

2026-08

Movenet on frame difference: a motion-aware pose-based preprocessing approach for real-world violence detection

Raisha, Shahla Sarmin

Independent University, Bangladesh (IUB)

<https://ar.iub.edu.bd/handle/11348/1576>

Downloaded from IUB Academic Repository



MOVE_{NET} ON FRAME DIFFERENCE: A MOTION-AWARE POSE-BASED PREPROCESSING APPROACH FOR REAL-WORLD VIOLENCE DETECTION

August 2026

Prepared by:

Shahla Sarmin Raisha

ID: 2211595

Department of Computer Science and Engineering

Independent University, Bangladesh

Supervised by:

Md Rashedur Rahman, D.Eng.

Assistant Professor

Department of Computer Science and Engineering

Independent University, Bangladesh

A Final Year Design Project (FYDP) Report submitted for the Degree of
Bachelor's of Computer Science & Engineering

Attestation

I am aware of the fact that plagiarism is strictly prohibited and it conflicts with my university's rules and regulations. I guarantee the authenticity of my work. I have used some copyrighted material and models of others in my project, which have been properly cited following the international standards and proper guidelines.

Author Name:

.....

Signature:

.....

Evaluation Committee

Supervisor

Name: _____ Signature: _____

Internal Examiner 1

Name: _____ Signature: _____

Internal Examiner 2

Name: _____ Signature: _____

External Examiner

Name: _____ Signature: _____

Acknowledgment

I would like to express my sincere gratitude to my supervisor, **Md Rashedur Rahman, D. Eng.**, for his valuable guidance, support, and feedback throughout this project. His advice and encouragement helped me stay focused and improve my work throughout the research.

I would also like to sincerely thank **Mr. Siam Tahsin Bhuiyan**, for his support, guidance, and assistance throughout the project. His suggestions and support were very helpful in carrying out and completing this work.

I am grateful to the **Medical Imaging Research and Analysis Wing (MIRAW), Center for Computational and Data Sciences (CCDS), Independent University, Bangladesh (IUB)**, for providing the computational resources and facilities necessary to conduct this research.

Finally, I acknowledge the use of Large Language Models (LLMs), including Claude by Anthropic, Gemini by Google, and ChatGPT by OpenAI, to improve the clarity, grammar, and language of my writing. These tools were used solely for editorial and language refinement, and were not used for developing research methodology, hypothesis, data analysis, experimental results, and interpretations.

Abstract

Automated violence detection from surveillance video can support continuous monitoring and reduce the burden of manual video surveillance. But in real-world videos, there are various real-world background obstructions, varying lighting conditions, occlusion, and irrelevant visual information, making the detection difficult. This study investigates whether preprocessing can improve violence detection by emphasizing motion and human body structure before classification. One motion- and pose-based preprocessing method is proposed, namely MoveNet on Frame Difference, in which first frame differencing is used to remove static background content and highlight content in motion, and then MoveNet multipose estimation is used to obtain the human skeletal structure from the frames of motion content. To examine the effectiveness of this approach and the importance of processing order, five preprocessing variants were evaluated on the RWF-2000 dataset: Frame Difference, MoveNet, MoveNet Without Background, MoveNet on Frame Difference, and Frame Difference on MoveNet. This effect of pre-processing was compared across all variants using the same EfficientNet-B0 and LSTM classification architecture, training configuration, and held-out test set used to evaluate all variants. The proposed method achieved the best overall performance, with 81.50% accuracy, 88.50% recall, 0.8271 F1-score, and 0.8661 AUC. It also outperforms the reverse-order ablation Frame Difference on MoveNet, which reached 74.00% accuracy, indicating that the order of motion and pose information is important for the detection performance. The results show that frame differencing prior to pose estimation gives a more informative representation, by using the two information sources of motion localization and human body structure, but without introducing background-related noises. Overall, the findings demonstrate that an appropriate preprocessing strategy can improve real-world violence detection without requiring a more complex classification architecture.

Contents

1	Introduction	10
1.1	Background	10
1.2	Motivation	12
1.3	Research Questions	13
1.4	Objective	15
1.5	Thesis Organization	15
2	Literature Review	17
2.1	Prior Studies	17
2.1.1	Traditional and Classical Methods	17
2.1.2	Early Deep Learning Methods	18
2.1.3	Advanced Deep Learning Methods	20
2.1.4	ROI-Guided and Human-Focused Methods for Violence Detection	21
2.2	Challenges Faced	23
2.3	Scope of Study	24
3	Dataset	25
3.1	The RWF-2000 Dataset	25
3.2	Data Collection and Composition	25
3.3	Dataset Partitioning and Leakage Prevention	26
3.4	Comparison with Prior Datasets	27
3.5	Dataset Split Used in This Study	28
4	Methodology	29
4.1	Overview	29
4.2	Preprocessing	30
4.2.1	Frame Difference	30

4.2.2	MoveNet	30
4.2.3	Proposed Method: MoveNet on Frame Difference	31
4.2.4	Ablation: Frame Difference on MoveNet	32
4.3	Classification Model	33
4.3.1	Frame Sampling	33
4.3.2	EfficientNet-B0: Spatial Feature Extraction	33
4.3.3	LSTM: Temporal Aggregation	34
4.3.4	Classification Head	35
4.4	Training Configuration	35
4.5	Evaluation	36
5	Results and Analysis	38
5.1	Overview	38
5.2	Main Results: Preprocessing Variants Compared	38
5.2.1	Raw Video Baseline	39
5.2.2	MoveNet	40
5.2.3	Frame Difference	40
5.2.4	MoveNet without Background	41
5.2.5	MoveNet on Frame Difference	42
5.3	Ablation Study: Order of Fusion	43
5.4	Summary of Key Findings	44
6	Discussion	46
6.1	Overview	46
6.2	Why MoveNet on Frame Difference Performed Well	46
6.3	Why Frame Difference on MoveNet Performed Poorly	47
6.4	What Works and What Does not	48
6.5	Comparing All Five Preprocessing Methods	48
6.6	Comparison with Previous Studies	49
6.7	Practical Significance	50
7	Conclusion	51
7.1	Summary	51
7.2	Research Questions Answered	52
7.3	Analysis of the Social Effects	53

7.4	Environmental Impact Analysis and Sustainability	53
7.5	Ethical Issues	54
7.6	Limitations and Future Work	54
7.7	Closing Thoughts	55
Bibliography		56

List of Figures

1.1	Manual vs Automated Violence Detection	11
2.1	Evolution of Violence Detection Methods	17
3.1	Sample Frames from the RWF-2000 Dataset.The Frames Show (a) a Frame from Fight Video, and (b) a Frame from Non-Fight Video.	25
3.2	Class Distribution of the RWF-2000 Dataset.	26
3.3	Dataset Partitioning Used in This Study.	28
4.1	Proposed Method Overview	31
4.2	Sample Output Frame from MoveNet on Frame Difference Preprocessing .	31
4.3	Ablation Method Overview	32
4.4	Sample Output Frame from Frame Difference on MoveNet Preprocessing .	32
4.5	EfficientNetB0 Achitecture Diagram	34
4.6	Long Short-Term Memory Model Diagram	34
5.1	Performance comparison of different preprocessing variants on the RWF- 2000 dataset.	39
5.2	Confusion matrix and ROC curve for the MoveNet variant.	40
5.3	Confusion matrix and ROC curve for the Frame Difference variant.	41
5.4	Confusion matrix and ROC curve for the MoveNet without Background variant.	42
5.5	Confusion matrix and ROC curve for the MoveNet on Frame Difference variant.	42
5.6	Performance comparison of preprocessing order in the ablation study. . . .	44
5.7	Confusion matrix and ROC curve for the Frame Difference on MoveNet variant.	45

