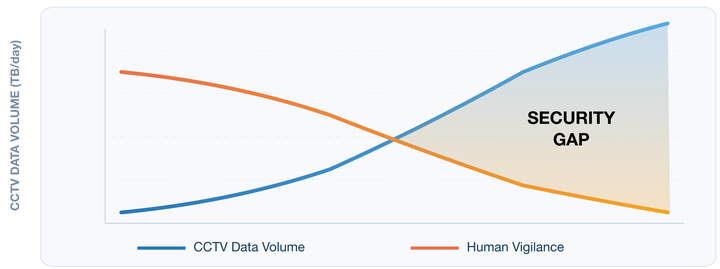
AI-Based Physical Aggression Detection for Video Surveillance Systems

Ashmaan Balan Shaikh | C00317323 | Supervisor: Rejwanul Haque | MSc. Applied Artificial Intelligence

SETU Ollscoil Teicneolaíochta an Oirdheiscirt — South East Technological University

1 INTRODUCTION

The Attention Gap



CHALLENGE
Exponential growth of CCTV footage vs. limited human attention span.

CONSEQUENCE
Operator fatigue leads to missed critical events and slow response times.

SOLUTION DRIVER
Automated, real-time detection of physical aggression.

2 RESEARCH GAP

CONTROLLED DATASETS
Hockey Fights | Movies

ACCURACY
96–99%

GAP
RESEARCH

REAL-WORLD ENVS.
Surveillance | Street CCTV

SEVERE DEGRADATION

Simple: 96-99%
Complex: 70–81%

The "Generalization Cliff": Performance drops **20–30%** in cross-dataset validation.

Computational Cost
Training energy: **240–570 Wh** — hindering practical deployment.

3 RESEARCH QUESTION & HYPOTHESIS

PRIMARY QUESTION

How to develop a deep learning model that is both reliable in identifying physical aggression and generalizes easily across diverse surveillance settings?

SECONDARY QUESTIONS

Data Diversity
Does multi-dataset training close the cross-domain accuracy gap?

Architecture Triage
Which model (Transformer vs. 3D-CNN vs. Hybrid) offers best trade-off?

Optimization
Can pruning & quantization reduce computational load without sacrificing precision?

CORE HYPOTHESIS

Generalization Improvement

Multi-dataset training will significantly improve cross-domain generalization, achieving 85–90% accuracy on unseen datasets vs. current 70–81% baseline.

BASELINE PERFORMANCE 70–81%
TARGET PERFORMANCE 85–90%

4 DATASETS

CONTROLLED
Boxing | Film | Cinematic Set

Training on controlled vs. chaotic environments

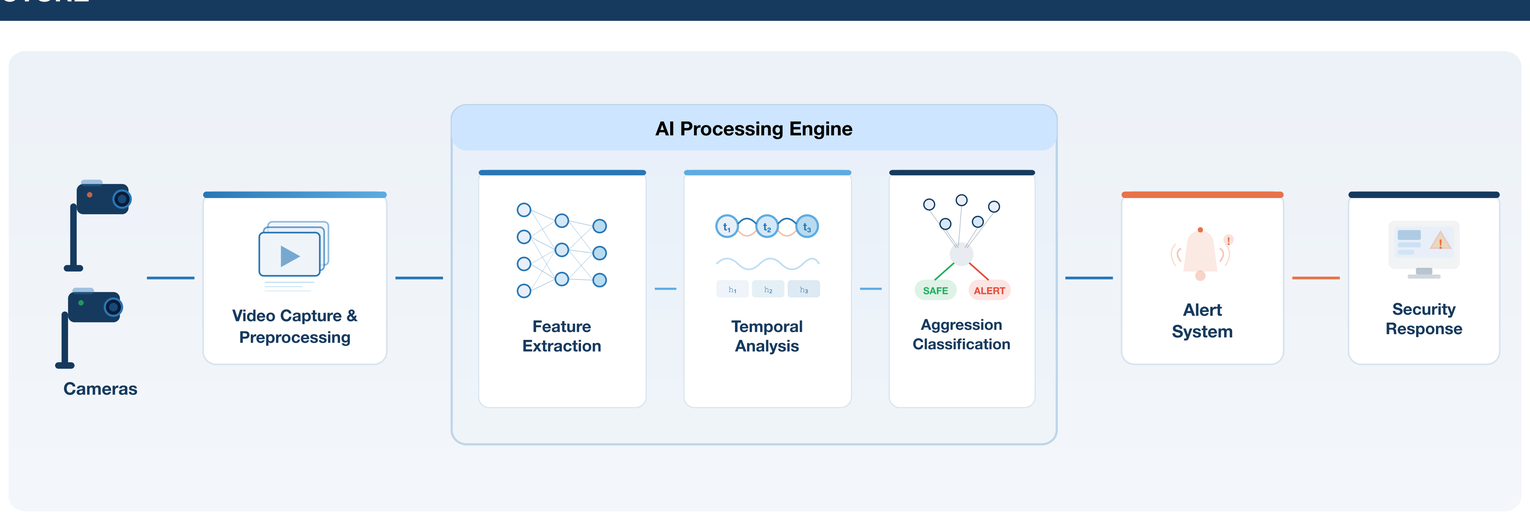
CHAOTIC ENVIRONMENTS
Street Footage | Low-Res CCTV | Phone Recordings | Real-World

