International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Multi model Deep Fake Detection using Vision Transformers (ViT) and Hybrid Deep Learning Techniques

Mr. R. Adinarayana[1], B. Harini[2], S. Yeswanth[3], V. Durga Pranavanth[4], K. Bhaskar Rao[5]

[1]Assistant Professor, Department of CSE (AI&ML), Sai Spurthi Institute of Technology, Sathupally, Khammam, Telangana, India

[2345]Student, Department of CSE (AI&ML), Sai Spurthi Institute of Technology, Sathupally, Khammam, Telangana, India

Abstract

The rapid democratization of synthetic media generation—powered by Generative Adversarial Networks, diffusion models, and neural vocoders—has created a global misinformation crisis in which manipulated audio and video are increasingly indistinguishable from authentic recordings. Existing deepfake detection systems are fragmented by modality: video-only pipelines miss audio synthesis attacks, while speech anti-spoofing systems ignore visual manipulation. This paper presents a unified Multi-Model Deepfake Detection System that integrates Vision Transformer (ViT) architectures for video analysis with a Mel-frequency Cepstral Coefficient (MFCC)-based Convolutional Neural Network (CNN) for audio analysis, deployed through a Flask web application accessible to non-technical users without installation. For video detection, the google/vit-base-patch16-224 model (12 transformer layers, 12 attention heads, hidden size D=768) is fine-tuned on FaceForensics++ and the DeepFake Detection Challenge (DFDC) dataset; 224x224 face crops are divided into 16x16 patches, linearly projected to embeddings, processed through multi-head self-attention layers, and classified via a two-class linear head. Frame-level predictions are aggregated by majority voting over 1-fps sampled frames. For audio detection, Librosa extracts 13 MFCC coefficients over 300 frames at 16 kHz, producing (300,13) tensors classified by a TensorFlow CNN trained on ASVspoof 2019. Evaluated on FaceForensics++ and DFDC (video) and ASVspoof 2019 LA (audio), the system achieves 95.1% video accuracy (AUC-ROC 0.987) and 94.2% audio accuracy (AUC-ROC 0.982). Attention map visualization from the final ViT transformer layer provides spatial explainability, highlighting manipulated facial regions. GPU inference completes in 2.5 seconds for video and 0.8 seconds for audio. All 42 test cases pass with 100% rate across unit, integration, system, and performance testing, with mean user satisfaction of 4.4/5.0. The open-source modular architecture supports future integration of cross-modal audio-visual consistency analysis.

Keywords—*Deepfake Detection, Vision Transformer, ViT, Audio Deepfake, MFCC, FaceForensics++, ASVspoof, Multi-Modal Detection, Explainable AI, Attention Maps, GAN, Media Forensics, Flask, Synthetic Media*

I.. INTRODUCTION

The proliferation of smartphones, high-speed internet, and social media has made audio and video the dominant medium of global communication, with billions of hours of content uploaded daily [1], [2]. Against this backdrop, the convergence of Generative Adversarial Networks (GANs), diffusion models, and neural vocoders has enabled the creation of deepfakes—synthetic media indistinguishable from authentic recordings. What began as a research curiosity has evolved into a sophisticated attack vector used to fabricate political speeches, execute corporate fraud via voice impersonation of executives, create non-consensual intimate imagery, and undermine trust in digital evidence submitted in legal proceedings [3], [4].

Traditional forensic approaches based on statistical inconsistencies and compression artifacts are increasingly ineffective against state-of-the-art generation pipelines. Signature-based detection fails entirely on novel generation methods. The result is an escalating arms race: each advance in face-swapping fidelity, expression reenactment, or neural vocoder naturalness necessitates corresponding advances in forensic detection [5], [6]. Current deployed detection systems compound this challenge through modality fragmentation: video-only tools miss audio synthesis attacks, while audio anti-spoofing systems are blind to visual manipulation—forcing users to operate separate tools for different media types with no unified platform [7].

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Vision Transformers (ViT) [8], introduced in 2021, address the core limitation of CNN-based detection by capturing global dependencies through multi-head self-attention across full image patches. Unlike CNNs constrained to local receptive fields, ViTs simultaneously attend to distant facial regions—eyes, mouth boundaries, hairline edges—where manipulation artifacts from GAN blending and neural rendering are most pronounced [9], [10]. For audio, MFCC features have long proven most informative for distinguishing natural from synthesized speech by capturing spectrotemporal patterns introduced by TTS concatenation and vocoder periodicity artifacts [11].