List of Tables

3.1	Comparison of violence detection datasets	27
4.1	Hyperparameters and training settings applied identically to all five preprocessing variants.	35
5.1	Comparative Results Across Preprocessing Methods	38
5.2	Ablation Study: Order of Sequential Fusion	43

Chapter-1: Introduction

1.1 Background

Violence is one of the major threats to public health around the world. According to the World Health Organization, injuries and violence together cause approximately 4.4 million deaths each year, accounting for around eight percent of all global mortality [1]. Physical assault, interpersonal conflict, and crowd-based violence are major causes of death and serious injury. Their effect go far beyond immediate injuries, often leading to long-term disability, psychological trauma, and loss of economic resources that impact families, communities, and health systems [1].

Violence also causes major economic losses by increasing healthcare costs, property damage, and loss of productivity. The 2026 Global Peace Index estimates that the global economic impact of violence amounted to more than twenty-one trillion US dollars annually [2]. This figure includes direct costs such as healthcare and law enforcement, as well as substantially larger indirect losses from reduced productivity and foregone investment [2]. These numbers are not abstract. They expose the truth about the pressure on governments and institutions to do more to prevent violence than just react to it after it happens.

There has been a lot of growth in surveillance infrastructure due to these pressures. In recent years, it has become commonplace to have cameras placed in schools, hospitals, banks, markets, transit stations, and public streets in cities around the world [3]. It's a well-known fact that anything that can be seen cannot be done, and when it is, the evidence recorded by the camera makes for a strong deterrent. However, in practice these two functions are used for different time frames. Evidence is gathered after a crime has been committed. Deterrence only is effective if potential offenders think they are being monitored. Both have no requirement for anyone to be on the lookout for the video in real time, and most deployments don't have anyone. The cameras record 24/7, records pile up and the vast majority of the footage is only seen when a problem has already occurred.

This leaves a big disconnect between surveillance infrastructure and coverage. The gap is a result of the fact that it is actually challenging to monitor video in real time for human operators. The video from hundreds of cameras could be streaming at one security desk. After minutes, let alone hours, there can be attention lapses. It is unlikely that an operator will be oriented towards the correct camera when a violent incident occurs and fatigue exacerbates the issue during a long shift [3]. But the job becomes even more difficult when you start to use real-world CCTV footage because of the variety of factors associated with the footage that the monitoring system has to deal with: varying lighting across the day

and night cycle, low spatial resolution of older installed equipment, motion blur due to high speed of movement, dynamic backgrounds and partial occlusion of one person by another are all common features of the footage processed by a monitoring system [4]. This is not a domain where performance is measured under controlled laboratory conditions.

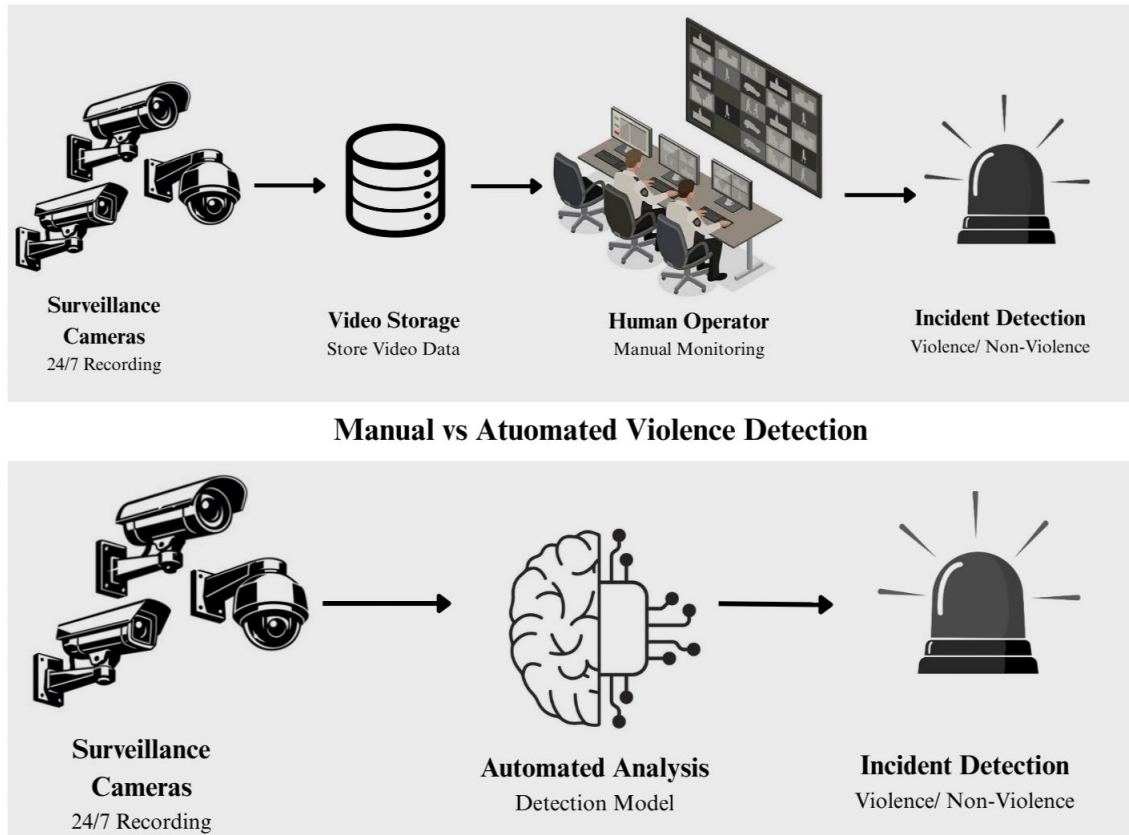


Figure 1.1: Manual vs Automated Violence Detection

This gap is directly addressed by automatic violence detection. The need for continuous human attention is eliminated by a software system that can analyze surveillance footage and spot violent conduct as it happens or in almost real time from a continuous stream. When a possible event is spotted, the system flags portions for human review instead of requiring an operator to monitor every frame. Operator workload decreases, response times improve, and coverage becomes truly continuous rather than only nominal [3].

Automated violence detection is a field in the larger area of video-based human action recognition. The aim is binary classification: For a short video segment, classify whether it contains a violent event or not. Initial attempts to solve this problem were based on manually created descriptors derived from sequences of pixels. Some of the most popular were spatio-temporal interest points (STIP), Motion Scale-Invariant Feature Transform (MoSIFT), and Histogram of Oriented Gradients (HOG) combined with Support Vector Machine classifiers or Bag-of-Words encoding schemes [3]. The methods could identify gross patterns of motion in controlled footage, but are brittle. But when the test conditions were very different from the training conditions (which is always the case when you switch from a training set to a real surveillance set), the performance took a turn for the worse without the slightest escape route.

