eISSN 2783-6851

VILNIUS TECH
Vilnius Gediminas Technical University

NEW TRENDS in COMPUTER SCIENCES

2025
Volume 3
Issue 2
Pages 126–139
https://doi.org/10.3846/ntcs.2025.25567

LIGHTWEIGHT DEEP MODELS FOR VIDEO ANOMALY DETECTION: A COMPARATIVE STUDY OF AUTOENCODERS AND MOBILENETV2 ON THE AVENUE DATASET

SeyedMohammad VAHEDI [1]✉, Pavel STEFANOVIČ [2]

[1]*Department of Information Technologies, Vilnius Gediminas Technical University, Vilnius, Lithuania*
[2]*Department of Information Systems, Vilnius Gediminas Technical University, Vilnius, Lithuania*

Article History:
- received 11 December 2025
- accepted 23 January 2026

Abstract. Video anomaly detection aims to identify unusual events in surveillance footage, yet many existing deep learning solutions remain too computationally heavy for real-time deployment on resource-limited hardware. This study presents a systematic comparison of three lightweight deep learning models for frame-level anomaly detection on the Avenue dataset, including a baseline 2D convolutional autoencoder, an enhanced reconstruction-based autoencoder with refined feature representation and decoding strategy, and a MobileNetV2-based supervised classifier fine-tuned for anomaly recognition. The baseline autoencoder achieves moderate detection performance, with an approximately AUC of 0.75. In contrast, the enhanced autoencoder improves reconstruction quality and raises the AUC to approximately 0.84 through more effective feature abstraction rather than increased architectural depth. The strongest results are obtained by the MobileNetV2 classifier, which achieves an AUC close to 0.99, high precision and recall, and a stable confusion matrix. These results demonstrate that lightweight architectures, when combined with appropriate training strategies and careful handling of class imbalance, can outperform more complex models. Overall, the study confirms that architectural efficiency and learning paradigm selection are more critical than model depth alone, making lightweight models well-suited to practical, real-time video anomaly detection scenarios.

Keywords: video anomaly detection, lightweight models, convolutional autoencoder, MobileNetV2, frame-level detection, avenue dataset.

✉Corresponding author. E-mail: ***seyedmohammad.vahedi@stud.vilniustech.lt***

1. Introduction

Video anomaly detection (VAD) focuses on identifying rare and abnormal events in long surveillance video streams, where the vast majority of observed activity corresponds to normal behavior. With the extensive deployment of CCTV systems in public and semi-controlled environments, continuous manual monitoring is neither scalable nor reliable. As a result, automatic detection of anomalous events has become an essential task in computer vision research and intelligent surveillance systems (Adam et al., 2008; Mahadevan et al., 2010). Despite substantial progress enabled by deep learning, video anomaly detection remains inherently challenging. Abnormal events are typically sparse, highly diverse in appearance, and often confined to small spatial regions of video frames. In addition, most surveillance datasets exhibit severe class imbalance, with anomalous frames representing only a small fraction of the data. These characteristics significantly complicate model training and generalization, particularly for data-driven approaches that rely on learning regular patterns from normal behavior (Hasan et al., 2016; Sultani et al., 2019).

The CUHK Avenue dataset exemplifies these challenges and is widely used as a benchmark for evaluating video anomaly detection methods. Anomalies in Avenue are short-term, spatially localized, and heterogeneous, while the training data consist almost exclusively of normal activities. Under such conditions, complex and highly parameterized deep learning models often struggle to generalize beyond the limited normal training data, leading to unstable detection performance and increased false alarms (Hasan et al., 2016; Mahadevan et al., 2010). In response to these limitations, recent research has increasingly emphasized lightweight and computationally efficient deep learning models for video anomaly detection. Multiple survey studies highlight that compact architectures can reduce the risk of overfitting, improve real-time applicability, and remain effective on datasets with limited training samples and strong class imbalance (Abdalla et al., 2024; Wu et al., 2024; Yadav & Kumar, 2022). Such models are particularly relevant for practical surveillance scenarios, where computational resources, latency constraints, and deployment feasibility play a critical role.

In this study, we investigate frame-level video anomaly detection on the Avenue dataset by systematically comparing three lightweight deep learning models. Rather than introducing a new architecture, the focus is on analyzing how different learning paradigms—unsupervised reconstruction, temporal prediction, and supervised classification—perform under identical experimental conditions. This enables a clear assessment of how supervision level, temporal modeling, and architectural simplicity influence detection robustness in a highly imbalanced surveillance setting.

The main contributions of this work are summarized as follows:

- A controlled comparative study of three lightweight architectures for frame-level video anomaly detection on the CUHK Avenue dataset, including a 2D convolutional autoencoder, an enhanced sequence 2D convolutional autoencoder, and a MobileNetV2-based supervised classifier.
- A comprehensive evaluation using metrics suitable for highly imbalanced data, including frame-level AUC and confusion-matrix analysis.
- An empirical analysis demonstrating how supervision level and model simplicity affect anomaly detection performance on datasets characterized by limited training data and localized abnormal events.
- Practical insights into the strengths and limitations of lightweight reconstruction-based and classification-based approaches for surveillance-oriented anomaly detection.

