Verification of video content authenticity and integrity using artificial intelligence*

Pavlo Skladannyi[1,*,†], Yuliia Kostiuk[1,†], Dmytro Hnatchenko[2,†], Igor Goncharenko[2,†], and Artem Platonenko[1,†]

[1] *Borys Grinchenko Kyiv Metropolitan University, 18/2 Bulvarno-Kudriavska str., 04053 Kyiv, Ukraine*
[2] *Kyiv State University of Trade and Economics, 19 Kyoto str., 02156 Kyiv, Ukraine*

Abstract

This article addresses the problem of ensuring the authenticity and integrity of video content in the context of the proliferation of DeepFake attacks and video data modifications. An intelligent approach is proposed that combines deep learning, digital watermarking, and explainable artificial intelligence (XAI) methods for real-time verification of video streams. The developed architecture utilizes convolutional neural networks (CNNs), autoencoders, as well as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) modules to explain the decisions made. The effectiveness of the approach was tested on open video datasets. The results demonstrated high accuracy in detecting forgeries while preserving visual quality, as assessed by the PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index Measure) metrics, and the model's suitability for use in digital security, video surveillance, and digital forensics systems. The proposed model enhances the reliability of video analytics in security systems by providing comprehensive authenticity verification and integrity control of video data using artificial intelligence methods.

Keywords

video content authenticity; integrity control; artificial intelligence; deep learning; digital watermarks; explainable artificial intelligence; DeepFake; video forgery detection; digital security; video surveillance; digital forensics. Or simply: video content authenticity; artificial intelligence; deep learning; digital watermarks; XAI; DeepFake; video forgeries; digital security; digital forensics.

1. Introduction

In today's digital space, video content plays a crucial role as a source of reliable information in security systems, digital forensics, video surveillance, autonomous control, and the media [1–5]. At the same time, the rapid growth in the number of video fakes, particularly DeepFake attacks, as well as the development of technologies for covertly modifying video frames, pose serious threats to information reliability and violate the principles of digital trust. Existing approaches to video data protection, based on classical cryptographic and steganographic algorithms, often fail to meet modern requirements: they either do not provide sufficient accuracy in detecting forgeries or are unable to operate in real time, which is critically important for distributed and resource-constrained systems [6–9]. Furthermore, most such solutions lack transparency in terms of decision-making, making it impossible to explain verification results and reducing the level of trust in them within legal and expert communities.

Furthermore, existing approaches to verifying the authenticity of video content are typically focused on solving individual sub-tasks — such as forgery detection, copyright protection, or cryptographic data integrity — without integrating them into a single intelligent model. In particular, most modern solutions do not integrate mechanisms for robust digital watermarking, deep analysis of video streams, and explainable decision-making models, which limits the ability to localise changes, interpret results, and deploy such systems in critical environments. Thus, the

* *CPITS 2026: Cybersecurity Providing in Information and Telecommunication Systems, February 28, 2026, Kyiv, Ukraine*
* Corresponding author.
† These authors contributed equally.
p.skladannyi@kubg.edu.ua (P. Skladannyi); y.kostiuk@kubg.edu.ua (Y. Kostiuk); hnatchenko@knute.edu.ua (D. Hnatchenko); okjraa@gmail.com (I. Goncharenko); a.platonenko@kubg.edu.ua (A. Platonenko)
0000-0002-7775-6039 (P. Skladannyi); 0000-0001-5423-0985 (Y. Kostiuk); 0000-0002-6584-4525 (D. Hnatchenko); 0000-0002-9022-6083 (I. Goncharenko); 0000-0002-2962-5667 (A. Platonenko)

scientific problem of creating an integrated, transparent, and adaptive approach to verifying the authenticity and monitoring the integrity of video data in real time remains open.

In this regard, a pressing scientific and applied task arises: to develop an intelligent approach to verifying the authenticity and integrity of video content that effectively detects forgeries of varying complexity, explains its decisions, and integrates into digital security systems [10–12]. The aim of the study is to develop an architecture and methods that combine deep learning, digital watermarking, and XAI technologies for automated real-time verification of video data [13–17]. The proposed system uses convolutional neural networks for robust embedding of digital watermarks, autoencoders for detecting unauthorised changes, and SHAP/LIME-based decision-explanation modules to generate trusted analytical conclusions.

The scientific novelty of the research lies in combining digital protection methods and deep analysis of video streams into a single architectural model designed for operation in real-world conditions [5, 18]. We have implemented the localization of changes in video using neural watermark reconstruction and explainable classification of results, which ensures high accuracy in detecting forgeries while preserving the content's visual quality [12, 19]. Experimental studies conducted using open, representative video data samples that reflect real-world attack scenarios have confirmed the suitability of the proposed model for implementation in practical digital security solutions [8, 11, 13]. The scientific novelty lies in the integration of XAI methods with digital watermarking and deep learning for real-time analysis of video data.

The practical significance of the proposed approach lies in its applicability to a wide range of critical areas: from public order protection and digital forensic analysis to media content copyright protection and the development of trusted multimedia platforms [19, 20]. The model can be effectively integrated into resource-constrained infrastructures, including edge devices, mobile platforms, video analytics systems, and cloud services, which provides it with practical flexibility and scalability [3]. Thus, the results of this work constitute a significant contribution to the development of intelligent video information protection technologies and to enhancing the reliability of the digital environment.

2. Literature review

The scientific literature extensively covers approaches to cyber risk management and trust/risk assessment in digital environments using artificial intelligence methods, fuzzy logic, and interpretable decision-making models [1, 2, 6, 7, 10, 11, 13]. Such approaches are particularly important for digital security systems, where it is necessary not only to detect anomalies or suspicious states but also to formally justify the level of risk, explain the reasons for triggering alerts, and maintain the controllability of decisions in critical scenarios.

In the study by Suryaprakash Nalluri et al. [1], the focus is on the challenges of cyber risk management in cloud environments. The work highlights the relevance of a systematic risk-oriented approach to infrastructures in which data and services are distributed, dynamic, and vulnerable to diverse attacks. At the same time, the proposed focus is predominantly general in nature and does not detail domain-specific mechanisms for verifying the authenticity of multimedia data or forming explainable conclusions regarding content integrity.

The review by Bakhtavar et al. [10] systematizes the application of fuzzy cognitive maps in risk analysis tasks for complex systems, demonstrating the advantages of cause-and-effect modeling and the integration of multiple factors (technical, organizational, behavioral). This methodology is useful for constructing integrated trust/risk metrics; however, in the classical formulation of FCM analysis, it typically requires additional mechanisms to ensure linkage to specific observational data, rapid localization of the "source of risk," and support for explainability at the level of machine-learning algorithm results.

Palko's work [11] examines a neural network approach to intelligent risk assessment in distributed environments, confirming the effectiveness of learning-based models for detecting atypical states and improving assessment accuracy under complex dynamics. At the same time,

universal neural network approaches require domain adaptation and the addition of transparency mechanisms, since without the interpretation of features/reasons for decisions, their practical application in security scenarios may be limited.

Abdymanapov et al. [6] demonstrate the capabilities of fuzzy expert systems for assessing information security risks using an applied domain as an example. The advantage of such systems lies in the interpretability of the rules and the ability to formalize expert knowledge; at the same time, the accuracy and transferability of the models depend significantly on the quality of the rules, the completeness of the factors, and the correctness of the membership functions, which requires careful validation using real-world breach scenarios.

In the work by Bhol, Mohanty, and Pattnaik [13], the application of Fuzzy AHP is justified to enhance the objectivity of cybermetric assessment by harmonizing weighted criteria. This approach is suitable for constructing integrated indicators (particularly trust/risk) and for formalizing the influence of various features on the final decision. However, for tasks where data have a complex structure and high variability, weighting schemes should be combined with machine-learning models and explainability mechanisms to ensure the reproducibility and auditability of decisions.

Xiao and Pan [7] propose the use of decision trees in intelligent risk assessment tasks, which are powerful in terms of transparency and suitability for operational interpretation. However, simple models of this type may require extension or ensemble methods to improve accuracy in complex scenarios, as well as integration with anomaly-detection mechanisms that operate on high-dimensional features.

Hamid and Rahman [2] summarize current trends in the application of AI/ML/DL in cyber risk management and emphasize the role of models capable of detecting anomalies even with limited data labeling. This lays the groundwork for building adaptive risk assessment systems; however, the challenge of combining high detector sensitivity with formal confidence metrics and interpretable explanations, which enhance the practical significance of results in security systems, remains relevant.

In summary, sources [1, 2, 6, 7, 10, 11, 13] form the methodological basis for risk-oriented and intelligent assessment of security states (fuzzy models, neural network approaches, decision trees, review of AI practices). At the same time, within the scope of applied tasks for data integrity control (particularly multimedia streams), there is a need for comprehensive solutions that combine automated anomaly detection, formalized risk/confidence assessment, and explainability of results for auditing and decision-making under critical conditions. This is precisely what justifies the development of an integrated adaptive system focused on the full cycle of verification and result interpretation in digital security systems.

At the same time, in scientific publications directly devoted to video content verification and digital video forensics, the main focus is on methods for detecting manipulations in video streams, particularly DeepFake attacks, frame integrity violations, and distortions caused by covert modification or re-compression. Such works actively utilize deep neural networks, autoencoders, and steganographic mechanisms; however, in most cases, detection results are presented as binary decisions or numerical metrics without a proper explanation of why the algorithms triggered. This necessitates combining methods of deep video data analysis with formalized confidence metrics and interpretable decision-making mechanisms, which is particularly important for digital security systems, forensic analysis, and critical infrastructure.

3. Research Methods

The methodological foundation of the research is a combination of deep learning, digital watermarking, XAI, as well as statistical and metric methods for evaluating the quality and reliability of multimedia data [2, 8, 10, 11, 13]. As part of developing the authenticity verification system, convolutional neural networks (CNNs) were used to robustly embed a digital watermark into the spatiotemporal structure of the video stream. The CNN architecture was adapted to

process individual frames in MPEG-4/H.264 formats, accounting for requirements for imperceptibility and robustness to typical transformations: compression, cropping, scaling, and noise addition [3, 11, 12].