Deep learning had a significant influence on the direction of the field. By doing so, convolutional neural networks showed themselves to be capable of learning local representations of

features from images without the need for feature engineering, and architectures for video started to emerge from the mid-2010s. Two-stream networks separately computed RGB frames and pre-computed optical flow in parallel at the architecture level. Convolutional LSTM [5] networks took the LSTM [5] memory mechanism and applied it to structured data, such as images, enabling the modelling of temporal dependencies without losing spatial structures. Three-dimensional convolutional networks were trained directly on the spatiotemporal features of stacked frame volumes. Recently, attention mechanisms and light-weight mobile architectures have enabled to achieve higher accuracy with lower computational costs. These have been carried over to the domain of violence detection, where the prevailing strategy is to use some form of deep spatial feature extractor, along with a temporal aggregation mechanism.

However, the problem is still not solved. The real world is a tough input modality, and there is not yet a complete gap between the performance on clean benchmark datasets and the deployment. If the camera system is installed in a ceiling mounted location, and the image is grainy, at night and with some obscuration, a model learned from broadcast-quality video of a hockey game or a Hollywood film will not generalize well to that type of video. This explains why datasets created from real surveillance footage rather than the manipulated footage, are relevant for proper evaluation.

1.2 Motivation

Raw RGB videos are high-dimensional signals. Most of that information is irrelevant to the task of identifying violent activity: the texture of the floor, the design of the wallpaper, the colors of the parked cars in the background, the static geometry of the environment. A classifier trained on unprocessed RGB videos has to learn from labeled examples alone how to discriminate between those irrelevant elements of visual information and the spatiotemporal patterns that actually distinguish a fight from a handshake. In a small training dataset, this dual learning task is more difficult than necessary.

Video pre-processing before classification represents a way to make the task simpler. If the representation of the input data can be transformed in such a way that the irrelevant content is filtered out even before the classifier gets it, the classifier is left with less noise to separate from the signal, thus making the relevant signal relatively more salient. Two kinds of pre-processing operations have been proposed in the literature on violence detection: each addressing one specific kind of irrelevant information.

The first approach is frame differencing. For any pair of adjacent frames $F(t)$ and $F(t-1)$, the absolute per-pixel difference $|F(t) - F(t-1)|$ yields an image in which the regions of static content become zero, while the regions of motion appear to be white. Applied to surveillance video, this transformation filters out the static background almost completely, leaving visible only what moved from frame to frame. In a violent interaction, that is, the people involved. In a non-violent video, the resulting difference frames are mostly dark and sparse. The classifier is fed with a representation in which motion distribution, rather than the appearance of the scene, constitutes the discriminating signal. Sani et al. [6] compared the performance of raw frame input versus frame-difference input on the UCF-Crime dataset [7] in terms of a ResNet50-LSTM [5] classifier and found that frame

differences improve the AUROC of the violence class from 0.89 to 0.91, as well as reduce the training time by approximately three percent.

The second approach is pose estimation. Pose estimation predicts the coordinates of anatomically-defined body keypoints in an image, yielding a stick figure skeleton that contains information about the geometric layout of the human body without representing its appearance. Aggressive interactions have certain characteristic body positions such as raised arms, extended elbows, fast limb rotations, and lack of upright posture. All of these are much easier to detect from a skeleton representation than from a raw frame, where clothing colors and skin tones do not provide any useful information. In particular, Bhuiyan et al. [8] used MoveNet pose estimation for processing the RWF-2000 dataset [4] and trained a ResNet50V2+LSTM [5] classifier on the resulting skeletons. They achieved an accuracy of 68.50 percent on raw video and of 77.08 percent on pose-estimated video, a relative improvement of 8.58 percentage points using the same dataset and the same classifier architecture.

Both results are encouraging, but neither of the two studies compared their preprocessing approach to alternatives. Bhuiyan et al. [8] compared raw video against pose-estimated video; they did not experiment with frame differencing and did not explore any combination of the two. Sani et al. [6] performed frame differencing on the UCF-Crime dataset; [7] they did not experiment with pose estimation and did not use the RWF-2000 dataset. Barapatre et al. [9] performed a classification of RWF-2000 [4] without using any preprocessing techniques whatsoever and achieved an accuracy of 68 percent. Garcia-Cobo and SanMiguel [10] designed a two-stream architecture which combined skeletal and change detection features into a single model, obtaining an accuracy of 90.25 percent on the RWF-2000 dataset. However, this fusion was architectural and not preprocessing-related; it was applied to an otherwise fixed architecture. Aung et al. [11] combined YOLO-NAS pose estimation with frame differencing in a dual-stream ConvLSTM model and obtained results across several datasets, but once again the fusion was architectural and not preprocessing-related.

Overall, these studies leave a specific question unanswered: what happens when you apply frame differencing and pose estimation sequentially, as a preprocessing pipeline feeding a single consistent classifier? And does the order matter? Frame differencing first produces motion-residual images that suppress static background; MoveNet [12] then annotates skeletal keypoints on those motion-residual images, potentially detecting pose more cleanly because the background noise has already been removed. The reverse order applies MoveNet [12] to full RGB frames first, then differences the resulting skeleton sequences; this preserves the full visual context for pose detection but then subtracts consecutive skeleton frames, which amplifies the small detection jitter that MoveNet [12] produces even on stationary subjects. These two orderings encode different assumptions about what information is most important to preserve, and there is no published comparison of them on the same dataset with the same classifier. This thesis was motivated by that gap.

1.3 Research Questions

The motivation outlined above raises a specific empirical question: whether combining frame differencing and pose estimation in a deliberate sequence produces a more useful

input for violence detection than any alternative, and whether that design choice holds up under both a broad comparative evaluation and a controlled ablation. This thesis is organized around three questions specific enough to be answered with evidence, each addressed by the results in Chapter 5 and interpreted in Chapter 6.

1. **RQ1:** Does applying frame differencing before MoveNet pose estimation, as a sequential preprocessing pipeline, produce a more effective input representation for violence detection than either technique applied independently?
2. **RQ2:** Across five distinct preprocessing strategies evaluated under a fixed classifier and identical training conditions on the RWF-2000 dataset, which input representation produces the strongest violence detection performance, and what does the ranking reveal about what information matters most for this task?
3. **RQ3:** Does the order in which frame differencing and MoveNet pose estimation are applied have a measurable and consistent effect on violence detection performance, and if so, why?

RQ1 is the foundational question on which the other two depend. Prior work has shown that frame differencing improves over raw RGB on UCF-Crime [17] and that MoveNet pose estimation improves over raw RGB on RWF-2000 [16], but neither study tested whether combining both operations in sequence produces a representation that captures what neither technique carries alone. RQ1 asks whether applying frame differencing first and pose estimation second closes both gaps simultaneously, outperforming every individually applied alternative. This is the purpose of Objective 1, answered in Section 5.2.5 and interpreted in Section 6.2.

RQ2 asks a broader question than RQ1. Even if the proposed method outperforms individual techniques, that finding alone does not establish where it sits relative to all plausible alternatives, nor does it reveal what the full ranking of five preprocessing configurations implies about which kinds of information most benefit the downstream classifier. Answering this requires holding the classifier, training configuration, dataset, and evaluation protocol identical across all five conditions so that preprocessing is the only variable. This is the purpose of Objective 2, answered in Section 5.2 and interpreted in Section 6.5.

