International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue5
Sep-Oct- 2026, PP 01-06

# Deepfake Detection Technology: A Review Of Lightweight Transfer-Learning Detectors, Artefact Cues And Generalization To Unseen Manipulations

**Thorat Sanika Santosh[1] Prof. M. C. Kolse[2] Prof. S. S. Bhor[3] Prof. S. K. Said[4]**
Department of Artificial Intelligence and Data Science
Jaihind College of Engineering
Pune, Maharashtra, India

**Abstract—Deepfake technology uses deep learning to create or manipulate images and videos so that a person appears to say or do something that never happened. Although it has legitimate uses, it can spread misinformation, enable fraud and damage reputations. This paper reviews deepfake detection technology with a focus on lightweight, deployable detectors and the generalization problem. It examines how manipulated faces are produced, the benchmark datasets used for training and testing, and detectors based on transfer learning, in which convolutional networks such as MobileNetV2 and EfficientNet pretrained on ImageNet are fine-tuned to classify face images or video frames as real or fake. It reviews artefact cues that detectors exploit, including blending boundaries, frequency-domain traces of up-sampling and fine-grained texture differences captured by attention, and the finding that detectors trained on one generator can transfer to others when trained with suitable augmentation. Detection of images from diffusion models is also discussed. A five-stage architecture comprising media input, face extraction and preprocessing, transfer-learning feature extraction, classification and a user interface built with a lightweight web framework is proposed, together with research gaps, objectives, a comparison of detector backbones and challenges such as compression, dataset bias and adversarial evasion. The review concludes that predictions should be treated as indications rather than proof and that robust generalization remains the main open problem.**

***Index Terms—Deepfake detection, transfer learning, MobileNetV2, EfficientNet, face forgery, frequency artefacts, generalization, digital forensics***

## I. Introduction

A short video of a public figure making a statement they never made can circulate widely before anyone checks it. Face-swapping and reenactment tools have made such manipulations easy, and the results are often convincing to ordinary viewers. Tolosana *et al.* [1] surveyed face manipulation methods, including identity swap, expression swap and entire face synthesis, and the corresponding detection approaches.

Many detection models achieve high accuracy on the dataset they are trained on but fail on new manipulation methods or compressed social-media videos. At the same time, practical tools for journalists, students and the public need to run on ordinary hardware. This paper reviews lightweight, transfer-learning-based detectors and the generalization challenge.

## II. Background and Fundamental Concepts

### 2.1 Types of Facial Manipulation and Detection Cues

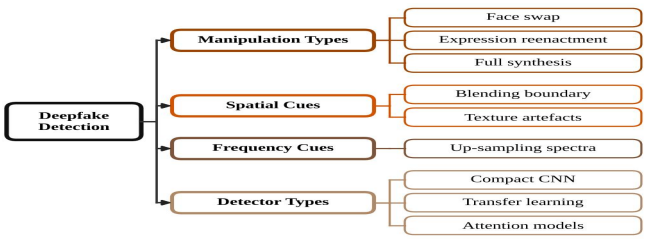



**Fig. 1 organizes deepfake types and the cues that detectors exploit.**

**Fig. 1. Deepfake types and detection cues.**

### 2.2 Transfer Learning

Transfer learning reuses knowledge from a source task to improve learning on a target task [2]. Networks pretrained on ImageNet [3] already capture edges, textures and shapes; fine-tuning their upper layers on real and fake faces requires far less data and computation than training from scratch.

## III. Literature Survey

**Datasets.** FaceForensics++ [4] provides videos manipulated with several methods at different compression levels. Celeb-DF [5] contains higher-quality deepfakes with fewer visible artefacts, and the DFDC dataset [6] offers large-scale, diverse data.

**Compact and transfer-learning detectors.** Afchar *et al.* [7] proposed MesoNet, a compact network focusing on mesoscopic image properties. MobileNetV2 [8] uses inverted residual blocks for efficient mobile inference, and EfficientNet [9] scales networks for better accuracy per parameter; both are widely used as pretrained backbones for detection.

**Artefact cues.** Li *et al.* [10] proposed Face X-ray, which detects the blending boundary created when a manipulated face is inserted, and generalizes to unseen methods. Durall *et al.* [11] showed that up-convolution in generative networks leaves spectral distortions detectable in the frequency domain. Zhao *et al.* [12] used multiple attention maps to capture subtle local artefacts.

