AN INTERPRETABLE FRAMEWORK FOR SYNTHETIC MEDIA DETECTION BASED ON MULTI-DOMAIN FORENSIC ANALYSIS

A. Primbetov[1], A. Alimbaeva[2]
Phd student, Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi[1]
Phd student, Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi[2]
https://doi.org/10.5281/zenodo.20625101

Abstract. *In this paper, we introduce an interpretable multi-domain forensic framework for media manipulation detection as an improvement over our previous work using a CNN-LSTM-Spline framework. While our previous research largely relied on spatial-temporal deep features extraction, the present framework also employs provenance/watermark authentication, audio forensics, facial-spatial inconsistency analysis, temporal-motion inconsistency analysis, and forensic analysis in frequency domain in an integrated weighted fusion framework. The performance testing carried out on 455 video clips from Celeb-DF and FaceForensics++ datasets revealed that the presented framework yielded 91.5% accuracy in detection, demonstrating its superiority to traditional single-domain detection frameworks.*

Keywords: *manipulated media detection, multimedia forensics, frequency-domain analysis, temporal consistency analysis, provenance verification, interpretable forensic framework.*

INTRODUCTION

Recently, deepfake technology has become one of the key dangers to digital trust, information integrity, and media authenticity. Using deep learning algorithms like Generative Adversarial Networks (GANs), autoencoders, and diffusion-based deep learning models, deepfakes allow creating highly accurate synthetic videos with controlled face animations and voice characteristics [1]. In spite of many useful implementations of such technologies in various spheres, there is no denying that deepfakes pose a number of risks related to manipulation, misinformation, cybercrime, etc. [2].

A fast development of deep learning models for deepfakes generation has significantly increased the level of visual authenticity and coherence between adjacent frames of a video. Contemporary deepfakes allow reducing artifacts and providing natural inter-frame transitions that look very similar to real-life situations [3]. The problem arises since not only humans but also traditional AI systems encounter difficulties in identifying deepfake content and distinguishing it from an authentic video.

Several recent studies have demonstrated that human perception alone is highly unreliable for detecting deepfakes. In a comprehensive meta-analysis conducted by Diel et al. (2024), 56 scientific studies involving more than 86,000 participants were analyzed to evaluate human performance in identifying AI-generated content [4].

Modality	Deepfake Accuracy (%)	Real Accuracy (%)
Audio	62.08	70.67
Image	53.16	68.46
Text	52	66.81
Video	57.31	68

Table 1: Human Accuracy in Detecting Deepfake Content.

These performance discrepancies also help to shed light on the problem at hand. According to the data provided in source [4], human identification success is 62.08%, 53.16%, 52%, and 57.31% for audio, image, text, and video deepfakes, respectively. This information shows that even visually sophisticated manipulations can easily deceive humans, especially when videos contain realistic face motion and high-quality graphics. Similar results were found by our team during an empirical investigation among 492 students of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, with deepfake detection accuracy estimated at roughly 57% [5]. The study showed that 46% of participants have never heard about deepfakes, which means that their knowledge about this form of AI-generated misinformation is not enough [5].

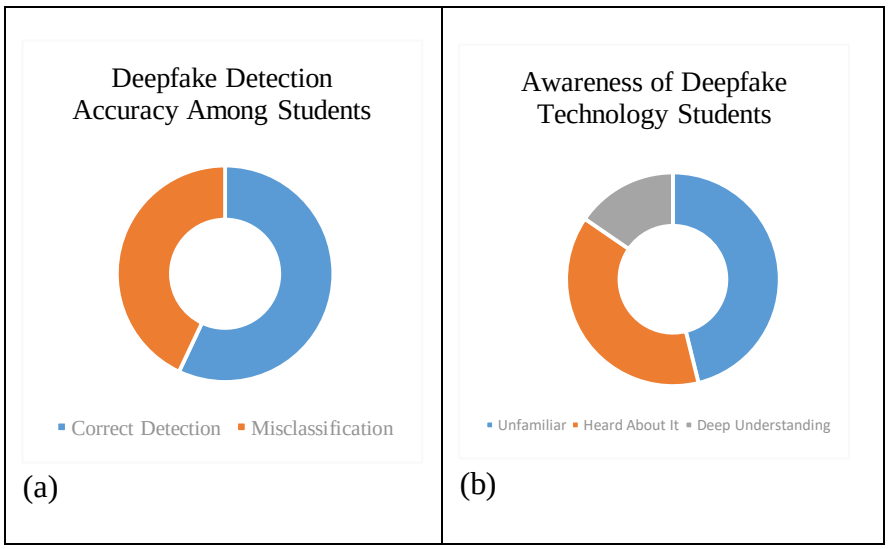


Figure 1: Deepfake detection accuracy and awareness of deepfake technology among students.

The results indicate that although the majority of participants correctly identified manipulated content, a substantial proportion of respondents still misclassified deepfake videos, demonstrating the limitations of human perception in distinguishing synthetic media from authentic content.

Furthermore, participants were asked about their prior awareness of deepfake technology. As shown in Figure 1 (a), 46.1% of respondents stated that they were unfamiliar with the concept of deepfakes, while only 15.4% claimed to possess detailed knowledge about AI-generated media manipulation [5]. These results show that despite the extensive use of technology, a significant number of students lack adequate knowledge regarding new methods of manipulation facilitated by artificial intelligence. These facts further emphasize the importance of building an automated system to detect deepfakes that are not visible to the naked eye.

In our earlier research, we have presented a deep hybrid framework based on ResNeXt and Long Short-Term Memory (LSTM) networks that can detect deepfake videos [6]. In our architecture, we used the ResNeXt network to extract spatial features and the LSTM network to capture the temporal dependencies in order to detect frame-level artifacts and inconsistencies between frames. Our experiments performed on various datasets such as DFDC, FaceForensics++, and Celeb-DF revealed that a combination of spatial and temporal analysis can increase the accuracy of detection in comparison to standard methods that focus only on frame analysis [6]. Nevertheless, as more advanced technologies of deepfake generation emerge, artifacts become less noticeable.

As revealed by recent research works, contemporary manipulated media tend to include invisible inconsistencies dispersed through various forensic dimensions instead of being associated with some visible inconsistencies in only one category of visual artifacts [7]. While the use of recent deep-learning-based models for generating manipulated content leads to an increase in spatial and temporal consistencies, retaining complete consistency related to metadata format, audio behavior, face structure, motion coherence, and spectrum is not easy. Some inconsistencies, which can be detected based on their origin, such as strange encoding formats, evidence of recompression, absence of authentication features, and damage to watermarks, can show up in media generation and processing procedures [8].

