https://www.sworldjournal.com/index.php/swj/article/view/swj33-01-083
DOI: 10.30888/2663-5712.2025-33-01-083

HYBRID AI FRAMEWORK FOR EFFICIENT ANOMALY DETECTION IN VIDEO SURVEILLANCE DATA

Bondarchuk Andrii
d.m.s., prof
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID: https://orcid.org/0000-0001-5124-5102
Bushma Oleksandr
d.m.s., prof
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID: https://orcid.org/ 0000-0003-1604-6129
Dovzhenko Tymur
Ph.D ., as.prof.
State University of Information and Communication Technologies, Kyiv, Ukraine
ORCID: https://orcid.org/0009-0006-9930-5091
Hashko Andrii
graduate student
ORCID: https://orcid.org/0009-0000-4103-8425
State University of Information and Communication Technologies, Kyiv, Ukraine

Abstract. *Contemporary video surveillance infrastructure produces substantial data streams, posing challenges for efficient real-time processing. Current automated anomaly detection techniques frequently demand extensive computational resources and function as opaque "black box" systems, constraining their deployment in critical domains including public security and safety monitoring. This work introduces an integrated methodology tackling two fundamental limitations: ineffective handling of superfluous visual data and lack of algorithmic transparency in artificial intelligence systems. The proposed framework merges an advanced informative frame selection technique with interpretable detection model processing. The initial phase employs a hybrid optimization approach integrating InceptionV3 convolutional neural networks with genetic algorithms, achieving 70-85% data reduction while preserving 98% recall performance. The subsequent phase delivers not only anomaly classification but also produces comprehensible decision explanations via explainable AI (XAI) integration, specifically utilizing Grad-CAM and guided backpropagation techniques. Experimental evaluation on benchmark datasets confirms the superiority of the proposed method over contemporary solutions. Results demonstrate 3-5% enhancement in classification precision coupled with reduced computational requirements. Additionally, the system generates visual decision rationalizations through heatmap representations, thereby increasing operational trustworthiness. This integrated framework facilitates the deployment of effective real-time video analysis systems that provide comprehensive decision transparency and operational accountability.*

Keywords: *artificial intelligence, video surveillance, information systems, genetic algorithm, modeling, computer vision, video surveillance data.*

Introduction.

Modern video surveillance systems generate vast amounts of data, making manual analysis practically impossible. Automated detection of anomalous events, particularly acts of violence, using artificial intelligence methods faces two key challenges: the

inefficiency of processing irrelevant data and the "black box" nature of deep neural networks' decision-making, where human operators cannot understand the machine's logic. This undermines trust in the system and complicates its deployment in critical domains such as public security and facility protection.

Analysis of recent studies on improving the efficiency of video surveillance systems shows significant progress in enhancing detection accuracy. However, key challenges in computational efficiency and result interpretability remain unresolved. A major step in addressing this problem has been made in the work of Salman, et al. [1].

Research in recent years can be broadly divided into three main categories. The first category comprises methods based on deep learning. Studies [2-4] demonstrate the high effectiveness of architectures based on 3D convolutions and transformers in anomaly detection tasks. However, these approaches require processing the complete video stream, leading to excessive computational costs. Models often operate as "black boxes," complicating their application in critical systems where decision-making transparency is essential.

The second category involves approaches to key frame selection. Works [5-6] propose methods for processing and compressing video data, particularly based on motion analysis. Despite reducing data volume, these methods often miss important frames with sudden anomalies unrelated to motion. Other studies propose machine learning methods for frame selection, but they fail to consider feature optimization at the individual frame level, limiting their effectiveness. The third category addresses the processing of massive data streams, real-time operation, and compliance with decision transparency requirements. This issue is examined in works [6-7]. However, these solutions often overlook optimization at the individual node level, leading to inefficient resource utilization.

The aim is to develop a comprehensive approach for automated anomaly detection in video recordings by combining an efficient key frame selection method with an interpretable deep learning model (XAI-Inv3), aimed at overcoming the limitations of modern video analytics systems regarding computational inefficiency and insufficient decision-making transparency.

Research Objectives:

1. To conduct a review of contemporary approaches to video anomaly detection, identifying their shortcomings, particularly the high computational costs associated with processing redundant data.
2. To define and implement a fitness function that ensures a balance between classification accuracy and the number of selected features.
3. To propose a frame comparison mechanism based on calculating the Euclidean distance between optimized feature vectors and adaptive threshold determination for identifying key frames.

Research Results.

To overcome these limitations, a methodology is proposed that utilizes an intelligent filter to select only the most important frames containing potential anomalies from the video stream, along with tools for analyzing the selected frames for both event classification and providing human-understandable explanations of the decisions made. This approach significantly reduces computational load and enhances the transparency of the system's operation.

