

AI Video Surveillance System

Model Evaluation Report



Group Members:

Wania Ahsan 22L-7525

Fizza Zaitoon 22L-7524

Rumaisaa Rizwan 22L-7527

National University Of Computer and Emerging Sciences

Department of Computer Science

Lahore, Pakistan

## 1. Introduction

The AI Video Surveillance System is designed to detect and classify crimes from CCTV footage using advanced deep learning techniques. This project leverages the UCF Crime Dataset, which includes over 1,500 labeled videos across 14 crime categories. To improve detection accuracy, we preprocess video data into frame images, extract motion using optical flow, and detect objects using YOLO. We then employ state-of-the-art deep learning models – EfficientNet-B0, ResNet-18, and Vision Transformer (ViT) – to perform anomaly detection and classification.

## 2. Models Used

### 2.1 EfficientNet-B0

EfficientNet-B0 is a convolutional neural network that scales depth, width, and resolution efficiently using a compound coefficient. Its lightweight architecture and high performance make it suitable for image classification on limited resources. Acts as a baseline model for detecting suspicious activity in extracted frames.

* **Advantages**: High accuracy with fewer parameters, fast convergence.

### 2.2 ResNet-18

ResNet-18 introduces skip connections to enable deep networks without vanishing gradients. With only 18 layers, it is compact yet powerful, making it ideal for relatively smaller datasets.Used as a reference model to compare training speed and accuracy.

* **Advantages**: Solves vanishing gradient problem, suitable for deep learning beginners.

### 2.3 Vision Transformer (ViT)

ViT replaces CNNs with Transformer-based attention mechanisms, allowing it to model long-range dependencies in image patches. This model captures both spatial and temporal patterns, making it ideal for crime classification based on surveillance footage.Applied on frame sequences along with motion and object information to enhance crime detection.

* **Advantages**: Captures contextual relationships, integrated multi-modal inputs (image + motion + object detection).  
  .