Other than audio distortions, altered videos are known to suffer from abnormal geometry of faces, asymmetrical landmarks, blurry faces, and inconsistent spatial blending because of faulty neural rendering and latent space transformation. Apart from that, temporal inconsistencies are another limitation of today's generation models, where flickering in the frames, unstable optical flow, motion inconsistencies, and temporal artifacts can be seen between adjacent frames. Furthermore, through the process of frequency analysis, it is found out that fabricated content includes many hidden spectral issues such as periodic frequency artifacts, texture repetitions, and Fourier distribution inconsistencies [9,10].

The primary goal of this research paper is to design a highly interpretable multi-domain forensic framework for detecting manipulated media by performing forensic analysis using multiple perspectives, including provenance, acoustic, spatial, temporal, and frequency domain perspectives. In this work, the designed forensic framework will include provenance and watermark verifications, audio forensics, face-spatial inconsistencies, temporal motion consistency analysis, and forensic analysis based on the frequency domain approach in a weighted fusion scheme.

METHODOLOGY

Proposed study offers a framework for interpretative detection of synthetic media using multi-domain forensic analysis. Unlike existing systems, the proposed system incorporates a combination of multiple forensic evidences to increase both robustness and interpretability of the detection algorithm.

Proposed approach builds on the previous works using the CNN-LSTM-Spline architecture for media detection; however, it incorporates new forensic analysis components that can analyze suspicious inconsistencies in multiple independent domains [6]. Whereas the previous detection models focused mainly on deep feature extraction from the spatial-temporal domain, the current approach explicitly incorporates forensic reasoning to generate interpretative evidence of manipulated media. As such, the suggested forensic analysis framework includes five independent

components, which are as follows: Provenance and Watermark Analysis, Audio Forensic Analysis, Facial and Spatial Analysis, Motion Analysis, and Frequency Domain Forensic Analysis.

2.1 Provenance and Watermarking Verification

The first module in the forensic process of the suggested framework will be provenance verification and digital watermark analysis. While conventional visual forensics is predominantly concerned with detecting visual artifacts, provenance forensics considers the origin, integrity, and consistent encoding of media items. The objective of the forensics module will be the detection of suspicious features in synthetic media introduced during synthesis, recompression, or unauthorized editing. State-of-the-art video synthesis techniques commonly alter the original encoding scheme of videos as part of the rendering, recompression, neural synthesis, or editing steps. This results in synthetic media having unusual metadata attributes, inconsistent encoding characteristics, and lack of authenticity credentials compared to camera-captured media.

In order to ensure authenticity, the proposed provenance verification mechanism extracts a range of diverse metadata forensic evidence features. Such features include resolution of the video, frame rate, codec, duration, encoding level, C2PA authenticity certificate, and watermark. The architectural design and flowchart of such verification mechanism within the international framework are shown in Figure 2.

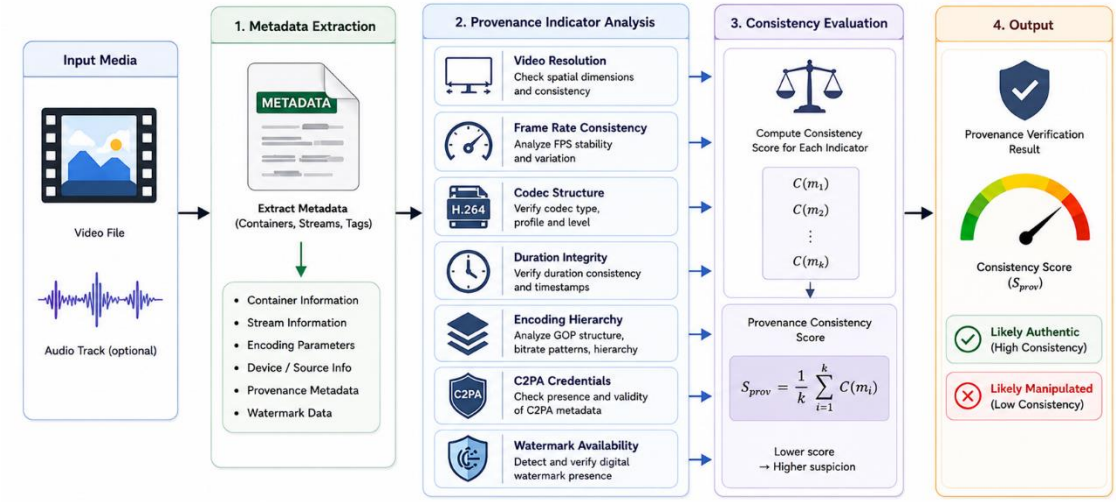


Figure 2: Provenance and watermarking verification pipeline

True video material generally exhibits consistency within its own acquisition characteristics in accordance with actual camera specifications and normal encoding practices. On the other hand, tampered video material frequently displays aberrations in encoding characteristics due to multiple compression cycles, manipulation through editing programs, neural renderings, artificial frame creation, or the intentional manipulation of metadata. Let: $M = \{m_1, m_2, ..., m_k\}$ represent the extracted metadata attribute set, where:

- m_i denotes the i^{th} metadata attribute,
- k corresponds to the total number of extracted forensic metadata features.

The provenance consistency score is estimated as:

$$S_{\text{prov}} = \frac{1}{k}\sum_{i=1}^{k} C(m_i)$$

where:

- $C(m_i)$ denotes the consistency evaluation function associated with metadata attribute m_i
- S_{prov} represents the overall provenance consistency score.

The consistency evaluation function measures the degree of similarity between observed metadata features and the expected properties of legitimate media acquisition chains. Therefore, the higher the inconsistency scores, the more likely it is that the content is fabricated or manipulated. In addition to the verification of structural metadata information, the suggested methodology includes watermark authenticity and provenance credential validation checks. With the advent of contemporary digital provenance standards, including those developed by the Coalition for Content Provenance and Authenticity (C2PA), cryptographically verified metadata signatures that allow tracing media provenance and edits have become available. Hence, the absence or inconsistency of provenance credentials constitute reliable signs of content manipulation.

