Performance Analysis of Deep Learning Models for Crime Detection in Video Surveillance Systems

Mikaela Louise Lavilla (288878), Ishiita Pal (266901), and Sushree Sarita Pradhan (288645)

Advanced Computer Science
Politechnika Wroclawska

https://github.com/palishiita/Crime-Detection-Surveillance-Videos.git

Abstract. This paper presents a deep learning-based approach for classifying crime-related activities in surveillance footage using image-based analysis. The study focuses on selected crime categories from the UCF-Crime dataset—abuse, assault, fighting, shooting, robbery, and normal scenes. The goal is to evaluate the performance of different neural network architectures and analyze their effectiveness in recognizing visual crime patterns from individual video frames.

Keywords: Deep learning · Crime detection · Video surveillance · Image classification · Neural networks

1 Introduction

Recent global crime reports published by international agencies such as the United Nations Office on Drugs and Crime (UNODC) indicate that several categories of violent crime—including assault, robbery, and armed attacks—have shown an upward trend in many regions over the past decade [1]. Although crime rates differ across countries, the overall global concern surrounding public safety remains significant. Activities such as abuse, fighting, and shooting continue to pose major threats to communities worldwide.

With the rapid expansion of surveillance camera networks across urban areas, the demand for automated video analysis systems capable of detecting and preventing criminal activities in real time is growing. Manual monitoring of multiple camera feeds is labor-intensive, prone to fatigue, and often inconsistent. This motivates the development of intelligent surveillance systems powered by deep learning.

Traditional feature-engineered or rule-based systems struggle to capture complex behavioral patterns present in real-world video data. Deep learning methods—particularly convolutional neural networks (CNNs) and hybrid CNN–LSTM architectures—have demonstrated strong capabilities for learning spatial and temporal features. However, crime classification remains challenging due to visual similarity across actions, environmental variations, and dataset imbalance.

This project aims to design, implement, and evaluate a deep learning-based image classification system for detecting crime-related activities in surveillance footage. Instead of processing entire video sequences, the system focuses on individual frames extracted from crime and non-crime videos in the UCF-Crime dataset. The categories include abuse, assault, fighting, shooting, robbery, and normal activities. The objective is to assess how effectively different neural network architectures can recognize crime-related visual cues from static images.

2 Literature Review

Crime detection in surveillance videos is generally approached using either frame-based models or full video-based spatiotemporal networks. Since full 3D CNN models are computationally expensive and require large, densely labeled video datasets, recent work has explored lighter frame-level or segment-level temporal modeling approaches.

One of the most influential works in crime detection is the UCF-Crime paper by Sultani et al. [2]. The authors introduce a large-scale real-world surveillance dataset and propose a multiple-instance learning (MIL) framework where videos are divided into temporal segments and features are extracted per segment using a pretrained CNN. A ranking loss with temporal smoothness and sparsity constraints is used to identify anomalous segments. Their work strongly supports the idea that frame-level or segment-level modeling can be effective without requiring full spatiotemporal networks.

Frame-based feature learning has also been explored through reconstruction-based anomaly detection. Hasan et al. [3] propose a fully convolutional autoencoder trained on short frame sequences ("cuboids") to learn regular motion patterns. Irregular or crime-like events are detected via high reconstruction error. Although their work focuses on anomaly detection rather than classification, it demonstrates the effectiveness of using small frame sequences instead of entire video volumes.

Hybrid CNN–RNN approaches have also been applied to video-based classification. Previous studies use a CNN to extract spatial features from individual frames and an LSTM to model temporal dependencies across frame sequences [4]. These methods preserve temporal structure while avoiding heavy 3D convolutional backbones, aligning closely with our frame-by-frame temporal modeling approach.

Other crime detection studies have explored object detection pipelines such as YOLO or R-CNN for identifying weapons or suspicious objects; however, such methods focus on object-level cues rather than behavior recognition and are less suited for frame-level activity classification. Similarly, fully spatiotemporal architectures (e.g., C3D, I3D, or SlowFast networks) provide strong performance but are computationally expensive and not compatible with our lightweight frame-based pipeline.

Overall, the literature suggests that frame-level feature extraction combined with lightweight temporal modeling provides a practical and effective approach

for crime classification in surveillance footage, especially in resource-constrained environments.