The remainder of this paper is organized as follows. Section 2 reviews related work on video anomaly detection, with an emphasis on deep learning and lightweight approaches. Section 3 describes the CUHK Avenue dataset, preprocessing steps, and the experimental setup. Section 4 presents the evaluated model architectures and experimental results. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related works

VAD has evolved significantly over the past decade, progressing from hand-crafted feature representations and statistical modeling to deep learning–based spatio-temporal architectures and, more recently, lightweight models designed for real-time deployment. Early works focused on modeling motion patterns and deviations using classical computer vision and

signal processing techniques, while more recent studies emphasize data-driven learning and scalable deployment (Adam et al., 2008; Mahadevan et al., 2010). Comprehensive surveys further systematize this evolution by reviewing benchmark datasets, evaluation protocols, and learning paradigms across different eras of VAD research (Abdalla et al., 2024; Keleko Teguede, 2022; Choudhry et al., 2023; Pang et al., 2022; Wu et al., 2024; Yadav & Kumar, 2022). These reviews consistently highlight persistent challenges, including severe class imbalance, a scarcity of anomalous events, and the need for fair, frame-level evaluation. Early research on video anomaly detection primarily relied on handcrafted motion descriptors, trajectories, and statistical models of normal behavior. Optical flow statistics, spatio-temporal interest points, and trajectory-based representations were commonly used to characterize regular crowd dynamics and detect deviations (Adam et al., 2008; Mahadevan et al., 2010). Sparse and low-rank reconstruction models extended this line of work by decomposing scenes into dominant regular components and rare abnormal patterns, showing effectiveness on early benchmarks but reduced robustness in complex environments (Cong et al., 2011; Gnouma et al., 2018). Later surveys note that although these methods are computationally efficient, their reliance on manual feature design limits scalability and generalization (Anoopa & Salim, 2022; Fernandes et al., 2019; Middha et al., 2024).

To improve motion sensitivity, several approaches explicitly incorporated optical flow into anomaly detection pipelines. These methods either treat flow maps as direct network inputs or integrate them as auxiliary cues for shallow convolutional or reconstruction-based models (Hasan et al., 2016; Liu et al., 2018). Survey studies emphasize that optical flow-based representations provide strong local motion information and are particularly effective for crowd anomalies and sudden motion changes, but they introduce significant computational overhead and are sensitive to flow estimation errors (Duong et al., 2023; Pathirannahalage et al., 2024). These limitations motivate the exploration of lighter architectures that learn motion implicitly from RGB frames. Deep autoencoders marked a significant shift in unsupervised VAD by enabling the learning of standard spatio-temporal patterns directly from data. Convolutional autoencoders trained on standard sequences identify anomalies by increasing reconstruction error, serving as a widely used baseline in the literature (Hasan et al., 2016; Zhao et al., 2017). Prediction-based approaches further enhance temporal modeling by forecasting future frames or features and using prediction error as an anomaly score. Liu et al. (2018) demonstrated that future-frame prediction can outperform pure reconstruction on several benchmarks. Subsequent studies incorporate ConvLSTM units and hierarchical temporal modeling to capture dynamic evolution more accurately (Medel & Savakis, 2016). Surveys confirm that reconstruction- and prediction-based methods remain the dominant paradigms for unsupervised deep VAD (Abdalla et al., 2024; Wu et al., 2024). Despite their effectiveness, these models are often computationally intensive and rely on long temporal windows, complicating deployment in real-time or resource-constrained scenarios.

To address the limited discriminative power of basic autoencoders, memory-augmented architectures store representative standard patterns in external memory modules and constrain reconstructions accordingly (Gong et al., 2019). Attention mechanisms further enhance performance by focusing on salient spatial or temporal regions, improving anomaly localization (Barbalau et al., 2023; Ristea et al., 2022). Weakly supervised methods reduce labeling

costs by using video-level annotations instead of frame-level labels. Multiple-instance learning frameworks enable scalable anomaly detection on long untrimmed videos, though often at the expense of precise temporal localization (Choudhry et al., 2023; Sultani et al., 2019). Continual learning approaches have also been explored to adapt models to evolving scenes, but they introduce additional complexity and stability challenges (Doshi & Yilmaz, 2020). As surveillance systems scale up, computational efficiency and deployability have become central concerns. Several reviews highlight the importance of lightweight architectures and edge computing for real-time anomaly detection (Li et al., 2025; Noghre et al., 2025; Patrikar & Parate, 2022). MobileNet-based architectures represent a key development in this direction. MobileNetV2 introduces inverted residuals and depthwise separable convolutions to achieve an efficient accuracy–latency trade–off (Sandler et al., 2018). Subsequent studies tailor MobileNet-family models for embedded platforms such as Raspberry Pi, demonstrating their feasibility for real-time video analytics (Glegoła et al., 2021). Lightweight CNNs have also been applied to surveillance-related tasks such as violence detection and safety monitoring, showing competitive performance with significantly reduced computational cost (Suba et al., 2022).