**Generalization.** Wang *et al.* [13] found that a classifier trained on images from one GAN with careful augmentation generalizes surprisingly well to other CNN generators. Corvi *et al.* [14] studied the detection of images generated by diffusion models and found that detectors trained on GAN images do not always transfer. Table I summarizes these works.

| Ref. | Authors | Year | Technique Used | Merits | Demerits |
|---|---|---|---|---|---|
| [7] | Afchar et al. | 2018 | MesoNet compact CNN | Fast detection | Weak on high quality |
| [8] | Sandler et al. | 2018 | MobileNetV2 | Efficient backbone | General vision |
| [4] | Rössler et al. | 2019 | FaceForensics++ | Multi-method benchmark | Limited diversity |
| [10] | Li et al. | 2020 | Face X-ray | Generalizes via blending cue | Needs blending step |
| [11] | Durall et al. | 2020 | Spectral analysis | Frequency artefacts | Removable by post-processing |
| [13] | Wang et al. | 2020 | Cross-generator detection | Augmentation aids transfer | CNN generators |
| [12] | Zhao et al. | 2021 | Multi-attention detection | Captures local artefacts | Heavier model |
| [14] | Corvi et al. | 2023 | Diffusion image detection | Transfer gaps revealed | Images only |

**TABLE I Summary of Literature Survey**

## IV. Research Gap

- Detectors often fail on manipulation methods not seen in training.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue5
Sep-Oct- 2026, PP 01-06

- Social-media compression removes fine artefacts.
- Lightweight models trade accuracy for speed.
- Diffusion-generated content requires new detection cues.
- Users may misinterpret probabilistic outputs as definitive proof.

### 4.1 Objectives

- **O1:** To review deepfake types and benchmark datasets.
- **O2:** To examine lightweight transfer-learning detectors.
- **O3:** To study artefact cues and generalization techniques.
- **O4:** To propose a deployable deepfake detection pipeline.
- **O5:** To discuss limitations and responsible interpretation of results.

## V. Proposed System Architecture

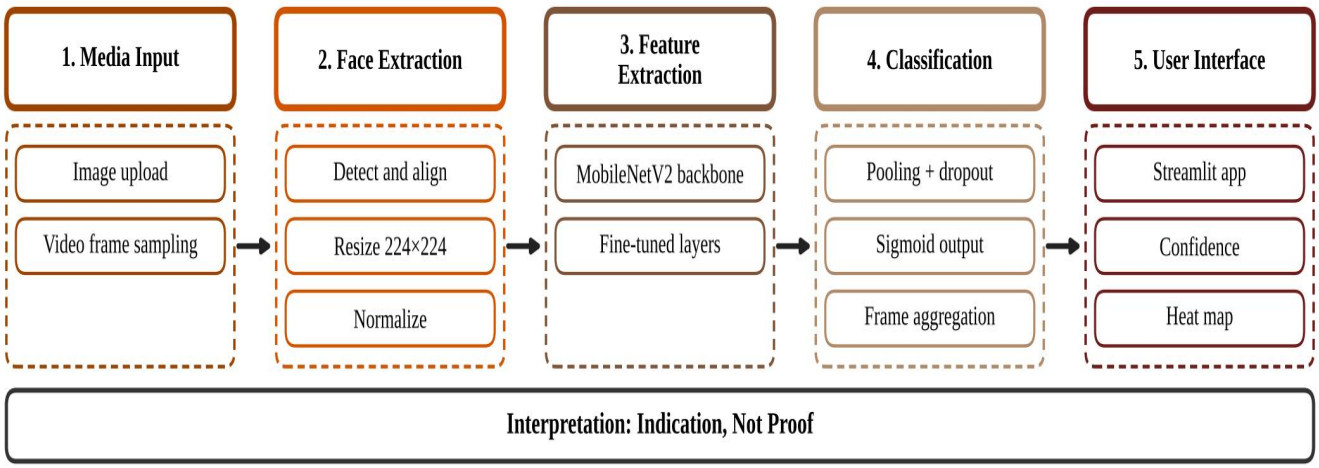



**Fig. 2 presents the proposed detection pipeline.**

**Fig. 2. Proposed lightweight transfer-learning deepfake detection pipeline.**

### 5.1 Media Input Stage

Users upload an image or video; video is sampled at a fixed frame rate.

### 5.2 Face Extraction and Preprocessing Stage

Faces are detected, aligned, cropped and resized to the backbone input size (for example 224×224) and normalized with ImageNet statistics.

*C. Feature Extraction Stage*

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue5
Sep-Oct- 2026, PP 01-06

A pretrained MobileNetV2 (or EfficientNet) backbone extracts features; upper layers are fine-tuned on real and fake faces with compression and blur augmentation.