RQ3 exists because a favorable answer to RQ2 would still leave open whether the proposed method’s advantage over its reversed-order counterpart is attributable to sequence specifically, rather than simply to the presence of both operations together. Since both Frame Difference on MoveNet and MoveNet on Frame Difference use identical techniques and differ only in ordering, any performance gap isolates the effect of sequence alone. RQ3 asks whether that gap exists, how large it is, and whether the reasoning in Section 4.2.4, that differencing first gives MoveNet a cleaner input while the reverse order amplifies its own estimation jitter, is confirmed empirically. This is the purpose of Objective 3, answered in Section 5.3 and interpreted in Section 6.3.

Together, these research questions define the key aspects of the investigation and provide the basis for formulating the specific objectives of this study.

1.4 Objective

This study pursues three specific objectives, corresponding to a proposed method, a comparative evaluation, and an ablation study.

1. **Propose MoveNet on Frame Difference** as a preprocessing pipeline for violence detection. The primary contribution of this thesis is a preprocessing method in which frame differencing is applied to the raw surveillance video first, and MoveNet multipose pose estimation [12] is then applied to the resulting frame-difference images. Frame differencing removes static background content by retaining only the regions of inter-frame pixel change, concentrating the visual signal on where motion occurred. MoveNet [12] then annotates human skeletal keypoints on those motion-residual images, producing a representation that encodes both the spatial location of movement and the geometric configuration of the human body during that movement. This sequential combination is proposed as a more informative input for violence classification than either operation applied alone or in the reverse order.
2. **Compare five preprocessing variants** on RWF-2000 [4] using a fixed classifier. To evaluate the contribution of preprocessing rather than the classifier, the same EfficientNet-B0 + LSTM [13] [5] architecture is trained independently on five different input representations derived from the same dataset: raw frame difference video, MoveNet [12] skeleton overlay on original RGB frame, MoveNet skeleton without background, and the proposed MoveNet on Frame Difference. A sixth condition, raw video, is taken from Bhuiyan et al. [8] as an external baseline. Performance across all conditions is measured using accuracy, precision, recall, F1-score, and AUC on the same held-out test set of 400 clips from RWF-2000 [4], giving a direct comparison of what each preprocessing choice contributes when the classifier is held constant.
3. **Conduct an ablation study on fusion order.** The proposed method applies frame differencing before pose estimation. To test whether this ordering is material, the reverse is also evaluated: MoveNet [12] is applied first to the original RGB frames, and frame differencing is then applied to the resulting skeleton video. This ablation isolates the effect of operation order from the effect of using both operations at all. If the two orderings produce similar results, order does not matter and either could be used in practice. If they diverge significantly, the result would identify which ordering extracts more useful information and explain why.

Overall, these objectives establish a systematic framework for developing, comparing, and validating the proposed preprocessing strategy, with particular emphasis on the contribution of preprocessing choice and operation order to violence detection performance.

1.5 Thesis Organization

This thesis is structured into seven chapters, with each chapter contributing to the development, evaluation, and interpretation of the proposed violence detection approach.

- **Chapter 2 (Literature Review)** reviews existing approaches to violence detection, covering traditional methods, deep learning techniques, pose estimation, and frame differencing, and identifies the research gap addressed in this study.
- **Chapter 3 (Dataset)** describes the RWF-2000 dataset, including its collection and composition, partitioning, leakage prevention, comparison with related datasets, and the final dataset split used in this study.
- **Chapter 4 (Methodology)** presents the proposed approach and experimental framework, including frame difference and MoveNet-based preprocessing, their sequential combinations, the EfficientNet-B0 and LSTM classification model, training configuration, and evaluation metrics.
- **Chapter 5 (Results and Analysis)** presents the experimental results of the different preprocessing variants, including the raw video baseline, individual preprocessing methods, the proposed approach, and the ablation study on preprocessing order.
- **Chapter 6 (Discussion)** interprets the findings, examining the effectiveness of the different preprocessing strategies, their comparative performance, and the practical significance of the results.
- **Chapter 7 (Conclusion)** summarizes the main findings, addresses the research objectives, and discusses the social, environmental, and ethical considerations, limitations, and potential directions for future work.

Overall, the thesis progresses from the background and existing research to the dataset, methodology, experimental findings, discussion, and concluding insights.

Chapter-2: Literature Review

2.1 Prior Studies

The method for automated detection of violence in video has progressed over the years and several methodological frameworks have been proposed: handcrafted feature descriptors combined with classical classifiers, early deep learning models where hand-designed features are replaced by learned convolutional representation, further deep learning models that incorporate attention mechanisms, efficient convolutions or transformer-based components, and finally methods that explicitly separate the human motion and human pose from the background appearance through frame differencing and pose estimation.

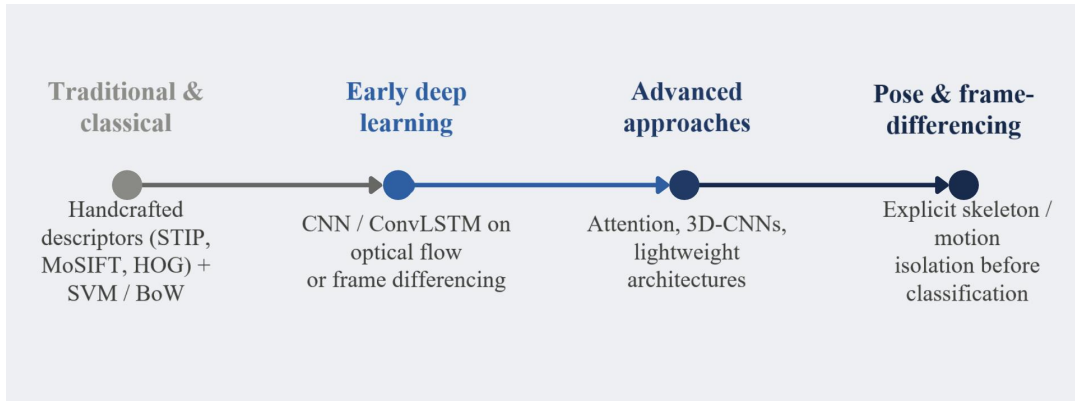


Figure 2.1: Evolution of Violence Detection Methods

The following subsections review representative works from each of these categories.

2.1.1 Traditional and Classical Methods

Before deep learning, handcrafted spatio-temporal descriptors were used along with classical machine learning classifiers for violence detection. One influential early study evaluated local spatio-temporal descriptors, Space-Time Interest Points (STIP) and Motion Scale-Invariant Feature Transform (MoSIFT), for detecting fights in both sports footage and movie scenes [14]. Features were encoded using a Bag-of-Words model obtained by k-means clustering over a visual vocabulary, and each video was represented as a histogram of visual words before being classified by SVM (Support Vector Machine) with a Histogram Intersection Kernel. STIP and MoSIFT, however, achieved a 91.7 percent accuracy rate on

the Hockey Fights dataset and a 90.9 percent accuracy rate on the Action Movies dataset, respectively, on the Hockey Fights dataset, while on the Action Movies dataset, MoSIFT significantly outperformed STIP, with an accuracy rate of 89.5 percent against 59.0 percent for STIP, demonstrating poor transferability of handcrafted descriptors between different visual settings.

