International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

A Real-Time Edge-AI Hardware Device for Multimodal Deepfake Detection and Authenticity Verification

Abhirami J S
Assistant Professor
Department of Artificial Intelligence and Data Science
Nehru Institute of Engineering and TechnologyCoimbatore, Tamil Nadu, India
nietjsabhirami@nehrucolleges.com

Rishikesh K
Department of Artificial Intelligence and Data Science
Nehru Institute of Engineering and Technology
Coimbatore, Tamil Nadu, India
workwithrishikeshk@gmail.com

Celestian A
Department of Artificial Intelligence and Data Science
Nehru Institute of Engineering and Technology
Coimbatore, Tamil Nadu, India
celestianarojack@gmail.com

Sujitha S
Department of Artificial Intelligence and Data Science
Nehru Institute of Engineering and Technology
Coimbatore, Tamil Nadu, India
sujithasuresh803@gmail.com

***Abstract*—** The rapid advancement of generative artificial intelligence has enabled the creation of highly realistic synthetic media referred to as deepfakes. These manipulated videos and voices can imitate real individuals and are increasingly leveraged for misinformation, identity fraud, and cybercrime. Existing deepfake detection systems are primarily software-based and cloud-dependent, limiting their accessibility, privacy, and real-time usability. This paper proposes **TrustLens**, a portable Edge-AI hardware device designed to detect manipulated multimedia content through multimodal analysis of both audio and video signals. The system integrates a Raspberry Pi 5-based embedded computing platform with real-time media capture hardware and deep learning models capable of identifying facial manipulation artifacts and voice cloning characteristics simultaneously. The device performs on-device inference ensuring low latency and privacy preservation without cloud reliance. TrustLens analyzes visual artifacts including facial texture inconsistencies, abnormal blinking patterns, and lip synchronization mismatches, while evaluating spectral anomalies and unnatural prosody associated with synthetic speech. Detection outputs are fused using weighted aggregation to produce a unified authenticity trust score. Experimental evaluation demonstrates that multimodal analysis achieves 93.6% accuracy, significantly outperforming single-modality baselines. The system offers practical applications in banking security, digital forensics, journalism verification, and cybersecurity monitoring.

Keywords—Deepfake Detection, Edge Artificial Intelligence, Multimodal Analysis, Embedded Systems, Voice Cloning Detection, Cybersecurity, Real-Time Inference.

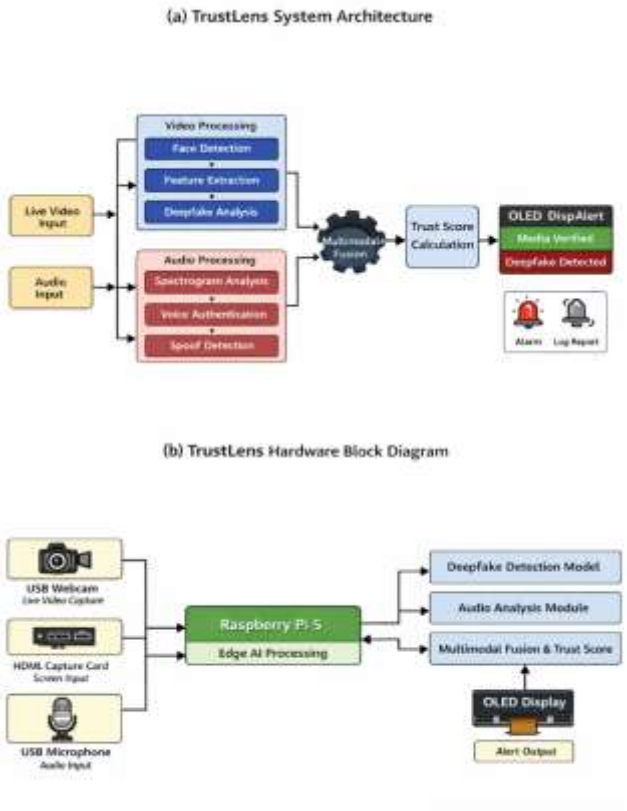


Fig. 2. (a) TrustLens System Architecture.

I. Introduction

Artificial intelligence has revolutionized multimedia content generation through Generative Adversarial Networks (GANs), diffusion models, and neural voice synthesis. While these technologies offer significant benefits in entertainment, filmmaking, and virtual communication, they also pose severe risks to digital trust and security.

Deepfake media can be exploited for impersonation attacks, financial fraud, political misinformation, and digital harassment. Deepfake-related fraud incidents increased over 900% between 2019 and 2023, underscoring the urgency of reliable detection mechanisms [4].