Although existing literature covers a wide range of anomaly detection paradigms, direct comparisons under strictly identical experimental conditions remain limited. In this study, a controlled comparison is conducted on the CUHK Avenue dataset using consistent preprocessing, training protocols, and evaluation metrics. Three lightweight models representing unsupervised reconstruction, temporal prediction, and supervised classification are evaluated to isolate the impact of supervision level, temporal modeling, and architectural simplicity in a highly imbalanced surveillance setting. The summary of some related research is presented in Table 1.

Table 1. The summary of related works

Authors	Core methodology	Key contributions	Main limitations
Adam et al. (2008), Mahadevan et al. (2010), Cong et al. (2011)	Hand-crafted motion features, trajectories, statistical modelling, sparse / low-rank reconstruction	Established early benchmarks for crowd anomaly detection; introduced motion- and trajectory-based modelling of normal behaviour	Strongly dependent on manual feature design; limited representation capacity; weak generalization to complex and crowded scenes
Hasan et al. (2016), Liu et al. (2018)	Optical flow maps as inputs or auxiliary cues to CNN / AE models	Provide explicit motion information; improve sensitivity to dynamic anomalies compared to purely appearance-based features	High computational cost for flow estimation; sensitive to flow noise and camera motion; less suitable for strict real-time constraints
Hasan et al. (2016), Zhao et al. (2017)	Convolutional and spatio-temporal autoencoders trained on normal data; anomalies detected via reconstruction error	Enable unsupervised learning of normal patterns; do not require anomaly labels; form strong baselines on standard datasets	Limited discriminative power when anomalies are subtle or visually similar to normal patterns; often rely on relatively heavy architectures

End of Table 1

Authors	Core methodology	Key contributions	Main limitations
Liu et al. (2018), Medel and Savakis (2016)	Future frame prediction; temporal regularity modelled via CNN / ConvLSTM networks	Provide stronger temporal cues by explicitly modelling frame-to-frame evolution; often outperform pure reconstruction on motion-centric anomalies	Computationally more demanding; sensitive to noise, camera motion, and temporal misalignments; less attractive for lightweight edge deployment
Gong et al. (2019), Ristea et al. (2022), Doshi and Yilmaz (2020)	External memory modules, attention mechanisms, and continual learning on top of AE / predictive backbones	Improve modelling of normal patterns; reduce trivial reconstruction of anomalies; better focus on salient regions and evolving scenes	Increased architectural and training complexity; higher memory footprint; harder to tune and deploy under resource constraints
Sultani et al. (2018)	Multiple Instance Learning with video-level labels; ranking or scoring of segments within long videos	Reduce annotation cost by avoiding frame-level labels; scale better to long, untrimmed videos	Temporal localization remains coarse; performance depends on assumptions in MIL formulation; label noise can degrade detection quality
Ristea et al. (2022), Barbalau et al. (2023), Pang et al. (2022)	Self-attention over long sequences; spatio-temporal graphs for object interactions	Capture long-range temporal dependencies and structured interactions; achieve strong performance on complex benchmarks	Very high computational and memory requirements; currently impractical for lightweight, edge-oriented settings on datasets like Avenue
Sandler et al. (2018), Glegoła et al. (2021), Patrikar and Parate (2022), Suba et al. (2022)	Lightweight CNN backbones (e.g., MobileNetV2), depthwise separable convolutions, edge-focused deployment strategies	Achieve favourable accuracy–latency trade-offs; enable real-time or near real-time inference on embedded / edge devices; directly relevant to practical CCTV systems	Often evaluated in isolation or on different tasks (e.g., violence detection); limited controlled comparisons with unsupervised baselines on standard VAD datasets

3. Background of the experimental investigation

The methodology of the experimental investigation is illustrated in Figure 1 and follows a structured pipeline consisting of preprocessing, optimization, training, and evaluation stages. First, each video sequence from the CUHK Avenue dataset is decomposed into individual RGB frames, which is a common practice in frame-level video anomaly detection to enable independent analysis of visual patterns. Second, all frames are resized to a fixed spatial resolution to satisfy the input requirements of the evaluated models, namely 128×128 pixels for the autoencoder-based architectures and 224×224 pixels for the MobileNetV2 classifier, in line with prior lightweight anomaly detection and CNN-based studies (Wu et al., 2024; Sandler et al., 2018).