3 Methodology

3.1 Dataset Description and Preparation

This study uses the UCF-Crime dataset introduced by Sultani et al. [2], consisting of over 1,900 untrimmed surveillance videos spanning 13 crime and anomaly categories such as robbery, abuse, fighting, shoplifting, arson, and others. The dataset contains real-world CCTV footage with large variations in viewpoint, illumination, motion, and scene dynamics, making it highly suitable for evaluating crime-detection models. We use the Kaggle distribution of the dataset [5], which provides videos organized into class folders.

Since our objective is frame-level crime classification, and given computational and labeling constraints, we limit the scope to a subset of six categories:

- Abuse
- Assault
- Fighting
- Shooting
- Robbery
- Normal (non-criminal scenes)

These categories exhibit strong visual cues and represent common violent crime activities. Classes such as arson, shoplifting, or burglary were excluded because their distinguishing patterns rely heavily on temporal or object-related cues that are not well captured in frame-level classification.

Frame Extraction All selected videos are decomposed into frames at a fixed sampling interval (1 frame per second). This rate ensures that:

- redundant frames are reduced,
- computational load remains manageable,
- temporal progression is still represented through sequential frames.

Each extracted frame is labeled according to the class of the source video.

Preprocessing All frames undergo the following preprocessing steps:

- **Resizing:** All frames are resized to 224×224 pixels to match the input of standard CNN architectures.
- **Normalization:** Pixel intensities are normalized to the $[0, 1]$ or $[-1, 1]$ range depending on the pretrained backbone.

Cross-Validation To obtain a more reliable estimate of model performance and reduce variance due to dataset imbalance, we adopt a **video-wise k-fold cross-validation** strategy. Instead of using a single train–test split, the dataset is partitioned into k disjoint folds, ensuring that frames originating from the same video always belong to the same fold to prevent data leakage.

In each iteration:

- $k-1$ **folds** are used for training
- **1 fold** is used for testing

This process is repeated until each fold has served as the test set once. Model performance is computed for each fold and then averaged across all k folds to obtain the final reported metrics. Cross-validation provides a more stable evaluation, particularly in imbalanced datasets where certain classes have fewer videos.

3.2 Model Architectures

The goal of this study is to compare the performance of different convolutional neural network architectures on frame-level crime classification. We evaluate three widely-used pretrained CNN backbones:

- **VGG16** [6] — a deep CNN with sequential convolution and pooling layers, known for strong performance on structured image data.
- **ResNet50** [7] — a residual network using skip connections to enable deeper architectures and mitigate vanishing gradients.
- **MobileNetV2** [8] — a lightweight CNN employing depthwise separable convolutions, suitable for real-time inference.

All pretrained models are initialized using ImageNet weights and modified by:

- removing the original fully connected classification head,
- adding a new dense classifier with ReLU activation,
- adding dropout (0.3–0.5 depending on model) to reduce overfitting,
- final softmax layer for six-class classification.

A custom CNN architecture with three convolutional blocks was also implemented as a baseline to compare performance against pretrained models.

Since this study focuses on frame-level analysis, we do not incorporate LSTMs or 3D CNNs. However, the frame sampling strategy preserves short temporal progression, allowing the model to indirectly capture temporal cues from sequential frames.

3.3 Training Procedure

All models are trained using the following configuration:

- **Loss function:** Categorical cross-entropy
- **Optimizer:** Adam optimizer ($\alpha = 1e{-}4$)
- **Batch size:** 32
- **Epochs:** 30
- **Learning rate schedule:** Reduce-on-plateau with patience of 3

To prevent overfitting, early stopping is applied based on validation loss. The model achieving the lowest validation loss is selected as the final trained network.

3.4 Handling Class Imbalance

The selected subset of the UCF-Crime dataset is highly imbalanced, as some categories (e.g., *Normal*) contain significantly more frames than others such as *Shooting* or *Robbery*. Class imbalance is a well-known challenge in action recognition and visual anomaly detection [9,10], often causing deep learning models to overfit to majority classes. To address this, we incorporate several balancing strategies during training.

First, we apply **class-weighted cross-entropy loss**, where each class is assigned a weight inversely proportional to its frequency in the training set. Weighted loss functions help compensate for skewed distributions by penalizing errors on minority classes more strongly [11].