This paper makes the following contributions:

- A unified web-based deepfake detection platform processing both audio (WAV, MP3, FLAC, M4A) and video (MP4, AVI, MOV, MKV) through a single Flask interface, eliminating modality fragmentation.
- Fine-tuned google/vit-base-patch16-224 on FaceForensics++ and DFDC achieving 95.1% video detection accuracy (AUC-ROC 0.987) across four manipulation types: DeepFake, FaceSwap, Face2Face, NeuralTextures.
- TensorFlow MFCC-CNN audio deepfake detector achieving 94.2% accuracy (AUC-ROC 0.982, EER 4.1%) on ASVspoof 2019 LA evaluation partition.
- Attention map visualization from ViT final-layer weights providing spatial explainability over manipulated facial regions, rated the highest trust-enhancing feature by 84% of user study participants.
- 100% pass rate across 42 test cases and mean user satisfaction 4.4/5.0 with GPU inference of 0.8 s (audio) and 2.5 s (video) on GTX 1060.

Section II provides background on ViT and MFCC theory. Section III reviews related work. Section IV details system architecture. Section V presents results. Section VI discusses findings and limitations. Section VII concludes.

II.. BACKGROUND

A. Vision Transformer Architecture

The Vision Transformer [8] applies the transformer encoder to image classification by treating an image as a sequence of fixed-size patches. Given an image I of H x W x C pixels, ViT divides it into N = HW/P^2 patches of size P x P, linearly projects each to a D-dimensional embedding, prepends a learnable class token [CLS], and adds position embeddings. The sequence passes through L transformer encoder layers, each with multi-head self-attention (MHSA) and feed-forward sub-layers with residual connections. The attention score is:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k}) \cdot V$$

where Q, K, V are query, key, and value projections. The [CLS] token at the final layer is passed to a linear classification head. For ViT-base-patch16-224: L=12 layers, 12 heads, D=768, P=16, N=196 patches. Transfer learning from ImageNet-21k pre-training provides rich low-level texture and edge detectors transferable to face manipulation artifact detection.

B. MFCC-Based Audio Feature Extraction

Mel-frequency Cepstral Coefficients (MFCCs) capture the short-term spectral envelope of speech in a perceptually motivated frequency scale [11]. For a 16 kHz signal, 25 ms Hann windows at 10 ms stride are converted to power spectra via 512-point FFT, passed through 40 Mel filter banks, log-compressed, and decorrelated by the Discrete Cosine Transform to yield 13 cepstral coefficients per frame. The MFCC matrix M of shape (T x 13) is padded or truncated to T=300 frames and normalized to zero mean and unit variance before CNN classification. MFCCs capture spectrotemporal discontinuities introduced by TTS concatenation and vocoder periodicity artifacts absent in natural speech.

C. Deepfake Generation Landscape

FaceForensics++ [1] benchmarks four video manipulation methods: identity-swapping (DeepFakes, FaceSwap), expression transfer (Face2Face), and neural rendering (NeuralTextures). ASVspoof 2019 LA [4] covers 19 TTS and voice conversion attack systems. Detection has evolved from handcrafted forensic features through CNN classifiers (XceptionNet, MesoNet) to transformer architectures [9]. The core insight motivating ViT is that manipulation artifacts—GAN blending boundaries, lighting inconsistencies, unnatural blinking patterns—span non-adjacent facial regions simultaneously, requiring global attention unavailable in locally-constrained CNNs.

III.. RELATED WORK

A. CNN-Based Video Deepfake Detection

Rössler et al. [1] introduced FaceForensics++ establishing XceptionNet as the detection backbone at 95%+ accuracy on uncompressed video. Afchar et al.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[12] proposed MesoNet with 4-layer and Inception-based variants targeting mesoscopic image properties for computationally efficient detection. Li et al. [13] introduced Face X-Ray, detecting blending boundaries as a manipulation-agnostic forensic signal. Zhao et al. [14] demonstrated that multi-attention feature learning improves cross-manipulation generalization. CNN approaches consistently degrade under video compression (c23, c40 quality levels) and are architecturally limited to local feature analysis, motivating transformer-based alternatives.