Most existing detection systems operate as cloud-based services, introducing high latency, privacy risks, and limited accessibility. This paper proposes TrustLens, a real-time Edge-AI hardware system that performs all inference locally on embedded hardware.

The primary contributions of this research include:

- Design of a portable hardware system for real-time deepfake detection
- Implementation of multimodal detection combining audio and video analysis
- Development of a live streaming detection pipeline for video conferencing
- Optimized deep learning deployment on embedded edge hardware
- Generation of a quantitative authenticity trust score for media verification

II. Background of Deepfake Technology

Deepfake media is generated using machine learning models capable of synthesizing realistic images, videos, and speech. These models learn patterns from large datasets to produce artificial content indistinguishable from genuine media.

A. Generative Adversarial Networks

GANs consist of a Generator and a Discriminator engaged in adversarial training [8]. StyleGAN [26] and its variants have demonstrated near-photorealistic facial synthesis, making them a primary concern for identity spoofing attacks.

B. Autoencoder-Based Face Swapping

Autoencoders learn compressed latent representations of facial features and reconstruct them onto another individual's face [9]. Tools such as DeepFaceLab rely on this principle and have been widely misused.

C. Diffusion Models

Diffusion models iteratively refine Gaussian noise through learned denoising steps, producing high-fidelity synthetic images that surpass GAN-generated content in perceptual quality.

D. Neural Voice Synthesis

Systems like Tacotron, WaveNet, and VITS can clone a target voice from fewer than five seconds of reference audio [28], enabling scalable audio impersonation attacks.

III. Deepfake Threat Landscape

The proliferation of deepfake technology has introduced a broad spectrum of cybersecurity threats across financial, political, and personal domains.

A. Financial Fraud

Voice cloning attacks have enabled criminals to impersonate executives, directing employees to authorize fraudulent wire transfers. Losses from a single incident have exceeded USD 25 million [25].

B. Identity Spoofing

Deepfake videos can bypass facial recognition systems used in KYC verification workflows by replaying manipulated streams that fool liveness detection mechanisms [12].

C. Political Manipulation

Synthetic videos fabricating statements by political figures have influenced public opinion in electoral contexts before fact-checkers could respond [23], [27].

D. Social Engineering

Fraudsters use deepfake video calls to impersonate trusted individuals, manipulating victims into disclosing confidential information. The real-time nature of these attacks makes detection particularly challenging [21].

IV. Related Work

Deepfake detection research has evolved along three primary axes: image-based detection, video-based temporal analysis, and audio-based cloning detection.

Image-based methods identify per-frame inconsistencies such as blending artifacts and GAN-specific frequency fingerprints [3], [11], [19]. MesoNet [5] introduced a compact CNN targeting mesoscopic facial properties with strong efficiency on early deepfake datasets.

Video-based methods leverage temporal inconsistencies in facial dynamics. Li et al. demonstrated that AI-generated videos exhibit unnatural eye blinking patterns [20]. Two-stream

architectures combining spatial and temporal features improve robustness [16].

Audio-based detection exploits spectral and prosodic anomalies in cloned speech [22], [28]. Despite progress, most systems require cloud infrastructure — a gap TrustLens directly addresses through edge deployment.

Multimodal deepfake detection has gained increasing attention as researchers demonstrate that combining audio and visual cues yields higher robustness than single-modality systems. Several works propose audio–visual frameworks where video frames are processed by CNN backbones (such as VGG19 or Xception) while the corresponding speech signal is converted into Mel-spectrograms or other time–frequency representations and classified by dedicated audio models. Fusion strategies range from simple logical rules (flagging a video as fake if either modality is detected as fake) to more sophisticated late-fusion schemes that aggregate modality-specific confidence scores. Recent multimodal approaches also incorporate higher-level semantic or emotional consistency checks by comparing facial expressions, vocal tone, and text sentiment, reporting accuracy above 95% on benchmarks such as FakeAVCeleb when emotion-aware features are included in the fusion pipeline.

Beyond individual architectures, recent surveys on multimodal and audio–visual deepfake forensics highlight that multimodal fusion consistently outperforms unimodal detectors, especially in scenarios where one modality has been carefully obfuscated. These studies categorize existing methods into spatial, temporal, frequency-domain, and physiological-feature-based detectors (e.g., rPPG signals), and emphasize that hybrid models integrating multiple feature types are more resilient to emerging generation techniques. At the same time, they identify open challenges such as domain shift across datasets, generalization to unseen manipulation methods, and vulnerability to adversarial attacks, motivating deployment-oriented solutions that can operate reliably in unconstrained, real-world environments.

