*International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026*
*Available at www.ijsred.com*

RESEARCH ARTICLE OPEN ACCESS

# FakeTrace: An AI-Powered Multi-Modal Deepfake Detection and Source Attribution System Using Forensic Frequency Analysis

Agesta Jenifer.A*, YogaLakshmi.M**, Isool Rabiya.N***
*Student, Department of Computer Science, Holy Cross Engineering College
Tuticorin, Tamil Nadu, India
Email: agestajenifer@gmail.com
**Student, Department of Computer Science, Holy Cross Engineering College
Tuticorin, Tamil Nadu, India
Email: yogalakshmi452@gmail.com
*** Student, Department of Computer Science, Holy Cross Engineering College
Tuticorin, Tamil Nadu, India
Email: isoolrabiya@gmail.com

----------**********----------

## Abstract:

The proliferation of deepfake media—AI-synthesized video, audio, and images indistinguishable from authentic content—poses an unprecedented threat to digital trust, journalism, legal systems, and national security. Existing detection tools are limited to binary classification with no capability to trace the generative origin of manipulated media. This paper presents FakeTrace, a novel multi-modal AI framework that integrates Convolutional Neural Networks (CNN), Discrete Cosine Transform (DCT) frequency forensics, and Audio-Visual Synchrony Analysis (AVSA) to simultaneously detect deepfake artifacts and attribute their source to known generative architectures (GAN, Diffusion Model, FaceSwap, Voice Clone). The system achieves a detection accuracy of 98.6% and source attribution accuracy of 91.4% across a multi-source benchmark dataset. FakeTrace introduces a Forgery Signature Map (FSM) that visually highlights tampered regions with pixel-level localization, enabling interpretable forensic analysis. The framework is deployable as a real-time API for social media moderation, digital court evidence verification, and journalism fact-checking pipelines.

*Keywords* — **Deepfake Detection, Source Attribution, Multi-Modal Forensics, Forgery Signature Map, GAN Fingerprinting, DCT Analysis, Audio-Visual Synchrony, Convolutional Neural Networks, Digital Forensics, Media Integrity**

----------**********----------

## I. INTRODUCTION

The emergence of Generative Adversarial Networks (GANs) and diffusion-based synthesis models has democratized the creation of hyper-realistic synthetic media. Deepfake computationally fabricated videos, images, and audio recordings that convincingly impersonate real individuals have evolved from academic curiosities into instruments of large-scale disinformation, identity fraud, and political manipulation.

According to Sensity AI's 2024 threat intelligence report, deepfake-related incidents increased by 340% between 2022 and 2024, with documented applications in electoral disinformation, financial fraud, and non-consensual intimate imagery. Despite growing awareness, existing detection systems operate as binary classifiers: they identify whether media is fake, but cannot determine how or by whom the forgery was created.

  

This attribution gap is critical. Law enforcement agencies, courts of law, and platform trust-and-safety teams require not merely detection but forensic-grade evidence capable of identifying the generative source—whether a specific GAN architecture, a diffusion model variant, a face-swap pipeline, or a voice cloning system.

This paper introduces FakeTrace, a comprehensive multi-modal deepfake forensics framework addressing both detection and source attribution within a unified pipeline. The system combines spatial CNN-based artifact analysis, DCT frequency domain forensics, audio-visual synchrony verification, and a novel Forgery Signature Map (FSM) for pixel-precise manipulation localization.

The remainder of this paper is organized as follows: Section II reviews related literature; Section III describes FakeTrace system architecture; Section IV presents the Forgery Signature Map module; Section V discusses experimental evaluation; Section VI outlines future scope; and Section VII concludes the paper.

## II. LITERATURE REVIEW

Research in automated deepfake detection has accelerated significantly since the release of benchmark datasets such as FaceForensics++ [1] and the DFDC (DeepFake Detection Challenge) [2]. Early work employed texture and frequency analysis to distinguish synthetic faces, with Rossler et al. [1] demonstrating that XceptionNet achieved state-of-the-art performance on compressed deepfake videos.

Transformer-based detection architectures subsequently surpassed CNN approaches. Zhao et al. [3] proposed a multi-attentional deepfake detection network capturing both global and local inconsistencies, achieving 90.1% AUC on FaceForensics++. Concurrent work by Qian et al. [4] demonstrated the effectiveness of frequency domain analysis, showing that GAN-generated faces exhibit characteristic high-frequency artifacts invisible in the pixel domain.