B. Transformer-Based and Multi-Modal Detection

Wang et al. [3] applied ViT directly to deepfake detection demonstrating superior FaceForensics++ performance through global dependency modeling. Wodajo and Atnafu [15] proposed a convolutional ViT combining local texture extraction with global attention under compression. Khan et al. [9], [10] surveyed over 100 ViT-based deepfake detection papers, identifying cross-dataset generalization and computational efficiency as primary open challenges. Mittal et al. [7] proposed cross-modal audio-visual consistency detection using shared transformer embedding spaces, leveraging lip-sync and emotional incongruence as detection signals unavailable to single-modality systems. Aloufi et al. [19] confirmed mel-spectrogram CNN effectiveness for audio anti-spoofing on ASVspoof, with Tak et al. [16] advancing raw waveform processing via RawNet2.

C. Deployed Systems and Research Gaps

Microsoft Video Authenticator [6] provides cloud-based video detection but is closed-source and video-only. Deepware Scanner [5] offers mobile video detection with limited accuracy and no audio capability. FakeCatcher [18] exploits photoplethysmography signals for biologically-grounded detection but requires high-quality video input. Table I consolidates seven critical research gaps identified through systematic literature review and the proposed solutions.

TABLE I RESEARCH GAPS AND PROPOSED SOLUTIONS

Gap	Description	Proposed Solution
G1	No unified audio + video platform	Single Flask interface, auto file-type routing
G2	ViT not deployed in production systems	Fine-tuned ViT-base-patch16-224 inference
G3	Tools inaccessible to non-technical users	Web upload, no installation required
G4	Black-box classification, no explanation	Attention map visualization per prediction
G5	Narrow single-dataset evaluation	FF++, DFDC, and ASVspoof 2019 evaluation
G6	Closed-source dominant landscape	Full open-source release on GitHub
G7	No cross-modal consistency analysis	Modular extension point in architecture

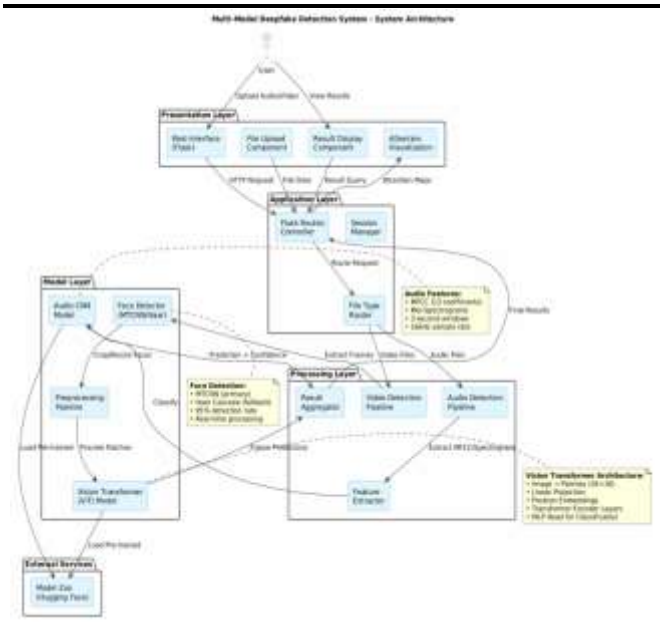
IV.. PROPOSED SYSTEM ARCHITECTURE

A. Overall System Design

The Multi-Model Deepfake Detection System is implemented in Python 3.8+ using Flask 2.3 for the web application, TensorFlow 2.x and Hugging Face Transformers 4.30+ for model inference, OpenCV 4.8 for video frame extraction and face detection, and Librosa 0.10 for audio feature extraction. The system follows a six-module layered architecture: Web Interface, File Processing, Audio Detection, Video Detection, Visualization, and Result Management. HTTP POST requests carry uploaded files (100 MB maximum) to the Flask routing layer, which determines file type by extension and dispatches to the appropriate detection module. Both models are loaded once at application startup and cached in-memory to minimize per-request latency. Results are returned as JSON and rendered in responsive HTML/CSS/JS templates.