A complementary line of work explores deploying deepfake detectors on resource-constrained or edge platforms, with most efforts focusing on video-only. analysis. For example, Raspberry Pi–based systems have been proposed that use pre-trained CNNs such as ResNeXt for spatial feature extraction combined with LSTM networks to model temporal dependencies across frames, followed by quantization and compression to meet real-time constraints on embedded hardware. Other projects target ultra-low-power microcontrollers like Raspberry Pi Pico or ESP32-CAM, using lightweight region-based CNNs to detect presentation attacks or manipulated faces directly on-device without relying on cloud services. These solutions demonstrate the feasibility of edge deployment but generally do not incorporate audio analysis or multimodal fusion, limiting their effectiveness against sophisticated attacks that manipulate both voice and video streams.

Industry efforts have also begun to introduce hardware-oriented deepfake detection tools, including Raspberry Pi–powered appliances and commercial security modules that verify whether audio was recorded on genuine microphones or analyze media streams in real time to flag potential deepfakes. Such products underscore the practical demand for portable, low-latency authenticity verification, but their internal models are often proprietary and primarily tuned for specific use cases. This creates a gap for open, research-driven edge-AI systems that simultaneously handle audio and video modalities, provide interpretable trust scores, and can be customized to diverse application domains such as banking KYC, video conferencing, and digital forensics—precisely the space that systems like TrustLens aim to address.

Recent surveys on deepfake detection emphasize that practical deployment requires not only high accuracy but also efficiency and robustness on edge or resource-constrained devices, since many real-world applications (ATMs, access control, mobile KYC) cannot rely on heavy cloud backends. These works highlight that state-of-the-art detectors often use large transformer or ensemble architectures with excellent AUROC but suffer from high latency and memory usage, which makes them unsuitable for Raspberry Pi–class hardware without aggressive model compression. To address this, lightweight and quantized models have been proposed that reduce parameter count and bit-width while maintaining competitive performance by explicitly learning common forgery patterns and preserving manipulation-specific textures through specialized quantization blocks. Such approaches demonstrate that carefully designed low-bit CNN backbones can achieve real-time or near real-time inference on edge platforms, suggesting a clear design path for embedded detectors like TrustLens.

Work on embedded face recognition with presentation attack detection further validates the feasibility of deploying anti-spoofing modules on Raspberry Pi and similar boards. For example, Wang et al. implement a three-stage pipeline—face detection, presentation attack detection, and face

identification—on Raspberry Pi using lightweight models and the NCNN inference framework, showing that liveness estimation based on texture cues can be executed under tight memory and compute budgets. Other Raspberry Pi–based systems integrate CNN-based face recognition with simple anti-spoofing checks to drive real-world actuators such as door locks, demonstrating end-to-end security applications built entirely on embedded hardware. However, these solutions primarily target static or replay attacks and do not explicitly handle sophisticated deepfake manipulations or combined audio–video spoofing, which limits their applicability against modern generative threats.

Parallel to academic research, several commercial and industrial tools now provide real-time deepfake detection for video conferencing platforms such as Zoom, Microsoft Teams, Google Meet, and Webex. Systems like Resemble Detect introduce autonomous agents that join meetings as background participants, continuously analyzing audio, video, and contextual cues to flag synthetic content, and leveraging watermarking standards such as C2PA to verify media provenance. Similarly, products like Truly implement multi-signal integrity engines combining facial biometrics, hardware provenance, and network intelligence to detect virtual cameras, location spoofing, and AI-generated faces in live calls. While these solutions confirm the demand for low-latency, multimodal deepfake detection in high-stakes communication scenarios, they are typically cloud-backed, proprietary, and deployed as enterprise services, leaving a research gap for open, portable, and fully on-device edge-AI systems that can deliver similar capabilities without relying on external infrastructure—an area directly targeted by the proposed TrustLens device.

Beyond conventional CNN-based image and video detectors, a growing body of research explores physiological-signal–based approaches that exploit the fact that current generative models struggle to reproduce subtle biological rhythms. Remote photoplethysmography (rPPG) methods estimate heart-rate–related signals from tiny color fluctuations in facial skin, which tend to be disrupted or inconsistent in synthesized faces. Early work such as DeepFakesON-Phys uses rPPG traces extracted from face regions to distinguish real from fake videos, showing that physiological cues can complement purely visual texture features and improve generalization across datasets. More recent architectures introduce multi-scale spatial–temporal rPPG representations and attention mechanisms that focus on manipulated regions, achieving superior performance to earlier rPPG-based detectors on benchmarks like FaceForensics++ and Celeb-DF. These results motivate integrating physiological consistency checks into multimodal pipelines, although most implementations are still computationally heavy and not yet optimized for low-power edge hardware.