Apart from encoding integrity, frame-wise compression consistency checks are carried out by means of estimating quantization irregularities:

$$Q_{err} = \frac{1}{N}\sum_{i=1}^{N} |q_i - \hat{q}_i|$$

where:

- $q_i$denotes the observed quantization coefficient,
- $\hat{q}_i$ represents the expected encoding coefficient,
- N corresponds to the total number of analyzed coefficients.

Abnormal quantization distributions frequently indicate multiple compression stages or recompression artifacts introduced during synthetic media generation and editing processes.

The provenance verification module provides interpretable forensic evidence regarding media origin, authenticity, and encoding integrity prior to deeper spatial and temporal analysis stages. Unlike purely appearance-based detection systems, provenance analysis enables the framework to identify suspicious manipulation traces even when visual artifacts become difficult to observe directly.

2.2 Audio Forgery Analysis

The analysis of audio is another vital forensic component of the suggested framework since the modern state-of-the-art systems generate extremely realistic synthesized voices that can imitate the natural human voice with remarkable accuracy. Although there have been significant advances in neural speech synthesis technologies, synthesized human voice often includes imperfections, such as spectral inconsistencies, vocoder artifacts, temporal inconsistencies, and distorted frequencies. Natural speech signals generated via acoustic processes usually retain all physiologic features that are not necessarily reproduced by synthesized speech generators that emphasize perceptual realism over physiologic realism. Thus, unnatural features of synthesized audio may include smooth spectra, lower-frequency variation, and irregular harmonics and timing transitions.

In order to identify acoustic discrepancies, the aforementioned auditory forensic model assesses a comprehensive range of acoustic features that complement one another. These acoustic features include spectral entropy, variation in the zero-crossing rate, the silence ratio, mel-spectrum consistency, high frequency energy distribution, and speech duration continuity. The design and processing flow of the aforementioned audio forensic system is illustrated in Figure 3.

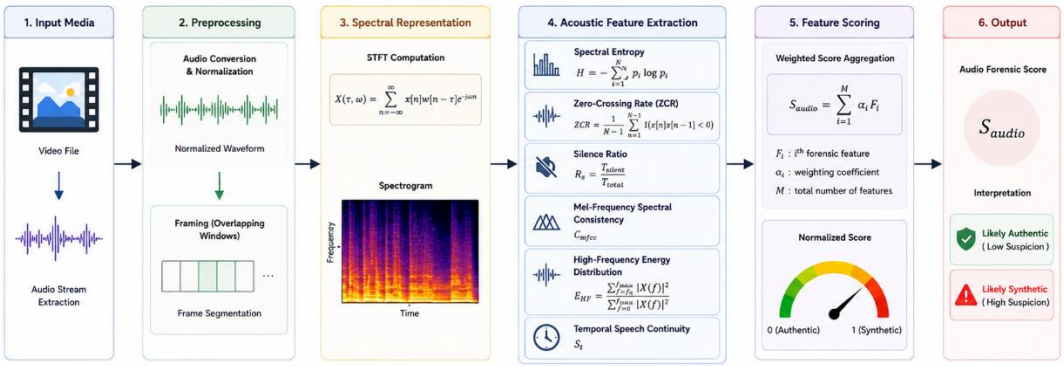


Figure 3: Audio forensic analysis pipeline

Initially, the audio stream is separated from the input media and converted into a normalized waveform representation. The signal is subsequently divided into overlapping temporal windows for frame-level acoustic analysis. Let: x[n] represent the discrete audio signal sampled from the input media. To analyze spectral properties, the Short-Time Fourier Transform (STFT) is applied:

$$X(\tau, \omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-\tau]e^{-j\omega n}$$

where:

- x[n] denotes the audio waveform,
- w[n] represents the temporal analysis window,
- τ corresponds to the temporal frame index,
- ω denotes the frequency component.

The generated spectrogram representation enables detailed examination of spectral irregularities frequently introduced during neural speech synthesis. To evaluate spectral complexity and randomness, spectral entropy is estimated as:

$$H = -\sum_{i=1}^{N} p_i \log p_i$$

where:

- p_i represents the normalized spectral probability distribution,
- N corresponds to the total number of spectral bins.

Authentic human speech generally exhibits naturally varying spectral entropy distributions due to physiological vocal tract dynamics and environmental acoustic conditions. In contrast, synthetic speech often demonstrates over-smoothed spectral structures resulting in reduced entropy variability. The proposed framework additionally computes the Zero-Crossing Rate (ZCR) to measure temporal signal oscillation characteristics:

$$ZCR = \frac{1}{N-1}\sum_{n=1}^{N-1} oI(x[n]x[n-1] < 0)$$

where:

- I(·) denotes the indicator function,
- N represents the number of audio samples.

Irregular ZCR fluctuations frequently indicate vocoder instability or synthesis artifacts associated with generated speech signals.To further evaluate high-frequency preservation, the spectral energy ratio is estimated:

$$E_{HF} = \frac{\sum_{f=f_h}^{f_{max}} |X(f)|^2}{\sum_{f=0}^{f_{max}} |X(f)|^2}$$

where:

- f_h denotes the high-frequency threshold,
- f_{max} corresponds to the maximum observable frequency component.

Synthetic speech generation pipelines frequently suppress high-frequency components due to neural smoothing operations and vocoder reconstruction limitations. Consequently, abnormal high-frequency energy distributions may indicate manipulated audio generation.

The final audio forensic score is computed by aggregating all extracted acoustic indicators:

$$S_{audio} = \sum_{i=1}^{M} \alpha_i F_i$$

where:

- F_i denotes the i[th] extracted forensic feature,
- α_i represents the corresponding weighting coefficient,
- M corresponds to the total number of audio forensic indicators.

The audio analysis module provides interpretable acoustic evidence regarding suspicious speech generation characteristics and contributes complementary forensic information to the overall multi-domain decision framework.

2.3 Facial and Spatial Inconsistency Analysis

Facial and spatial analysis is among the main forensic modules of the proposed framework due to the fact that manipulated visual media often include minor inconsistencies that deviate from the natural geometry of the facial features. Even though current deep learning methods can generate images with visually similar appearances, they are unable to reproduce fine spatial relationships, facial symmetry, and geometric consistency of facial features.