In a later study, principles from fluid mechanics were used to describe motion in crowded scenes by computing a significant derivative of dense optical flow, which was able to isolate both local and temporal motion changes from convective and spatial motion changes [15]. Both of these types of motion were modeled as a Bag-of-Words, and the combination of both as a single descriptor was classified using an SVM with Histogram Intersection Kernel for the whole video classification and a Latent Dirichlet Allocation model for the detection of the abnormal event times. The approach was designed to work on dense crowds and was heavily reliant on the quality of the underlying optical flow estimate, achieving good accuracy on a dataset of violent movies (96.89 percent accuracy) and strong area-under-curve scores on prison riot and public panic recordings, though on a dataset of crowd footage it achieved 85.43 percent accuracy on the Violence in Crowds benchmark.

In another study, motion was first segmented using optical flow, with a Local Histogram of Oriented Gradient computed from RGB frames, followed by the combination of the two descriptors with a further study of building up handcrafted descriptors [16]. The two types of descriptors were encoded independently with the Bag-of-Words model; the encoded words were fused at a late stage and trained by SVM with a Radial Basis Function kernel. A hybrid appearance and motion method was more effective on Hockey Fight (95.1 percent), BEHAVE (100 percent), and Crowd Violence (94.31 percent), but optical flow computation made it less suitable for real-time applications, and the fixed spatial block grid was less effective when the scale of the subjects was suddenly changed in a scene.

Throughout this period, a pattern can be seen. On benchmarks with more limited domains like hockey games and brief movie clips, handcrafted descriptors coupled with Bag-of-Words encoding and SVM classification could get to high accuracy levels, but their success depended heavily on hand-crafting features that aligned nicely with the visual aspects of the data sets studied, thus favoring the use of learned features in later work.

2.1.2 Early Deep Learning Methods

The first step towards using deep learning for the problem of detecting violence involved using automatically learned convolutional features instead of handcrafted motion descriptors, but in many cases, motion information from optical flow or frame differencing was still used as an input signal. An early instance of estimating dense optical flow was to use the Farneback algorithm to estimate it, and then generate a Violent Flow descriptor from the flow magnitudes and input this descriptor to a convolutional neural network that learned discriminative patterns in its own convolution and pooling layers, but not through a hand-designed histogram [17]. On the Crowd, Hockey, and Movies data sets, this CNN-based Violent Flow network achieved 96 percent, 98 percent, and 99 percent accuracy, respectively, and the authors explained that simple handcrafted flow descriptors were not enough for more complex violent behavior, which is why the learned CNN stage is proposed.

Later work has adopted frame differencing as a more common and lower-cost motion signal than full optical flow. Another work used a pre-trained VGG16 network with a ConvLSTM layer to extract spatial motion information at a lower computational cost by replacing the optical flow input with the adjacent-frame difference between two frames [18]. The ConvLSTM layer, which had 256 filters and 3x3 kernels, learned temporal dependencies across the sequence, followed by batch normalization and fully connected layers that finally classified the sequence. This method achieved high accuracy on various tests, such as 92.4 percent on the RWF-2000 dataset [4], 99.1 percent on the Hockey Fights data set, and 100 percent on Movies, although the authors admitted that frame differencing is only approximate optical flow and could be less effective at capturing fine motions in low-contrast or slow-motion videos.

In a two-stream architecture, the frames were first suppressed in the background, and then spatio-temporal associations were modeled by a pre-trained MobileNetV2 and a Separable Convolutional LSTM (SepConvLSTM) [19] for extracting spatial features from each stream of the background-suppressed frames. The features from the two streams were combined using multiplication, concatenation or addition prior to the classification, and the most effective combination achieved 89.75 percent accuracy for the RWF-2000 dataset [4], 99.0 percent for the Hockey data set, and 100 percent for the Movies data set, with significantly fewer parameters and floating-point operations than other similar models, though the authors reported that the model still failed with ambiguous body movement and partial occlusion.

Convolutions were replaced by depthwise separable and grouped convolutions that are inspired by MobileNet [20] in lightweight architecture design, while adding 3-dimensional convolutions for the temporal dynamics and 2-dimensional convolutions for the spatial depth. Despite the reduced computational complexity, the authors reported 100 percent accuracy on Movies, 94.71 percent on Hockey, 91.41 percent on Crowd Violence, and 81.98 percent on RWF-2000 [4], and the number of parameters was reduced to 0.35 million, and the speed of inference on CPU exceeded 128 frames per second.

A similar 3-stage architecture was proposed with a time-distributed U-Net with a MobileNetV2 encoder for extracting spatial features per frame, followed by the LSTM to capture motion over 30 consecutive frames, and dense layers for classification into a binary class [21]. While it performed well at 96.1 percent accuracy on Hockey Fights and 99.5 percent on Movie Fights, on the more complex RWF-2000 [4] benchmark, it hit 82.0 percent accuracy, still struggling to match the accuracy of simpler datasets, and it was unable to differentiate between various types of violent events due to its binary output.

In addition to fully labeled training data, Weak Supervision was explored to train the person detector by providing examples of a full spatiotemporal action tube, summarized as a Dynamic Image summarizing 16 frames of motion [22]. The temporal features were extracted using a 3D-CNN, the spatial features using a 2D-CNN, followed by a Channel Fusion and Attention Mechanism, and finally, the video-level classification using Multiple Instance Learning, which requires only video-level labels instead of manually drawn bounding boxes. The method performed well on Hockey Fight with 97.30 percent accuracy, RLVS 92.88 percent, RWF-2000 [4] 88.71 percent, and achieved a 31.92 percent localization error on the UCFCrime2Local data set but suffered from a loss of accuracy in crowded scenes, when occlusion was present, and when the body position was unusual.

2.1.3 Advanced Deep Learning Methods

More recent work has introduced attention mechanisms, efficient video-native architectures, and specialized multi-branch designs to push violence detection performance further while often also targeting real-time deployment. One such model, built around frame differencing to emphasize motion, used a ResNet50 backbone for spatial feature extraction, a ConvLSTM layer for temporal modeling, a Multi-Head Self-Attention module to highlight important spatiotemporal regions, and a second ConvLSTM to refine these attended features before final classification [23]. This attention-augmented architecture reached an AUC of 97.48 percent on UCF-Crime [7] and 96.21 percent accuracy on RWF-2000 [4], outperforming several prior state-of-the-art methods, though the authors noted that the model did not use raw RGB appearance, audio, or explicit body-part detection, any of which might further improve performance.

Efficiency-focused work also explored adapting video-native three-dimensional convolutional networks originally trained for general action recognition. One study fine-tuned an X3D-M model, originally pre-trained on the Kinetics-400 action recognition dataset, comparing a fully fine-tuned version against a transfer-learned version that froze the base network and trained only three dense layers on top [24]. Input clips were downsampled to sixteen frames at 256 by 256 resolution, and the fully fine-tuned model reached around 99.5 percent accuracy on Crowd Violence and 91 percent on RWF-2000 [4] with only about 2.98 million parameters, remaining accurate even under heavy video compression, though the model generalized poorly to crowd scenes it had not specifically been trained on, and sudden ordinary movements such as running sometimes triggered false alarms.