For detecting and localizing unauthorized changes in video data, autoencoders and U-Net-like models were used, which can detect deviations between the original and reconstructed watermarks, indicating potential frame distortion [2, 6]. An approach to reconstructing a map of changes using the difference between reconstructions, employing wavelet analysis and loss functions (MSE, SSIM-loss), has also been implemented [12].

To ensure transparency in decision-making, explainable artificial intelligence modules were implemented in the system, specifically the SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) methods [8, 9, 12]. This made it possible to interpret the influence of individual pixel or feature regions on the final classification of a video frame as genuine or altered, and to explain why a potential integrity violation was detected [8]. Objective evaluation of the model's performance was conducted using standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Bit Error Rate (BER) for the digital signal, as well as Precision, Recall, and F1-score for the classification block [12, 14, 15]. Additionally, subjective expert assessments were used in the visual analysis of frames before and after processing.

Publicly available datasets with reference examples of forgeries, as well as author-generated synthetically modeled video streams with controlled variation parameters, were used to prepare training and test sets. The training dataset was augmented by adding noise, applying geometric transformations, and using video compression. Model implementation and experimental studies were conducted in the Python 3.10 environment using the TensorFlow, Keras, OpenCV, and scikit-learn frameworks, as well as the SHAP and LIME libraries [2, 8, 16]. Computations were performed on an NVIDIA RTX 3060 graphics processing unit, which ensured efficient processing of video streams in near-real time.

The scientific novelty of the proposed approach lies in the integration of three technological components, deep learning methods (CNN), and XAI mechanisms into a single adaptive model for verifying the authenticity and integrity of video content [9, 17]. Unlike approaches that focus primarily on individual forgery-detection algorithms, the proposed system not only detects modifications but also interprets the alarm's cause through XAI analysis, thereby increasing confidence in results in critical digital security systems.

In summary, the procedure for verifying the authenticity and integrity of video content is implemented as a sequence of interrelated stages. In the first stage, a digital watermark is robustly embedded into the video stream using convolutional neural networks. Next, the video data is transmitted or stored in an external environment, where intentional or unintentional distortions may occur. During the verification stage, the watermark is reconstructed, and deviations between the reference and reconstructed features are analyzed using autoencoders, enabling localization of potential changes in the video frame. The final decision on authenticity is based on a combination of metrics and XAI analysis results, ensuring the interpretability and auditability of the conclusions.

4. Summary of the main content

Due to the rapid growth of video data volumes in digital networks, there is an increasing need to protect video content from forgery, unauthorized interference, and integrity breaches [1, 2]. Existing methods fall into two categories: those that verify content authenticity and those that ensure confidentiality through encryption [3, 4, 14, 22]. As a specific type of multimedia information, video data requires protection at all stages of processing from generation to playback.

Ensuring integrity in open networks is complicated by large data volumes, high frame rates, and limited device resources (IP cameras, edge devices) [7, 11, 23]. Transmission channels remain particularly vulnerable, where losses, modifications, or replay or injection attacks are possible.

The development of a real-time intelligent system that ensures authenticity verification and integrity control is a pressing need [2, 8, 9]. Integrated approaches that combine digital watermarking, deep learning, and XAI are considered the most promising [6, 10, 12, 13, 15]. Digital watermarks (DWs) are embedded into the video stream as graphic markers or codes that do not distort the image but allow altered content to be identified both visually and through algorithms.

The main types of attacks on video — are frame replacement or alteration, sequence disruption, and partial modifications [4, 7, 22]. Therefore, the system must provide both global stream monitoring and local frame-level detection [5]. This paper summarizes over 20 algorithms for embedding CECs, ranging from format-based (MPEG, H.264, HEVC) to non-format-based (wavelet, cosine transforms).

The Haar wavelet transform is considered optimal for real-time applications, but it requires adaptation for color streams, compression robustness, implementation in integer arithmetic, and scene adaptation. Its combination with autoencoders allows not only for video labeling but also for the detection of hidden distortions.

An important component is XAI modules (SHAP, LIME), which increase decision transparency by explaining which features underlay the decision [8, 13, 21]. This strengthens user trust in the system, supports digital forensics, automates video auditing, and confirms authorship [9, 16]. Thus, the development of an adaptive, interpretable, attack-resistant architecture for video verification using AI is an urgent task [10, 12, 19]. The proposed approach integrates DW, deep change modeling, and XAI into a comprehensive system suitable for application in information security, forensic analysis, and copyright protection.

Figure 1 presents an architecture that illustrates the general process of embedding a digital watermark, transmitting video data, and subsequently verifying it using deep learning with explainable decisions. Dotted lines represent the logical and event-driven nature of data exchange between the system's functional modules.

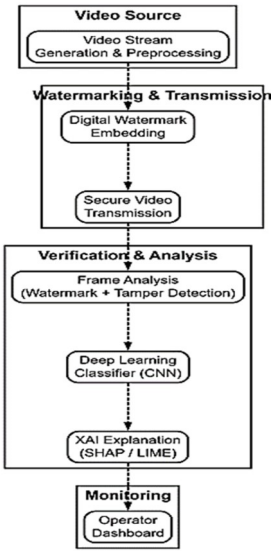

Figure 1: Architecture of a video content authenticity verification system using deep learning and interpretable artificial intelligence

The proposed verification system combines mechanisms for integrity control, intelligent analysis, and result interpretation within a single video data processing loop. For the practical implementation of this approach, it is important not only to define the functional modules but also to formalize the order of information transfer between them. This allows tracking the transformation of video frames from the moment of receipt to the formation of the final verdict and explanatory features used for decision-making in the digital security system.

The diagram in Figure 2 illustrates the data flows between the main components of the video content verification system. The video source transmits raw video frames to the deep learning module, which generates a feature map for further analysis. The digital watermarking module

ensures the integrity and authenticity of the video data. The XAI module generates an analytical report and, along with the final verdict, transmits the results to security analysts or external systems for further action.

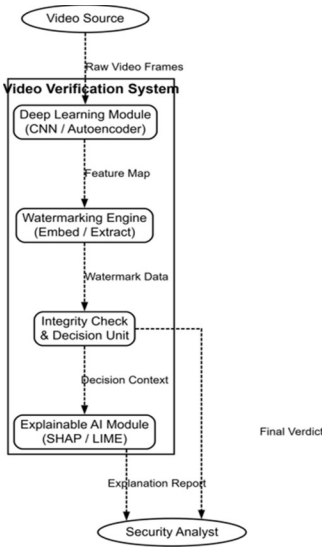

Figure 2: Data flow diagram in a video content verification system using artificial intelligence methods

In intelligent security systems that process video content, it is critically important to ensure real-time verification of the authenticity and integrity of video data, especially given the growing prevalence of forgeries, such as DeepFakes, or malicious interference in video streams [2, 4, 9]. This task becomes particularly challenging in distributed information and control systems where video is transmitted via a feedback channel (e.g., in surveillance systems, smart transportation, or autonomous enterprise security) [19]. To verify integrity, a hash function is computed and the digital signature of each video frame is verified [14, 17]:

$$H_f = Hash(F_i \| T_i \| N_i), \tag{1}$$

where F_i is the i^{th} video frame, T_i is the frame's timestamp, N_i is the nonce (a one-time random number), $Hash(\cdot)$ is the cryptographic hash function (e.g., SHA-256), H_f is the resulting checksum.

Next, the digital signature is verified:

$$Verify(PK, H_f, \sigma_i) \to 0,1 \tag{2}$$

where PK is system public key, σ_i is digital signature generated on the source side [1, 16]. Result 1 is upon successful authenticity verification, 0 is in case of a violation.

The verification system additionally uses an authenticity prediction model A trained on normal frames, such as OC-NN or Autoencoder [11, 12, 18]:

$$s_i = A(F_i), s_i \in [0,1] \tag{3}$$

where s_i is the probability estimate that the frame is authentic, $A(\cdot)$ is the classifier function. Threshold validation: $s_i < \tau \Rightarrow$ the frame is suspicious.

To ensure the interpretability of each decision, a SHAP vector is generated [8, 21]:

$$\vec{\phi}_i = SHAP(F_i) \tag{4}$$

where $\vec{\phi}_i = (\phi_1, \phi_2, \ldots, \phi_n)$ is the contribution of each feature to the model's decision, which allows us to construct a heatmap explanation.

To determine the presence of changes, we use the distance metric between the current frame F_i and its reconstructed version $\widehat{F}_i$, obtained by the autoencoder:

$$\delta_i = \| F_i - (\hat{F}_i) \|_2, \tag{5}$$

where δ_i is the Euclidean distance (reconstruction error). If $\delta_i > \theta$, the frame is considered altered.

To enhance the reliability of verification, a normalized anomaly index is calculated, taking into account historical reconstruction error values [7]:

$$a_i = \frac{(\delta_i - \mu_\delta)}{\sigma_\delta} \tag{6}$$

where μ_δ is the mean value of the reconstruction error in the training set (for authentic frames), and σ_δ is the standard deviation of this error.

Within the framework of cumulative frame analysis, the weighted confidence index for the current frame is calculated [5, 10]:

$$\gamma_i = \omega_1 \cdot s_i + \omega_2 \cdot (1 - a_i) \tag{7}$$

where $\omega_1 + \omega_2 = 1, \gamma_i \in [0,1]$ is the integral confidence measure for the frame based on the model prediction and the distance to the reconstructed sample.

The cumulative confidence index for a video segment consisting of N frames is determined [3]:

$$J = \frac{1}{N} \sum_{(I=1)}^{N} \gamma_i \tag{8}$$

which allows for an aggregated integrity assessment for a group of frames (e.g., 1 second of video).