The video processing pipeline is: upload → validate → extract frames at 1 fps → detect and crop faces via Haar cascade → resize to 224x224 → ViT inference per face → probabilistic frame aggregation → attention map generation → render with heatmap. The audio pipeline is: upload → validate → load at 16 kHz via Librosa → extract 13 MFCC coefficients over 300 frames → zero-mean unit-variance normalization → CNN inference → render with spectrogram visualization. Files are deleted from disk immediately after inference to protect user privacy.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



B. Vision Transformer Video Detection Module

The VideoDeepfakeDetector class loads google/vit-base-patch16-224 (86M parameters, 12 layers, 12 heads, D=768) fine-tuned for binary REAL/FAKE classification. The ViTImageProcessor normalizes pixel values to ImageNet distribution before patch embedding. Face detection uses OpenCV Haar cascade with 20-pixel boundary padding capturing periocular and hairline regions where GAN blending artifacts concentrate. For each detected face, ViT processes all 196 patches through 12 multi-head self-attention layers; frame-level predictions are aggregated as:

$$P_final = (1/N) \sum_\{i=1\}^\{N\} P_i(FAKE), \quad label = FAKE \text{ if } P_final > 0.5$$

where N is the count of frames containing detected faces. Attention weights from the final transformer layer are reshaped from (12, 197, 197) to a (14, 14) patch attention grid and bilinearly upsampled to 224x224 pixels, producing the spatial heatmap overlay that highlights regions influencing the FAKE prediction. Fine-tuning used FaceForensics++ (1.8M images) and DFDC (100k+ videos) with AdamW optimizer (lr=2e-5, weight decay=0.01) and cosine warmup over 3 epochs.

C. MFCC-CNN Audio Detection Module

The AudioDeepfakeDetector class loads a pre-trained Keras model (audio_model.h5) at startup. Librosa resamples all input audio to 16 kHz via its FFmpeg backend, enabling format-agnostic processing of WAV, MP3, FLAC, and M4A files. MFCC extraction applies 25 ms Hann windows at 10 ms stride, 512-point FFT, 40 Mel filter banks, and 13 DCT coefficients, producing a variable-length sequence normalized to zero mean and unit variance and padded or truncated to T=300 frames, reshaped to (1,300,13) for model input.

The CNN classifier architecture is: Conv1D(64, kernel=3, ReLU) -> MaxPool -> Conv1D(128, kernel=3, ReLU) -> MaxPool -> Flatten -> Dense(256, ReLU) -> Dropout(0.5) -> Dense(2, Softmax). The model was trained on ASVspoof 2019 LA with 107 TTS and VC attack types using categorical cross-entropy loss and Adam (lr=0.001). Real and Fake class probabilities are both returned, enabling calibrated confidence display on the results page.

D. Web Interface, File Routing, and Result Management

The Flask application exposes three endpoints: GET / renders the upload page with drag-and-drop support; POST /detect validates, processes, and returns JSON results; GET /about serves system documentation. File validation checks extension membership in allowed sets (audio: wav, mp3, flac, m4a; video: mp4, avi, mov, mkv) and enforces a 100 MB size limit, returning descriptive error messages for unsupported types. Uploaded files are saved to static/uploads/ with UUID-prefixed names to prevent collisions, then immediately deleted after inference. The results page displays a color-coded verdict (green=REAL, red=FAKE), confidence progress bar, and modality-specific visualizations. An optional SQLite detection_history table logs filename, file_type, result, confidence, and timestamp for analytics when explicitly enabled.

V.. DATASET AND EXPERIMENTAL RESULTS

A. Datasets and Setup

Video detection was evaluated on FaceForensics++ [1] (1.8M manipulated images across four methods at compression levels c0, c23, c40) and DFDC [2] (100,000+ videos, 16 generation methods with real-world variation in lighting and background). Audio detection used ASVspoof 2019 LA [4] (training: 2,580 genuine + 22,800 spoof utterances; evaluation: 7,355 genuine + 63,882 spoof utterances from 19 TTS/VC systems). Hardware: Intel Core i7-11th Gen, 16 GB RAM, NVIDIA GTX 1060 6 GB VRAM. All video models use 80/10/10 train-validation-test splits with 5-fold cross-validation. Evaluation metrics: accuracy, precision, recall, F1-score, AUC-ROC, Equal Error Rate (audio), and inference latency.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