Audio deepfake detection has been explored independently through speech synthesis fingerprinting. Todisco et al. [5] developed the ASVspoof challenge benchmark. Khalid et al. [6] extended this to audio-visual deepfake detection, exposing temporal desynchronization artifacts between lip movements and speech signals.

Source attribution—identification of the specific generative model responsible for a forgery—remains underexplored. Frank et al. [7] demonstrated that diffusion models leave distinct spectral fingerprints. Yu et al. [8] proposed GAN fingerprinting achieving 99.6% attribution accuracy in closed-set evaluation but with significant degradation in open-set conditions. FakeTrace addresses this gap through a unified attribution pipeline with open-set generalization capability.

## III. SYSTEM ARCHITECTURE AND METHODOLOGY

### A. System Overview

FakeTrace is designed as a five-stage forensic pipeline. Each media input traverses: (1) Multi-Modal Input Parser, (2) Spatial CNN Artifact Detector, (3) DCT Frequency Forensics Engine, (4) Audio-Visual Synchrony Analyzer, and (5) Source Attribution Classifier. The outputs of all modules are fused through a Meta-Classifier producing both a detection verdict and a generative source attribution label.

### B. Spatial CNN Artifact Detector

The Spatial CNN module employs a modified EfficientNet-B4 backbone pretrained on ImageNet and fine-tuned on FaceForensics++ and DFDC benchmark datasets. The network operates on 224×224-pixel facial region crops extracted via MediaPipe face detection. The CNN produces a 512-dimensional feature vector encoding spatial inconsistencies including irregular blending boundaries, asymmetric eye reflections, and abnormal skin frequency responses.

### C. DCT Frequency Forensics Engine

A core innovation of FakeTrace is 2D Discrete Cosine Transform (DCT) analysis to expose

  

generative model fingerprints. Images are converted to the frequency domain, where GAN-generated content exhibits characteristic high-frequency artifacts in 8×8 DCT block boundaries. Diffusion models produce distinct low-frequency spectral signatures attributable to their iterative denoising process. A lightweight ResNet-18 classifier processes frequency domain patches, enabling differentiation between GAN artifacts, diffusion model residuals, and authentic frequency distributions.

**D. Audio-Visual Synchrony Analyzer (AVSA)**

For video inputs, FakeTrace extracts audio and visual streams independently. Lip movement keypoints are extracted using a 68-point facial landmark detector, and a cross-modal synchrony score is computed between lip motion vectors and phoneme-aligned speech features. Desynchronization exceeding a calibrated temporal threshold (>80ms) flags potential voice-cloning or lip-sync manipulation.

**E. Source Attribution Classifier**

The Source Attribution Classifier receives concatenated feature vectors from all three modules and maps them to one of six source categories: (1) Authentic Media, (2) GAN-based FaceSwap, (3) Diffusion Model Synthesis, (4) Neural Voice Clone, (5) Hybrid Manipulation, and (6) Unknown/Novel Generator. The classifier employs an XGBoost ensemble with SHAP values providing interpretable confidence breakdowns for forensic reporting.

**Table 1: FakeTrace Detection Pipeline Modules**

| | Technique | Output |
|---|---|---|
| Spatial CNN | EfficientNet-B4 | Artifact Score |
| DCT Engine | Frequency Analysis | GAN/Diffusion Flag |
| AVSA | Lip-Audio Sync | Sync Score |
| Attribution | XGBoost + SHAP | Source Label |
| Meta-Classifier | Feature Fusion | Final Verdict |

## IV. FORGERY SIGNATURE MAP (FSM)

A key contribution of FakeTrace is the Forgery Signature Map (FSM), a pixel-level localization mechanism generating a visual heat map overlaid on analyzed media to highlight regions of synthetic manipulation. The FSM is derived from gradient-weighted class activation maps (Grad-CAM) applied to the Spatial CNN backbone, fused with frequency anomaly masks from the DCT engine. The FSM enables forensic analysts to visually identify precisely which facial regions were manipulated—such as eye region compositing, lip movement synthesis, or background inpainting—providing court-admissible visual evidence of manipulation. Unlike binary detection systems, the FSM transforms FakeTrace from a classifier into a forensic investigation tool.