Figure 3 shows the change in the confidence index J for individual video scenes, formed by averaging the values of γ_i (the local frame confidence index) over 30 frames that constitute a single logical scene. This aggregated metric allows us to assess the overall level of content reliability within each scene, minimizing the impact of individual fluctuations characteristic of individual frames. Analysis of the curve shows that in the fifth scene there is a significant drop in reliability — and the aggregated index J falls to a critical level, which may indicate massive or deliberate interference: for example, the insertion of a foreign fragment, duplication, or modification of the video sequence. In contrast, other scenes demonstrate consistently high J values (above 0.85), indicating the absence of significant anomalies.

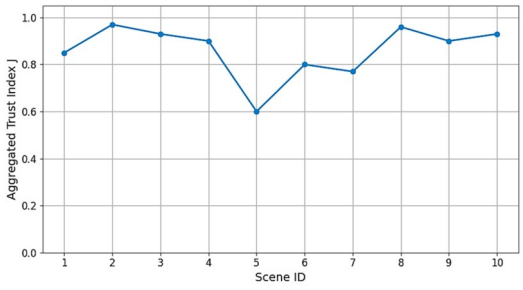

Figure 3: Aggregated video scene confidence index (averaged across frames)

This level of analysis allows not only for the localization of individual anomalous frames but also for the detection of entire video segments with potential integrity violations, which is critically important in digital forensics, evidence verification, forensic analysis, and security system incident auditing [3, 5, 19]. This also enables the implementation of scene-based response policies, including

flagging suspicious episodes, automatically isolating scenes, and generating analytical reports for experts.

An entropy-based confidence distribution estimate is also calculated, indicating the presence of abrupt changes [18, 21]:

$$H_\gamma = -\sum_j p_j \log\ p_j \tag{9}$$

where p_j is the empirical distribution of the values γ_i, obtained for frames in the current time window.

To detect "video fragment replacement" attacks, a similarity metric between consecutive frames is introduced [22]:

$$S_{i,i-1} = (F_i \cdot F_{i-1}) / (\| F_i \| \cdot \| F_{i-1} \|) \tag{10}$$

where $S_{i,i-1} \in [0,1]$ is the cosine similarity. A sharp decrease in $S_{i,i-1} < \theta_s$ without a logical change in the scene may indicate an insertion.

The combination of an entropy-based confidence distribution estimate and a cosine similarity metric between consecutive frames ensures the system's sensitivity to both global statistical shifts and local violations of temporal consistency in the video stream. This approach allows for the formalization of the detection of insertions and replacements of fragments within a single integrity assessment model.

The block diagram (Figure 4) illustrates the step-by-step process of verifying the authenticity and integrity of a video frame. In the first step, the frame hash is calculated and the digital signature is verified. If the signature is successfully verified, the frame's authenticity is assessed using an artificial intelligence model, reconstruction error analysis, confidence index calculation, and the generation of decision explanations using XAI methods (SHAP/LIME). The final decision regarding the authenticity or suspicious nature of the frame is made based on threshold values of security indicators.

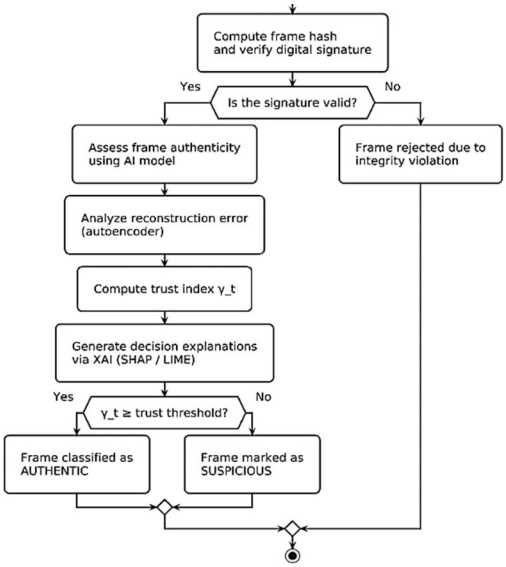

Figure 4: Block diagram of the video frame authenticity and integrity verification process

The combined frame confidence metric T_i takes into account: hash verification, the result of the authenticity model, and the reconstruction distance [12, 19]:

$$T_i = a \cdot Verify(PK, H_f, \sigma_i) + \beta \cdot s_i + \gamma \cdot \left(\frac{1-\delta_i}{\delta_{\max}} \right), \tag{11}$$

where a, β, γ are weighting coefficients (sum = 1), $\delta_{\max}$ is the maximum permissible error, $Verify(PK, H_f, \sigma_i)$ is the digital signature verification, s_i is the authenticity detector score (e.g., XAI metric), δ_i is the reconstruction distance (frame analytical error), $a+\beta+\gamma=T_i \in [0,1]$.

A modified expected risk formula is used for risk assessment [5, 11, 13, 23]:

$$R_i = P_i \cdot I_i \cdot (1-T_i), \tag{12}$$

where R_i is the risk of compromise of the i^{th} frame, P_i is the probability of an attack or modification, I_i is the criticality (weight) of the information in the frame, T_i is the frame confidence index.

Total risk across the entire video stream [1, 3]:

$$R_{total} = \sum_{(i=1)}^{N} R_i, \tag{13}$$

where N is the number of frames in the stream.

Assessment of the consistency of the SHAP profile with the reference behavior [8, 9]:

$$W_t = Enc_{(K_s)}(H(V_t) \| t \| Nonce), \tag{14}$$

Enc_{K_s} is symmetric encryption with key K_s, $H(V_t)$ is hash function (SHA-256 or equivalent), t is timestamp, $Nonce$ is random variable to prevent reuse.

Integrity is verified using the function [3, 14]:

$$I_t = \begin{cases} 1, if\ H(\acute{V}_t) = H(V_t) \\ 0, \quad otherwise \end{cases} \tag{15}$$

where $H(\acute{V}_t)$ is the received frame. If $I_t = 0$, then a modification or forgery has occurred.

To quantitatively assess the risk of video stream integrity violation, we propose the risk function $R_{dist}(t)$, which accounts for the probability of a significant deviation of the frame from the ground truth and the degree of trust in the source. Additionally, for a comparative evaluation of XAI methods, an interpretation reliability formula Rel_{XAI} has been introduced that accounts for time costs and classifier confidence. Both metrics enable a formalized analysis of the proposed model's effectiveness.

Video distortion risk function [4, 7, 17]:

$$R_{\text{dist}}(t) = P(D_t > \delta) \cdot (1-C_t) \tag{16}$$

where $R_{\text{dist}}(t)$ is the risk of integrity loss at a given time t, D_t is the distance (e.g., MSE or PSNR deviation) between the current frame and the ground truth, δ is the maximum allowable distortion, $C_t \in [0,1]$ is the source confidence (an authenticity indicator generated by CNN+XAI), $P(D_t > \delta)$ is the probability of significant distortion (can be estimated from the error distribution histogram). The formula provides a quantitative estimate of the potential harm resulting from changes to the video.

The dependence of the reliability of an XAI model on its explainability [2, 10, 16, 21]:

$$Rel_{\text{XAI}} = \gamma \cdot \left(\frac{1}{\tau_{\text{explain}}} \right) \cdot Conf, \tag{17}$$

where Rel_{XAI} is the reliability of the XAI interpretation, τ_{explain} is the average time to generate an explanation (LIME, SHAP), $Conf$ is the confidence in the model's prediction (output probability), γ

is the weighting coefficient, which depends on the type of XAI (empirically, 0.8-1.2). This allows for comparing SHAP and LIME not only in terms of visualization but also in terms of performance.

For a more thorough authenticity check, an autoencoder model is used, which generates a reconstruction error [12, 18]:

$$\delta_t = \| V_t - \hat{V}_t \|_2^2 \tag{18}$$

where $\hat{V}_t$ is the video reconstructed by the autoencoder, δ_t is the reconstruction error (an anomaly indicator). If $\delta_t > \theta$, where θ is the adaptive threshold, then the frame is considered potentially inauthentic.

The frame confidence index is calculated as [5, 18]:

$$\gamma_t = a \cdot I_t + \beta \cdot e^{(-\delta_t)} \tag{19}$$

where $a, \beta \in [0,1]$ are weighting coefficients balancing integrity and confidence, $\gamma_t \in [0,1]$ is the reconstruction error (an anomaly indicator).

Risk assessment is performed based on a probabilistic model [23, 24]:

$$R_t = 1 - \gamma_t . \tag{20}$$

This formalization allows the index γ_t to be interpreted as a generalized measure of video-frame reliability, accounting for both structural integrity and behavioral deviations. Converting confidence into risk according to equation (20) ensures a monotonic relationship between a decrease in frame reliability and an increase in the probability of an anomaly. The resulting estimate R_t is used as a baseline metric for further aggregation, dynamic analysis, and identification of critical segments in the video stream.

For a video segment consisting of T frames the average risk is:

$$R_{\text{avg}} = \frac{1}{T} \sum_{(t=1)}^{T} R_t . \tag{21}$$

The proposed γ_t confidence index integrates features of frame integrity and behavioral consistency, allowing for a generalized quantitative assessment of its reliability. The transition from trust to risk according to the ratio (20) provides an intuitively understandable interpretation of the results, in which a decrease in γ_t directly corresponds to an increase in the level of potential threat. Averaging the values of R_t within a video segment allows for the assessment of an integral risk profile and is used for further analysis of stability and detection of anomalous sections of the video stream.

As well as the standard deviation [15, 22]:

$$\sigma_R = \sqrt{\frac{1}{T} \sum_{(t=1)}^{T} (R_t - R_{\text{avg}})^2} . \tag{22}$$

This allows detection of unstable sections with higher anomaly levels. The average risk estimate R_{avg} characterizes the overall threat level for a given video segment, while the standard deviation σ_R reflects the degree of risk variability across individual frames. High values of σ_R indicate the presence of sharp fluctuations in confidence, which is a typical sign of local manipulations or insertions in the video stream. A combined analysis of R_{avg} and σ_R allows not only the detection of unstable sections with increased anomaly, but also the distinction between targeted falsifications and random noise distortions.