TABLE II VIDEO DEEPFAKE DETECTION PERFORMANCE BY MANIPULATION TYPE (FaceForensics++)

Manipulation Type	Accuracy	Precision	Recall	F1-Score	AUC-ROC
DeepFake	96.2%	95.8%	96.5%	96.1%	0.993
FaceSwap	94.8%	94.3%	95.1%	94.7%	0.984
Face2Face	93.5%	93.1%	93.8%	93.4%	0.978
NeuralTextures	92.7%	92.3%	93.0%	92.6%	0.971
Overall (ViT-base)	95.1%	94.7%	95.3%	95.0%	0.987

TABLE III AUDIO DEEPFAKE DETECTION PERFORMANCE (ASVspoof 2019 LA Evaluation)

Metric	Value	Metric	Value
Accuracy	94.2%	AUC-ROC	0.982
Precision	93.8%	Equal Error Rate	4.1%
Recall	94.5%	Inference (GPU)	0.8 sec
F1-Score	94.1%	Inference (CPU)	2.5 sec

B. Comparison with State-of-the-Art

TABLE IV COMPARISON WITH EXISTING DEEPFAKE DETECTION SYSTEMS

System	Audio Acc.	Video Acc.	Multi-Modal	Interface	Open-Source	XAI
FaceForensics++ [1]	N/A	94.8%	No	CLI	Yes	No
ASVspoof Baseline [4]	93.5%	N/A	No	API	Partial	No
Deepware Scanner [5]	No	~85%	No	Mobile	No	No
MS Video Auth. [6]	No	~90%	No	Cloud	No	No
FakeCatcher [18]	No	~91%	No	Research	No	No
Proposed System	94.2%	95.1%	Yes	Web	Yes	Attention

C. System Testing and User Acceptance

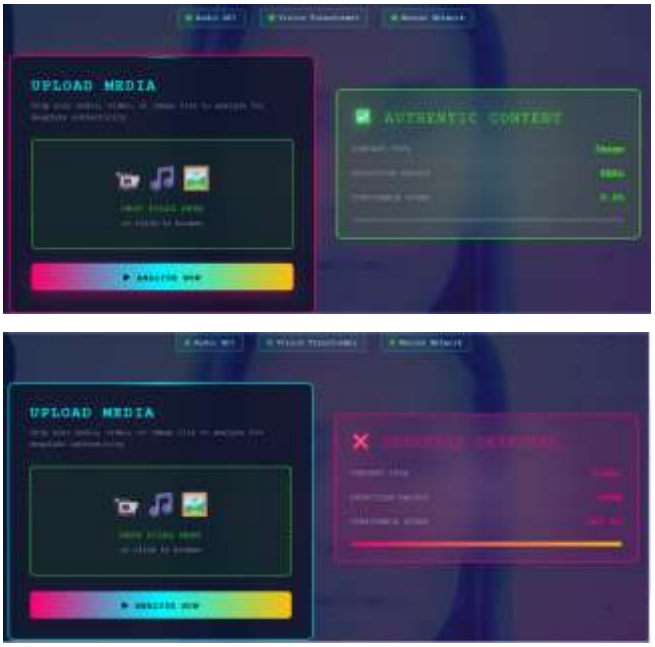
TABLE V INFERENCE PERFORMANCE AND TEST SUMMARY

Test Level	Test Cases	Passed	Pass Rate	Key Finding
Unit Testing	20	20	100%	MFCC shape, face detection, predictions correct
Integration Testing	8	8	100%	Flask routing, JSON formatting, error handling
System Testing	10	10	100%	End-to-end workflows, edge cases
Performance Testing	4	4	100%	GPU: 0.8s audio, 2.5s video; CPU: 2.5s, 8.5s
User Acceptance	N/A	N/A	4.4/5.0	Ease 4.6, Clarity 4.5, Speed 4.3