In contrast to true facial image creation that is created via the process of camera-based physical capture, facial parts synthesis pipelines employed by artificial media generation are prone to independent recreation of facial segments using latent space transformation, neural interpolation, and face-swap. This leads to the appearance of facial structure inconsistencies like facial asymmetry, improper landmark geometry, blurry boundaries, unusual texture of the face, smooth jawline, and blending artifacts. The facial forensic method detects such inconsistencies by evaluating geometric landmarks, facial symmetry, and facial structure edges.

Initially, facial regions are extracted from individual video frames using RetinaFace-based face localization. The detected facial region is subsequently normalized and forwarded to the landmark extraction module. Let:

$F_i \in R^{H \times W \times 3}$

represent the extracted facial frame, where:

- H and W denote spatial dimensions,
- F_i corresponds to the i[th] facial frame.

Facial landmark localization is performed using MediaPipe Face Mesh, which extracts dense facial keypoints describing geometric facial structure:

$$L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

where:

- L denotes the facial landmark set,
- n represents the total number of detected landmarks.

To evaluate geometric symmetry, the proposed framework computes average bilateral landmark deviation relative to the facial center axis

$$A_{sym} = \frac{1}{N}\sum_{i=1}^{N} \left| d_i^L - d_i^R \right|$$

where:

- d_i^L and d_i^R denote corresponding left and right landmark distances,
- N corresponds to the total number of symmetric landmark pairs.

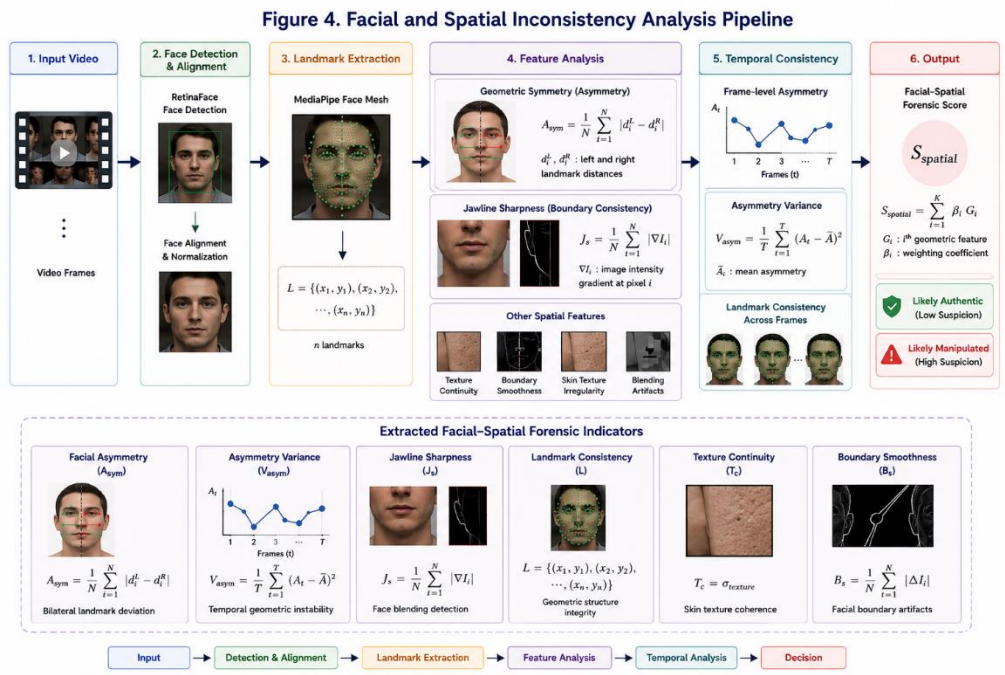


Figure 4: Facial and spatial inconsistency analysis pipeline

Authentic facial structures generally demonstrate naturally balanced geometric relationships, whereas manipulated content frequently introduces asymmetrical distortions caused by imperfect neural blending or latent-space misalignment. To further evaluate boundary consistency, jawline sharpness analysis is performed using spatial gradient estimation:

$$J_s = \frac{1}{N}\sum_{i=1}^{N} \left| \nabla I_i \right|$$

where:

- ∇I_i denotes the image intensity gradient at pixel i,
- J_s represents the estimated jawline sharpness score.

Edge softening techniques in face swapping and neural fusion operations often smoothen facial edges to avoid any visual artifacts that might be created by the process, hence making the edges less sharp at the facial edges. Moreover, geometric variability for frame asymmetry is determined to assess consistency in time sequences as follows:

$$V_{asym} = \frac{1}{T}\sum_{t=1}^{T}(A_t - \overline{A})^2$$

where:

- A_t denotes frame-level asymmetry estimation,
- $\overline{A}$ represents the temporal mean asymmetry value,
- T corresponds to the total number of analyzed frames.

Manipulated videos frequently demonstrate unstable geometric relationships across neighboring frames due to temporal inconsistency during synthesis generation. The final facial-spatial forensic score is computed through weighted aggregation of all extracted geometric indicators:

$$S_{spatial} = \sum_{i=1}^{K}\beta_i G_i$$

where:

- G_i denotes the i^{th} geometric forensic feature,
- β_i represents the corresponding weighting coefficient,
- K corresponds to the total number of spatial forensic indicators.

The facial and spatial analysis module provides interpretable geometric evidence regarding suspicious facial inconsistencies and contributes one of the strongest discriminative components within the proposed multi-domain forensic framework.

2.4 Temporal and Motion Consistency Analysis

Temporal consistency evaluation plays an important role in forensic media authentication, since videos with tampering often contain abnormal motion dynamics and inter-frame inconsistencies, which differ from real human motion dynamics. While current generative models greatly improve the static visual realism of videos, consistent motion between frames continues to pose a significant problem during synthetic video generation. Real videos show physically constrained motions caused by human movement laws, camera movements, lighting conditions, and temporal interactions among objects.