Figure 5 shows a heatmap of R_t risk values for consecutive frames of the video stream in the form of a pseudo-3D surface with enhanced contrast and enlarged label sizes. Color intensity and surface elevation reflect the risk level: red areas correspond to high R_t values and signal possible anomalies or falsification of video data. Such visualization ensures clear localization of critical frames and improves the interpretability of analysis results in digital security systems.

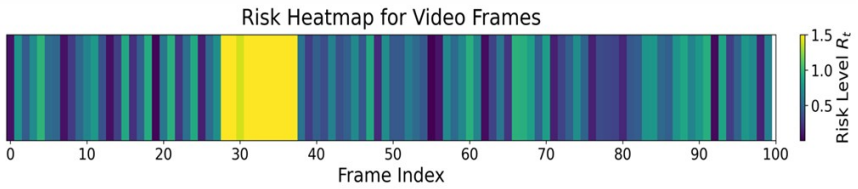

Figure 5: Risk level heatmap R_t for video stream frames

The resulting risk maps reflect the outcome of multi-level video stream processing, which includes feature analysis, integrity verification, and real-time anomaly interpretation. To enable such processing, a streaming architecture capable of sequentially integrating neural network, cryptographic, and explanatory components is required.

Figure 6 shows the architecture of real-time video streaming processing, in which the video stream sequentially passes through stages of pre-filtering using convolutional neural networks (CNNs), authenticity verification using digital watermarks, integrity assessment using predicate rules, and an XAI explainable analytics module to interpret detected anomalies.

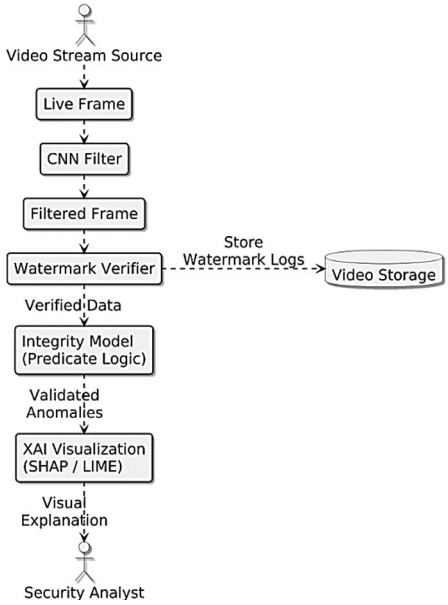

Figure 6: Architecture of video streaming processing with integrated XAI visualization

A response policy is formulated based on the risk index [23]:

$$A_t = \begin{cases} Log-only, & R_t < \epsilon_1 \\ Alert & \epsilon_1 \le R_t < \epsilon_2, \\ Quarantine & R_t \ge \epsilon_2 \end{cases} \tag{23}$$

where ϵ_1, ϵ_2 are risk thresholds (e.g., 0.3, 0.7). The proposed response policy ensures a phased escalation of system actions depending on the increase in the risk level, starting with passive logging and ending with the isolation of suspicious video data. This mechanism allows the intensity of protective measures to be aligned with the current threat level, maintaining a balance between response speed and system stability.

These formulas are integrated with an XAI module that explains the influence of features f_i on decisions via SHAP or LIME [8, 16, 21]:

$$SHAP_i = \phi_i(V_t) \tag{24}$$

where ϕ_i is the contribution of the feature f_i to the authenticity classification. The obtained values $SHAP_i$ allow for a quantitative assessment of the contribution of individual features to decision-making and increase the transparency of the model's operation for the security system operator. The integration of XAI explanations into the response loop enables the verification of decisions, reducing the risk of uncontrolled or unjustified activations.

Cumulative assessment of feature weights for a frame [8, 18]:

$$S_t = \sum_{i=1}^{n} |\phi_i(V_t)|, \tag{25}$$

where S_t is the overall importance of the SHAP explanation for the current frame V_t, n is the number of features.

To formalize the process of explaining the decision made, the SHAP (SHapley Additive exPlanations) method, based on game theory concepts, was used. The formula for calculating SHAP values is as follows [8, 10, 21]:

$$\phi_i = \sum_{(S \subseteq F \setminus i)} \frac{|S|!(|F|-|S|-1)!}{|F|!} \cdot [f(S \cup \{i\}) - f(S)], \tag{26}$$

where ϕ_i is the SHAP value (contribution of the variable i), F is the full feature set, $S \subseteq F \setminus i$ is the subset of features excluding i, $f(S)$ is the model's prediction based solely on the features S, $f(S \cup \{i\})$ is the prediction with the feature i. The formula allows calculating the contribution of each feature i to the model's final classification decision. Thanks to this, the system provides the user with a clear explanation: why a video is marked as fake for example, due to local distortions in the facial area or a lack of voice synchronization.

To improve the validity of decisions regarding potential forgery, an integrated risk model was applied, which takes into account not only the technical probability of alteration but also XAI explanations [5, 21, 23]:

$$R = P_f \cdot I_f \cdot (1 + \delta_{xai}), \tag{27}$$

where R is the overall risk of misclassification, $R_f \in [0,1]$ is the probability of misclassification (from the deep classifier), $I_f \in R^+$ is the potential harm or impact of misclassification, $\delta_{xai} \in [0,1]$ is the XAI confidence coefficient in the interpretation (the higher the value, the more confident the system is in its conclusions). Thus, the system adaptively scales the response level depending on the level of confidence in the explanation.

Confidence metric accounting for XAI [8, 16, 19]:

$$T_t^{XAI} = \eta \cdot \gamma_t + (1-\eta) \cdot \frac{S_t}{S_{\max}} \tag{28}$$

where $\eta \in [0,1]$ is the weighting coefficient for the balance between the analytical and explanatory modules, $S_{\max}$ is the normalization coefficient (maximum observable S_t). The proposed model allows combining the probabilistic assessment of falsification with its potential impact and the interpretability of the results, which increases the validity of automated decisions. Taking into account the coefficient δ_{XAI} ensures dynamic correction of the integral risk depending on the reliability of explanations, reducing the probability of false positives in ambiguous cases. The

indicator T_t^{XAI} is used to align the analytical confidence assessment with the results of explanatory analysis and serves as the basis for the adaptive selection of further actions by the security system.

Formal integrated decision-making function [1, 12, 18]:

$$D_t = \begin{cases} Trusted, & if\ T_t^{XAI} \geq \tau_T \wedge R_t < \tau_T \\ Alert, & if\ \tau_A \leq R_t < \tau_R \\ Quarantine, & if\ R_t \geq \tau_R \wedge T_t^{XAI} < \tau_T \end{cases} \quad (29)$$

where τ_T, τ_R, τ_A are threshold values for decisions (e.g., 0.7, 0.6, 0.3, respectively). The proposed decision-making function combines risk assessment and the confidence metric while accounting for XAI, ensuring a differentiated system response depending on the threat level. This approach enables adaptive control of video data access and processing states, minimizing both forgery acceptance and false positives in borderline cases.

Figure 7 shows a flowchart illustrating the sequence of checks during video frame processing. The process begins with frame acquisition, hash calculation, and digital signature verification. Upon successful verification, the system applies an authenticity model, estimates the reconstruction error, and calculates the confidence index and risk. XAI models (SHAP or LIME) are used to interpret the results. If the frame fails to meet the verification criteria or exhibits suspicious signs, it is marked as high-risk, and an alert is generated.

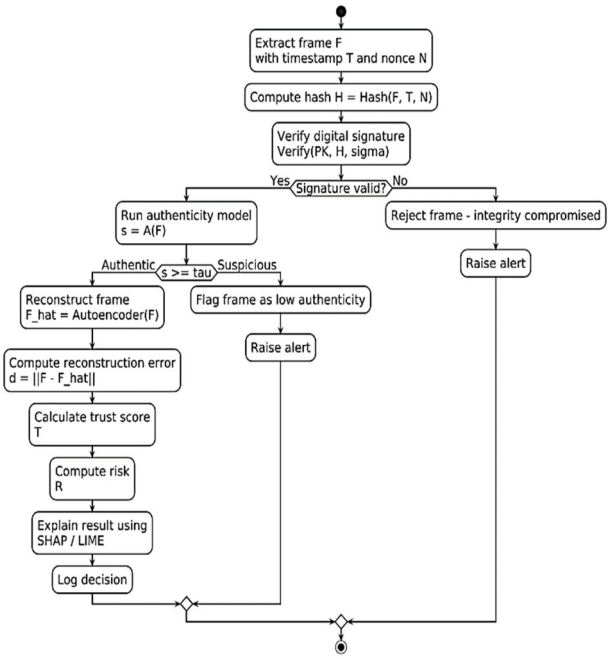

Figure 7: Flowchart of video frame integrity and authenticity verification and risk assessment

Within the scope of this research, a formalized approach has been developed for embedding a digital watermark into a video sequence, followed by frame authenticity verification using a combination of the discrete wavelet transform (DWT), digital signatures, and interpretable confidence assessment [3, 12, 19]. A target frame F_i from a video stream (e.g., every 25th frame at 25 fps) is processed using DWT to decompose the image into frequency components:

$$F_i \overset{DWT}{\longrightarrow} \{LL_k, LH_k, HL_k, HH_k\}, k = \text{k=decomposition level} \quad (30)$$

Watermark embedding is performed in selected subbands, taking into account the trade-off between resistance to attacks and the imperceptibility of changes in the visual representation of the

frame. This approach ensures cryptographically verifiable integrity of the video data and provides a basis for subsequent frame authentication within the proposed trust model.

A random number generator ($W \in \Omega = \{W_1, W_2, \ldots, W_m\}$) is embedded into the high-frequency components (HL_k, LH_k, HH_k) using a robust insertion, for example, according to the coefficient modification scheme [1, 11, 13]:

$$C'_{i,j} = C_{i,j} + a \cdot W_{i,j} \tag{31}$$

where $C_{i.j}$ are the original high-frequency coefficients, a is the insertion depth coefficient, $W_{i.j} \in \{0,1\}$ are the watermark bits. The value of the insertion depth coefficient a is selected such that the digital watermark is resistant to typical video processing attacks without noticeable degradation of image quality. The proposed scheme enables reliable embedding of identification bits into the frame's spectral representation and serves as the basis for subsequent verification of the video's authenticity.