Third, image normalization is applied to ensure numerical stability during optimization and to align the input distribution with the assumptions of each learning paradigm. For the autoencoder-based models, pixel intensities are linearly scaled to the [0,1] range, a widely

adopted normalization strategy in reconstruction-based anomaly detection that stabilizes pixel-wise loss functions and facilitates convergence (Hasan et al., 2016; Zhao et al., 2017). Given an input image I with pixel values $I(x, y, c) \in [0,255]$, the normalized image $\hat{I}$ is computed as:

$$\hat{I}\left(x,y,c\right)=\frac{I\left(x,y,c\right)}{255}. \tag{1}$$

In contrast, the MobileNetV2 model employs the standard input preprocessing defined in its original architecture to preserve compatibility with ImageNet-pretrained weights. Specifically, each RGB channel is mapped to the [–1,1] range according to:

$$\hat{I}\left(x,y,c\right)=\frac{I\left(x,y,c\right)}{127.5}-1. \tag{2}$$

This normalization centers the input distribution around zero and matches the statistical assumptions under which the convolutional filters were originally trained, which has been shown to be critical for stable fine-tuning and optimal performance in lightweight CNNs (Sandler et al., 2018; Glegoła et al., 2021). Following preprocessing, the training subset is used to optimize model parameters by minimizing a task-specific loss function. For the autoencoder models, optimization aims to minimize the reconstruction error between the input frame I and its reconstructed output $\tilde{I}$, commonly formulated as a mean-squared error loss:

$$\mathcal{L}_{AE}=\frac{1}{N}\sum_{i=1}^{N}\| I_i-\tilde{I}_i \|_2^2, \tag{3}$$

which is standard in unsupervised video anomaly detection to capture deviations from learned normal patterns (Hasan et al., 2016; Gong et al., 2019).

For the MobileNetV2 classifier, parameter optimization is performed using binary cross-entropy loss between the predicted anomaly probability p_i and the ground-truth label y_i:

$$\mathcal{L}_{CLS}=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log\left(p_i\right)+\left(1-y_i\right)\log\left(1-p_i\right)\right], \tag{4}$$

which is appropriate for highly imbalanced binary classification problems commonly encountered in frame-level anomaly detection (Johnson & Khoshgoftaar, 2019).

Finally, the optimized models are evaluated on the testing subset using frame-level anomaly scores and classification outputs, as summarized in Figure 1. No temporal stacking or sequence aggregation is employed, as the objective of this study is strictly frame-level anomaly detection, allowing a focused comparison of learning paradigms without introducing additional temporal modeling complexity.

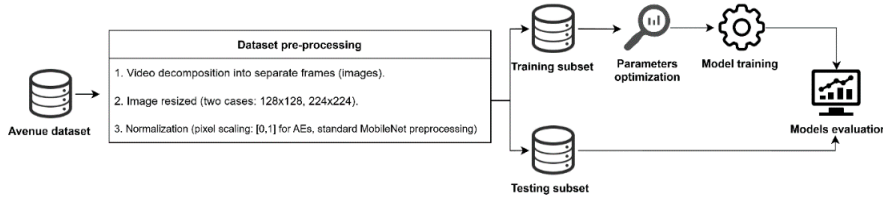

Figure 1. The methodology of experimental investigation

3.1. Dataset analysed

The experiments in this study are conducted on the Avenue dataset, a popular benchmark for video anomaly detection. The dataset contains short surveillance clips recorded in an outdoor campus environment, where anomalies include actions, such as running, throwing objects or moving in unusual directions. An important challenge is that anomalous events appear suddenly and often cover only small regions of the frame which makes detection difficult (Gnouma et al., 2018; Mahadevan et al., 2010). The dataset provides frame-level annotations through a MATLAB file that lists the exact frame intervals where anomalies occur. These intervals are short and unevenly distributed, resulting in a strong class imbalance. This issue is common in real-world surveillance datasets and can affect both reconstruction-based models and CNN classifiers (Johnson & Khoshgoftaar, 2019). Because of this imbalance, metrics such as AUC and precision–recall curves provide a more reliable evaluation than accuracy alone (Davis & Goadrich, 2006). The sample of the dataset is presented in Figure 2.

For all models, frames were resized to a fixed resolution and normalized. The autoencoder models used pixel values scaled to the [0, 1] range while MobileNetV2 used the standard preprocessing recommended in the original paper (Sandler et al., 2018). A small validation set was created from the normal training frames to monitor overfitting, following common practice in reconstruction-based anomaly detection (Wu et al., 2024).

Figure 2. Sample of the dataset used in experiments

3.2. Lightweight model architectures evaluated in this study

This section describes the three lightweight model architectures evaluated in this study for frame-level video anomaly detection on the CUHK Avenue dataset. The comparative experimental investigation was performed using three different types of models. The first model is a simple 2D Convolutional Autoencoder used as the baseline for this study. Autoencoders are widely used in video anomaly detection because they learn to reconstruct normal patterns and produce higher errors on abnormal frames (Hasan et al., 2016). The baseline model includes a small encoder–decoder structure with a few convolutional and transpose convolutional layers. It is intentionally kept simple to show the starting performance of a lightweight reconstruction model on the Avenue dataset. This model does not use temporal information.