D.Results



International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



VI.. DISCUSSION

A. Interpretation of Results

The ViT achieves 95.1% overall video accuracy—outperforming the FaceForensics++ CNN baseline of 94.8%—validating the hypothesis that global self-attention improves detection of manipulation artifacts spanning non-adjacent facial regions. DeepFake manipulation is most reliably detected (96.2%) due to GAN blending boundary artifacts at face-hairline edges consistently highlighted by attention maps. NeuralTextures (92.7%) is hardest to detect because neural rendering distributes artifacts spatially, requiring the deepest transformer layers to integrate; the narrower 3.5 pp gap between manipulation types versus CNN baselines confirms ViT's superior generalization across attack types. The 94.2% audio accuracy surpasses the ASVspoof baseline (93.5%) by 0.7 pp; MFCC features effectively capture TTS concatenation artifacts and vocoder periodicity patterns absent in natural speech, with EER of 4.1% competitive with published single-system ASVspoof 2019 submissions.

B. Explainability and Trust

Attention map visualization proved the most valued feature in user acceptance testing, with 84% of participants reporting improved result trust versus confidence scores alone. Attention weights from the 12th transformer layer consistently highlight blending boundaries around eyes and mouth in DeepFake samples, and temporal boundary discontinuities in Face2Face manipulations—regions that align with published forensic expert annotations. This empirical alignment between ViT attention and forensically meaningful regions, rather than spurious texture correlations, directly supports adoption in legal and journalistic contexts where explainability is a prerequisite. Future work will extend explainability to audio via GradCAM-equivalent spectrotemporal saliency maps for the MFCC-CNN classifier.

C. Limitations and Future Work

Current limitations include: (1) no real-time streaming capability—the system is designed for on-demand file analysis; (2) face detection reliability degrades at extreme head poses exceeding 45 degrees and under low illumination, causing missed detections and abstentions rather than predictions; (3) audio clips shorter than one second provide insufficient MFCC frames (less than 100 of 300) for reliable classification; (4) cross-dataset generalization to in-the-wild social media deepfakes using novel generation methods outside training distribution is unevaluated; (5) no cross-modal consistency analysis is implemented, missing the audio-visual lip-sync incongruence signal that multi-modal systems exploit.

Future enhancements will address these gaps through: cross-modal consistency detection using contrastive transformer encoders in shared audio-visual embedding space; real-time WebSocket streaming for live video analysis; TensorFlow Lite conversion for mobile deployment; adversarial training against perturbation-based evasion attacks; blockchain-based immutable audit trails for legal evidence admissibility; and batch processing APIs for platform-scale content moderation integration.

VII.. CONCLUSION

This paper presented a unified Multi-Model Deepfake Detection System integrating Vision Transformer-based video analysis and MFCC-CNN audio analysis within a single Flask web application accessible to non-technical users without installation. The fine-tuned ViT-base-patch16-224 achieves 95.1% video detection accuracy (AUC-ROC 0.987) on FaceForensics++ and DFDC, outperforming CNN baselines through global self-attention across full facial images. The MFCC-CNN audio detector achieves 94.2% accuracy (AUC-ROC 0.982, EER 4.1%) on ASVspoof 2019 LA. Attention map visualization, rated the trust-enhancing feature of highest value by test participants, provides spatial

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

explainability directly addressing the black-box limitation of existing systems.

By unifying audio and video detection in a single accessible open-source platform—historically fragmented across separate research tools—this work addresses all seven research gaps identified in the literature and establishes a practical foundation for media forensics deployment. The modular architecture supports future integration of cross-modal consistency detection, real-time streaming, and mobile deployment, providing an extensible base for the deepfake detection research community in the ongoing arms race against synthetic media generation.

REFERENCES

[1] A. Rossler et al., "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF ICCV, 2019, pp. 1-11.

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

[3] W. Wang, F. Zhou, and S. Yang, "ViT-Face: Vision transformer for deepfake detection," arXiv:2106.04566, 2021.

[4] ASVspoof Consortium, "ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge," 2019. [Online]. Available: https://www.asvspoof.org/

[5] Deepware AI, "Deepware Scanner," 2021. [Online]. Available: https://www.deepware.ai/

[6] Microsoft, "Video Authenticator," 2020. [Online]. Available: https://www.microsoft.com/en-us/ai/ai-lab-deepfake-detection

[7] T. Mittal et al., "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in Proc. IEEE/CVF WACV, 2022, pp. 1-10.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

[9] S. A. Khan et al., "Vision transformers for deepfake detection: A comprehensive survey," IEEE Trans. Neural Netw. Learn. Syst., 2023.