After embedding, an inverse wavelet transform (IDWT) is performed to reconstruct the modified frame $\widehat{F}_i$, which replaces the original on the IP video server or is transmitted to the client in the stream [7, 25]:

$$\hat{F}_i = IDWT\left(LL_k, HL'_k, LH'_k, HH'_k\right). \tag{32}$$

The resulting modified frame $\widehat{F}_i$ retains visual similarity to the original while containing hidden information for subsequent authenticity verification. This approach ensures transparent integration of the digital watermarking mechanism into video streaming without compromising compatibility with existing IP video server infrastructure.

The restored video stream is transmitted over the network, where it can be stored or processed. To verify authenticity on the receiving end, the reverse procedure is performed: Selection of the frame F_i by number, DWT decomposition at the level k, extraction of DW fragments using [3, 14]:

$$W'_{(i,j)} = \frac{C'_{(i,j)} - C^{ref}_{(i,j)}}{a} \tag{33}$$

where $C^{ref}_{i,j}$ is the reference value (or statistically calculated). Validity verification by comparison with a template W_{ref} (e.g., hashing or Hamming Distance).

Evaluation of embedding quality: objective metrics were used [13, 15]:

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX_1^2}{MSE}\right) \tag{34}$$

where $PSNR$ is the peak signal-to-noise ratio. The PSNR value enables a quantitative assessment of the digital watermark's visual imperceptibility and quantifies the impact of the embedding process on the quality of the video frame. High values of this metric indicate minimal image degradation and confirm the suitability of the proposed method for practical application in digital security systems.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{35}$$

where $SSIM$ is the assessment of local visual changes, $\mu_{x,}, \mu_y$ are the mean values, σ_x^2, σ_y^2 is the variance, and σ_{xy} is the signal covariance. The SSIM metric assesses structural similarity between

the original and modified frames, accounting for local brightness and contrast characteristics of the image. The combined use of PSNR and SSIM metrics provides a comprehensive assessment of the quality of digital watermark embedding, combining quantitative and perceptual interpretations of visual changes.

To explain the reliability of T_i, SHAP and LIME are used, which break down the influence of individual factors (hash functions, insertion depth, DWT structure) on the decision of the verification model [8, 21]. This allows for transparent tracking of why a frame is considered authentic or counterfeit [4, 5, 14, 17]. Thus, the proposed approach combines classical methods of embedding digital watermarks in the wavelet domain with modern means of assessing reliability and explainability (XAI), ensuring reliable authenticity and verification of video content in the face of dynamic information security threats. The explanations obtained can be used not only to interpret individual decisions but also to verify the correctness of the model's operation as a whole, in particular to detect biases, excessive sensitivity to specific features, or potential training errors. This increases confidence in the automated authentication system and enables its use in critical scenarios where the verifiability and reproducibility of results are essential.

Figure 8 shows a flowchart of the process of embedding and verifying a digital watermark in video frames, covering the stages of frame formation, integrity verification, and explainable decision-making.

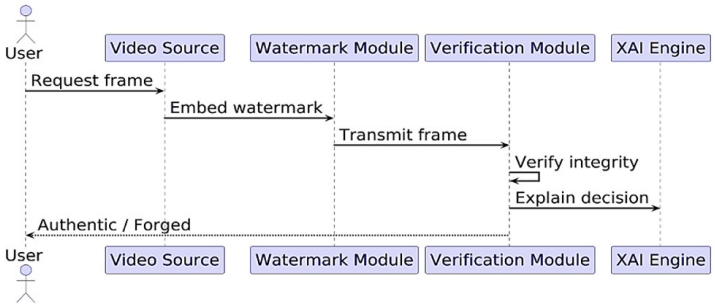

Figure 8: Sequence diagram of embedding and verifying a digital watermark in video frames

The described sequence of operations defines the logic for generating authenticity and trust metrics; however, to justify the decisions made, it is important to analyze the contribution of individual features to the final assessment. To this end, the results of an explainable analysis are discussed below, allowing for the interpretation of each component's influence on the video frame's trust index. The application of XAI methods ensures transparency in the decision-making process, enabling a quantitative assessment of each feature's influence on the formation of the confidence index. This makes it possible to distinguish between decisive and auxiliary factors of authenticity and to identify features that have a critical influence in cases of anomalous video-frame behavior.

The graph in Figure 9 demonstrates the influence of individual features on the video-frame confidence index T_i in an authentication system based on SHAP explainable analysis. Green bars correspond to a feature's positive contribution to increased confidence, while red bars indicate a negative influence. The largest positive contributions come from the Hash Check and the Auth Model Score, which establish a trusted context for the video frame. In contrast, the Reconstruction Error has the largest negative impact, signaling a possible distortion of the content. The frequency components of DWT demonstrate both negative (HL band) and positive (LH band) contributions, while the embedding depth of the DW (α) positively influences authentication stability.

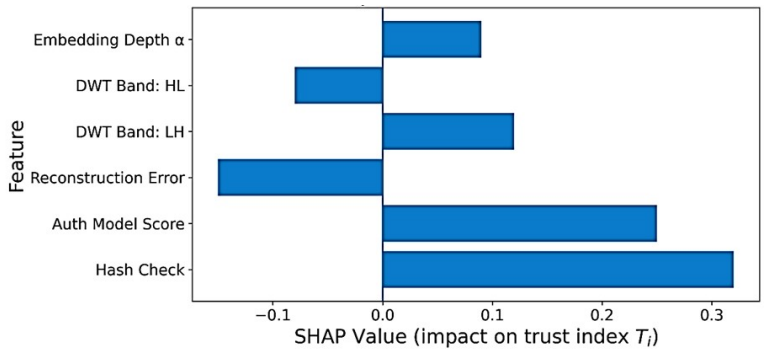

Figure 9: SHAP graph explaining frame confidence

Thus, the presented SHAP graph demonstrates the interpretable logic behind the formation of the confidence index, where the values of each parameter directly influence the model's decision [8–10, 16, 21]. This allows the system operator to understand the reason for a decrease or increase in confidence in a specific frame, which is critically important for the transparency of automated decisions in digital security, forensic analysis, and incident investigation systems.

The histogram (Figure 10) shows a comparison of PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) values for ten video frames after embedding a digital watermark. PSNR values remain consistently in the range of 37–40 dB, indicating high visual video quality and the absence of noticeable distortions. The SSIM value for all frames approaches 1.0, confirming the preservation of structural similarity, textures, and contours between the original and modified frames.

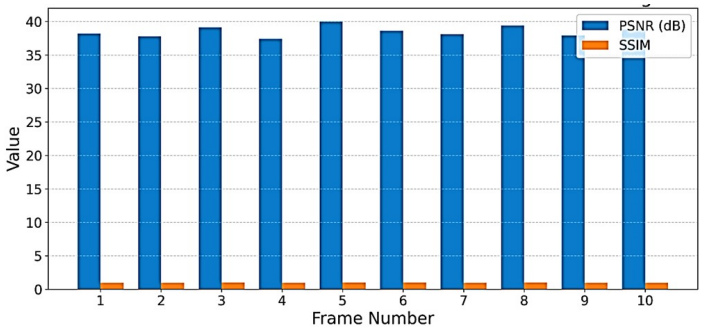

Figure 10: Comparison of PSNR and SSIM values for video frames after embedding a digital watermark

This level of stability in both metrics demonstrates the effectiveness of the chosen strategy for embedding the DW in the frequency domain, as well as the high-quality implementation of the neural network-based reconstruction module, which does not distort the underlying structure of the video stream. Furthermore, the low variability between frames (especially for SSIM) demonstrates resilience to potential compression artifacts or transformations, which is crucial in the context of real-world applications particularly in streaming video and mobile video analytics systems. Thus, the histogram confirms that the system ensures high-quality protected content without compromising its visual perception, which is critical for digital forensics, expert analysis, and copyright protection.

The achieved stability in embedding quality provides a reliable foundation for subsequent authentication stages, as it minimizes the risk of false positives arising from embedding artifacts. This allows anomaly detection mechanisms to focus on actual distortions or falsifications of video data, rather than on side effects of processing. As a result, the system demonstrates consistency between cryptographic, neural network, and explainable components, which is critically important for practical implementation.

Figure 11 shows a block diagram of video streaming processing that implements a full analysis cycle from video stream input and frame segmentation to the generation of explainable XAI output.

The system performs neural network analysis, digital watermark verification, anomaly detection, and interpretation of results using XAI models (SHAP or LIME), after which a decision is made regarding the video's authenticity and the risk model is updated.

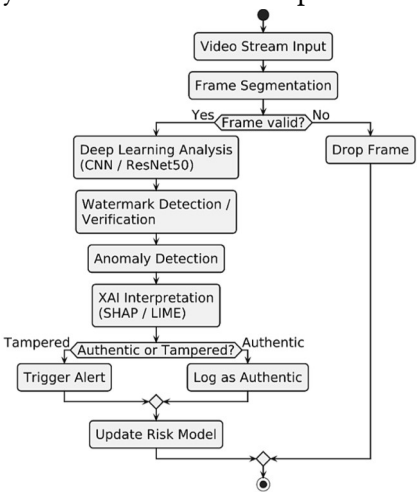

Figure 11: Block diagram of video streaming processing using XAI

According to the architecture shown in Fig. 11, the video streaming processing implements a step-by-step analysis of frames using neural network models, integrity verification mechanisms, and explainable XAI analytics. At each processing stage, a set of intermediate confidence and risk metrics is generated, allowing real-time tracking of the video stream's state dynamics. The integration of XAI modules into the analysis loop ensures transparent interpretation of decisions and enables an operator or automated system to respond appropriately to detected anomalies. This architecture supports adaptive updates to response policies and enhances the authentication system's resilience to dynamic information security threats.