Each frame is processed independently. Because of its shallow structure, it tends to miss subtle anomalies and often produces blurred reconstructions on fast or irregular motions. These limitations are consistent with observations in previous works using basic autoencoders for anomaly detection (Zhao et al., 2017).

The second model expands the baseline design by adding deeper convolutional blocks, larger feature maps and normalization layers. These changes help the autoencoder learn richer representations of normal frames and reduce reconstruction noise. Deeper reconstruction-based models have been shown to improve stability and reduce false positives in similar tasks (Wu et al., 2024). Although the structure is still lightweight, the enhanced model produces sharper reconstructions and responds more clearly to unusual objects or motion patterns. This improvement explains the higher AUC achieved by this model compared to the baseline autoencoder (2DConv AE, AUC ≈ 0.75 and Enhanced 2DConv AE, AUC ≈ 0.84).

The third model uses MobileNetV2 as a feature extractor followed by a small classification head. MobileNetV2 is designed for efficient inference on limited hardware and relies on inverted residual blocks with depthwise separable convolutions (Sandler et al., 2018). These design choices significantly reduce computation while preserving strong feature quality, making MobileNet suitable for lightweight anomaly detection. In this study, MobileNetV2 is fine-tuned on frame-level labels derived from the ground truth of the Avenue dataset. Because the model directly predicts whether a frame is normal or anomalous, it avoids the limitations of autoencoder reconstruction. Prior work has shown that MobileNet variants can achieve strong accuracy even on embedded devices (Glegoła et al., 2021), supporting their use as practical alternatives to heavier CNN architectures.

The selection of models in this study was guided by the objective of conducting a controlled and fair comparison of representative lightweight learning paradigms for frame-level video anomaly detection under identical experimental conditions. Rather than exhaustively evaluating all existing VAD approaches, this work focuses on isolating the impact of supervision level and architectural simplicity on detection performance.

Accordingly, three models were deliberately chosen to represent distinct yet widely adopted paradigms in the literature. The baseline 2D convolutional autoencoder serves as a canonical unsupervised reconstruction-based method and provides a commonly used reference point in video anomaly detection studies. The enhanced autoencoder extends this baseline to examine how improved feature representation within the same reconstruction paradigm affects anomaly sensitivity, without introducing fundamentally new mechanisms. Finally, the MobileNetV2-based classifier provides lightweight, supervised feature-based detection, selected for its strong accuracy–efficiency trade-off and relevance to real-time and edge-oriented deployment scenarios.

More complex architectures, such as memory-augmented models, attention-based networks, or long-term temporal sequence models, were intentionally excluded, as they introduce additional architectural complexity, higher computational overhead, and multiple confounding factors. Including such models would obscure the core objective of this study, which is to analyze the performance gap between reconstruction-based and classification-based lightweight approaches under controlled conditions. This deliberate scope restriction ensures clarity of comparison and supports reproducible conclusions.

4. Experimental investigation results

All experiments were conducted in JupyterLab using a Quadro T2000 (graphic card) as GPU. The Avenue dataset was processed by extracting every frame from the training and testing videos. Only normal frames from the training set were used to train the two autoencoder models while the MobileNetV2 classifier was trained using frame-level labels derived from the ground truth of the test set. The baseline autoencoder and the enhanced autoencoder were trained using the Mean Squared Error loss. The MobileNetV2 classifier was trained with a binary cross-entropy loss. All models were optimized using Adam with a fixed learning rate. Batch sizes were selected based on the available GPU memory to maintain stable training. Model performance was evaluated using frame-level AUC, precision–recall curves and confusion matrices. These metrics are preferred for imbalanced datasets where accuracy alone does not provide a reliable measure of the model's behavior (Davis & Goadrich, 2006). The hyperparameters for each model (ID: 1, 2, 3) training is presented in Table 2.

Table 2. The training parameters for each model

	Model 1	Model 2	Model 3
Architecture	Conv + BN + ReLU (4 layers) ConvTranspose (4 layers)	Conv + ReLU (3 layers) ConvTranspose (3 layers)	Frozen MobileNetV2 features Custom classifier head
Input Size	128×128×3	128×128×3	224×224×3
Train Data	Normal Only	Normal Only	Normal + Anomalous
Testing Data	15%	15%	20%
Batch Size	16	16	16
Epochs	40	50	50
Loss Function	MSELoss	MSELoss	BCE with LogistLoss
Optimizer	Adam, lr = 1e-3	Adam lr = le-3	Adam lr = le-3
Bottleneck Channels	256	256	_
Trainable Parameters	77.987	2.1M	2.2M
Output Type	Reconstructed frame	Reconstructed frame	Anomaly probability