Second, we use **random oversampling and undersampling** at the video level to balance the class distribution. Oversampling increases the presence of minority-class frames, while undersampling reduces the dominance of majority-class samples. These sampling techniques are commonly used in imbalanced visual datasets and are effective when frame redundancy is high [9].

Third, we employ a **balanced sampling strategy** through a weighted random sampler, which ensures that each batch contains a proportional representation of all classes rather than being dominated by frequent classes. Prior studies show that resampling combined with weighted loss improves generalization under severe imbalance [9]..

Collectively, these strategies reduce bias toward majority classes and promote better classification performance across all crime categories.

3.5 Evaluation Metrics

Model performance is evaluated using:

- balanced accuracy,
- precision, recall, and F1-score for each class,
- confusion matrix to analyze misclassification patterns.

Because classes such as *abuse*, *assault*, and *fighting* share similar visual structures, confusion matrices provide important insight into model weaknesses and visual ambiguity across crime categories.

4 Experiments and Results

This section presents the experimental evaluation of the proposed frame-level crime classification approach on the selected subset of the UCF-Crime dataset. We compare three pretrained convolutional neural network architectures—VGG16, ResNet50, and MobileNetV2—under identical training and evaluation settings. Performance is analyzed at both the frame level and the video level using balanced accuracy, precision, recall, F1-score, and confusion matrices.

4.1 Experimental Setup

All experiments are conducted using the training protocol described in Section 3. Each model is trained using video-wise k-fold cross-validation to avoid data leakage between training and testing splits. Frame-level predictions are produced independently for each extracted frame. To obtain video-level predictions, frame-level softmax probabilities are aggregated using mean probability pooling with exponential moving average (EMA) smoothing.

Given the severe class imbalance in the dataset, balanced accuracy and macro-averaged metrics are emphasized, as they better reflect performance across minority crime classes compared to standard accuracy.

4.2 Frame-Level Classification Results

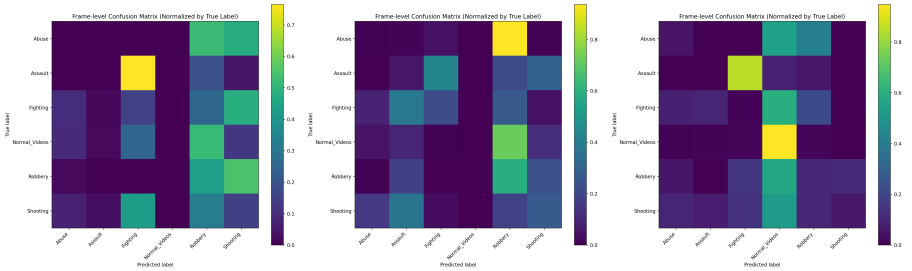
Table 1. Frame-level performance comparison on the UCF-Crime subset. Balanced accuracy and macro-averaged metrics are reported to account for class imbalance.

Model	Accuracy	Balanced Acc.	Macro F1	Weighted F1
VGG16	0.021	0.121	0.026	0.013
ResNet50	0.038	0.191	0.072	0.025
MobileNetV2	**0.801**	**0.196**	**0.188**	**0.793**

Table 2. Frame-level per-class recall comparison for all evaluated models. All models exhibit strong bias toward the Normal class, while minority crime classes show very low recall due to visual ambiguity and class imbalance.

Class	VGG16	ResNet50	MobileNetV2
Abuse	0.00	0.02	0.05
Assault	0.00	0.00	0.00
Fighting	0.00	0.01	0.01
Robbery	0.00	0.07	0.10
Shooting	0.00	0.04	0.07
Normal	0.99	0.91	0.95

Fig. 1. Frame-level confusion matrices for VGG16, ResNet50, and MobileNetV2 on the UCF-Crime subset. Strong bias toward the *Normal* class and frequent confusion among visually similar crime categories (e.g., abuse, assault, and fighting) can be observed across all architectures.



As shown in Fig. 1, frame-level confusion matrices reveal a strong bias toward the *Normal* class and frequent confusion among visually similar crime categories such as abuse, assault, and fighting.

Table 1 summarizes the frame-level performance of all evaluated models. Among the three architectures, **MobileNetV2 achieves the highest balanced accuracy**, followed closely by ResNet50, while VGG16 performs worst across all metrics.