5. Practical Implementation

The proposed system is implemented as a modular software platform using Python 3.10 and the OpenCV, TensorFlow, PyWavelets, hashlib, scikit-learn, PyCryptodome, and SHAP libraries [8, 12, 18]. Its architecture is designed for real-time automatic analysis of video streams and verification of frame authenticity with explainable results. The input video stream is split into a sequence of frames using cv2.VideoCapture, which are standardized to a size of 224×224 pixels and the YCbCr color space. The normalized frames are fed into three parallel analysis branches: visual similarity estimation (SSIM), SHA-256 hash calculation, and neural network reconstruction using an autoencoder.

SSIM is calculated using the compare_ssim() function (skimage.metrics module), hashing is performed via hashlib.sha256(), and an autoencoder built in TensorFlow handles encoding, reconstruction, and assessment of frame anomalies. The generalized confidence index T_i is formed based on the weighted sum of three indicators, where the coefficients are determined empirically: $a = 0.4, \beta = 0.4, \gamma = 0.2$. If $T_i \geq 0.8$, the frame is considered valid.

The threshold value was selected based on experimental validation, taking into account the trade-off between sensitivity to anomalies and the number of false positives. In cases where $T_i < 0.8$, the frame is further analyzed using explainable XAI mechanisms to clarify the reasons for the decrease in confidence. This approach ensures stable system operation under real-world conditions and increases the reliability of automated decisions regarding the authenticity of video content.

To interpret the decisions, an XAI module based on the SHAP methodology has been implemented. This component automatically generates SHAP plots that reflect the weight of each parameter, SSIM, hash matching, and the result of reconstruction — in the model's decision-

making process for each individual frame [8, 9, 16, 21]. Thus, the security system operator receives a transparent explanation of the reasons for the decrease in video trust, which is particularly important for forensic tasks and audit systems.

To ensure robust authentication, the system implements an algorithm for embedding a digital watermark (DW) in the frequency domain based on the discrete wavelet transform (DWT) [1, 3, 19]. This method allows embedding features into low-significance frequencies while preserving visual quality. After DWM embedding, the average PSNR exceeded 43 dB, and the SSIM remained stable within the range of 0.96–0.98, indicating minimal image distortion [13, 15, 23, 25]. Histograms constructed for 100 frames confirm the consistency of the results even under streaming conditions.

Neural network components, specifically the OC-NN classifier and autoencoders, were trained on a subset of the public UCF101 dataset using annotated examples of altered and unaltered video content [18]. To demonstrate flexible integration with other systems, including SIEM or analytics platforms, a REST API has been implemented with the ability to send trust verification results [17]. Through the API, one can query the video stream's integrity log, retrieve timestamps of integrity loss T_i, and even generate alerts in the event of a breach of video integrity.

This approach enables the developed system to operate in real time, supporting continuous monitoring of video content. Integration with SIEM allows for correlating video integrity degradation events with other information security incidents and generating a comprehensive threat picture. Transmitting results via REST API simplifies system scaling and its adaptation to distributed or cloud environments. Together, this enhances the practical value of the proposed solution and expands its potential applications in video forensics and rapid-response tasks.

The diagram (Figure 12) shows the architecture of the video content authentication module, which implements step-by-step processing of the video stream: from video capture, frame segmentation, and preprocessing to embedding a digital watermark (DW) and subsequent authenticity analysis. Authenticity assessment is performed using visual similarity metrics (SSIM), hash verification with the SHA-256 function, and anomaly analysis with a neural network autoencoder. A combined confidence index is generated based on the obtained metrics and stored in a confidence log. An XAI component that generates SHAP graphs is used to interpret the results, while a REST API provides external access to the trust assessment logs. The VCA is implemented in the frequency domain using the discrete wavelet transform (DWT).

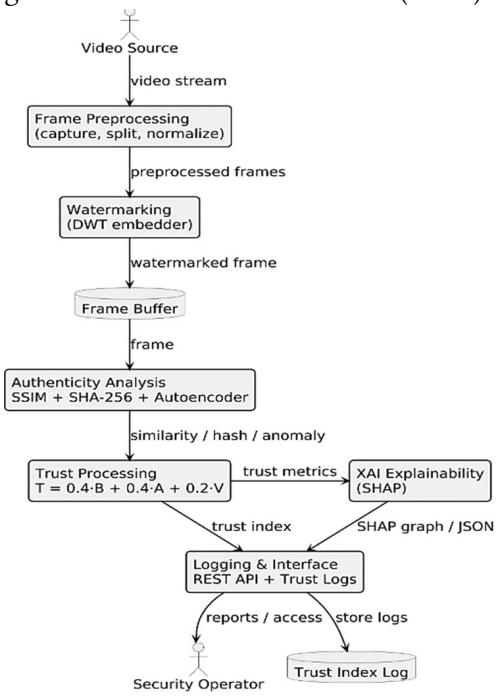

Figure 12: Architecture of the video content authentication module

The described architecture of the authenticity verification module defines the internal logic for video data processing and the generation of confidence metrics at the individual-frame level. For the practical implementation of this approach in the context of streaming processing and integration with external services, a comprehensive software architecture is required that ensures the scalability and manageability of components. Such an architecture generalizes the module's functional blocks and defines their interactions as a service-oriented system.

Figure 13 shows the software system's architectural diagram for verifying the authenticity of video content. The Video Ingestor component captures and preprocesses frames. The Watermark Embedder embeds digital watermarks. The Analysis Service performs SSIM analysis, hashing, and anomaly detection. The Trust Evaluator generates a trust index, which is supplemented by an explanation from the XAI Explainer. Interaction with external systems is carried out via a REST API, and all results are stored in the Trust Logs Database.

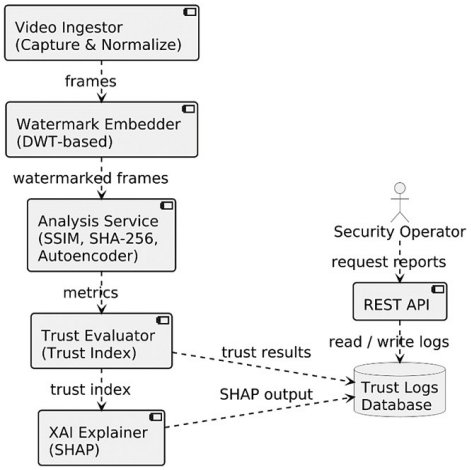

Figure 13: Architecture diagram of the video content authenticity verification system

The described software system architecture defines the composition of the main components and their interaction within the internal video content trust verification loop. To integrate the system with external clients, analytics platforms, or other security modules, a unified mechanism for accessing its functionality is required. This mechanism is the REST API, which provides standardized interaction with the system's key services and access to analysis results.

The diagram in Figure 14 illustrates the key REST API requests for interacting with the video content trust verification module. The POST /api/upload request is used to upload video frames along with metadata; GET /api/trust-index retrieves the trust index for a given frame; GET /api/shap-report returns an explanation of the decision based on SHAP; and GET /api/logs provides access to the trust assessment log for the selected time interval.

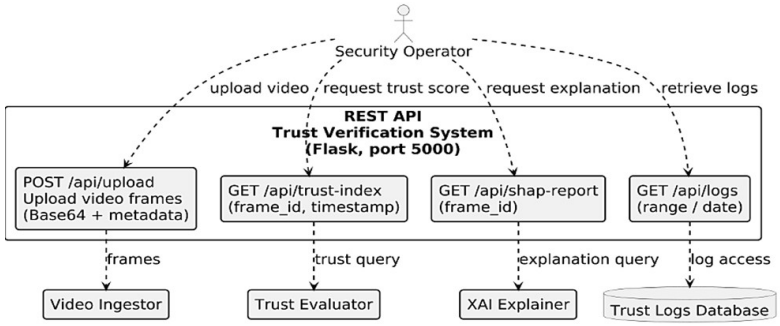

Figure 14: Diagram of the REST API for the video content trust verification module

To comprehensively evaluate the approaches implemented in this study, a comparative table was created that outlines the key characteristics of each method: effectiveness, potential limitations, and specific applications in the context of verifying the authenticity and integrity of video content. Deep learning methods (CNN/DL) demonstrate high accuracy in detecting

DeepFake-type forgeries [11], but require significant amounts of training data [2, 18]. Digital watermarking technology ensures reliable embedding of verification markers [19], although in some cases it may affect image quality [25]. XAI provides visual interpretations of model decisions, which is critically important for analysis in security systems [8, 21], but requires additional processing resources. Integrity modeling based on predicate logic allows for formalizing authenticity verification but requires access to ground-truth data [16]. Ultimately, the mechanism for detecting DeepFake attacks is implemented as a synergy of several approaches, ensuring high accuracy even in complex scenarios involving changes in the video stream. Table 1 summarizes the effectiveness, potential limitations, and comments regarding the application of CNNs, digital watermarking, XAI, and other methods.

Table 1.
Evaluation of the effectiveness of the implemented methods

Method	Effectiveness (rating)	Goal / Problem	Comments on implementation
CNN/Deep Learning	High	Counterfeit detection	Implemented using ResNet50 + classifier
Digital watermarking	Medium-High	Video integrity verification	Dynamic insertion of markers into the video stream
XAI (SHAP, LIME)	High	Model decision interpretation	SHAP plots based on frame features
Integrity modeling	High	Calculation of losses/changes	Formalized via checksum verification
DeepFake detector	High	Detection of generative fakes	Comparison with examples from the anomaly database

For clarity, the comparative characteristics of the main approaches to video content verification are presented in the form of a graphical visualization (Fig. 15). The graph compares four methods — CNN, CNN with a digital watermark, XAI (SHAP/LIME), and DeepFake Detector — based on four key metrics: effectiveness, interpretability, computational cost, and resistance to attacks. The proposed visualization allows for an assessment of the trade-offs between detection accuracy, decision-making transparency, resource costs, and the level of security for each approach. The XAI method deserves special attention, as it combines high performance with the highest level of interpretability, which is critically important for trusted video content verification tasks.