Model 1 (Baseline Autoencoder) and Model 2 (Enhanced Autoencoder) were trained exclusively on normal frames. Their ability to reproduce unseen test frames forms the basis of anomaly detection. Model 1 shows reasonable reconstruction capabilities but struggles with fine spatial structures. In contrast, Model 2 produces sharper and more stable reconstructions, indicating stronger feature learning. For both autoencoders, reconstruction error is consistently higher for anomalous frames than for normal frames, confirming that the models successfully capture the regular motion patterns present during training. Reconstruction-based error distributions were compared for normal and anomalous frames in Models 1 and 2. Normal frames show a compact distribution with low error values, while anomalous frames exhibit wider ranges and higher mean errors. Although both models can

separate normal from anomalous behavior, Model 2 provides clearer separation and more stable error patterns than Model 1. The sample of reconstruction is presented in Figure 3.

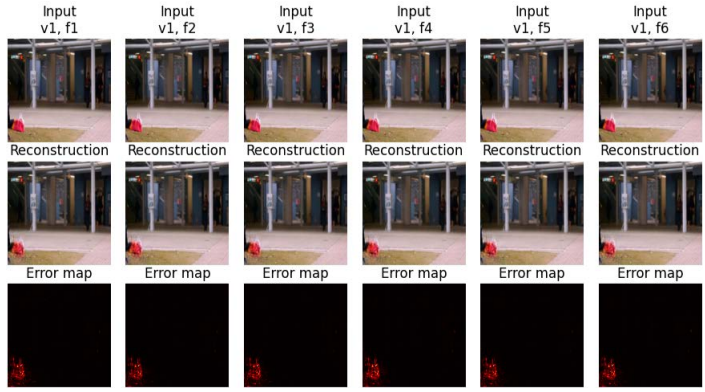

Figure 3. Sample input frames, reconstructions and error maps produced by the autoencoder models

Unlike the autoencoder-based models Model 3 (MobileNetV2) is a supervised classifier trained directly on labeled normal and anomalous frames. Using ImageNet-pretrained weights, class-balancing and lightweight fully connected layers, this model achieved the strongest performance among all models. MobileNetV2 demonstrates excellent discrimination between normal and anomalous frames, producing a steep ROC curve and highly confident predictions. The model generalizes well despite class imbalance, primarily due to its robust feature extraction and balanced training strategy. As we can see in Figure 4, the fame-level AUC values clearly highlight the performance differences.

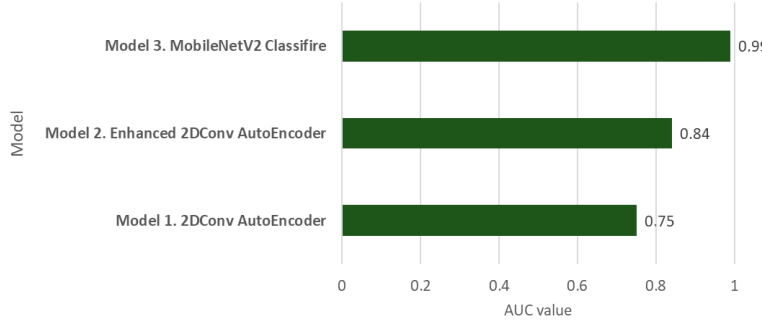

Figure 4. Frame-level AUC comparison across all models

The improvements from Model 1 to Model 2 reflect the benefits of deeper feature extraction and more stable reconstruction (Figure 5). Model 3's significant lead demonstrates the effectiveness of supervised learning for this dataset when labels are available.

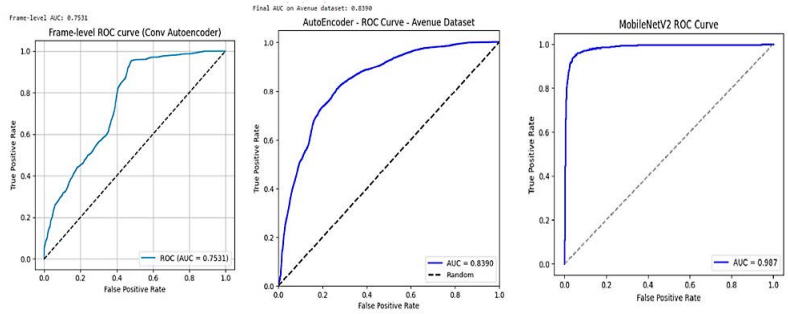

Figure 5. ROC curves for all three models on the Avenue dataset

In Figure 6, the confusion matrix shows that the model correctly identified 985 out of 1,036 anomalous frames (TP), while misclassifying only 51 anomalies as normal (FN). For normal frames, it produced 1,923 true negatives with 106 false positives. This balance between Type-I and Type-II errors indicates a stable and reliable decision boundary. The quantitative metrics further confirm this behavior: MobileNetV2 reached an accuracy of 0.9488, precision of 0.9028, recall of 0.9508, and an F1-score of 0.9262. Its AUC of 0.9865 demonstrates excellent separability between normal and anomalous behavior, outperforming both baseline models by a clear margin. These results can be attributed to MobileNetV2's efficient architecture which uses inverted residual blocks and depthwise separable convolutions to extract expressive features while maintaining low computational cost. Additional design choices in this study such as using pretrained ImageNet weights, standard RGB normalization, 224×224 input resolution, and class-imbalance weighting further improved the model's robustness.