### 5.3 Classification Stage

A global pooling layer, dropout and a sigmoid output produce a fake probability per frame; video scores are aggregated across frames.

### 5.4 User Interface Stage

A lightweight web interface (for example, Streamlit) shows the prediction, confidence and a heat map, with a note that the result is an indication, not proof.

## VI. Methodology and Working

Fig. 3 describes the workflow. The classifier is trained with binary cross-entropy,

$$L = -[y \cdot \log p + (1 - y) \cdot \log(1 - p)] \quad (1)$$

and the video-level score averages frame probabilities,

$$P_video = (1/N) \Sigma_i p_i \quad (2)$$

Evaluation uses accuracy, AUC and cross-dataset testing to measure generalization.

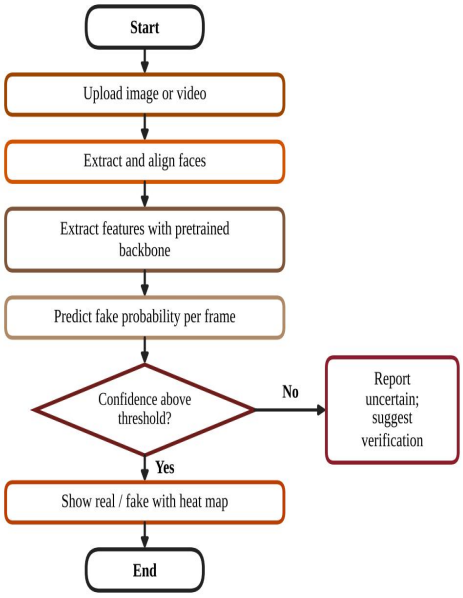



**Fig. 3. Deepfake detection workflow.**

## VII. Comparative Analysis

Table II compares detector backbones. Compact models are fast enough for web and mobile use, while larger models and attention-based designs give higher accuracy at greater cost.

| Backbone | Parameters | Speed | Typical Accuracy |
|---|---|---|---|
| MesoNet | Very few | Very fast | Moderate |
| MobileNetV2 | ≈3.5 M | Fast | Good |
| EfficientNet-B0 | ≈5.3 M | Fast | Good–High |
| Xception | ≈23 M | Moderate | High |
| Attention-based | Large | Slow | High |

**TABLE II Comparison of Detector Backbones**

## VIII. Challenges and Future Scope

Table III summarizes challenges. Future directions include temporal video-level models, audio-visual consistency checks, content provenance standards and continual learning for new generators.

| Challenge | Mitigation |
|---|---|
| Unseen methods | Artefact-focused and augmented training |
| Compression | Train with compressed samples |
| Dataset bias | Diverse, balanced datasets |
| Adversarial evasion | Robust training; ensembles |
| Over-trust | Confidence display; human verification |

**TABLE III Challenges and Mitigation**

## IX. Conclusion

Deepfake detection is an important tool for digital forensics and media literacy. The reviewed literature shows that transfer-learning detectors with backbones such as MobileNetV2 can be accurate and lightweight, and that artefact-focused methods improve generalization. Because new generators keep appearing, predictions should be treated as indications, combined with provenance and human verification.

## 9.1 Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

# References

1. R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “DeepFakes and beyond: A survey of face manipulation and fake detection,” Inf. Fusion, vol. 64, pp. 131–148, 2020.
2. S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, 2010.
3. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE CVPR, 2009, pp. 248–255.
4. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in Proc. IEEE/CVF ICCV, 2019, pp. 1–11.

5. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A large-scale challenging dataset for DeepFake forensics,” in Proc. IEEE/CVF CVPR, 2020, pp. 3207–3216.
6. B. Dolhansky et al., “The DeepFake Detection Challenge (DFDC) dataset,” arXiv:2006.07397, 2020.
7. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” in Proc. IEEE WIFS, 2018.
8. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE/CVF CVPR, 2018, pp. 4510–4520.
9. M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.
10. L. Li et al., “Face X-ray for more general face forgery detection,” in Proc. IEEE/CVF CVPR, 2020, pp. 5001–5010.
11. R. Durall, M. Keuper, and J. Keuper, “Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions,” in Proc. IEEE/CVF CVPR, 2020, pp. 7890–7899.
12. H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, “Multi-attentional deepfake detection,” in Proc. IEEE/CVF CVPR, 2021, pp. 2185–2194.
13. S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “CNN-generated images are surprisingly easy to spot... for now,” in Proc. IEEE/CVF CVPR, 2020, pp. 8695–8704.
14. R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, “On the detection of synthetic images generated by diffusion models,” in Proc. IEEE ICASSP, 2023.