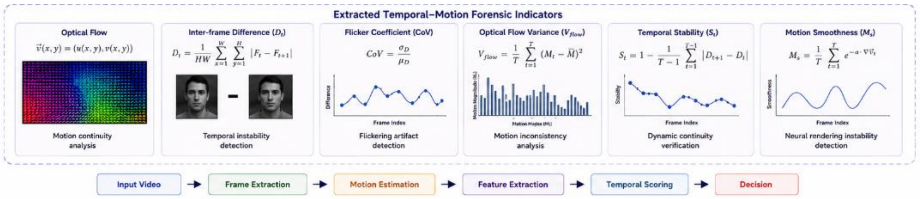


Figure 5: Temporal and motion consistency analysis pipeline

Conversely, tampered videos tend to display some specific temporal weaknesses, such as frame flickering, facial jitter, inconsistent motion flow, non-linear motion changes, temporal smoothing errors, and chaotic frame differences. This may be due to single-frame synthesis, poor motion estimation, unstable neural rendering, and temporal misalignment issues. The motion forensic system evaluates the mentioned discrepancies through motion consistency evaluation, optical flow estimation, inter-frame inconsistency measurement, and temporal consistency

evaluation. The execution process of the temporal verification approach is shown in Figure 5.Initially, consecutive video frames are extracted and normalized for temporal analysis. Let:

F_t, F_{t+1} represent two consecutive video frames. To estimate motion continuity between adjacent frames, dense optical flow analysis is performed:

$$v(x,y) = \big(u(x,y), v(x,y)\big)$$

where:

- u(x,y) and v(x,y) denote horizontal and vertical motion components,
- v(x,y) represents the optical flow vector at pixel location (x,y).

Inter-frame temporal variation is computed as:

$$D_t = \frac{1}{HW}\sum_{x=1}^{W}\sum_{y=1}^{H}|F_t(x,y) - F_{t+1}(x,y)|$$

where:

- H and W denote frame dimensions,
- D_t represents the frame-difference score.

The proposed framework additionally evaluates optical flow variance across temporal sequences:

$$V_{flow} = \frac{1}{T}\sum_{t=1}^{T}(M_t - \bar{M})^2$$

where:

- M_t denotes motion magnitude estimation at frame t,
- $\bar{M}$ represents the average motion magnitude,
- T corresponds to the total number of analyzed frames.

Videos containing synthetic manipulations generally exhibit unstable optical flow distributions and irregular temporal behavior compared to authentic videos with naturally smooth motion continuity. The final temporal forensic score is computed through weighted aggregation of temporal indicators:

$$S_{temporal} = \sum_{i=1}^{P}\gamma_i T_i$$

where:

- T_i denotes the i^{th} temporal forensic feature,
- γ_i represents the weighting coefficient,
- P corresponds to the total number of temporal indicators.

The temporal-motion module contributes interpretable dynamic evidence regarding abnormal motion behavior and frame-level inconsistencies within manipulated media.

2.5 Frequency-Domain Forensic Analysis

Frequency Domain Analysis is another critical forensic element of the framework due to the presence of hard-to-notice spectral anomalies that are created by modern generative systems. While the manipulated videos can seem highly plausible from a visual perspective for human perception, their synthetic pipelines can generate anomalous distributions of frequencies, periodic noise, and inconsistent spectra caused by upsampling processes, GANs, diffusion models, and re-compressions.

While unlike the actual videos obtained using physical imaging devices, the synthesis process of video forgery necessarily leads to changes to the original statistical distribution of spatial frequencies. Generally, natural images contain frequency properties physically consistent with their optical acquisition, sensor noise, illumination differences, and the continuity of texture properties. However, the frequency inconsistencies found in manipulated images include unusual high-frequency filtering, periodicity in spectra, repetitive texture patterns, GAN fingerprints, unauthentic Fourier spectra distribution, and wavelet behavior. The reason for such inconsistencies in the frequency domain lies in the fact that generative models pay attention to the visual authenticity of an image in the spatial domain but fail to accurately reproduce its frequency properties. The described frequency-domain analysis module examines these inconsistencies using Fourier spectrum analysis, frequency energy estimation, wavelet decomposition, and frequency consistency measures. The operation process of this frequency domain analysis pipeline is illustrated in Figure 6.

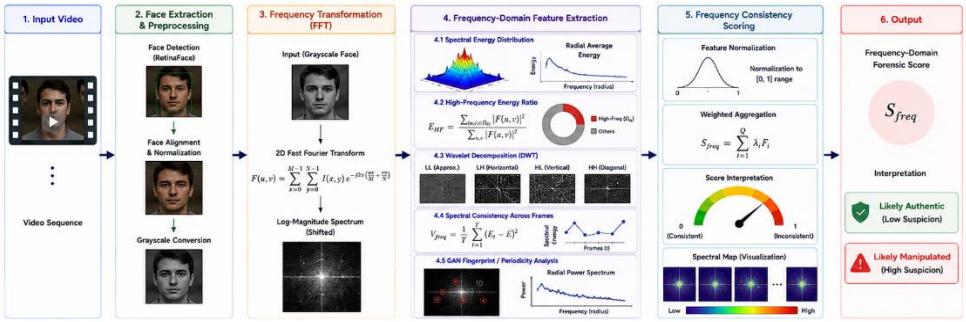


Figure 6: Frequency-domain forensic analysis pipeline

Initially, facial regions extracted from video frames are converted into grayscale frequency representations prior to spectral analysis. Let: I(x,y) represent the spatial-domain image intensity function. To analyze spectral characteristics, the two-dimensional Fast Fourier Transform (FFT) is applied:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x,y) e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}$$

where:

- I(x,y) denotes image intensity values,
- F(u,v) represents the frequency-domain spectrum,
- M and N correspond to image dimensions.

Authentic images generally demonstrate naturally distributed spectral energy patterns, whereas manipulated images frequently contain abnormal periodic frequency concentrations caused by neural generation artifacts. To evaluate spectral energy consistency, the proposed framework computes high-frequency energy distribution:

$$E_{HF} = \frac{\sum_{(u,v)\in\Omega_H} |F(u,v)|^2}{\sum_{u,v} |F(u,v)|^2}$$

where:

- Ω_H denotes the high-frequency spectral region,
- E_{HF} represents the high-frequency energy ratio.

Synthetic media frequently suppresses high-frequency details due to smoothing effects introduced during neural rendering and image reconstruction processes. Consequently, abnormal

high-frequency energy distributions may indicate manipulated content generation. To further analyze localized spectral irregularities, discrete wavelet decomposition is additionally performed:

$$W(a,b) = \frac{1}{\sqrt{a}} \int x(t) \psi \left(\frac{t-b}{a} \right) dt$$

where:

- a denotes the scaling parameter,
- b represents the translation parameter,
- $\psi(\cdot)$ corresponds to the mother wavelet function.