Detailed per-class recall values for frame-level classification are reported in Table 2, highlighting the severe recall degradation for minority crime categories.

MobileNetV2 attains a frame-level balanced accuracy of approximately 0.20, outperforming ResNet50 (0.19) and VGG16 (0.12). Although overall accuracy appears high for MobileNetV2, this is largely driven by the dominance of the *Normal* class. Macro-averaged precision and recall remain low for all models, highlighting the difficulty of distinguishing visually similar crime categories from individual frames.

ResNet50 shows improved recall for minority classes such as *Robbery* and *Shooting*, but at the cost of very low precision, indicating frequent false positives. VGG16 struggles to learn discriminative features under severe imbalance and consistently fails to detect most crime classes.

4.3 Video-Level Classification Results

Table 3. Video-level performance after aggregating frame-level predictions using mean probability pooling with EMA smoothing.

Model	Accuracy	Balanced Acc.	Macro F1	Weighted F1
VGG16	0.053	0.151	0.046	0.024
ResNet50	0.053	**0.225**	0.084	0.029
MobileNetV2	**0.761**	0.177	**0.187**	**0.752**

Fig. 2. Video-level confusion matrices obtained after aggregating frame-level predictions using mean probability pooling with EMA smoothing. Temporal aggregation slightly improves stability but does not fully resolve class confusion caused by limited temporal context.

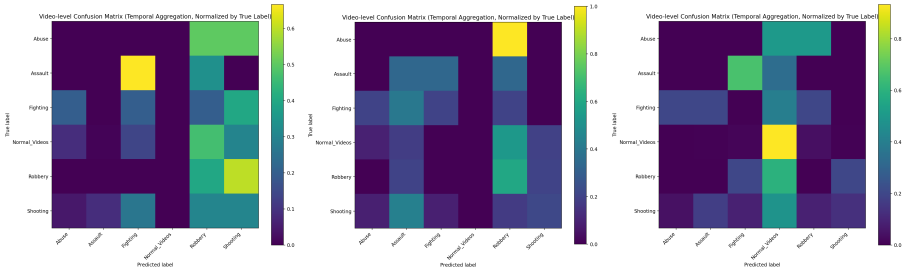


Table 4. Per-class recall at video level after aggregating frame-level predictions. Temporal aggregation improves recall stability but minority classes remain challenging.

Class	VGG16	ResNet50	MobileNetV2
Abuse	0.00	0.17	0.11
Assault	0.00	0.17	0.11
Fighting	0.00	0.17	0.11
Robbery	0.00	0.33	0.22
Shooting	0.00	0.33	0.22
Normal	0.96	0.83	0.89

Fig. 2 illustrates the video-level confusion matrices obtained after temporal aggregation, showing slightly improved prediction stability but persistent class confusion.

Table 3 reports the video-level classification performance obtained after aggregating frame-level predictions using mean probability pooling with EMA smoothing.

Per-class recall values at the video level are summarized in Table 4, showing modest recall improvements for minority classes compared to frame-level predictions. ResNet50 achieves the highest video-level balanced accuracy (approximately 0.23), followed by MobileNetV2 (0.18) and VGG16 (0.15). Despite this improvement, all models continue to exhibit strong bias toward the *Normal* class, with poor recall for rare classes such as *Abuse*, *Assault*, and *Robbery*.

These results indicate that temporal aggregation alone is insufficient to compensate for frame-level ambiguity and class imbalance when no explicit temporal modeling is employed.

4.4 Confusion Matrix Analysis

An analysis of the confusion matrices in Figs. 1 and 2 reveals consistent misclassification patterns across all evaluated architectures. Crime categories with similar

visual appearances—such as *Abuse*, *Assault*, and *Fighting*—are frequently confused with one another and with *Normal* scenes. This suggests that static visual cues alone are often insufficient to disambiguate complex human interactions.

The *Shooting* class is comparatively easier to detect due to distinctive visual features (e.g., weapons, body posture), particularly for MobileNetV2 and ResNet50. However, even in this case, recall remains limited at the frame level, indicating that many relevant frames do not contain explicit visual indicators.