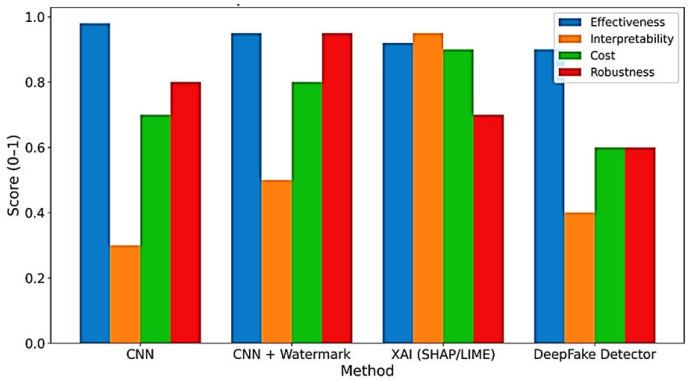

Figure 15: Comparison of the performance, interpretability, computational costs, and robustness of AI-based video content verification methods

Overall, a functional prototype with a REST architecture was developed, in which the main modules interact via HTTP requests using standard JSON interfaces. For validation purposes, a test environment was created that emulates requests, eliminating the need to fully disclose the source code. The main processing stages are implemented as separate endpoints, including CNN filtering modules [11], digital watermark embedding [19, 25], and XAI interpretation [8, 20]. The system's architecture allows for easy integration with existing solutions without requiring source code, while adhering to licensing restrictions.

Thus, the proposed system implements a comprehensive approach to verifying the authenticity and integrity of video content in security systems, combining methods of cryptographic control,

deep learning [18], and explainable artificial intelligence [8]. Its application is possible in the context of digital forensics [5], the protection of cloud video services [3], incident monitoring in critical infrastructure [17, 23], and cybercrime investigations [4], where trust in video is of critical importance.

6. Discussion

The results of implementing the proposed system demonstrated its ability to perform reliable real-time analysis of the authenticity of video content [1, 3, 19]. The use of a multi-level verification approach (combining SSIM, SHA-256 hash functions, and autoencoder reconstruction) reduces the probability of false decisions regarding the authenticity of individual frames [24]. The analysis of confidence in a video frame is performed using a generalized confidence index that formally accounts for the weight of each criterion through weighting and normalization functions.

The use of the XAI module (SHAP) in the proposed architecture ensured decision transparency, as confirmed by the construction of interpretive heatmaps on video frames with detected anomalies [8, 21]. The REST API interface allows for easy integration of the system with external platforms, including SIEM systems, cybersecurity microservices, and cloud video storage [17, 20]. This ensures the architecture's scalability and suitability for industrial applications.

However, the results also revealed several challenges. First, the performance of the autoencoder depends on the architectural configuration and pre-training on relevant data [18]. Second, detection accuracy decreases significantly in cases of video with severe compression or processing artifacts [22]. In some cases, the system exhibits sensitivity to lighting variations, which can lead to a falsely low SSIM. This indicates the need to further optimize normalization functions and expand the training dataset.

Despite its high performance, the system has several limitations. First, the autoencoder's performance is domain-sensitive: models trained on one type of video (e.g., street cameras) may exhibit reduced accuracy on another (e.g., monitor screens) [2, 11]. Second, SSIM and hash functions are sensitive to compression, noise, and post-processing, which complicates the detection of manipulations in heavily compressed or transformed videos [7]. Furthermore, SHAP explanations can be computationally expensive at high frame rates.

Compared with classical approaches to verifying the integrity of video content, which are typically limited to hash checks or embedded digital watermarks, the proposed system offers greater flexibility and reliability for detecting manipulations. Experimental results demonstrate a performance advantage over baseline methods of [18–25]% on F1-score and Recall metrics, confirming the system's increased sensitivity to complex types of tampering. Unlike commercial solutions with closed-source verification logic, the proposed approach additionally ensures transparency in decision-making by integrating an XAI module that generates explanations for detection results (Figure 16). Thus, the system combines the advantages of deep learning methods, classical similarity metrics, and interpretable analysis logic.

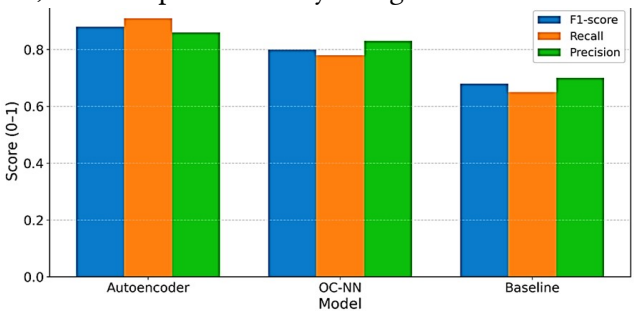

Figure 16: Comparative evaluation of counterfeit detection accuracy using different models

The comparison presented demonstrates the advantages of the proposed approach at fixed decision thresholds; however, for a more comprehensive assessment of classification quality, it is advisable to analyze the model's behavior at variable thresholds. Such an analysis allows us to

evaluate the system's ability to distinguish between authentic and modified video clips regardless of the specific threshold configuration. For this purpose, ROC analysis is used, which is a standard tool for evaluating the discriminative properties of classifiers. The ROC curve visualizes the trade-off between the model's sensitivity and specificity across the entire range of threshold values. The area under the curve (AUC) serves as an integral measure of classification performance and is used to quantitatively compare the effectiveness of the proposed approach with that of baseline models.

To evaluate the quality of classifying video clips as authentic or modified, a ROC (Receiver Operating Characteristic) curve was constructed, which reflects the relationship between the model's sensitivity (True Positive Rate) and the false positive rate (False Positive Rate) at various decision thresholds. The ROC curve in Fig. 17 demonstrates the proposed model's high ability to detect forgeries, as confirmed by the area under the curve (AUC) value > 0.9, which indicates high classification accuracy.

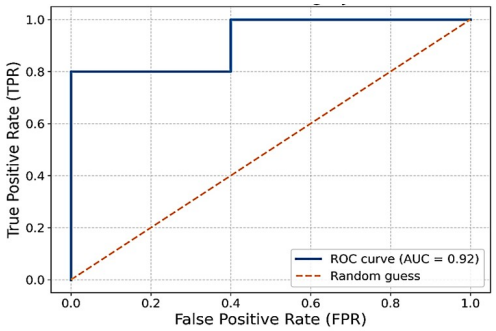

Figure 17: ROC curve for video change classification

Although ROC analysis allows for an assessment of the model's overall discriminative capability regardless of threshold values, it does not reflect the spatio-temporal structure of detected anomalies in the video stream. For practical video forensics tasks, it is important not only to determine the fact of tampering but also to localize the frames and features that led to a decrease in confidence. To this end, it is advisable to supplement the quantitative classification assessment with a visual analysis of the distribution of anomalies. Such visual analysis provides an interpretable identification of the video stream sections most vulnerable to manipulation.

A spatio-temporal heatmap allows us to correlate peaks of abnormality with specific frames and features, which facilitates a deeper understanding of the nature of the detected changes. This, in turn, provides a basis for a well-founded selection of decision-making thresholds that depend on the application context. As a result, the system combines a global assessment of classification quality with localized anomaly analysis, increasing the practical value of the results obtained.

In addition to the ROC curve and performance metrics, the three-dimensional heatmap of anomalies (Figure 18) illustrates the spatiotemporal distribution of probable forgeries within the video stream. The visualization demonstrates the intensity of anomalous changes for individual frames and features, allowing for the localization of video segments potentially altered by DeepFake or other attacks. This approach not only enables a decision regarding authenticity but also provides an interpretable localization of the most suspicious sections, thereby increasing user trust in the system.

Spatio-temporal localization of anomalies allows for the identification of critical segments of the video stream; however, for practical use of the system, it is necessary to define the parameters under which the final decision regarding forgery is made. In particular, the choice of a threshold value directly affects the balance between detection sensitivity and the rate of false positives. Therefore, the next step is to analyze how key classification quality metrics depend on the decision threshold.

Such an analysis allows us to determine the system's operating mode based on the requirements of a specific application scenario, with an emphasis either on maximizing forgery detection or on minimizing false positives.

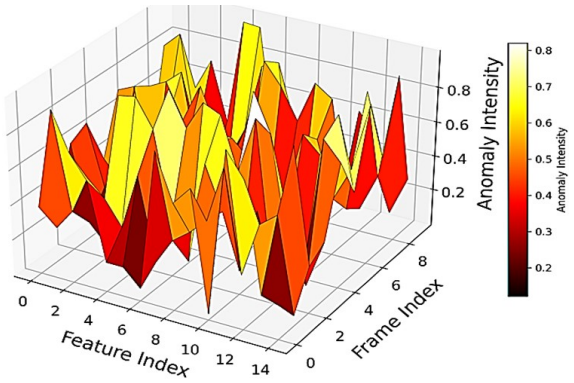

Figure 18: Heatmap of anomalies in the video stream

Comparing precision, recall, and the F1 score at different thresholds allows for a quantitative assessment of the trade-off between these metrics. Particular attention is paid to the range of threshold values where the best balance among them is achieved. This ensures a well-founded selection of the classification threshold for stable system operation under real-world conditions.

The graph in Figure 19 allows selection of the optimal classification threshold based on the trade-off among precision, recall, and the F1 score. It visualizes the dependence of these metrics on the threshold value, which determines the system's sensitivity to detecting counterfeits. Thanks to this graph, it is possible to determine the balance point at which the system demonstrates the best trade-off between minimizing false positives and maintaining a high level of change detection. In particular, the F1-maximum is observed in the threshold range of 0.45–0.55, indicating the model's effective performance at moderate classifier confidence.