Surveillance, personal recordings, varying quality, complex scenarios

Dataset Overview

DATASET	ENVIRONMENT TYPE	KEY FEATURE
RLVS	Real-World Surveillance	Low-res, chaotic, authentic
Hockey Fights	Sports	Fast motion, single context
Violence in Movies	Cinematic	Professional lighting, staged
RWF 2000	Benchmark	Real world situations
SCVD	City CCTV Surveillance	Benchmark dataset

5 PROPOSED SYSTEM ARCHITECTURE



6 METHODOLOGY

PHASE 1 Preprocessing
Input Video Frames
↓
Normalization & Augmentation
↓
32-Frame Temporal Clips
↓
Class Balancing & Quality Verification

PHASE 2 Multi-Dataset Training

Three Training Strategies:

1. Pooled Combined Training
2. Sequential Fine-Tuning
3. Domain Adversarial Learning

Gradient reversal for learning dataset-invariant features across all five benchmark datasets.

PHASE 3 Architecture Search

Comparative Evaluation:

TimeSformer Transformer
VS.
3D-CNN Convolutional Neural Network
LSTM Hybrid Long Short Term Memory

Evaluated on: Accuracy, F1-Score, ROC-AUC, and Inference Latency across all five datasets with LODO.

PHASE 4 Optimization

Techniques:

Knowledge Distillation → **Network Pruning** → **Reduced Power Draw**

Additional: Attention pruning, quantization, and MobileNet-3D variants for edge deploy.

Target: NVIDIA Jetson
Resource-constrained edge hardware deployment testing

7 EVALUATION PROTOCOL

Data Split

70% Train | 15% Val | 15% Test

Balanced stratification across all datasets

Cross-Validation: LODO
Leave-One-Dataset-Out protocol to simulate real-world deployment shock.

Evaluation Metrics

- Accuracy
- F1-Score
- ROC-AUC
- Precision and Recall
- Confusion Matrices
- Inference Latency

8 TECHNOLOGIES

NumPy · pandas · scikit-learn · PyTorch · TensorFlow · einops · ultralytics YOLOv8 · OpenCV · kaggle · seaborn

9 EXPECTED CONTRIBUTIONS & IMPACT

IMPROVED GENERALIZATION
Domain randomization (multi-dataset training) is more critical than model complexity for real-world deployment.

POLICY IMPLICATIONS
Reliable, unbiased AI surveillance that reduces human monitoring fatigue and ensures faster public safety intervention.

NEXT STEPS

01 Deployment testing on resource-constrained hardware (e.g., NVIDIA Jetson).

02 Integration of explainability (XAI) to visualize why a clip was flagged as violent.

KEY REFERENCES

- Bertasius, G. et al. (2021). TimeSformer: Is Space-Time Attention All You Need for Video Understanding? *Proceedings of ICML.*
- Negre, P. L. S. et al. (2024). Deep Learning for Violence Detection: A Comprehensive Review. *Neural Computing and Applications, 36(15), pp. 8417–8444.*
- Li, H. et al. (2025). IDG-ViolenceNet: Identity-Preserved Violence Detection Using Graph-Based Deep Learning. *Expert Systems with Applications, 238, p. 121847.*
- Chen, W. et al. (2024). ViolenceNet: Dense Multi-Head Self-Attention with Bidirectional ConvLSTM for Detecting Violence. *Electronics, 13(8), p. 1523.*
- Zhang, Y. et al. (2024). CUE-Net: Violence Detection Video Analytics with Spatial Cropping, Enhanced Unified Memory, and Network Pruning. *IEEE Transactions on Industrial Informatics, 20(8), pp. 10234–10244.*