A three-module design separated spatial motion extraction, short-term temporal modeling, and long-term temporal modeling into distinct stages [25]. The first module removed background using frame differencing and morphological operations, the second module combined three consecutive frames into a single temporal image processed by MobileNetV2, and the third module used global average pooling and dense layers to capture long-range temporal patterns before classification. This staged approach reached 88.50 percent accuracy on RWF-2000 [4], 89.29 percent on a newly introduced surveillance dataset collected from CCTV cameras in Peru, 98.20 percent on Hockey, and 100 percent on Movies, running at 30 frames per second, though its accuracy on complex or heavily occluded scenes remained below that of larger transformer-based alternatives.

Reflecting a continued push toward practical deployment, one recent study built a complete real-time system around MobileNetV3Small, training on both RWF-2000 [4] and the Real-Life Violence Dataset with resizing, normalization, and augmentation, then deploying the trained model through a web application with automated alerting [26]. The system reached 95 percent accuracy, 88 percent precision, and a 92 percent F1-score while processing roughly 35 frames per second at about 80 milliseconds of latency per frame, though the authors acknowledged that dataset bias limited generalization to all real-world conditions and that false positives could still occur in crowded or visually complex scenes. A separate practical framework combined a MobileNet-SSD person detector, a 3D-CNN for spatiotemporal feature extraction over 50-frame sequences, and a SoftMax classifier, reaching 92.5 percent accuracy on RWF-2000 [4] and 97.2 percent on Hockey Fight while processing 72 frames per second, though the design currently relies on single-camera visual data alone, without audio or multi-view input that might help under difficult lighting conditions [27].

A simpler CNN-LSTM baseline sampled twenty frames per clip, resized and normalized them, extracted spatial features with a custom CNN, and modeled temporal dependence with an LSTM before a sigmoid-activated dense layer produced the final classification [9]. This more conventional architecture reached 68 percent accuracy, 72 percent precision, and 68 percent recall on RWF-2000 [4], a comparatively modest result that the authors attributed partly to the fixed twenty-frame sequence length, which may fail to capture longer violent events, and partly to sensitivity to extreme lighting change, heavy occlusion, and low-resolution footage, leading them to suggest that attention mechanisms or variable-length sequence modeling could improve results in future work.

2.1.4 ROI-Guided and Human-Focused Methods for Violence Detection

Violence detection in surveillance videos primarily involves analyzing interactions and movements between people. Therefore, focusing on the human subjects and their associated motion can help reduce the influence of irrelevant background information and appearance variations. In this context, human-focused approaches aim to extract information related to the regions, poses, and movements that are most relevant to the violent activity. Pose estimation represents the human body using keypoints or skeletal structures, while frame differencing emphasizes regions of temporal change between consecutive frames. These approaches can also be fused to combine complementary information from human pose and motion, potentially providing a more informative representation for violence detection.

2.1.4.1 Pose estimation

In one recent study, the specific contribution of pose estimation toward violence detection was explored by presenting two identical ConvLSTM models, one via RGB frames and the other via MoveNet [12] pose-estimated skeletal frames from the RWF-2000 dataset [4]. In all metrics measured, the pose-estimated model achieved better performance than the RGB baseline, with a higher F1 score, precision, and specificity (14 percent higher), and a higher recall (2 percent higher), showing a decrease in false positives and a general improvement in the balance of precision and recall. The authors wrote that errors in pose estimation (e.g., missing or inaccurate keypoints) may lead to misrepresentation of the resulting motion and impact classification results, and that the method was best suited for human-to-human violence, but was less effective when objects were part of the violence [8].

Transformer-based keypoint extraction has also been investigated for pose estimation instead of convolutional pose models. Another work adopted a transformer architecture called TokenPose to extract human keypoints and generate heatmaps of the keypoints, and then input them into a 3D-CNN that learned the motion patterns of the frames for final classification [28]. This transformer-based pipeline achieved an accuracy of 89.45 percent on the RWF-2000 validation set, outperforming a standard ConvLSTM baseline; a C3D model, and a 3D-CNN model using optical flow, with accuracies of 77.0 percent, 82.75 percent and 87.25 percent, respectively; while also achieving an accuracy of 92.7 percent on the authors' own school violence set, although the authors noted that transformer-based pose models can fail to reliably detect keypoints when multiple subjects overlap in a single

frame.

2.1.4.2 Frame differencing

The UCF-Crime dataset [7] was used in a comparison, where 20 frames were sampled regularly for the classic frame sampling approach, and 20 frames were computed as the difference between two consecutive frames for the motion-based approach; both methods were fed into the same ResNet50-LSTM classifier [6]. The frame difference approach achieved a higher micro-average AUROC of 0.91 compared to the standard frame approach with AUROC 0.89 and used approximately 2.9 per cent less training time and 4.2 per cent less testing time, indicating that the frame difference approach is both more accurate and potentially more practical for real-time use in this context. The authors noted that the method is highly dependent on observable movement and it is not effective in situations of subtle, slow or low movement violence, nor does it account for contextual information like weapons or other objects.

2.1.4.3 Fusion of pose and motion information

A study [10] proposed a two-stream network with human skeletons extracted by pose estimation and a change detection stream to achieve a trade-off of detection performance and computational complexity by using ConvLSTM layers in the change detection stream. With 62,583 trainable parameters, it achieved a maximum validation accuracy of 90.25 percent on RWF-2000 [4], which is quite a small model compared to most similar architectures, but nevertheless was dependent on the accuracy of the pose extraction step, and errors in keypoint extraction may be carried over to the final classification.

In this more elaborate fusion architecture, the authors extracted human pose using YOLO-NAS pose estimation against a black background, and the motion information was extracted by frame differencing, which was processed by ConvLSTM respectively and then fused by weighted sum with ReLU and Tanh activation functions and then classified by FC layers [11]. This dual-stream fusion achieved an accuracy of 95.05 percent on RWF-2000 [4], and performed competitively on other benchmarks, such as Hockey (98.50 percent), Movie (99.50 percent), Crowd (95.40 percent), RLVS (96.80 percent), and SCFD (91.33 percent), with an average F1-score of 96.12 percent and an inference time of approximately 0.33 seconds per clip. The authors stated that the method did not use any RGB appearance information, which can make it difficult to identify truly violent actions from similar-looking but non-violent actions like rough play, and that it had been primarily tested on the RWF-2000 dataset [4], and not a wider range of other datasets.

Finally, a method was proposed that fused pose estimation with a thin RGB stream, but did not eliminate the appearance information. This approach involves pre-training a lightweight 2D CNN for the primary classification of violence using human skeleton keypoints extracted by YOLO-Pose, and then a Weight Selection Module added the keypoints' confidence scores from the skeleton with the RGB features, followed by a Keyframe Weight Assignment module to highlight the most informative frames before making the final prediction [29]. It achieved a Hockey Fights accuracy of 97.95 percent and a RWF-2000 [4] accuracy of 93.28 percent at 33 frames per second with a 40 percent

clip length, with accuracy dropping slightly as clip length increased, but the authors stated that when the number of keypoints detected was below 9, due to occlusion or part of the person being outside the frame, the network resorted to using RGB features alone, which reduced accuracy.

Overall, prior studies demonstrate a progression from handcrafted features and classical classifiers to deep learning, attention-based architectures, and human-focused motion and pose-based methods. Despite the strong performance of recent approaches, challenges remain due to background interference, pose estimation errors, limited observable motion, and computational complexity. These limitations motivate further investigation into effective combinations of motion and pose information for robust violence detection in real-world surveillance videos.