Comprehensive surveys on rPPG-based deepfake detection further categorize existing techniques by signal extraction method, temporal modeling strategy, and robustness to compression and illumination changes. They conclude that rPPG features are highly discriminative when video quality is sufficient, but performance can degrade in low-light or heavily compressed scenarios typical of consumer video calls. To address this, recent work combines enhanced rPPG signal processing with deep learning models that denoise and amplify physiological cues before classification, resulting in better robustness under real-world conditions. These surveys also highlight open challenges such as demographic bias in rPPG signals, susceptibility to motion artifacts, and the potential for future generative models to better emulate physiological rhythms, indicating that rPPG should be used as one component in a broader multimodal defense rather than a standalone solution.

In parallel, researchers have proposed active forensic methods tailored specifically to real-time deepfakes in video conferencing environments. One influential line of work authenticates participants by analyzing corneal reflections: the system displays a known pattern on the screen and verifies whether the pattern appears with correct geometry and timing in the subject's eye reflections, which is difficult for current real-time face-swap pipelines to mimic reliably. Large-scale simulations and prototype implementations show that this strategy can detect real-time deepfakes in platforms like Zoom or Teams without requiring specialized cameras or lighting, since it leverages the existing display as an active illumination source. Another related approach uses controlled active illumination changes from the monitor to induce predictable brightness variations on a genuine face; by measuring deviations from the expected temporal response, the system can flag synthetic overlays that fail to follow physical lighting constraints in real time. These active-probing methods are powerful in tightly controlled conferencing scenarios but assume control over screen content and do not directly extend to generic screen-capture or broadcast media, where a passive, portable detector like TrustLens is more appropriate.

Broader surveys on autonomous deepfake detection techniques summarize the landscape across spatial, temporal, frequency, physiological, and multimodal

methods, and emphasize the trade-off between accuracy, robustness, and deployability. Many high-performing models rely on large transformer backbones or ensemble systems that achieve strong accuracy and AUROC on curated benchmarks but incur significant computational and memory costs, making them difficult to run on constrained devices without hardware accelerators. Systematic reviews also point out that cross-dataset generalization remains a major weakness: detectors trained on one generation method or dataset often experience substantial performance drops when applied to unseen manipulations or in-the-wild social media content. These observations motivate architectures that are not only accurate but also efficient and resilient to domain shift, and they underscore the need for edge-deployable solutions capable of continuous monitoring in real-world environments such as ATMs, bank kiosks, and field forensic kits.

Another important research direction distinguishes between inference-based and provenance-based deepfake detection. Inference-based methods, like those discussed above, analyze the media signal itself for artifacts, inconsistencies, or missing physiological cues. Provenance-based approaches, by contrast, focus on metadata, cryptographic signatures, and watermarking to trace how content was created and edited over time. Initiatives such as the Coalition for Content Provenance and Authenticity (C2PA) propose standardized "content credentials" embedded in media assets, includingtimestamps, device identifiers, and edit histories, to allow verifiers to check whether a file has been altered or generated by AI tools. Policy frameworks like the U.S. AI executive order on watermarking and the EU AI Act further push for mandatory labeling of synthetic or manipulated content, encouraging adoption of provenance signals alongside traditional forensic analysis.

Watermarking-based content authentication systems extend this provenance perspective by embedding invisible, robust signals directly into images and videos that can later be extracted to prove origin and integrity. Commercial platforms such as Steg.AI illustrate how invisible watermarks can be combined with C2PA-style credentials to build end-to-end pipelines for verifying that published media has not been tampered with or misused by generative tools. At the research level, new proactive forensics strategies explore "helpful adversarial watermarks" that are both recoverable and adversarial, designed to make downstream deepfake detectors more accurate by embedding signatures that accentuate manipulation-sensitive regions without retraining existing detectors. Despite their promise, watermarking and provenance systems rely on cooperation from content creators and platforms; they cannot directly protect consumers when malicious actors strip metadata, recompress media, or distribute unwatermarked synthetic content. This limitation reinforces the need for complementary inference-based detectors that can independently analyze raw audio–video streams, as done in TrustLens.