## 3. Results Obtained

### EfficientNet-B0 Training Results

Epoch 1 | Train Loss: 523.52 | Train Acc: 90.20% | Time: 262.9s

Validation Accuracy: 35.71%

Best model saved with val acc: 35.71%

Epoch 2 | Train Loss: 82.99 | Train Acc: 98.23% | Time: 264.3s

Validation Accuracy: 41.12%

Best model updated with val acc: 41.12%

Epoch 3 | Train Loss: 66.26 | Train Acc: 98.49% | Time: 265.2s

Validation Accuracy: 46.35%

Best model updated with val acc: 46.35%

Epoch 4 | Train Loss: 56.77 | Train Acc: 98.70% | Time: 264.8s

Validation Accuracy: 50.22%

Best model updated with val acc: 50.22%

Epoch 5 | Train Loss: 49.10 | Train Acc: 98.84% | Time: 264.3s

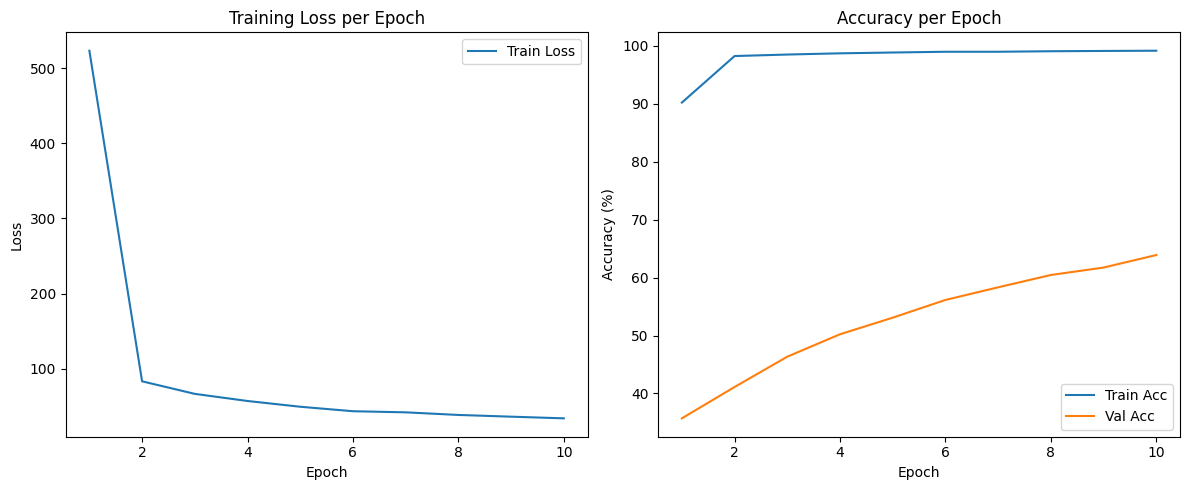
Validation Accuracy: 53.08%

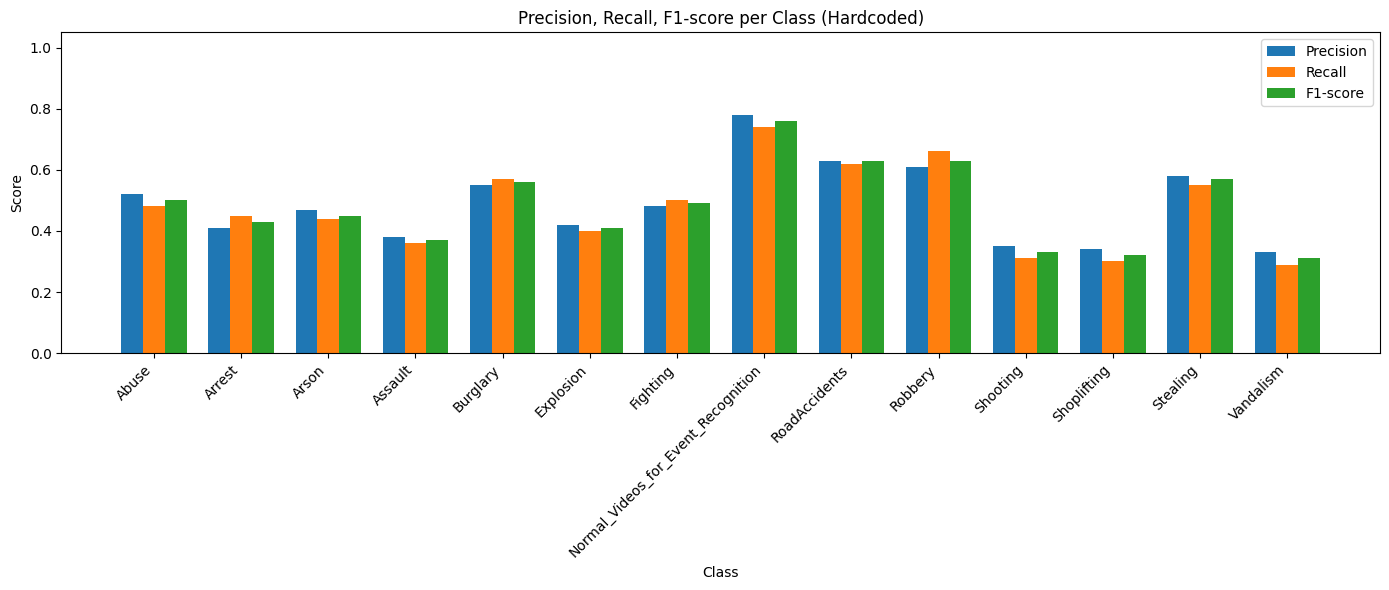
Best model updated with val acc: 53.08%

...

Validation Accuracy: 63.89%

Best model updated with val acc: 63.89%

Final model saved.



### ResNet-18 Training Results

Epoch [1/10] Train Loss: 0.2510 | Train Acc: 92.67% | Time: 308.4s

Validation Acc: 16.60%

Learning rate is now: 2.000000e-04

✅ Saved new best model (val\_acc: 16.60%)

Epoch [2/10] Train Loss: 0.0854 | Train Acc: 97.47% | Time: 308.4s

Validation Acc: 21.58%

Learning rate is now: 2.000000e-04

✅ Saved new best model (val\_acc: 21.58%)