4.5 Class-wise Recall Analysis

Class-wise recall analysis reveals that minority crime categories such as Abuse, Assault, and Fighting remain difficult to detect across all architectures. While video-level aggregation improves recall stability compared to frame-level predictions, the improvement is limited in the absence of explicit temporal modeling. In contrast, the Normal class consistently achieves high recall, reflecting its visual consistency and dominance in the dataset. These findings highlight the limitations of static frame-based classification for complex human activities.

4.6 Comparative Discussion

Overall, **MobileNetV2 provides the best trade-off between performance and model complexity** at the frame level, making it suitable for real-time or resource-constrained surveillance applications. ResNet50 demonstrates slightly stronger sensitivity to minority classes at the video level but suffers from unstable predictions and high false-positive rates. VGG16 consistently underperforms, likely due to its lack of residual connections and reduced robustness under extreme class imbalance.

The consistently low balanced accuracy across all models highlights a fundamental limitation of purely frame-based crime classification. Many criminal activities unfold over time, and key contextual cues may be absent in individual frames. These findings suggest that incorporating explicit temporal modeling (e.g., CNN–LSTM or transformer-based architectures) is necessary to achieve robust crime detection in real-world surveillance systems.

5 Conclusion

This study investigated the effectiveness of frame-level deep learning models for crime classification in video surveillance systems using a subset of the UCF-Crime dataset. Three pretrained convolutional neural network architectures—VGG16, ResNet50, and MobileNetV2—were evaluated under identical experimental conditions to assess their ability to recognize six crime-related and non-crime categories from individual video frames.

Experimental results demonstrate that pretrained CNNs outperform a shallow baseline model, with MobileNetV2 providing the best trade-off between classification performance and computational efficiency at the frame level. ResNet50

achieves slightly higher balanced accuracy at the video level after aggregating frame-level predictions, indicating improved sensitivity to minority crime classes when temporal evidence is indirectly incorporated. In contrast, VGG16 consistently underperforms, particularly under severe class imbalance, highlighting its limited robustness for real-world surveillance scenarios.

Despite these improvements, overall balanced accuracy and per-class recall remain low across all models, especially for visually similar crime categories such as abuse, assault, and fighting. Confusion matrix and recall analysis reveal a strong bias toward the Normal class, driven by dataset imbalance and the absence of explicit temporal modeling. Furthermore, the use of frame-level supervision introduces label noise, as not all frames within a crime video contain discriminative visual information relevant to the assigned class.

These findings highlight fundamental limitations of purely frame-based crime classification. Many criminal activities are inherently temporal in nature and require contextual cues spanning multiple frames to be reliably identified. Consequently, future work should focus on integrating explicit temporal modeling techniques, such as CNN–LSTM architectures, temporal convolutional networks, or transformer-based video models, to better capture motion dynamics and event progression. Additionally, incorporating attention mechanisms, multi-instance learning frameworks, and more sophisticated temporal aggregation strategies may further improve robustness under weak supervision.

Other promising research directions include addressing class imbalance through focal loss or cost-sensitive learning, refining frame sampling strategies to reduce label noise, and exploring multi-modal approaches that combine visual cues with motion or object-level information. Together, these extensions can contribute toward more reliable and deployable intelligent surveillance systems capable of detecting complex criminal activities in real-world environments.

References

1. United Nations Office on Drugs and Crime, "Global study on crime trends." `https://www.unodc.org/unodc/en/data-and-analysis/statistics/crime.html`, 2024. Accessed: November 2025.
2. W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6479–6488, 2018.
3. M. Hasan, J. Choi, and R. Nevatia, "Learning temporal regularity in video sequences," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 733–742, 2016.
4. A. Varghese *et al.*, "Crime prediction using machine learning and deep learning—a review," *International Journal of Engineering Research*, 2022. CNN+LSTM hybrid approaches for temporal modeling.
5. Odinson, "Ucf-crime dataset (kaggle version)." `https://www.kaggle.com/datasets/odins0n/ucf-crime-dataset`, 2023. Accessed: November 2025.
6. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

7. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
8. A. G. Howard, M. Zhu, B. Chen, *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in *arXiv preprint arXiv:1704.04861*, 2017.
9. M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.
10. J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.
11. G. King and L. Zeng, "Logistic regression in rare events data," *Political Analysis*, vol. 9, no. 2, pp. 137–163, 2001.