ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2. pp. 518–521
DOI: https://doi.org/10.5281/zenodo.24362434
Journal website: https://cajournal.kz

UDC 004.8

INTELLIGENT SYSTEM FOR DETECTING SYNTHETIC IMAGES AND VIDEO FRAGMENTS (DEEPFAKE DETECTION AI)

Bektursun Kamila, Bilimzhankyzy Aisha, Anvarova Muslima, Bektursyn Arailym
Satbayev University, Almaty, Kazakhstan

Bektursun Kamila — Fourth-year student, Department of Cybersecurity, Satbayev University, Almaty, Kazakhstan.

Bilimzhankyzy Aisha — Fourth-year student, Department of Cybersecurity, Satbayev University, Almaty, Kazakhstan.

Anvarova Muslima — Fourth-year student, Department of Cybersecurity, Satbayev University, Almaty, Kazakhstan.

Bektursyn Arailym — Fourth-year student, Department of Cybersecurity, Satbayev University, Almaty, Kazakhstan.

Scientific supervisor: Yubuzova Khalicha — PhD, Professor, Department of Cybersecurity, Satbayev University, Almaty, Kazakhstan.

© Bektursun Kamila, Bilimzhankyzy Aisha, Anvarova Muslima, Bektursyn Arailym, 2026

Abstract: This paper examines current approaches to the automated detection of synthetic images and video fragments generated using artificial intelligence. Visual, frequency-domain, and temporal indicators of manipulation are analyzed, together with machine-learning methods for classifying digital content. Based on a review of scientific studies, a modular Deepfake Detection AI architecture is proposed that combines preprocessing, feature extraction, frame-level analysis, and video sequence consistency assessment. Public datasets, evaluation metrics, and an experimental validation plan are discussed. The study is analytical and design-oriented; actual accuracy metrics for the proposed system remain to be established through experiments.

Keywords: deepfake; synthetic images; video fragments; computer vision; machine learning; convolutional neural networks; forgery detection; digital forensics.

Introduction

The development of generative adversarial networks (GANs), diffusion models, and face-swapping technologies has expanded the ability to create convincing digital content. Alongside legitimate uses in film, education, and design, these technologies create risks of identity impersonation, disinformation, and fraud. The visual realism of modern synthetic images often makes them difficult to distinguish from genuine photographs by visual inspection alone [1–3].

Video analysis poses a particular challenge: the quality of individual frames varies due to compression, lighting changes, camera movement, and loss of detail. A detector that performs well on one dataset may become less effective when analyzing content produced by an unseen generator or reposted on social media [3, 4]. A practical system must therefore consider not only image features but also the robustness of its decisions to data transformations.

The aim of this study is to justify the structure of an intelligent system for detecting synthetic images and video fragments and to develop a reproducible procedure for assessing its

effectiveness. To achieve this aim, scientific publications are reviewed, detection approaches are compared, public datasets are selected, and software architecture requirements are formulated.

1. Review of Existing Research and Problem Statement

Rössler et al. introduced the FaceForensics++ dataset and evaluation procedures for facial manipulation detectors, including under compression [1]. The Celeb-DF study by Li et al. demonstrated the importance of testing methods on more realistic manipulated videos [2]. DeepfakeBench brings together different methods, datasets, and standardized testing scenarios, enabling reproducible comparisons of algorithms [3]. GenImage focuses on the detection of images entirely generated by artificial intelligence and on evaluating detector generalization [4].

Detection approaches can be divided into four broad groups. The first analyzes visual artifacts, such as unnatural skin texture, blending boundaries around faces, and geometric inconsistencies. The second relies on spatial features extracted by convolutional neural networks (CNNs) and transformer-based models. The third examines frequency-domain patterns introduced by generation, resizing, and compression. The fourth uses temporal consistency across adjacent frames, for example anomalous changes in facial expression or movement [1–3, 5].

The problem is formulated as binary classification of content as authentic or synthetic, accompanied by an estimated probability of manipulation. For video, the final prediction should be based on multiple frames rather than on a single randomly selected image. In addition to a class label, the system should provide a numerical confidence estimate and preserve information about the analysis performed.

Table 1 – Comparison of the Main Deepfake Detection Approaches

Approach	Main features	Advantages	Limitations
Visual-artifact analysis	Face boundaries, textures, lighting	Interpretability	Artifacts become less noticeable as generation improves
CNN / spatial analysis	Local image features	Automatic feature extraction	Sensitivity to domain shifts
Frequency-domain analysis	Spectrum, high-frequency traces	Additional generation cues	Dependence on compression and postprocessing
Temporal analysis	Consistency between consecutive frames	Captures video-specific characteristics	Higher computational cost

2. Materials and Methods

The methodology consists of a comparative review of publications [1–6], modular system design, and a plan for controlled experiments. To compare methods fairly, the use of consistent dataset splits, identical preprocessing procedures, and evaluation metrics calculated on a test set excluded from training is proposed.

FaceForensics++ and Celeb-DF are considered as experimental sources for manipulated videos, while GenImage is considered for images synthesized by generative models [1, 2, 4]. DeepfakeBench can serve as a reference for standardizing the evaluation protocol [3]. To avoid data leakage, samples should be separated by original video, identity, or source wherever dataset structure permits, preventing related frames from appearing simultaneously in training and test sets.