Wavelet analysis enables localized examination of texture irregularities and multi-scale spectral inconsistencies commonly introduced during synthetic image generation. The proposed framework additionally evaluates spectral variance across neighboring frames to measure temporal frequency stability:

$$V_{freq} = \frac{1}{T} \sum_{t=1}^{T} (E_t - \overline{E})^2$$

where:

- E_t denotes frame-level spectral energy estimation,
- $\overline{E}$ represents the average spectral energy value,
- T corresponds to the number of analyzed frames.

Manipulated videos frequently exhibit unstable spectral distributions across temporal sequences due to frame-wise generation inconsistencies and recompression artifacts. The final frequency-domain forensic score is computed through weighted aggregation of spectral indicators:

$$S_{freq} = \sum_{i=1}^{Q} \lambda_i F_i$$

where:

- F_i denotes the i^{th} frequency-domain forensic feature,
- λ_i represents the weighting coefficient,
- Q corresponds to the total number of spectral indicators.

The frequency-domain analysis module provides interpretable spectral evidence regarding hidden synthesis artifacts and contributes complementary forensic information that may remain undetectable in purely spatial or temporal analysis frameworks.

2.6 Multi-Domain Forensic Fusion and Decision Mechanism

The proposed framework perform localized forensic analysis on detected anomalies in the input media. Each component produces its own confidence score, which is then combined in the final stage of analysis to calculate the overall media manipulation probability. The proposed framework architecture is illustrated in Figure 7.

Firstly, the input multimedia content is preprocessed through a holistic process that includes extraction of video frames, audio extraction, localization of the face region, meta-data extraction, and region of interest (ROI) detection. Thereafter, the extracted features are simultaneously fed to several forensic analysis modules working independently in parallel.

The modular approach allows for the simultaneous evaluation of suspicious behavior from multiple forensic perspectives, thus not relying on one type of feature only. For each forensic module, a suspiciousness score is estimated as:

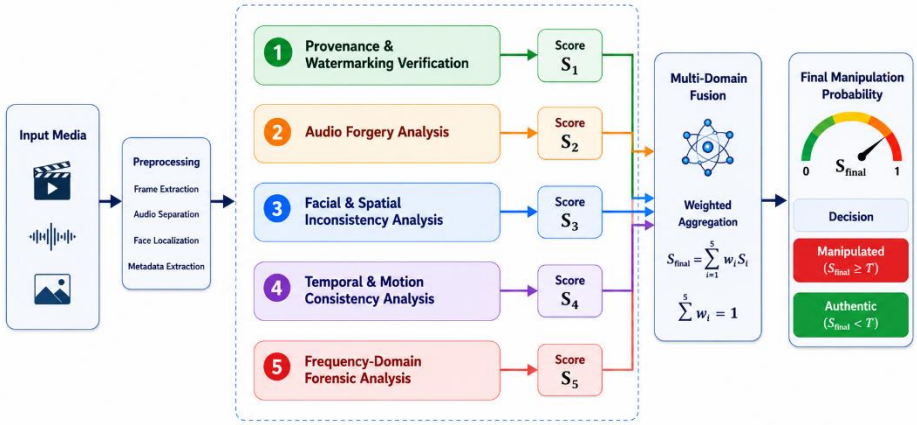


Figure 7: Proposed multi-domain forensic analysis framework for synthetic media detection.

$$S_i = f_i(X)$$

where:

- S_i denotes the forensic score produced by the i^{th} module
- $f_i$represents the corresponding forensic analysis function,
- X denotes the input media sample.

The individual forensic scores are aggregated using weighted multi-domain fusion:

$$S_{final} = \sum_{i=1}^{N} w_i S_i$$

where:

- S_{final} represents the final manipulation score,
- $w_i$denotes the weighting coefficient assigned to each forensic domain,
- N corresponds to the total number of forensic modules.

The weighting coefficients satisfy the normalization constraint:

$$\sum_{i=1}^{N} w_i = 1$$

In light of the fact that the framework proposed framework is mostly concerned with visual forensic analysis, greater weightage was accorded to facial-spatial and temporal-frequency analysis modules, while provenance and audio analyses provided complementary indicators for forensics. The weights assigned through experimentation are as follows:

Forensic Module	**Weight Coefficient**
Provenance & Watermarking Verification	(w_1 = 0.10)
Audio Forgery Analysis	(w_2 = 0.15)
Facial & Spatial Inconsistency Analysis	(w_3 = 0.30)
Temporal & Motion Consistency Analysis	(w_4 = 0.25)
Frequency-Domain Forensic Analysis	(w_5 = 0.20)

Table 2. Weight coefficients assigned to individual forensic analysis modules

This high contribution to facial-spatial analysis and temporal analysis is attributed to the high discriminating ability of geometry and motion distortions that are often introduced into

synthetic images. Additionally, frequency domain analysis provides useful spectral evidence pertaining to synthetic media creation by means of GANs and diffusion. On the other hand, provenance analysis and audio analysis constitute secondary evidence based on metadata and sound. Finally, the classification score is calculated using a threshold-based score interpretation method:

$$y = \begin{cases} Manipulated, S_{final} \geq T \\ Authentic, S_{final} < T \end{cases}$$

where:

- T denotes the experimentally selected decision threshold,
- y represents the final classification label.

Unlike standard black-box architectures, the proposed method will retain meaningful forensic traces that are independently produced by each module. As such, this method does more than merely predict the presence of media manipulation; it also identifies the forensic features whose importance was greatest in making this prediction.

3. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

The proposed multi-domain forensic framework was empirically tested on 455 videos, which were sampled from the Celeb-DF and FaceForensics++ (FF++) datasets. The sampling process was conducted such that the test set includes real videos, as well as videos that have been manipulated through different synthesis methods, facial expressions, illumination conditions, compression schemes, and motion events. The performance of the proposed framework was evaluated using standard classification metrics including accuracy, precision, recall, and F1-score. The classification accuracy is computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

while precision and recall are defined as:

$$Precision = \frac{TP}{TP + FP} \; Recall = \frac{TP}{TP + FN}$$

The harmonic mean between precision and recall is calculated using the F1-score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where: TP denotes true positives, TN represents true negatives, FP corresponds to false positives, FN denotes false negatives.

Table 3 summarizes the comparative performance of individual forensic modules and the proposed multi-domain fusion framework.