2.2 Challenges Faced

Real-world surveillance footage differs substantially from the curated benchmarks that came before it in this field, and that difference is a recurring theme across the studies reviewed above. Lighting, camera angle, resolution, and the subject’s distance from the camera all vary widely from one deployment to another, and a model trained under one set of these conditions frequently loses accuracy when tested under another. This transfer problem shows up as early as the handcrafted-feature era, where STIP and MoSIFT descriptors that reached over 90 percent accuracy on Hockey Fight fell to 59 to 89 percent on the more visually varied Movies dataset [14], and it persists into the deep learning era, where several of the pose- and motion-based methods reviewed above report clear performance drops on crowded or heavily occluded footage relative to their results on cleaner benchmarks [25] [10] [28] [29].

Motion-based preprocessing is only as informative as the movement it can detect. Optical flow and frame differencing both depend on visible displacement between frames, so they tend to under-represent violent behavior that is subtle, slow, or largely static, such as a stand-off that has not yet escalated into physical contact. Pose-based preprocessing carries a related but distinct weakness: it depends on the pose estimator correctly locating each person’s joints, and that estimation degrades when people are occluded, overlap with one another, or fall partly outside the camera frame, conditions that are common rather than rare in genuine CCTV footage. Several of the studies reviewed in Section 2.1.4 report exactly this failure mode, with keypoint errors under occlusion or crowding propagating directly into misclassification [8] [10] [29].

A further challenge concerns what is lost when RGB appearance is removed in favor of motion or pose alone. More than one study notes that appearance carries information neither motion nor pose fully replaces, since a rough but non-violent physical activity, playful shoving between friends, contact sport, can produce a motion or pose signature close to that of an actual fight, and without visual context to distinguish the two, a model built purely on motion or skeletal structure has less to work with. The authors of the weighted pose-and-motion fusion method make this point directly, noting that the absence of RGB appearance made it difficult to separate genuinely violent actions from visually similar but non-violent ones [11]. This tension runs through most of the pose- and

motion-based literature and is a factor this study has to weigh directly, since every method it proposes or compares, MoveNet on Frame Difference included, works from motion and pose signals rather than raw appearance.

Finally, accuracy and computational cost continue to pull against each other. Architectures accurate enough to handle difficult real-world footage, deeper backbones, multi-stream fusion, attention modules, are frequently too heavy for real-time deployment on the kind of hardware a surveillance system would actually run on, while the lighter architectures built for real-time use tend to give up some accuracy on harder benchmarks such as RWF-2000. This trade-off is visible across nearly every study surveyed in Section 2.1, from the parameter counts reported alongside accuracy figures to the explicit design goals stated by the authors of the more lightweight architectures [20] [25] [26] [27].

2.3 Scope of Study

This study does not retrain a raw-RGB baseline on RWF-2000. Two prior studies already cover that ground. Bhuiyan et al. [8] compared a background-suppressed MoveNet pose model against a raw RGB baseline on RWF-2000 and found the pose model ahead on every metric measured, including a 14-point gain in specificity. Sani et al. [6] compared frame-differenced input against standard RGB frame sampling under a shared ResNet50-LSTM architecture and found frame differencing ahead on AUROC, though that comparison was run on UCF-Crime [7] rather than RWF-2000 [4]. Taken together, these two studies establish that a background-suppressed pose signal and a frame-differenced motion signal each individually outperform raw RGB, which is the premise this study starts from rather than re-tests.

What neither of those studies, nor any other work reviewed in Section 2.1.4, tests directly is what happens when the two signals are combined, and specifically whether the order in which they are combined matters. This is the gap this study is built to fill. It proposes MoveNet on Frame Difference, a preprocessing method that applies MoveNet pose estimation [12] to already motion-differenced video, on the reasoning that differencing first concentrates the pose estimator’s attention on the regions of the frame where movement is actually occurring, rather than asking it to locate people against a full, unfiltered scene.

To evaluate this proposal, the study compares MoveNet on Frame Difference against four other preprocessing methods under one fixed EfficientNet-B0 + LSTM classifier [13] [5], so that any difference in outcome can be attributed to the preprocessing rather than to the model. Frame Difference and MoveNet Without Background serve as the two single-signal comparison methods already shown elsewhere to beat raw RGB, MoveNet with the original background retained serves as a check on whether background removal is actually necessary or whether the skeleton alone is doing the work, and Frame Difference on MoveNet, the same two operations applied in reverse order, serves as the study’s ablation: since it uses exactly the same two ingredients as MoveNet on Frame Difference and differs only in which operation runs first, any gap between the two variants isolates the effect of processing order specifically, separate from the effect of using motion and pose information at all.

Chapter-3: Dataset

3.1 The RWF-2000 Dataset

Cheng et al. proposed a large-scale benchmark for violence detection, the RWF-2000 (Real-World Fighting) dataset [4], which was specially designed for the task of violence detection in real-world surveillance scenarios. The dataset was designed to overcome several limitations of previous benchmarks, which often were plagued by one or more of the following issues: low image quality, limited number of clips, very long video clips, and source videos that were not representative of violence captured by surveillance cameras. In order to address all these problems at once, RWF-2000 [4] was put together.

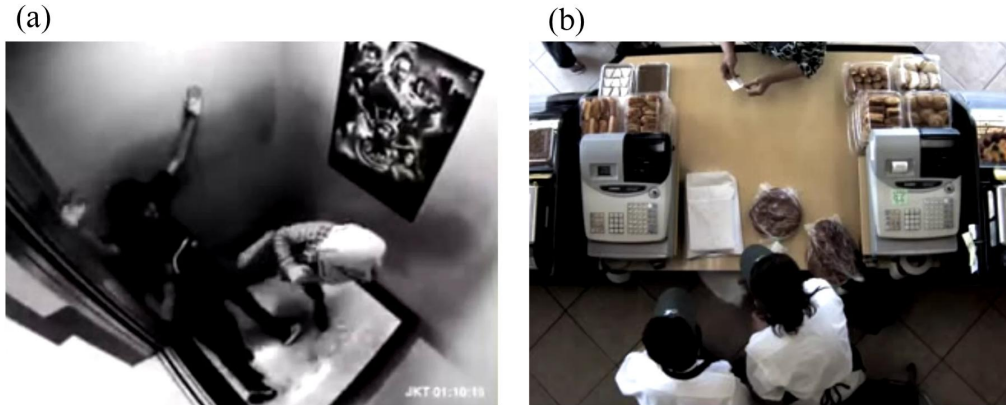


Figure 3.1: Sample Frames from the RWF-2000 Dataset. The Frames Show (a) a Frame from Fight Video, and (b) a Frame from Non-Fight Video.

3.2 Data Collection and Composition

The entire 2,000 clips in the dataset were obtained from the video-sharing site YouTube by using the keywords related to violence in various languages, such as English, Chinese, German, French, Japanese, and Russian. The multilingual approach was intentionally implemented to promote geographic and cultural diversity and to avoid skewing the

collection based on footage from one particular area. The downloaded videos were checked for any unwanted content, and then each video was divided into defined 5-second clips at 30 FPS [4]. Clips that were infeasible and/or not from a legitimate source camera were removed in a subsequent manual review. All other clips were labeled at the video level as Violent or Non-Violent.

There are actually 2,000 such clips in the dataset, of which 1000 are Violent, and 1000 are Non-Violent [4].

