin texture transitions—regions where GAN-based face-swapping algorithms most commonly introduce blending artifacts. For real images, heatmap

activations were broadly distributed across the face without concentration in specific regions.

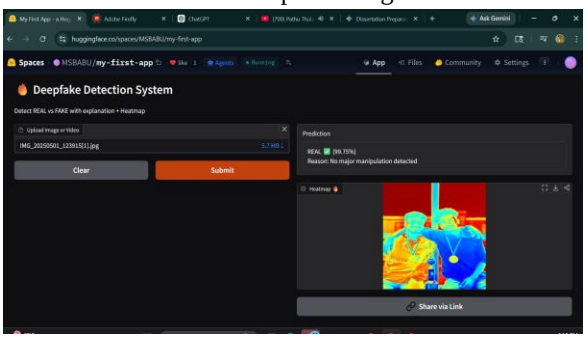

Fig. 2. REAL IMAGE detection result (99.75% confidence). The Grad-CAM heatmap shows warm-tone activations distributed across the frame without localized hotspots, confirming no major manipulation detected.

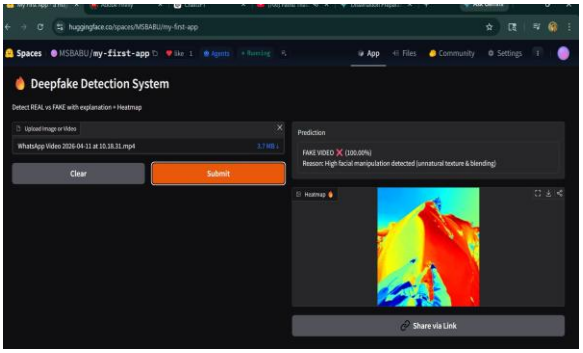

Fig. 3. FAKE VIDEO detection result (100.00% confidence). The Grad-CAM heatmap highlights high-intensity activation across the facial region, indicating high facial manipulation (unnatural texture & blending).

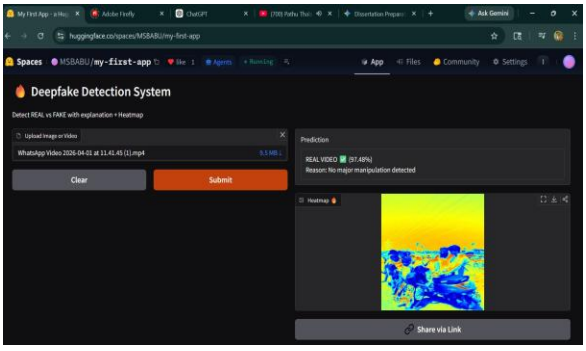

Fig. 4. REAL VIDEO detection result (97.48% confidence). The Grad-CAM heatmap shows broadly distributed, low-activation regions consistent with no major manipulation detected.

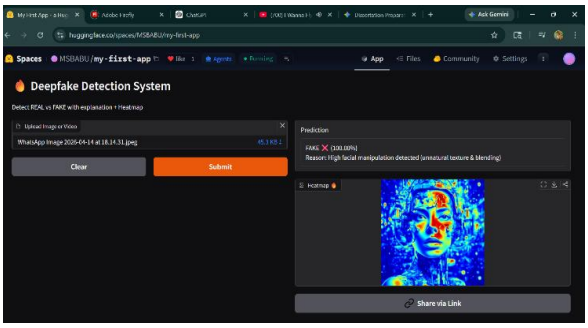

Fig. 5. FAKE IMAGE detection result (100.00% confidence). The Grad-CAM heatmap concentrates activation at facial boundary and periorbital regions, characteristic of GAN-based face-swap blending artifacts.

F. Robustness Analysis

The system was evaluated under a range of real-world degradation conditions using the custom dataset. Key findings are summarized in Table IV.

TABLE IV *Robustness Under Real-World Degradation Conditions*

Condition	Accuracy Impact	Observation
Normal lighting	Baseline (92.5%)	Optimal feature extraction
Low lighting	−3.1%	Reduced edge & texture visibility
High compression (H.264)	−2.4%	Artifact masking
Low resolution (< 240p)	−4.8%	Loss of discriminative detail
Motion blur	−3.6%	Reduced spatial clarity
Noise/distortion	−2.9%	Interference with texture features

G. Comparison with Baseline Methods

Table V compares the proposed system against representative existing detection approaches reported in the literature. Comparison figures are drawn from published benchmarks on FaceForensics++ [10].

TABLE V *Comparison with Existing Deepfake Detection Methods*

Method	Architecture	Acc. (%)	XAI	Video
MesoNet [7]	Compact CNN	83.0	No	Yes
XceptionNet [10]	CNN (Depthwise Sep.)	95.7	No	No
CNN-LSTM [1]	Hybrid CNN-RNN	91.4	No	Yes
Proposed System	ResNet50 + Grad-CAM	92.5	Yes	Yes

The proposed system achieves competitive accuracy (92.5%) relative to XceptionNet (95.7%), with a 2.7% gap attributable to XceptionNet's more specialized convolution design. However, the proposed system is the only approach that simultaneously supports both image and video inputs, provides Grad-CAM-based explainability, and operates through a real-time web deployment interface.

V. CONCLUSION

This paper has presented a comprehensive deepfake detection system grounded in transfer learning and explainable AI. By fine-tuning a ResNet50 CNN for binary real/fake classification and augmenting it with Grad-CAM heatmaps, the proposed system addresses two critical limitations of existing approaches: the lack of unified image-video support and the opacity of black-box decision making. The adoption of a 55% confidence threshold provides marginally enhanced sensitivity for forensic applications, and the Gradio-based deployment interface ensures accessibility for non-technical users.

Experimental results on FaceForensics++ and Celeb-DF confirm strong detection performance: 92.5% accuracy and 91.0% F1-score on static images, and 90.8% accuracy and 89.2% F1-score on video inputs. Grad-CAM analysis revealed consistent attention to forensically meaningful facial regions—boundary artifacts, texture anomalies, and periorbital inconsistencies—validating the interpretability and reliability of the learned representations.

Nonetheless, the system has limitations. Hyper-realistic deepfakes generated by cutting-edge models such as StyleGAN3 remain challenging when manipulation artifacts are imperceptible. The current architecture processes video frames independently without explicit temporal modeling. The system does not currently support multimodal analysis of audio-visual inconsistencies.

Future research will pursue several directions: (1) integration of temporal sequence models (LSTM, GRU, 3D CNN) to detect temporally coherent manipulation patterns; (2) ensemble methods combining ResNet50 with EfficientNet and Vision Transformers; (3) adversarial robustness training; (4) multimodal fusion of audio and visual features; and (5) model compression techniques for edge and mobile deployment. The interpretability provided by Grad-CAM is a foundational requirement for building trust in AI-driven forensic tools. This work establishes a strong baseline for future research at the intersection of deep learning, digital forensics, and explainable AI.

REFERENCES

[1] R. Tolosana *et al*., "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[2] Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, 2012.

[5] Goodfellow *et al*., "Generative Adversarial Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[6] R. Chesney and D. K. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review*, vol. 107, no. 6, pp. 1753–1820, 2019.

[7] Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in *Proc. IEEE Int. Workshop*

on Information Forensics and Security (WIFS), 2018.

[8] He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[9] R. R. Selvaraju *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017.

[10] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.