Table 2 – Proposed Datasets and Their Roles in the Study

Dataset	Content type	Purpose
FaceForensics++ [1]	Manipulated-face videos	Baseline training and compression robustness evaluation
Celeb-DF [2]	Realistic deepfake videos	Generalization testing on other video data
GenImage [4]	Real and AI-generated images	Image detection beyond face swapping
DeepfakeBench [3]	Tools and evaluation protocols	Comparable model benchmarking

© Central Asian Scientific Journal, 2026

3. Proposed Deepfake Detection AI Architecture

The proposed intelligent system follows a modular design. A user uploads an image or video. The preprocessing module validates the file format, normalizes resolution, and, when necessary, extracts representative frames. For video content, face detection and tracking are additionally performed so that features correspond to the same subject throughout the sequence.

The core module analyzes spatial features using a pretrained CNN model. A frequency-domain branch and a temporal consistency module are envisaged as extensions; their outputs can be combined at the probability level. The final stage produces a classification decision, estimates confidence, and generates a report describing the analyzed material.



Figure 1 – Conceptual Data Processing Pipeline of Deepfake Detection AI

For images, a baseline configuration with a single spatial model is sufficient. For videos, the manipulation probability is estimated from several selected frames. If p_i is the model prediction for frame i, a simple aggregated score can be calculated as follows:

$$P_{video} = (1 / N) \cdot \Sigma\, p_i, \quad i = 1, \ldots, N, \qquad (1)$$

where N is the number of analyzed frames. The arithmetic mean is convenient as a baseline, but when only a few frames are strongly manipulated, more robust aggregation, such as the median or a learned temporal classifier, may be required.

A decision threshold τ is used: if $P \geq \tau$, the content is flagged as potentially synthetic; otherwise, it is classified as presumably authentic. The threshold τ should be determined on a validation set while accounting for an acceptable false-positive rate. The system output must be interpreted as a probabilistic assessment rather than definitive proof of forgery.

4. Results of the Analytical Study

The comparison of scientific studies indicates that detecting face manipulation in videos and identifying fully synthesized images involve different sources of discriminative features and data distributions [1–4]. Consequently, deploying a single detector without testing it across multiple generation methods does not provide sufficient evidence of universal performance.

The analytical stage yielded four functional requirements: R1 — support for both images and videos; R2 — modular processing with a replaceable classifier; R3 — aggregation of frame-level results; and R4 — a reproducible report specifying probability, classification threshold, and processing parameters. The conceptual architecture shown in Figure 1 was designed around these requirements.

The comparison in Table 1 suggests that a spatial CNN is a practical starting point for a prototype: it supports an operational classification pipeline without requiring a complex video transformer. Frequency-domain and temporal features are best treated as incremental extensions whose effectiveness must be tested independently. This is a design conclusion drawn from the literature review, not a measurement of the proposed model's accuracy.

5. Experimental Evaluation Procedure and Discussion

The proposed empirical evaluation includes a baseline experiment with a single spatial classifier and two additional scenarios: testing on a different dataset and evaluating the effect of JPEG compression or resolution changes. For videos, predictions are calculated using the same number of frames per file, and final metrics are measured at the video level to prevent videos with many frames from disproportionately influencing the evaluation.

© Central Asian Scientific Journal, 2026

The principal metrics are Accuracy, the proportion of correct predictions; Precision, the proportion of true manipulated samples among those flagged by the system; Recall, the proportion of manipulated samples detected; F1-score, the harmonic mean of Precision and Recall; and ROC-AUC, the ability to rank authentic and synthetic samples by predicted risk. A confusion matrix, per-file analysis time, and throughput on the hardware used should also be recorded [3].

Table 3 – Proposed Evaluation Plan for the System

Experiment	Property evaluated	Expected analysis output
In-domain test	Baseline classification performance	Accuracy, F1, ROC-AUC
Cross-dataset test	Generalization ability	Metric changes without retraining
JPEG compression / resizing	Robustness to processing	Performance degradation or retention
Frame-based video analysis	Value of frame aggregation	Video-level metrics and processing time

When interpreting results, potential training-set bias must be considered: the model may learn codec properties, resolution, or signatures of specific generators instead of generalizable manipulation features. Data augmentation, independent external testing, and analysis of misclassified files are recommended to mitigate this risk. Explainability can be improved by visualizing regions that most strongly influence the prediction; however, such visualization is not itself proof of manipulation.

At the present stage, experimental Accuracy, F1-score, and ROC-AUC values are not reported, since training and testing of a custom prototype are not described in this study. The outcome consists of an evidence-informed component selection, a system architecture, and a validation plan.

Conclusion

Current approaches to detecting synthetic images and video fragments were reviewed, including methods based on visual, spatial, frequency-domain, and temporal features. A conceptual Deepfake Detection AI architecture was developed, comprising preprocessing, feature analysis, estimation of synthetic-content probability, and report generation. Public datasets and reproducible evaluation metrics were identified. Future work will involve implementing the software prototype, training the baseline model, performing independent testing, and comparing extended analysis configurations.

References:

1. Rössler A., Cozzolino D., Verdoliva L., Riess C., Thies J., Nießner M. FaceForensics++: Learning to Detect Manipulated Facial Images // Proceedings of ICCV. 2019.

Dataset for DeepFake Forensics // Proceedings of CVPR. 2020.

3. Yan Z. et al. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection // Advances in Neural Information Processing Systems. 2023.

4. Zhu M. et al. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image // Advances in Neural Information Processing Systems. 2023.

5. Wang S.-Y., Wang O., Zhang R., Owens A., Efros A. A. CNN-Generated Images Are Surprisingly Easy to Spot... for Now // Proceedings of CVPR. 2020.

6. Tolosana R., Vera-Rodriguez R., Fierrez J., Morales A., Ortega-Garcia J. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection // Information Fusion. 2020. Vol. 6. pp. 131–148. DOI: 10.1016/j.inffus.2020.06.014.

© Central Asian Scientific Journal, 2026