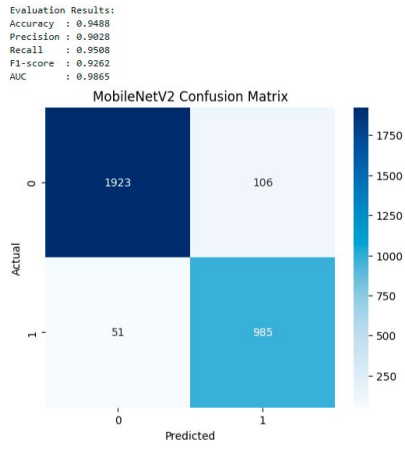

Figure 6. Confusion matrix for the MobileNetV2 classifier on the Avenue dataset

5. Conclusions

This study compared three lightweight deep-learning models for frame-level anomaly detection on the Avenue dataset: a simple 2D convolutional autoencoder, an optimized reconstruction-based autoencoder and a MobileNetV2-based classifier. The results highlight clear differences in how each model interprets visual anomalies and how architectural choices shape detection performance. The simple convolutional autoencoder, although computationally efficient, achieved only moderate discrimination (AUC ≈ 0.75), confirming the limitations of shallow reconstruction-based models in capturing the diverse and subtle motion patterns present in real-scene surveillance footage. The optimized autoencoder showed better reconstruction stability and improved anomaly separation (AUC ≈ 0.84), suggesting that deeper encoding layers and more expressive feature extraction can reduce error-map ambiguity and enhance anomaly sensitivity.

Among all models, MobileNetV2 demonstrated the highest and most stable performance (AUC ≈ 0.99), supported by strong precision, recall, and F1-score values. Its confusion matrix revealed balanced detection with both false positives and false negatives substantially lower than in the reconstruction-based approaches. These results indicate that, for frame-wise anomaly detection without temporal modeling, feature-based classification with lightweight pretrained architectures can outperform pixel-space reconstruction methods, especially when the dataset contains diverse lighting conditions, cluttered backgrounds, and subtle anomaly patterns. Overall, the results indicate that MobileNetV2 can serve as a strong and practical baseline for lightweight anomaly detection in datasets that share the characteristics of the Avenue dataset.

So, the gap between reconstruction-based and classification-based performance highlights opportunities for future research using hybrid models, temporal reasoning, or self-supervised feature learning.

References

Abdalla, M., Javed, S., Radi, M. Al, Ulhaq, A., & Werghi, N. (2024). *Video anomaly detection in 10 years*: *A survey and Outlook*. arXiv. https://doi.org/10.48550/arXiv.2405.19387

Adam, A., Rivlin, E., Shimshoni, I., & Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(3), 555–560. https://doi.org/10.1109/TPAMI.2007.70825

Anoopa, S., & Salim, A. (2022). Survey on anomaly detection in surveillance videos. *Materials Today*: *Proceedings*, *58*, 162–167. https://doi.org/10.1016/j.matpr.2022.01.171

Barbalau, A., Ionescu, R. T., Georgescu, M.-I., Dueholm, J., Ramachandra, B., Nasrollahi, K., Khan, F. S., Moeslund, T. B., & Shah, M. (2023). *SSMTL++*: *Revisiting self-supervised multi-task learning for video anomaly detection*. arXiv. https://doi.org/10.48550/arXiv.2207.08003

Choudhry, N., Abawajy, J., Huda, S., & Rao, I. (2023). A comprehensive survey of machine learning methods for surveillance videos anomaly detection. *IEEE Access*, *11*, 114680–114713. https://doi.org/10.1109/ACCESS.2023.3321800

Cong, Y., Yuan, J., & Liu, J. (2011). Sparse reconstruction cost for abnormal event detection. In *CVPR 2011* (pp. 3449–3456). IEEE. https://doi.org/10.1109/CVPR.2011.5995434

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning – ICML '0*, (pp. 233–240). Association for Computing Machinery. https://doi.org/10.1145/1143844.1143874

Doshi, K., & Yilmaz, Y. (2020). Continual learning for anomaly detection in surveillance videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*) (pp. 1025–1034). IEEE. https://doi.org/10.1109/CVPRW50498.2020.00135

Duong, H.-T., Le, V.-T., & Hoang, V. T. (2023). Deep learning-based anomaly detection in video surveillance: A survey. *Sensors*, *23*(11), Article 5024. https://doi.org/10.3390/s23115024