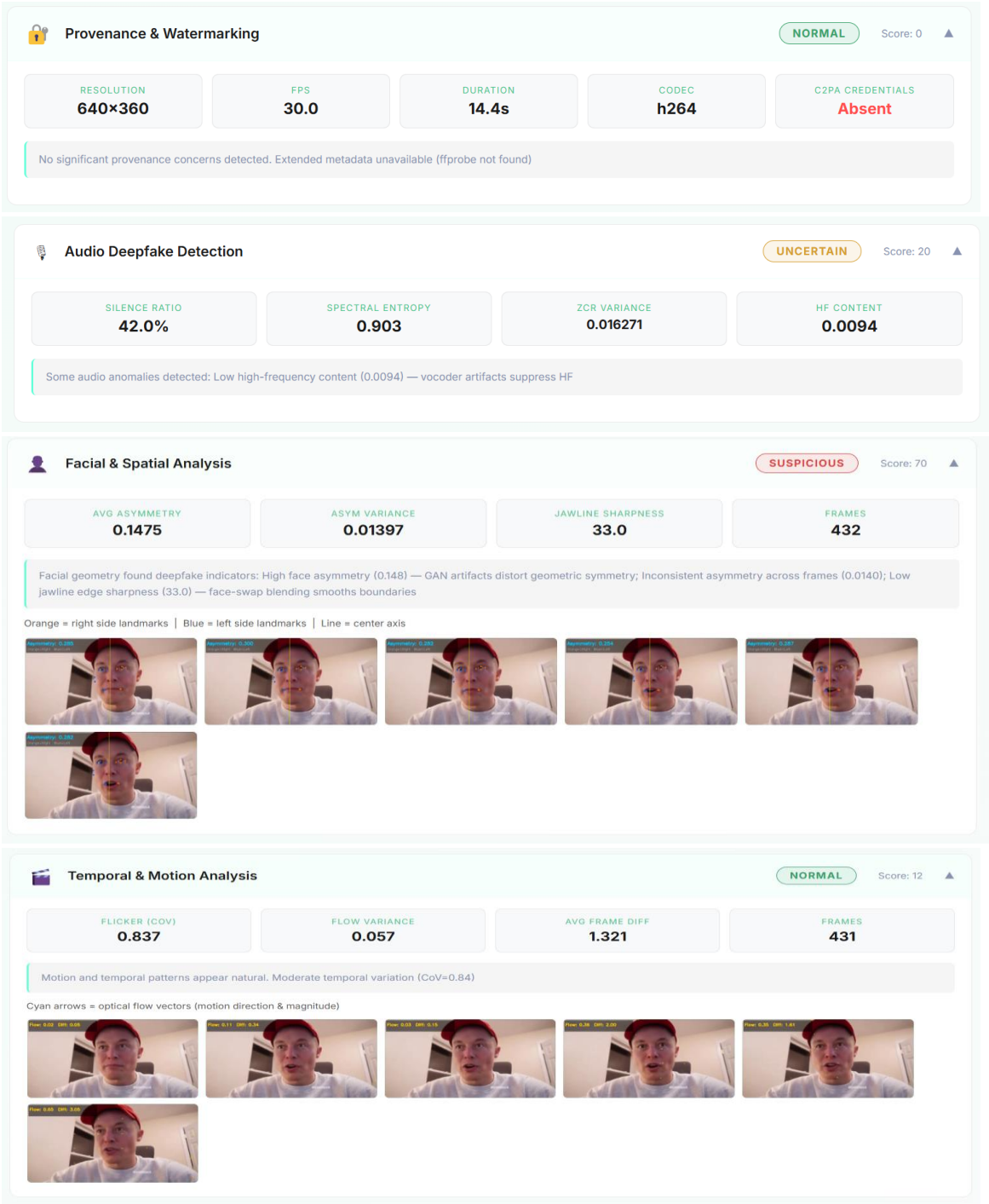
`	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Provenance & Watermarking	81.3%	80.7%	82.1%	81.4%
Audio Forensic Analysis	84.6%	83.9%	85.2%	84.5%
Facial & Spatial Analysis	89.7%	89.1%	90.4%	89.7%
Temporal & Motion Analysis	88.4%	87.8%	89.1%	88.3%
Frequency-Domain Analysis	90.3%	89.8%	90.9%	90.2%
Proposed Multi-Domain Fusion	91.5%	91.0%	92.1%	91.4%

Table 3: Comparative performance of individual forensic modules and the proposed multi-domain fusion framework.

From the experimental results, the multi-domain forensic fusion scheme proposed by us clearly outperforms independent forensic schemes. Forensic schemes based on the combination of

different forensic evidences show their strength in defending manipulated videos from visual realism. As for single forensic module, frequency domain forensic scheme was able to provide the best performance, as it is capable of detecting invisible artifacts caused by neural network-based manipulation process. Facial-spatial forensic scheme also performed excellently because manipulated videos usually have irregular landmarks and asymmetric faces.

In addition to frequency and spatial domain forensic schemes, there were many pieces of forensic evidence generated from temporal domain forensic scheme as it detects the flickering effect and the inconsistency pattern among consecutive frames. Figure 8 presents example forensic outputs generated from each analysis module within the proposed framework.



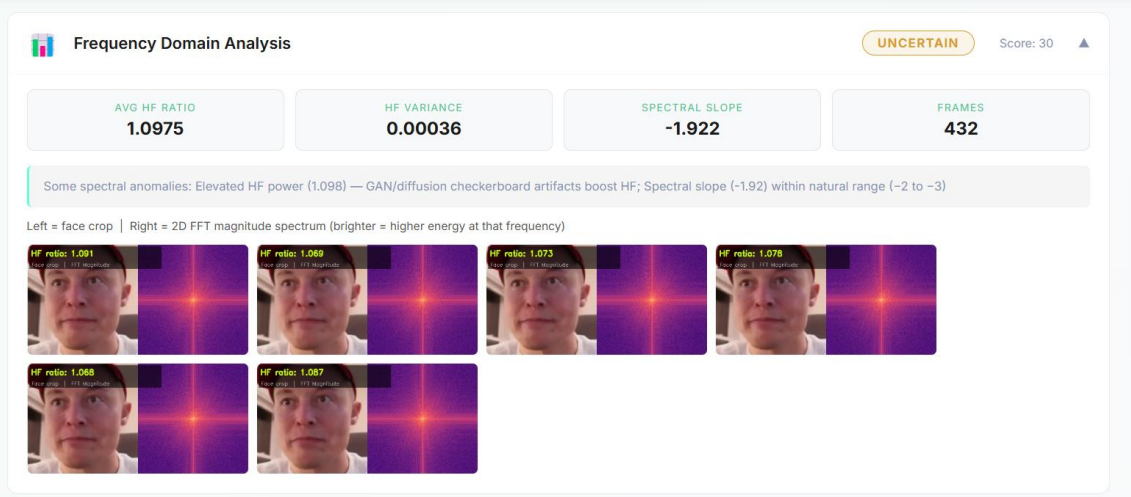


Figure 8: Forensic outputs generated from each analysis module within the proposed framework.