This methodology combines the power of the InceptionV3 convolutional neural network for extracting high-level features from each frame and a genetic algorithm for optimizing the selection of the most relevant features. The genetic algorithm iteratively evolves the set of features, maximizing accuracy while minimizing their quantity. The final stage involves calculating the Euclidean distance between consecutive frames based on the selected features and selecting those frames whose distance exceeds a dynamic threshold, indicating a significant change in the scene.

The system model is built on the InceptionV3 architecture, but its key aspect is modification for interpretability (eXplainable AI, XAI). It integrates gradient-based methods such as Grad-CAM (Gradient-weighted Class Activation Mapping) and guided backpropagation. These methods allow for the visualization, in the form of "heat maps," of the image regions that most influenced the model's prediction (e.g., a fist, a weapon, chaotic motion), providing the operator with a visual and understandable explanation of why an event was classified as anomalous.

In this methodology, the genetic algorithm begins by creating an initial population. Each individual in this population, called a "chromosome," is a binary vector. The length of the vector corresponds to the total number of features extracted from the frame by the InceptionV3 convolutional neural network. Each bit in the chromosome (gene) indicates the inclusion (1) or exclusion (0) of a specific feature from subsequent analysis. Thus, the chromosome represents a compactly encoded subset of candidate features.

Each chromosome is evaluated using a fitness function. This function is the most critical component as it guides the direction of evolution. It not only maximizes classification accuracy but also incorporates a penalty coefficient (λ) for an excessively large number of selected features. Mathematically, this is expressed as:

$$F\ (c) = A\ (c) - \lambda * D\ (c)$$

where,

$A\ (c)$ is the classification accuracy demonstrated by the model (e.g., XAI-Inv3) when using not the full set of features but only the subset encoded in the chromosome as input. It measures how well the model can detect anomalies using only those features for which the value in chromosome c is set to 1.

$D\ (c)$ is the total number of features selected (activated) in chromosome c. This is the count of ones (1) in the binary vector of the chromosome. For example, if a chromosome has the form [1, 0, 0, 1, 1, 0], then $D\ (c) = 3$.

This ensures a search for a compromise between high accuracy and efficiency. Based on the obtained fitness scores, selection occurs: chromosomes with a higher value have a greater probability of being chosen for "reproduction" using a "roulette wheel" or tournament selection mechanism.

The selected "parent" chromosomes undergo genetic operators. The crossover operator exchanges parts between two chromosomes, creating new "offspring" with a combination of parental features. The mutation operator randomly changes individual bits in the chromosomes (with a low probability), introducing new genetic information into the population and preventing stagnation in local optima. This cycle (evaluation, selection, crossover, mutation) repeats for a specified number of generations until a

suboptimal set of features is found that maximizes the value of the fitness function.

After the genetic algorithm has determined the optimal subset of features for each frame, the next step is to quantitatively measure the similarity between consecutive frames. For this purpose, the Euclidean distance is used. Let two vectors of optimized features, determined by the genetic algorithm for frames i and j, be denoted as:

$$F_i = \left(f_i^{(1)}, f_i^{(2)}, f_i^{(d)}\right), F_j = \left(f_j^{(1)}, f_j^{(2)}, f_j^{(d)}\right),$$

where, d — the number of features selected in the chromosome.

The Euclidean distance D_{ij} between these vectors is calculated by the formula:

$$D_{ij} = \sqrt{\sum_{k=1}^{d} (f_i^{(k)} - f_j^{(k)})^2}.$$

where,, $f_i^{(k)}$ and $f_j^{(k)}$— are the values of the k th feature for frames i and j respectively.

For each feature k the difference $(f_i^{(k)} - f_j^{(k)})$, is calculated and then squared to eliminate negative values and amplify large discrepancies.

A critical aspect is determining the threshold value that separates key frames from redundant ones. A fixed threshold is ineffective due to the variability in recording conditions. An adaptive approach can be employed: the threshold is calculated dynamically based on the statistics of the first N frames of the video (e.g., 1000). The average Euclidean distance for this sample is computed, and then the threshold value is set as a multiple of this average. This allows the system to automatically adapt to the specific video.

The final classification process is relatively simple yet effective. For each pair of consecutive frames, the Euclidean distance between their optimized feature vectors is calculated. If this distance exceeds the computed dynamic threshold, the second frame in the pair is marked as a key frame. All frames with distances below the threshold are considered redundant and discarded. This drastically reduces the volume of data input to the model while preserving information about all important events.

Traditional deep neural networks, such as the standard InceptionV3, demonstrate high accuracy but operate as "black boxes." This means that we only see the input data (frame) and the result (e.g., "anomaly"), but the internal decision-making process remains opaque. For critical applications, such as security, this is unacceptable because an operator cannot trust a system that does not explain its conclusions and may be prone to errors due to imbalanced data or artifacts [7].

The model's task is not only to classify the frame but also to generate human-understandable explanations: specifically which areas of the image and which visual features most influenced the decision. This transforms the model from a "black box" into a "glass" or "transparent box."