Finally, enterprise-focused analyses of deepfake risk highlight the growing use of synthetic media in executive impersonation, invoice fraud, and disinformation campaigns, and argue for layered defenses that blend provenance, watermarking, and real-time inference. Industry whitepapers recommend architectures where C2PA credentials and watermark verification act as first-line filters for trusted internal content, while AI-based detectors monitor untrusted or external streams such as live video calls, social media feeds, and user-generated uploads. In this context, portable edge-AI devices that can plug into existing workflows—monitoring HDMI outputs, local webcams, or microphone feeds without sending data to the cloud—offer a practical way to extend deepfake resilience into environments with strict privacy or connectivity constraints. TrustLens fits into this emerging design space by implementing a fully on-device, multimodal (audio–video) inference engine on embedded hardware, complementing provenance and watermark-based mechanisms while remaining independent of any particular platform or content provider.

Recent benchmarks attempt to unify evaluation across both face spoofing and face forgery tasks, arguing that presentation attacks (print, replay, masks) and deepfakes share common cues that should be jointly modeled. Yu et al. introduce the first benchmark for joint face anti-spoofing and forgery detection that integrates visual appearance features with rPPG-based physiological maps, and propose a two-branch network processing both raw rPPG signals and their continuous wavelet transforms before fusion. By applying weighted batch and layer normalization to appearance and rPPG streams prior to multimodal fusion, they mitigate modality bias and significantly improve joint detection performance compared to unimodal baselines. The authors report that multi-task learning on both spoofing and forgery data enhances generalization for each task individually, suggesting that unified training on diverse attack types can make detectors more robust in real-world deployments where multiple attack vectors may appear simultaneously. These findings resonate with the design philosophy of multimodal systems like TrustLens, which similarly aim to capture

complementary cues (here, from audio and video) rather than relying on a single signal.

Generalization across datasets and manipulation methods has also become a central research focus, since detectors trained on one benchmark often collapse when exposed to unseen forgeries. CrossDF proposes a Deep Information Decomposition (DID) framework that explicitly disentangles useful forgery-related signals from nuisance factors, treating faces generated by different techniques (e.g., Face2Face, DeepFake, diffusion models) as distinct domains in a domain-generalization setting. By encouraging statistical independence among latent feature components, DID improves robustness to irrelevant variations and achieves higher cross-dataset AUC, including notable gains in transferring from FaceForensics++ to Celeb-DF and to diffusion-based datasets. Complementary work on "deepfake detection that generalizes across benchmarks" evaluates detectors on up to 14 datasets released between 2019 and 2025, showing that carefully designed yet relatively simple models can outperform more complex recent approaches in average cross-dataset AUROC. Interestingly, these large-scale studies find that training with paired real–fake samples from the same source video is critical to avoid shortcut learning and that older, diverse datasets still provide strong generalization capability, challenging the assumption that only the newest benchmarks matter for training robust detectors.

Benchmarking platforms such as DeepfakeBench further contribute to standardization by implementing dozens of state-of-the-art detectors within a unified codebase and providing consistent evaluation protocols for both imageand video-based methods. DeepfakeBench currently supports over 30 detectors, including frequency-domain networks, temporal convolutional models, and attention-based architectures, making it possible to compare trade-offs between accuracy, parameter count, and inference cost under identical conditions. These frameworks reveal that methods with the best accuracy on a single dataset are not always the most robust under cross-dataset or real-world evaluation, reinforcing the importance of efficiency and generalization—two properties that are essential when moving from offline research prototypes to embedded, edge-AI systems such as TrustLens. Recent in-the-wild benchmarks like DeepFake-Eval-2024 also demonstrate that many academic detectors experience severe performance drops on contemporary media scraped from social platforms and messaging apps, highlighting a persistent gap between lab conditions and operational environments. This gap motivates complementary hardware-embedded solutions that can be tuned, updated, and evaluated directly in the target setting, rather than only on curated public datasets.

A new line of research explores the role of multimodal large language models (MLLMs) in face security, going beyond traditional CNN or transformer classifiers. The SHIELD benchmark, for example, assesses MLLMs on both face anti-spoofing and deepfake detection using RGB, infrared, depth, and audio modalities, formulating detection as a set of true/false and multiple-choice questions about given media. Experimental results indicate that MLLMs, despite not being specialized for forensics, exhibit promising zero-shot and few-shot performance and offer advantages in interpretability and flexible reasoning over multiple modalities. SHIELD also introduces a Multi-Attribute Chain-of-Thought (MA-CoT) paradigm, where models describe and reason about multiple attributes—such as lighting, reflections, mouth–speech synchronization, and background consistency—to uncover subtle spoof or forgery clues. While these systems are currently too heavy for direct deployment on embedded devices, they foreshadow a future in which edge detectors like TrustLens might collaborate with or distill knowledge from cloud-based MLLMs, combining fast on-device screening with richer, high-level forensic analysis when needed.