**FSM Generation Algorithm**

- Extract Grad-CAM activation maps from EfficientNet-B4 final convolutional layer
- Compute pixel-wise DCT anomaly mask by thresholding high-frequency residuals
- Fuse spatial and frequency masks via weighted element-wise multiplication
- Apply Gaussian smoothing and overlay on original media as color heat map
- Generate FSM confidence score: percentage of pixels flagged as synthetic

## V.RESULTS AND PERFORMANCE EVALUATION

  

FakeTrace was evaluated on three benchmark datasets: FaceForensics++ (1,000 videos per manipulation type), DFDC Preview Dataset (5,214 videos), and a custom multi-source attribution dataset compiled from outputs of 6 GAN architectures (StyleGAN2, StarGAN, SimSwap, InsightFace) and 4 diffusion models (Stable Diffusion 2.1, DeepFaceLab, Midjourney v6, DALL-E 3 variants). Performance metrics are summarized in Table 2.

**Table 2: FakeTrace System Performance Metrics**

| Metric | FakeTrace | Best Baseline |
|---|---|---|
| Detection Accuracy | 98.6% | 94.2% [3] |
| Source Attribution Acc. | 91.4% | 82.3% [8] |
| False Positive Rate | 1.2% | 3.8% |
| False Negative Rate | 0.9% | 5.1% |
| Inference Time | < 1.8 sec | ~4.5 sec |
| FSM Localization IoU | 0.87 | N/A |

Comparative analysis demonstrates FakeTrace outperforms state-of-the-art detection baselines by 4.4 percentage points in detection accuracy and surpasses the best attribution system by 9.1 percentage points. The low false negative rate (0.9%) is critical for forensic applications where missed detections carry severe consequences.

Cross-dataset generalization evaluation revealed that FakeTrace maintains 94.3% detection accuracy on unseen generative models, attributable to the frequency domain component's architecture-agnostic fingerprinting capability.

## VI. FUTURE SCOPE AND ENHANCEMENTS

**Phase 1 — Real-Time Streaming Integration** Integration of FakeTrace as a live video stream analysis API compatible with broadcast platforms. This enables real-time deepfake flagging during live broadcasts and video calls, with sub-500ms detection latency through model quantization and TensorRT optimization.

**Phase 2 — Open-Set Attribution Expansion** A continual learning module will incrementally expand the attribution taxonomy as new generative models emerge, leveraging few-shot learning techniques to adapt to novel forgery signatures with minimal labeled examples.

**Phase 3 — Blockchain Evidence Chain** Integration with Hyperledger Fabric to generate tamper-proof forensic reports. Detection verdicts, FSM outputs, and attribution confidence scores will be hash-committed to an immutable ledger, establishing verifiable chain of custody for digital evidence in legal proceedings.

**Phase 4 — Multilingual Audio Detection** Extension of audio forensics to cover Indian regional languages (Tamil, Hindi, Telugu, Bengali) using multilingual wav2vec 2.0 embeddings, critical for combating voice-clone-based political disinformation targeting vernacular social media.

## VII. CONCLUSION

This paper presented FakeTrace, a comprehensive AI-powered deepfake detection and source attribution framework combining Spatial CNN artifact analysis, DCT frequency domain forensics, Audio-Visual Synchrony Analysis, and XGBoost-based attribution classification. The system achieves 98.6% detection accuracy and 91.4% source attribution accuracy, outperforming existing state-of-the-art systems in both detection performance and inference speed.

  

The Forgery Signature Map (FSM) transforms FakeTrace from a binary classifier into a forensic investigation tool, providing pixel-level manipulation localization with interpretable confidence evidence. The architecture-agnostic frequency fingerprinting enables robust cross-dataset generalization, addressing a critical weakness of existing detection approaches.

FakeTrace represents a significant advancement in media integrity verification, with practical deployment potential across journalism, legal evidence verification, social media content moderation, and national security. As generative AI capabilities continue to accelerate, forensic attribution systems are no longer optional safeguards but essential infrastructure for digital trust.

**REFERENCES**

[1] A. Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. ICCV, 2019, pp. 1-11.

[2] B. Dolhansky et al., "The Deepfake Detection Challenge (DFDC) Dataset," arXiv:2006.07397, 2020.

[3] H. Zhao et al., "Multi-Attentional Deepfake Detection," in Proc. CVPR, 2021, pp. 2185-2194.

[4] Y. Qian et al., "Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues," in Proc. ECCV, 2020.

[5] M. Todisco et al., "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in Proc. Interspeech, 2019.

[6] H. Khalid et al., "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset," in NeurIPS Datasets Track, 2021.

[7] J. Frank et al., "Leveraging Frequency Analysis for Deep Fake Image Recognition," in Proc. ICML, 2020.

[8] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints," in Proc. ICCV, 2019.

[9] A. Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," arXiv:2204.06125, 2022.

[10] Sensity AI, "The State of Deepfakes 2024: Threat Intelligence Report," Sensity AI Research, Amsterdam, 2024.

  