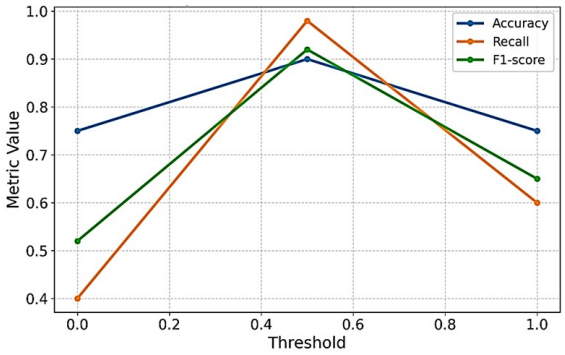

Figure 19: Graph of precision, recall, and F1 score as a function of threshold

In the future, it would be advisable to expand the system to support multi-frame cumulative verification, adaptive tuning based on content type, and integration with DeepFake-type manipulation-detection mechanisms. It is also worth considering introducing a confidence index at the frame-sequence or scene level, which would enable detection of more complex manipulation scenarios that do not manifest in individual frames.

Another area of development is the use of context-aware models that account for semantic and temporal relationships among frames. This will enhance the system's resilience to slow or masked changes in video content, which are characteristic of modern forgery methods. The application of transformer architectures and self-learning methods to improve the model's generalization ability is also promising. Collectively, these approaches will contribute to enhancing the system's accuracy, scalability, and practical applicability in the face of evolving information security threats.

Conclusions

As a result of the research, an intelligent system for verifying the authenticity and integrity of video content has been developed, combining deep learning, digital watermarking, cryptographic

verification, and XAI. A multi-level verification of video frame authenticity in real time has been implemented while maintaining high image quality and transparency of decisions.

Thanks to the integration of autoencoder reconstruction modules, digital watermarking in the frequency domain, confidence models, and interpretable SHAP/LIME analysis, the system ensures robust detection of forgeries even under conditions of compression, noise, and lighting changes. Formalized metrics for confidence, risk, and reconstruction error allow for the quantitative assessment of content authenticity and the detection of anomalous segments. Experimental results demonstrated an 18-25% increase in forgery detection accuracy (based on F1-score) compared to traditional approaches.

The system features a REST API for integration into external environments, including SIEM platforms and cloud services, as well as a rule-based module that enables adaptive updates to response logic. This makes the solution scalable and suitable for use in digital forensics, copyright protection, and video surveillance security.

Despite the results achieved, a number of limitations have been identified: the sensitivity of XAI interpretations to video volume and the domain specificity of the training data, as well as the dependence of the autoencoder's accuracy on scene type. Future research should focus on adaptive learning, constructing multi-frame confidence profiles, and optimizing explainable AI performance for real-time processing.

In the future, the model can be extended to support scene aggregation, new attack types (in particular, DeepFake), response policies, and use in a multisensory environment. This work establishes a scientifically sound foundation for creating modern, transparent digital verification solutions that address information security challenges amid growing multimedia threats.

Declaration on Generative AI

While preparing this work, the authors used the AI programs Grammarly Pro to correct text grammar and Strike Plagiarism to search for possible plagiarism. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] N. Suryaprakash, et al. Cybersecurity Risk Management in Cloud Computing Environment. Int. J. Sci. Res. Arch., 10 (01), 2023, 1062-1068. doi:10.30574/ijsra.2023.10.1.1127

[2] I. Hamid, M. Rahman, AI, Machine Learning, and Deep Learning in Cyber Risk Management. Discov Sustain 6, 389, 2025. doi:10.1007/s43621-025-01012-3

[3] K. Singh, R. Saxena, B. Kumar, AI Security: Cyber Threats and Threat-Informed Defense, In: 8th Cyber Security in Networking Conference (CSNet), Paris, France, 2024, pp. 305-312, doi:10.1109/CSNet64211.2024.10851770

[4] S. Jawhar, C. Kimble, J. Miller, Z. Bitar, Enhancing Cyber Resilience with AI-Powered Cyber Insurance Risk Assessment, In: IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2024, pp. 0435-0438, doi: 10.1109/CCWC60891.2024.10427965

[5] P. Kieseberg, et.al., Security Considerations for the Procurement and Acquisition of Artificial Intelligence (AI) Systems, In: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Padua, Italy, 2022, pp. 1-7, doi:10.1109/FUZZ-IEEE55066.2022.9882675

[6] S. Abdymanapov, M. Muratbekov, S. Altynbek, A. Barlybayev, Fuzzy Expert System of Information Security Risk Assessment on the Example of Analysis Learning Management Systems, In: IEEE Access, vol. 9, pp. 156556-156565, 2021, doi: 10.1109/ACCESS.2021.3129488

[7] Z. Xiao, S. Pan, Analysis of Intelligent Information Security Risk Assessment Based on Decision Tree, In: IEEE 4th Int. Conf. on Information Systems and Computer Aided Education (ICISCAE), Dalian, China, 2021, pp. 506-509, doi: 10.1109/ICISCAE52414.2021.9590795

[8] D. Humphreys, et al. AI hype as a Cyber Security Risk: The Moral Responsibility of Implementing Generative AI in Business. AI Ethics 4, 791-804, 2024. doi:10.1007/s43681-024-00443-4

[9] Y. Kostiuk, et al. A System for Assessing the Interdependencies of Information System Agents in Information Security Risk Management using Cognitive Maps. In: Cyber Hygiene & Conflict Management in Global Information Networks, 3925, 249-264 (2024).
[10] E. Bakhtavar, et al. Fuzzy Cognitive Maps in Systems Risk Analysis: A Comprehensive Review. Complex Intell. Syst. 7, 621-637, 2021. doi:10.1007/s40747-020-00228-2
[11] D. Palko, Intelligent Risk Assessment Models in Distributed Systems Based on the Neural Network Approach. Cybersecur. Educ. Sci. Tech., 3, 2025, doi:10.28925/2663-4023.2025.27.764
[12] P. Shetty, AI and Security, From an Information Security and Risk Manager Standpoint, In: IEEE Access, pp. 77468-77474, 2024. doi:10.1109/ACCESS.2024.3408144
[13] S. Bhol, J. Mohanty, P. Pattnaik, Enhancing Cybersecurity Metrics Evaluation Through the Application of Fuzzy AHP Methodology. In: Bandyopadhyay, S., Balas, V.E., Biswas, S.K., Saha, A.K., Thounaojam, D.M. (eds) Intelligent Computing Systems and Applications. ICICSA 2023. Lecture Notes in Networks and Systems, vol. 1010. Springer, Singapore. doi:10.1007/978-981-97-5412-0_10
[14] Y. Kostiuk, P. Skladannyi, V. Sokolov, S. Rzaieva S. Intelligent System for Simulation Modeling and Research of Information Objects. Proceedings of the 1st Workshop on Software Engineering and Semantic Technologies (SEST 2025), In: 15th Int. Scientific and Practical Programming Conf. (UkrPROG 2025), May 13-14, 2025, Kyiv, Ukraine, pp. 237-251.
[15] D. Parmar, et al., Blockchain-Enhanced AI Security and Trust Through Auditable Decision-Making and Data Integrity in Modern Industry, In: IEEE 2nd Int. Conf. on Innovations in High Speed Communication and Signal Processing (IHCSP), Bhopal, India, 2024, pp. 1-8, doi:10.1109/IHCSP63227.2024.10960006
[16] D. Evans, M. Coole, Artificial Intelligence in Security Technologies: Levels of Intelligent Autonomy and Risk, In: Int. Carnahan Conf. on Security Technology (ICCST), Hatfield, United Kingdom, 2021, pp. 1-6. doi:10.1109/ICCST49569.2021.9717405
[17] M. Mylrea, N. Robinson, Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an Entropy Lens to Improve Security, Privacy, Ethical AI. Entropy, 25 (10), 2023. doi:10.3390/e25101429
[18] Y. Kostiuk, P. Skladannyi, V. Sokolov, O. Zhyltsov, Y. Ivanichenko, Effectiveness of Information Security Control using Audit Logs. In: Proceedings of the Workshop on Cybersecurity Providing in Information and Telecommunication Systems, 2025, pp. 524-538.
[19] C. Park, J. Lee, Y. Kim, J.-G. Park, H. Kim, D. Hong, An Enhanced AI-based Network Intrusion Detection System using Generative Adversarial Networks, In: IEEE Internet of Things Journal, no. 3, pp. 2330-2345, 2023, doi:10.1109/JIOT.2022.3211346
[20] A. Tyagi, S. Kumari, Richa, Artificial Intelligence-Based Cyber Security and Digital Forensics, In: Artificial Intelligence-Enabled Digital Twin for Smart Manufacturing, Wiley, 2024, pp.391-419, doi:10.1002/9781394303601.ch18
[21] S. Jawhar, J. Miller, Z. Bitar, AI-Driven Customized Cyber Security Training and Awareness, In: IEEE 3rd International Conference on AI in Cybersecurity, Houston, TX, USA, 2024, pp. 1-5, doi:10.1109/ICAIC60265.2024.10433829
[22] L. Tian, Y. Meng, J. Zhang, Y. Liu, Research on Constructing Network Security Defense Systems for Research Institutions Based on Secure AI, In: 3rd International Conference on Artificial Intelligence and Computer Information Technology, Yichang, China, 2024, pp. 1-5, doi:10.1109/AICIT62434.2024.10730713
[23] G. Chakkappan, A. Morshed, M. Rashid, Explainable AI and Big Data Analytics for Data Security Risk and Privacy Issues in the Financial Industry, In: IEEE Conference on Engineering Informatics (ICEI), Melbourne, Australia, 2024, pp. 1-9, doi: 10.1109/ICEI64305.2024.10912422
[24] M. Chu, Y. Song, Analysis of Network Security and Privacy Security Based on AI in the IoT Environment, In: IEEE 4th Int. Conf. on Information Systems and Computer Aided Education (ICISCAE), Dalian, China, 2021, pp. 390-393, doi:10.1109/ICISCAE52414.2021.9590786
[25] B. Saurabh, et al., Generative AI Enabled Actionable Decision Support in Cyber Security Operations for Enterprise Security, In: ITU Kaleidoscope: Innovation and Digital Transformation for a Sustainable World, New Delhi, India, 2024, pp. 1-8, doi:10.23919/ITUK62727.2024.10772892