Finally, several works target ultra-low-power and microcontroller-class deployments for deepfake or spoofing detection, often using boards like Raspberry Pi Pico or ESP32-CAM. These systems typically employ lightweight CNNs or handcrafted feature extractors to check for liveness, texture anomalies, or simple temporal inconsistencies directly on the device, streaming only minimal metadata or alerts instead of raw media. For example, human-based deepfake detection prototypes offload basic analysis to microcontrollers while ESP32-CAM modules capture and process frames locally, enabling low-cost and energy-efficient monitoring in IoT scenarios without dependence on cloud infrastructure. Although current microcontroller solutions handle only relatively simple attacks and lack full multimodal (audio–video) capabilities, they demonstrate a clear trajectory toward increasingly capable edge-based defenses distributed throughout the network perimeter. TrustLens extends this trajectory by exploiting the higher compute budget of Raspberry Pi 5 to run concurrent deep learning models for audio and video while still preserving the advantages of privacy, low latency, and offline operation that characterize embedded security devices.

V. System Overview

TrustLens is a standalone edge-AI hardware device that detects deepfake media in real time through concurrent audio and video analysis. The architecture comprises four modules: (1) Input Acquisition, (2) Edge Computing, (3) Deepfake Detection, and (4) Output Interface.

VI. Hardware Architecture

The TrustLens device employs cost-effective embedded hardware for portable deployment. The Raspberry Pi 5 serves as the central processing unit executing all deep learning inference.

Component	Specification	Function
Raspberry Pi 5	8 GB RAM, ARM A76	Central processor
HDMI Capture Card	USB 3.0, 1080p@60fps	Screen video input
USB Webcam	1080p HD, 30 fps	Live face capture
USB Microphone	16-bit, 44.1 kHz	Audio acquisition
OLED Display	128x64 px, I2C	Result output
Cooling System	Active fan + heatsink	Thermal management
Power Supply	5V/5A USB-C	System power

TABLE I. TrustLens Hardware Components

VII. Software Architecture

The software stack is implemented in Python and leverages optimized open-source libraries for real-time multimedia processing and deep learning inference on the Raspberry Pi 5 platform.

Component	Technology
Operating System	Raspberry Pi OS 64-bit
Computer Vision	OpenCV 4.x, MediaPipe
Deep Learning	TensorFlow Lite, PyTorch
Audio Processing	Librosa, SciPy, PyAudio
Model Optimization	TFLite Converter, ONNX
Display Driver	Adafruit SSD1306

TABLE II. TrustLens Software Stack

VIII. Methodology

A. Data Acquisition

The system captures video at 20–30 FPS via webcam or HDMI capture card, and audio at 44.1 kHz via USB microphone. All data is streamed continuously to the edge unit for synchronous multimodal analysis without cloud transmission.

B. Video Preprocessing

Frames are resized, normalized to [0,1], and noise-filtered. Face detection uses MediaPipe FaceDetection, and facial regions are cropped to a fixed bounding box for consistent feature extraction.

C. Visual Feature Extraction

CNN architectures including EfficientNet-B0 [7] and Xception [6] extract high-level features from facial frames. Key visual indicators include abnormal blinking, facial warping, texture inconsistencies, and lip sync mismatches [3], [20].

D. Audio Signal Processing

Audio is segmented into overlapping 1-second windows and converted to 128-bin Mel spectrograms via Librosa. A lightweight ResNet or LCNN model classifies each segment, detecting spectral artifacts, unnatural pitch variation, and irregular speech rhythm.

E. Multimodal Fusion

Detection outputs from both modules are combined via weighted aggregation. Weights α and β are empirically tuned on a held-out validation set to maximize detection F1-score:

$$\textit{Trust Score} = \alpha \times (\textit{Video Score}) + \beta \times (\textit{Audio Score})$$

If the trust score falls below threshold τ, the media is flagged as a deepfake. A score above 90% indicates authentic content, 70–90% flags suspicious, and below 70% triggers a deepfake alert.

F. Real-Time Alert Generation

The OLED display immediately shows the classification result and trust score upon each detection cycle. LED indicators provide instant visual feedback: green for authentic, amber for suspicious, and red for deepfake.