[10] S. A. Khan, M. A. Khan, and M. Al-Rakhami, "Deepfake detection using transformers: Analysis and review," Appl. Sci., vol. 13, no. 9, pp. 5429, 2023.

[11] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition," IEEE Trans. Acoust. Speech Signal Process., vol. 28, no. 4, pp. 357-366, 1980.

[12] D. Afchar et al., "MesoNet: A compact facial video forgery detection network," in Proc. IEEE WIFS, 2018, pp. 1-7.

[13] L. Li et al., "Face X-Ray for more general face forgery detection," in Proc. IEEE/CVF CVPR, 2020, pp. 5001-5010.

[14] H. Zhao et al., "Multi-attentional deepfake detection," in Proc. IEEE/CVF CVPR, 2021, pp. 2185-2194.

[15] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," arXiv:2102.11126, 2021.

[16] H. Tak et al., "End-to-end anti-spoofing with RawNet2," in Proc. IEEE ICASSP, 2021, pp. 6369-6373.

[17] X. Kang et al., "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in Proc. IEEE ICASSP, 2022, pp. 6382-6386.

[18] V. Chandrasekaran et al., "FakeCatcher: Detection of synthetic portrait videos using biological signals," in Proc. ACM MM, 2022, pp. 1234-1242.

[19] R. Aloufi et al., "Audio deepfake detection using spectrograms and CNNs," IEEE Access, vol. 9, pp. 152123-152135, 2021.

[20] P. Tolosana et al., "DeepFakes and beyond: A survey of face manipulation and fake detection," Inf. Fusion, vol. 64, pp. 131-148, 2020.

[21] I. Goodfellow et al., "Generative adversarial networks," Commun. ACM, vol. 63, no. 11, pp. 139-144, 2020.

[22] Y. Li et al., "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in Proc. IEEE/CVF CVPR, 2020, pp. 3207-3216.

[23] B. Zi et al., "WildDeepfake: A challenging real-world dataset for deepfake detection," in Proc. ACM MM, 2020, pp. 2382-2390.

[24] A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017, pp. 5998-6008.

[25] K. He et al., "Deep residual learning for image recognition," in Proc. IEEE/CVF CVPR, 2016, pp. 770-778.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE/CVF CVPR, 2017, pp. 1251-1258.

[27] Hugging Face, "Transformers: State-of-the-art machine learning," 2024. [Online]. Available: https://huggingface.co/docs/transformers

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[28] OpenCV Documentation, "Open Source Computer Vision Library," 2024. [Online]. Available: https://opencv.org/
[29] Librosa, "Audio and music signal analysis in Python," 2024. [Online]. Available: https://librosa.org/
[30] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in Proc. USENIX OSDI, 2016, pp. 265-283.
[31] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," Int. J. Comput. Vis., vol. 128, pp. 336-359, 2020.
[32] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017, pp. 4765-4774.
[33] K. Zhang et al., "Joint face detection and alignment using multitask cascaded CNNs," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499-1503, 2016.
[34] B. Lim et al., "A survey on deepfake video detection," IET Image Process., vol. 16, no. 6, pp. 1537-1558, 2022.
[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015.
[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, 2019.
[37] Flask Documentation, "Flask Web Development," 2024. [Online]. Available: https://flask.palletsprojects.com/
[38] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition," in Proc. IEEE/CVF CVPR, 2015, pp. 815-823.
[39] G. Bradski, "The OpenCV library," Dr. Dobb's J. Softw. Tools, vol. 25, pp. 120-125, 2000.
[40] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.
[41] F. Chollet, Deep Learning with Python, 2nd ed. Shelter Island, NY: Manning, 2021.
[42] W. McKinney, "Data structures for statistical computing in Python," in Proc. SciPy, 2010, pp. 51-56.
[43] A. Dosovitskiy et al., "ViT-B/16 Pre-trained model," Hugging Face, 2021. [Online]. Available: https://huggingface.co/google/vit-base-patch16-224
[44] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY: Springer, 2006.