Interpretability is critically important for building trust and practical implementation. When the system detects an anomaly, the operator receives not just an alarm signal but visual confirmation: a heatmap that highlights the conflict area, a suspicious object, or an unusual action. This allows the operator to quickly verify the validity of the alarm and make a decision, avoiding false alarms. Thus, a bridge is formed between the high accuracy of the algorithm and human understanding.

Grad-CAM is one of the key methods in XAI-Inv3. It works based on gradient analysis. When the network makes a prediction (e.g., "fight"), Grad-CAM calculates which pixels in the last convolutional layer are most "responsible" for this prediction. Technically, the method computes a weighted sum of the feature maps of the last convolutional layer, where the weights are determined by the gradient of the target class with respect to these feature maps. The result is a heatmap—a semi-transparent overlay on the original image where "hot" colors (red, yellow) indicate the most important areas.

The Guided Backpropagation method complements Grad-CAM. It takes a more detailed approach to interpretation. This method also analyzes gradients but propagates backward through the network down to the input pixel level. Its key feature is filtration: it preserves only those gradients that have a positive influence on the predicted class and discards negative ones. As a result, a clear, highly detailed image is generated, where specific contours and textures (e.g., the silhouette of a weapon, outlines of

people in a fight) that contributed to the classification are prominently highlighted.

Together, these two methods provide a powerful and multi-level interpretation tool. Grad-CAM offers a general understanding of event localization, showing "where" the anomaly occurred (e.g., a group of people in the left corner of the frame). Guided Backpropagation elaborates on this by detailing "what" exactly in that area attracted the network's attention (individual objects). This combination ensures the most comprehensive and understandable explanation of the model's decision.

XAI-Inv3 is fundamentally based on the time-tested InceptionV3 architecture. This means it retains all its advantages: Inception modules for efficient multi-scale feature extraction, dimensionality reduction techniques to combat overfitting, and auxiliary classifiers to improve convergence during deep network training. This foundational structure ensures high performance in image classification tasks, forming the backbone of the entire system.

Conclusions and Future Research Directions.

It can be concluded that the proposed approach serves as a highly efficient preprocessing mechanism that eliminates informational noise and drastically reduces computational costs. This enables XAI-Inv3 to process only relevant data, enhancing both speed and potentially classification accuracy by focusing on significant events. Together, they form an integrated pipeline optimized for both processing speed and analytical quality.

This research is valuable for the development of security systems and automated video monitoring, as well as for the design of smart cities. By ensuring high accuracy, efficiency, and, most importantly, interpretability of algorithmic operations, the proposed methodology lays the groundwork for closer and more effective human–artificial intelligence collaboration in critical domains. Future research may focus on adapting the methodology for real-time operation and implementation in embedded systems.

References

1. Salman, M., Abbas, N., Rahman, S. I. U., Rehman, A., Alamri, F. S., Elyassih, A., & Saba, T. (2025). Enhancing surveillance anomaly detection with keyframes and

explainable inception model. Egyptian Informatics Journal, 31, 100769. https://doi.org/10.1016/j.eij.2025.100769

2. Masud U, Sadiq M, Masood S, Ahmad M, Abd El-Latif AA. LW-DeepFakeNet: a lightweight time distributed CNN-LSTM network for real-time DeepFake video detection. SIViP 2023;17(8):4029–37.

3. Theobald O. Machine learning: make your recommender system; build your recommender system with machine learning insights. Packt Publishing Ltd 2024

4. Bensakhria, Ayoub. “Leveraging Real-time Edge AI-Video Analytics to Detect and Prevent Threats in Sensitive Environments.” (2023).

5. Skladannyi, P., Kostiuk, Y., Rzaieva, S., Bebeshko, B., & Korshun, N. (2025). Adaptive Methods for Embedding Digital Watermarks to Protect Audio and Video Images in Information Systems. https://ceur-ws.org/Vol-4016/

6. Bondarchuk, A., Dibrivniy, O., Grebenyk, V., & Onyshchenko, V. (2021). Motion Vector Search Algorithm for Motion Compensation in Video Encoding. In 2021 IEEE 8th International Conference on Problems of Infocommunications, Science and Technology (pp. 345-348). IEEE. doi: 10.1109/PICST54195.2021.9772109

7. Buinytska, O., & Smirnova, V. (2024). Artificial intelligence technologies in research activities: overview and application. Continuing Professional Education: Theory and Practice, 81(4), 31–46. https://doi.org/10.28925/2412-0774.2024.4.2

8. Chemerys, O., Bushma, O., Lytvyn, O., Belotserkovsky, A., & Lukashevich, P. (2021). Network of Autonomous Units for the Complex Technological Objects Reliable Monitoring. In Reliability Engineering and Computational Intelligence (pp. 261-274). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-74556-1_16

9. Zhurakovskyi, B., Poltorak, V., Toliupa, S., Pliushch, O., & Platonenko, A. (2024). Processing and Analyzing Images based on a Neural Network. https://ceur-ws.org/Vol-3654/paper11.pdf