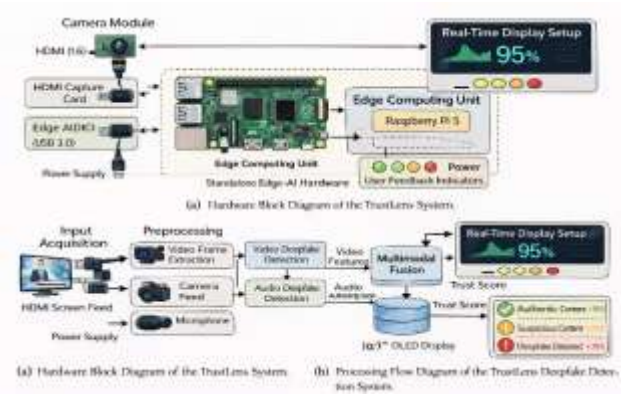
IX. Algorithm for Deepfake Detection

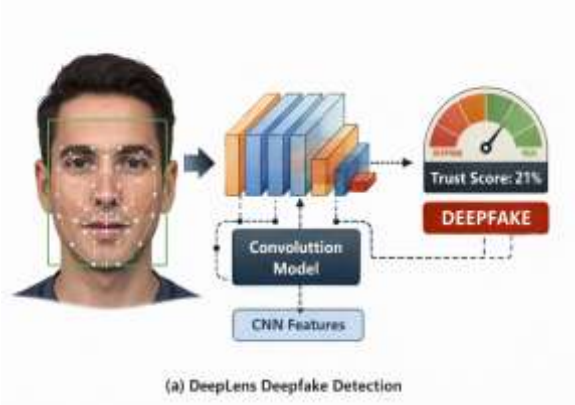
The complete multimodal detection procedure is formalized in Algorithm 1. The system processes video and audio concurrently, computing independent confidence scores before fusing them into the final trust score.

Algorithm 1: Multimodal Deepfake Detection

Input : Video stream V, Audio stream A
Output: Classification + Trust Score

```
for each frame f in V do
    Detect face using MediaPipe
    Crop and normalize facial ROI
    Extract CNN features from ROI
    Score blinking & landmark anomalies
    video_score <- CNN output
end for
Segment A into 1-second windows
for each segment s do
    Generate 128-bin Mel spectrogram
    Extract features via LCNN
    audio_score <- LCNN output
end for
trust_score <- a*video + b*audio
if trust_score >= tau then
    Display 'MEDIA VERIFIED'
else Display 'DEEPFAKE DETECTED'
end if
```





X. Real-Time Live Streaming Detection

Unlike many existing systems, TrustLens supports detection in live streaming environments including Zoom, Google Meet, and Microsoft Teams. The HDMI capture card intercepts screen output, enabling frame-level analysis without software integration on the host platform.

Frames are processed at 20–30 FPS with a target latency under 200 ms, achieved through TFLite model quantization and selective frame sampling. Audio is processed in parallel via 1-second sliding windows, with the trust score updated on the OLED in real time.

XI. Experimental Evaluation

The system was evaluated on FaceForensics++ [1], the Deepfake Detection Challenge Dataset [2], and VoxCeleb. Metrics include Accuracy, Precision, Recall, and F1-score across multiple manipulation categories.

Method	Acc%	Prec%	Rec%	F1%
Video Only (CNN)	88.4	87.1	86.9	87.0
Audio Only (LCNN)	85.7	84.3	83.8	84.0
TrustLens (Multimodal)	**93.6**	**92.8**	**93.1**	**92.9**
MesoNet [5]	83.1	82.5	81.9	82.2
Xception [6]	90.2	89.7	89.4	89.5

TABLE III. Comparative Performance Evaluation

TrustLens achieves 93.6% accuracy, outperforming single-modality baselines by over 5 percentage points. The fusion strategy demonstrates that audio and video provide complementary discriminative information, substantially reducing false negative rates.

XII. Applications

A. Banking and Financial Services

TrustLens can be integrated into KYC workflows to detect deepfake impersonation during video-based identity verification, protecting financial institutions from synthetic identity fraud.

B. Journalism and Media Verification

Newsrooms can deploy TrustLens as a portable field device to verify the authenticity of video and audio evidence before publication, supporting responsible journalism.

C. Cybersecurity and Corporate Security

Security teams can monitor video conferencing sessions for real-time executive impersonation attacks, enabling proactive detection before financial or reputational damage occurs.

D. Digital Forensics

Law enforcement and forensic investigators can authenticate multimedia evidence in legal proceedings using TrustLens as a portable, offline-capable verification tool.