As can be seen from the forensic outputs of manipulation, the fact is that the manipulated videos have been consistently found to show inconsistencies in several forensic domains. The abnormalities associated with frequency domain analysis include the inconsistent spectrum and the presence of periodic patterns, while in temporal domain analysis, the video shows instabilities in motion flow. Furthermore, the facial-spatial analysis shows inconsistency as well.

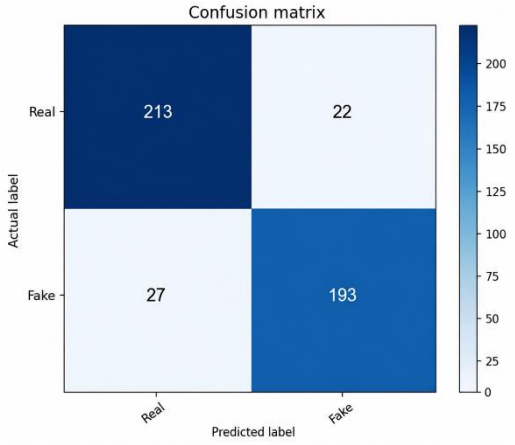


Figure 9: Confusion matrix obtained using the proposed multi-domain forensic fusion framework.

The confusion matrix clearly shows that the proposed model is able to effectively classify most of the genuine as well as manipulated examples with reasonably low rates of false positives and false negatives. The use of heterogeneous forensic features significantly diminishes reliance upon the individual artifact type.

In conclusion, the results obtained during the experiments clearly indicate that the combination of the five above-mentioned components is highly effective for contemporary media manipulation detection.

CONCLUSION

This study has introduced a multi-domain forensic framework towards detection of manipulated media based on the CNN-LSTM-Spline detection model. While traditional frameworks focus on using spatial deep learning features for media manipulation detection, the presented framework employs a combination of provenance and watermark verification, audio forensic analysis, facial-spatial inconsistency, temporal-motion inconsistency, and frequency-

domain forensic techniques in one weighted fusion architecture. The presented multi-domain forensic framework is aimed at detecting the presence of suspicious anomalies associated with the creation process of synthetic media, such as inconsistent encoding structure, frequency-domain abnormalities, geometric distortions, motion anomalies, and temporal inconsistencies.

The experiment, performed with a sample of 455 videos taken from Celeb-DF and FaceForensics++ data sets, revealed a detection accuracy rate of 91.5%. In addition, the proposed framework showed high resilience to realistic media manipulation. As the experimental results show, the combination of different forensic indications significantly enhances detection performance as well as makes the forensic process more interpretable compared to individual forensic modules. Overall, the proposed framework proves itself to be a powerful and interpretable technique for modern media manipulation detection.

REFERENCES

1. M. S. Rana, M. N. Islam, S. M. Shatabda, and A. H. M. R. Imon, “Deepfake detection: A systematic literature review,” *IEEE Access*, vol. 10, pp. 25494–25513, 2022.
2. D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in *Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1–6.
3. E. Altuncu, V. N. L. Franqueira, and S. Li, “Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review,” *Frontiers in Big Data*, vol. 7, p. 1400024, 2024.
4. A. Diel et al., “Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers,” *Computers in Human Behavior Reports*, vol. 16, p. 100538, 2024.
5. S. Maxmudjanov, A. Primbetov, and A. Alimbaeva, “Human perception versus artificial intelligence in deepfake video detection: An empirical study,” *Al-Farg’oniy Avlodlari*, vol. 1, no. 4, pp. 83–87, 2025.
6. S. Maxmudjanov, A. Primbetov, and A. Naimov, “Deepfake detection using a hybrid ResNeXt and LSTM architecture,” *Al-Farg’oniy Avlodlari*, vol. 1, no. 2, pp. 87–94, 2025.
7. J. Wu, K. Feng, X. Chang, and T. Yang, “A forensic method for deepfake image based on face recognition,” in *Proc. 4th High Performance Computing and Cluster Technologies Conf. and 3rd Int. Conf. Big Data and Artificial Intelligence*, 2020, pp. 104–108.
8. S. Sohail, S. M. Sajjad, A. Zafar, Z. Iqbal, Z. Muhammad, and M. Kazim, “Deepfake detection using deep learning: A unified forensic approach to detect AI-generated images and videos with fusion of eye, nose, and mouth landmarks,” *Computer Science and Mathematics*, Feb. 2025.
9. S. H. Silva, M. Bethany, A. M. Votto, I. H. Scarff, N. Beebe, and P. Najafirad, “Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models,” *Forensic Science International: Synergy*, vol. 4, p. 100217, 2022.
10. G. O. Y. L. F. Lundström-Imanov and N. Abdullayeva, “Verifiable provenance and watermarking for generative AI: An evidentiary framework for international operational law and domestic courts,” *arXiv preprint arXiv:2605.21002*, 2026.
11. F. Haugg, M. Elgendi, and C. Menon, “GRGB rPPG: An efficient low-complexity remote photoplethysmography-based algorithm for heart rate estimation,” *Bioengineering*, vol. 10, no. 2, p. 243, 2023.
12. Q. He et al., “GazeForensics: DeepFake detection via gaze-guided spatial inconsistency learning,” *Neural Networks*, vol. 180, p. 106636, 2024.

13. T. Jung et al., “DeepVision: Deepfakes detection using human eye blinking pattern,” *IEEE Access*, vol. 8, pp. 83144–83154, 2020.
14. J. Cech and T. Soukupova, “Real-time eye blink detection using facial landmarks,” *Center for Machine Perception, Czech Technical University in Prague*, pp. 1–8, 2016.
15. Google Research, “MediaPipe Face Mesh,” 2023. [Online]. Available: https://mediapipe.dev/
16. A. Susi, V. Akila, and V. Govindasamy, “Utilizing rPPG signal synchronization and deep learning techniques for deepfake video detection,” *IEEE Access*, 2025.