Fernandes, G., Rodrigues, J. J. P. C., Carvalho, L. F., Al-Muhtadi, J. F., & Proença, M. L. (2019). A comprehensive survey on network anomaly detection. *Telecommunication Systems*, *70*(3), 447–489. https://doi.org/10.1007/s11235-018-0475-8

Glegoła, W., Karpus, A., & Przybyłek, A. (2021). MobileNet family tailored for Raspberry Pi. *Procedia Computer Science*, *192*, 2249–2258. https://doi.org/10.1016/j.procs.2021.08.238

Gnouma, M., Ejbali, R., & Zaied, M. (2018). Abnormal events' detection in crowded scenes. *Multimedia Tools and Applications*, *77*(19), 24843–24864. https://doi.org/10.1007/s11042-018-5701-6

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., & Van Den Hengel, A. (2019). Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *2019 IEEE/CVF International Conference on Computer Vision* (*ICCV*) (pp. 1705–1714). IEEE. https://doi.org/10.1109/ICCV.2019.00179

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 733–742). IEEE. https://doi.org/10.1109/CVPR.2016.86

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), Article 27. https://doi.org/10.1186/s40537-019-0192-5

Keleko Teguede, A. (2022). *Deep Learning for anomaly detection in industry 4.0* [Doctoral thesis]. Institut National Polytechnique de Toulouse. https://theses.hal.science/tel-04248257v1

Li, Z., Yan, Y., Wang, X., Ge, Y., & Meng, L. (2025). A survey of deep learning for industrial visual anomaly detection. *Artificial Intelligence Review*, *58*(9), Article 279. https://doi.org/10.1007/s10462-025-11287-7

Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6536–6545). IEEE. https://doi.org/10.1109/CVPR.2018.00684

Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 1975–1981). IEEE. https://doi.org/10.1109/CVPR.2010.5539872

Medel, J. R., & Savakis, A. (2016). *Anomaly detection in video using predictive convolutional long short-term memory networks*. arXiv. https://doi.org/10.48550/arXiv.1612.0039

Middha K, Goyal S, Malhotra A, & Jain N. (2024). Anomaly detection in CCTV surveillance. *International Journal for Multidisciplinary Research*, *6*(1). https://doi.org/10.36948/ijfmr.2024.v06i01.12750

Noghre, G. A., Pazho, A. D., & Tabkhi, H. (2025). *A survey on video anomaly detection via deep learning*: *Human*, *vehicle*, *and environment*. arXiv. https://doi.org/10.48550/arXiv.2508.14203

Pang, G., Shen, C., Cao, L., & Hengel, A. Van Den. (2022). Deep learning for anomaly detection. *ACM Computing Surveys*, *54*(2), 1–38. https://doi.org/10.1145/3439950

Pathirannahalage, I., Jayasooriya, V., Samarabandu, J., & Subasinghe, A. (2024). A comprehensive analysis of real-time video anomaly detection methods for human and vehicular movement. *Multimedia Tools and Applications*, *84*(10), 7519–7564. https://doi.org/10.1007/s11042-024-19204-w

Patrikar, D. R., & Parate, M. R. (2022). Anomaly detection using edge computing in video surveillance system: review. *International Journal of Multimedia Information Retrieval*, *11*(2), 85–110. https://doi.org/10.1007/s13735-022-00227-8

Ristea, N. C., Madan, N., Ionescu, R. T., Nasrollahi, K., Khan, F. S., Moeslund, T. B., & Shah, M. (2022). Self-supervised predictive convolutional attentive block for anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 13566–13576). IEEE. https://doi.org/10.1109/CVPR52688.2022.01321

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520). IEEE. https://doi.org/10.1109/CVPR.2018.00474

Suba, N., Verma, A., Baviskar, P., & Varma, S. (2022). Violence detection for surveillance systems using lightweight CNN models. In *7th International Conference on Computing in Engineering & Technology* (*ICCET 2022*) (pp. 23–29). IET. https://doi.org/10.1049/icp.2022.0587

Sultani, W., Chen, C., & Shah, M. (2019). *Real-world anomaly detection in surveillance videos*. arXiv. https://doi.org/10.48550/arXiv.1801.04264

Wu, P., Pan, C., Yan, Y., Pang, G., Wang, P., & Zhang, Y. (2024). *Deep learning for video anomaly detection*: *A review*. arXiv. https://doi.org/10.48550/arXiv.2409.05383

Yadav, R. K., & Kumar, R. (2022). A survey on video anomaly detection. In *2022 IEEE Delhi Section Conference* (*DELCON*) (pp. 1–5). https://doi.org/10.1109/DELCON54057.2022.9753580

Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., & Hua, X. S. (2017). Spatio-temporal AutoEncoder for video anomaly detection. In *MM 2017 – Proceedings of the 2017 ACM Multimedia Conference* (pp. 1933–1941). Association for Computing Machinery. https://doi.org/10.1145/3123266.3123451