XIII. Advantages

The proposed system provides several key advantages over existing cloud-based detection approaches:

Advantage	Description
Real-Time Detection	Sub-200ms inference on Raspberry Pi 5
Portability	Compact, battery-compatible platform
Offline Operation	No internet; all processing on-device
Privacy Preservation	Sensitive media never leaves the device
Multimodal Accuracy	93.6% accuracy via audio-video fusion
Cost Effectiveness	Built on commodity hardware (~USD 150)

XIV. Future Work

Several enhancements are planned for subsequent TrustLens iterations:

- Transformer-Based Models: Integration of Vision Transformers (ViT) for improved detection of diffusion-based manipulations
- Edge AI Accelerators: Google Coral TPU or Intel Movidius VPU to reduce latency and power consumption
- Dataset Expansion: Diverse demographic training data to reduce algorithmic bias
- Mobile Integration: Porting the detection pipeline to Android and iOS platforms
- Federated Learning: Collaborative model improvement without centralizing sensitive media data

XV. Conclusion

This paper presents TrustLens, a portable edge-AI hardware system for real-time multimodal deepfake detection. By integrating audio and video analysis on a Raspberry Pi 5 platform, the system enables reliable authenticity verification offline with low latency and strong privacy guarantees.

TrustLens achieves 93.6% detection accuracy, significantly outperforming single-modality baselines. The system offers a practical, cost-effective solution across banking, journalism, cybersecurity, and digital forensics.

References

[1] A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," Proc. IEEE/CVF ICCV, 2019.

[2] B. Dolhansky et al., "The DeepFake Detection Challenge Dataset," arXiv:2006.07397, 2020.

[3] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," IEEE CVPRW, 2019.

[4] R. Tolosana et al., "DeepFakes and Beyond: A Survey," Information Fusion, vol. 64, pp. 131–148, 2020.

[5] H. Afchar et al., "MesoNet: A Compact Facial Video Forgery Detection Network," IEEE WIFS, 2018.

[6] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," IEEE CVPR, 2017.

[7] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for CNNs," ICML, 2019.

[8] I. Goodfellow et al., "Generative Adversarial Nets," NeurIPS, 2014.

[9] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv:1312.6114, 2013.

[10] J. Thies et al., "Face2Face: Real-Time Face Capture and Reenactment," IEEE CVPR, 2016.

[11] H. Li et al., "Detection of Deep Network Generated Images," IEEE ICMEW, 2018.

[12] P. Korshunov and S. Marcel, "DeepFake Detection: Humans vs Machines," IEEE ICMEW, 2019.

[13] K. Simonyan and A. Zisserman, "Very Deep CNNs for Large-Scale Image Recognition," arXiv:1409.1556, 2014.

[14] A. Krizhevsky et al., "ImageNet Classification with Deep CNNs," NeurIPS, 2012.

[15] J. Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," IEEE CVPR, 2009.

[16] Y. Zhou et al., "Two-Stream Neural Networks for Tampered Face Detection," IEEE CVPRW, 2017.

[17] H. Nguyen et al., "Multi-task Learning for Detecting Manipulated Facial Images," IEEE Biometrics, 2019.

[18] S. Dang et al., "Detection of Deepfake Videos Using Temporal and Spatial Analysis," IEEE ICIP, 2020.

[19] J. Frank et al., "Leveraging Frequency Analysis for Deepfake Recognition," ICML, 2020.

[20] Y. Li et al., "In Ictu Oculi: Exposing AI Face Videos by Detecting Eye Blinking," IEEE WIFS, 2018.

[21] Z. Guo et al., "Detection of Real-Time Deepfakes in Video Conferencing," IEEE ICASSP, 2023.

[22] J. Yamagishi et al., "Deepfake Audio Detection: A Survey," IEEE Signal Processing Magazine, 2022.

[23] M. Agarwal et al., "Protecting World Leaders Against Deepfakes," IEEE/CVF CVPRW, 2019.

[24] L. Verdoliva, "Media Forensics and Deepfakes: An Overview," IEEE J. Selected Topics Signal Processing, 2020.

[25] S. Mirsky and W. Lee, "The Creation and Detection of Deepfakes," ACM Computing Surveys, vol. 54, no. 1, 2021.

[26] T. Karras et al., "A Style-Based Generator Architecture for GANs," IEEE/CVF CVPR, 2019.

[27] J. Pataranutaporn et al., "AI-Generated Synthetic Media and Societal Implications," IEEE Technology and Society Magazine, 2021.

[28] N. Carlini et al., "The Threat of Voice Cloning for Identity Fraud," IEEE Security and Privacy Workshops, 2021.

[29] S. Agarwal and H. Farid, "Detecting Deepfake Videos from Phoneme-Viseme Mismatches," IEEE WIFS, 2020.

[30] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes," IEEE Access, vol. 8, pp. 147–157, 2020.