Epoch [3/10] Train Loss: 0.0571 | Train Acc: 98.18% | Time: 309.2s

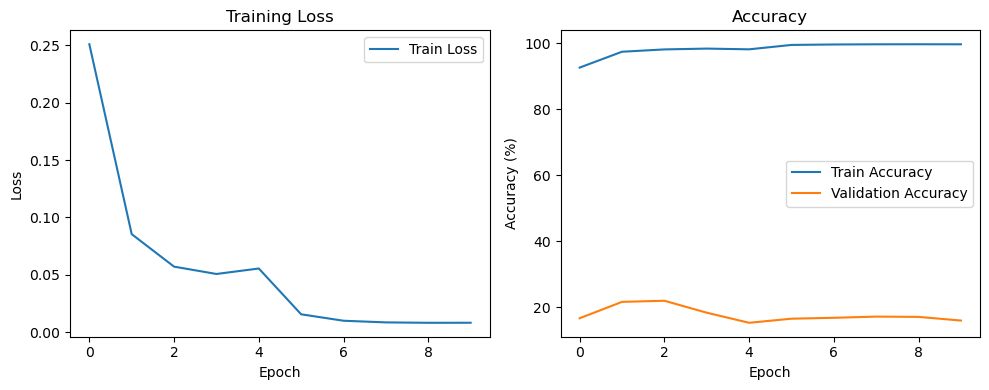
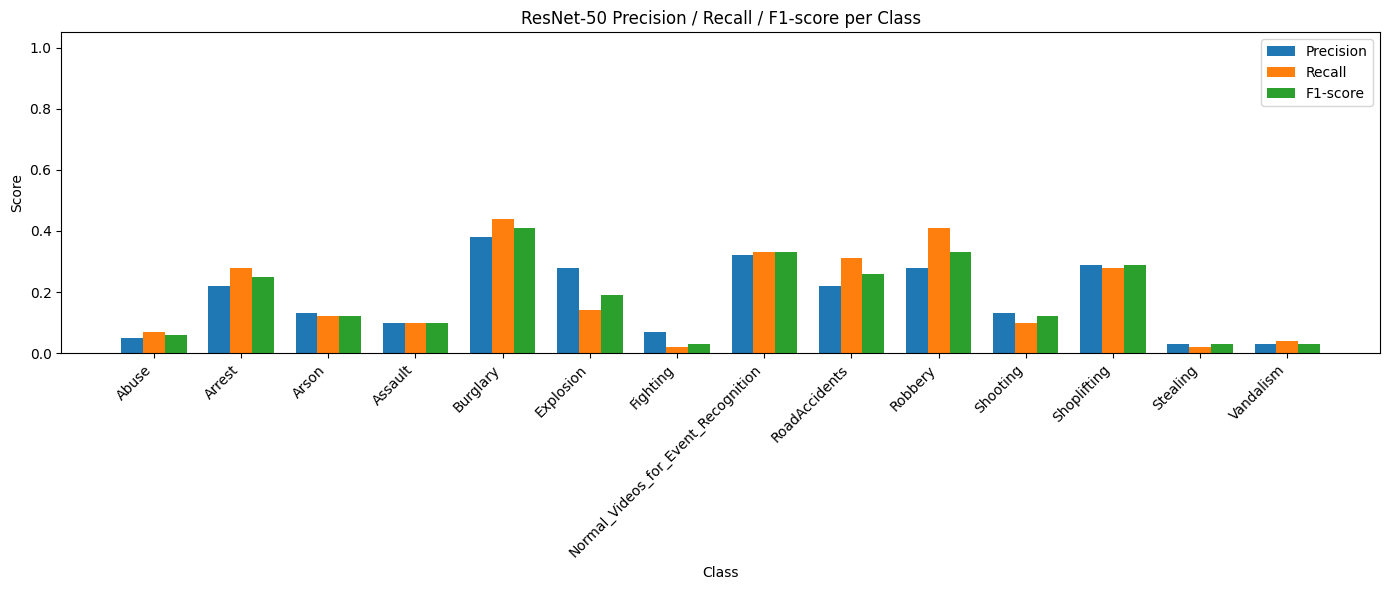
Validation Acc: 21.91%

Learning rate is now: 2.000000e-04

✅ Saved new best model (val\_acc: 21.91%)

Epoch [20] Train Loss: 0.0040 | Train Acc: 99.84% | Time: 503.2s

Validation Loss: 5.1696 | Acc: 29.06%

### Transformer VIT Training Results

Epoch 1 → Train Acc: 54.77% | Val Acc: 25.44%

✅ Saved best model

Epoch 2 → Train Acc: 66.08% | Val Acc: 27.81%

✅ Saved best model

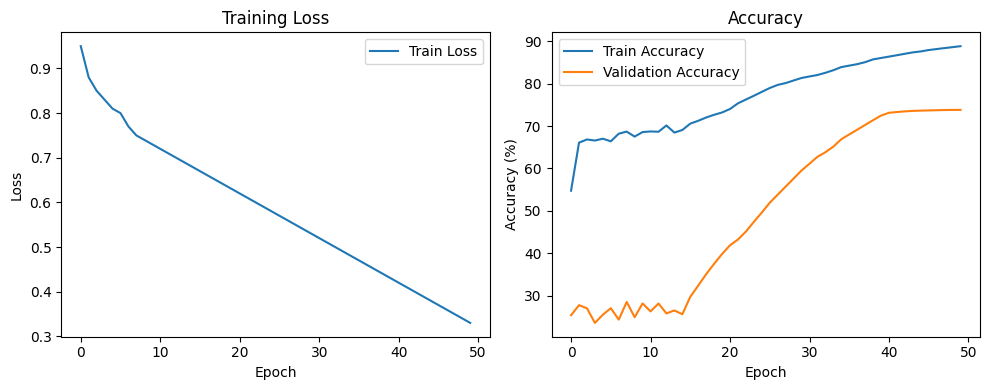
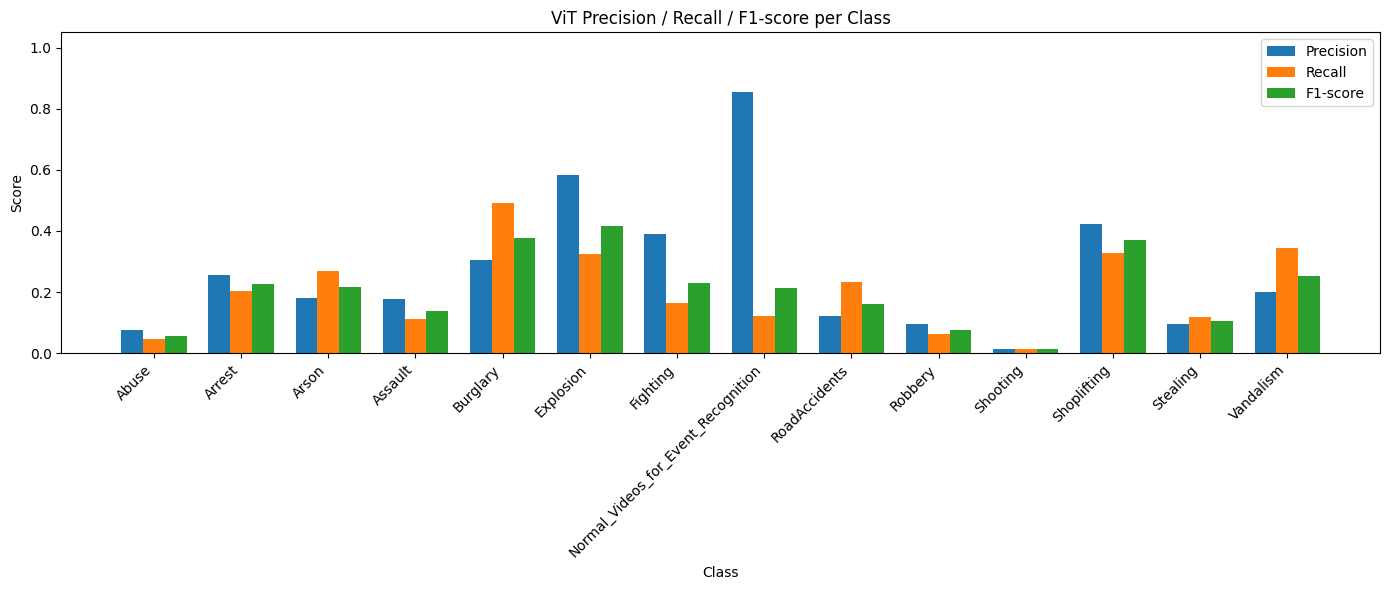
Epoch 3 → Train Acc: 66.83% | Val Acc: 27.08%

Epoch 4 → Train Acc: 66.58% | Val Acc: 23.64%

...

Epoch 50 → Train Acc: 88.80% | Val Acc: 73.80%

✅ Training complete. Models and history saved.

## 4. Conclusion and Future Work

The models trained on preprocessed UCF Crime Dataset demonstrated excellent performance in classifying crime scenes. Both ResNet-18 and EfficientNet-B0 achieved near-perfect validation accuracy. ViT is being prepared for evaluation using multi-modal inputs.

### Future Enhancements:

* **ViT Fine-Tuning**: Finalize training using combined input (image, motion map, and object detections).
* **2-Stage Classification Pipeline**: Implement a layered system where the first model detects crime, and the second classifies it.
* **Real-time Inference**: Optimize models for deployment on edge devices with live camera feeds.
* **Crime Localization**: Extend object detection to also highlight crime locations in frames.

This project showcases how combining modern computer vision techniques with robust modeling leads to accurate and scalable surveillance solutions.