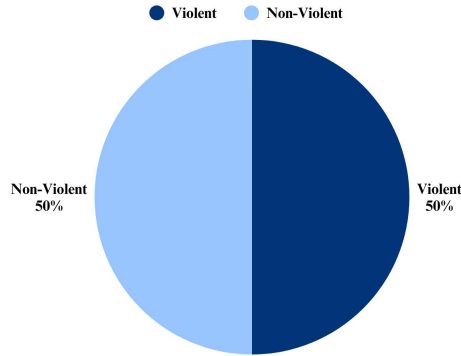


Figure 3.2: Class Distribution of the RWF-2000 Dataset.

All clips are authentic from surveillance cameras and not altered by post-processing. No restriction was placed on the type of violence; any subjectively identified violence, including fighting, robbery, explosion, shooting, or assault, was eligible for the selection criteria. This resulted in a range of violent incidents from personalised physical altercations to larger-scale disturbances of crowds, from indoor to outdoor environments and in a variety of lighting.

The clips cover a wide spectrum of video resolutions, including significant numbers in 240p, 360p, 720p, and 1080p, and the diversity of camera hardware used in real surveillance deployments [4]. Consistently, the balance between Violent and Non-Violent clips is maintained across all resolution levels, with no significant clustering of clips at any single level. The wide range of resolutions and light levels, coupled with the variety of types of scenes, make RWF-2000 [4] suitable for training and evaluating models for realistic deployment, not controlled laboratory conditions.

3.3 Dataset Partitioning and Leakage Prevention

About 1,000 separate surveillance video recordings were used to produce 2,000 clips [4]. Since several clips may be based on the same source video, the authors have taken special care to prevent data leakage from one partition to another. An automated script was used to place the clips in the partitions while ensuring that no two clips from the same source video were placed on either side of the training and test clips. To further validate, a color histogram was first computed for each clip, and pairwise cosine similarity was computed for the clips in the same partition. The top 30 percent of pairs in similarity score were examined manually to ensure that no near duplicate clips were split across partitions.

Following these procedures, the dataset was split into a training set of 1,600 clips (80 percent) and a test set of 400 clips (20 percent), while keeping the same class distribution for each set [4].

3.4 Comparison with Prior Datasets

When it was released, RWF-2000 [4] was the largest publicly available dataset of surveillance videos that were specifically gathered for the purpose of violence detection. Previous benchmarks like Hockey Fight, Movies Fight, and Crowd Violence were limited in scope, limited to scenes, or non-surveillance sources such as sports broadcasts or edited films. The UCF-Crime dataset [7] was significantly larger, but consisted of real surveillance data with videos of up to 10 minutes and only video-level annotations, making the task of fine-grained detection challenging. RWF-2000 [4] overcomes these problems by providing 2,000 video-level, clearly annotated five-second video clips from real footage.

Table 3.1 presents the comparison of RWF-2000 with previously established violence detection datasets, reproduced from Cheng et al. [4].

Table 3.1: Comparison of violence detection datasets

Dataset	Data Scale	Length/Clip (sec)	Resolution	Annotation	Scenario
BEHAVE [30]	4 videos (171 clips)	0.24–61.92	640 × 480	Frame-Level	Acted Fights
RE-DID [31]	30 videos	20–240	1280 × 720	Frame-Level	Natural
VSD [32]	18 movies (1,317 clips)	55.3–829.4	Variable	Frame-Level	Movie
CCTV-Fights [33]	1,000 clips	5–720	Variable	Frame-Level	Natural
HockeyFight [34]	1,000 clips	1.6–1.96	360 × 288	Video-Level	Hockey Games
MoviesFight	200 clips	1.6–2	720 × 480	Video-Level	Movie
CrowdViolence [35]	246 clips	1.04–6.52	Variable	Video-Level	Natural
SBU Kinect Interaction [36]	264 clips	0.67–3	640 × 480	Video-Level	Acted Fights
UCF-Crime [7]	1,900 clips	60–600	Variable	Video-Level	Surveillance
RWF-2000 [4]	2,000 clips	5	Variable	Video-Level	Surveillance

The comparison in Table 3.1 shows that the datasets used for violence detection differ substantially in scale, video duration, annotation type, and recording conditions. In contrast to datasets collected from acted scenes, movies, or specific sports environments, RWF-2000 provides a consistent surveillance-based setting with video-level annotations. This makes it suitable for evaluating preprocessing techniques under a common experimental framework. Therefore, RWF-2000 was selected as the primary dataset for the experiments conducted in this study.

3.5 Dataset Split Used in This Study

In this thesis, a partitioning scheme different from the two-way split used in the benchmark paper is used. The original release partitions the dataset into 80% data used for training and 20% for testing, but this study proposes a three-way partitioning, allowing for correct validation while training the model.

The test partition was fixed to the 400-clip benchmark paper and was not modified at all, and was completely withheld from all training and tuning operations. The rest of the clips (800) were split again into an 80:20 ratio, resulting in a training set and a validation set. In the complete 2,000-clip dataset, this gives the following distribution:

1. Training set: 64 percent of the total dataset (1,280 clips), used to optimize model parameters through backpropagation.
2. Validation set: 16 percent of the total dataset (320 clips), used to monitor training progress, detect overfitting, and guide any hyperparameter adjustments during the training process.
3. Test set: 20 percent of the total dataset (400 clips), held out completely and used only once for the final evaluation of the trained model.

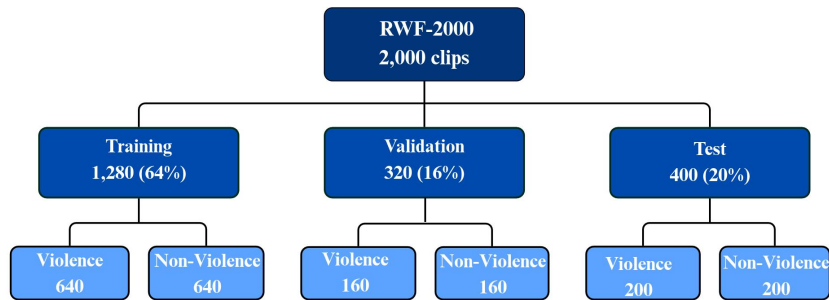


Figure 3.3: Dataset Partitioning Used in This Study.

This 3-way split makes it possible to separate the data the model learns from the data to guide training decisions, and the data on which the final result is reported. The figures displayed for the test set in Chapter 5 are thus based on footage that is not part of the model’s direct or indirect experience from any previous point.

In summary, RWF-2000 provides a diverse and challenging benchmark for evaluating violence detection under realistic surveillance conditions. Its balanced class distribution, fixed five-second clips, varied environments, and surveillance-based footage make it well suited to the objectives of this study. The dataset was further divided into training, validation, and test sets while keeping the original benchmark test set completely isolated from model development. This partitioning ensures that model training and tuning are conducted independently of the final evaluation data, thereby reducing the risk of data leakage and providing a reliable basis for comparing the different motion- and pose-based preprocessing approaches investigated in this thesis.

Chapter-4: Methodology

4.1 Overview

The methodology of this study is organized around a single controlling question: does the choice of preprocessing applied to surveillance video before classification change how well violence can be detected, and if two preprocessing operations are combined, does the order in which they are applied matter? To answer this cleanly, every experiment in the study shares the same classifier architecture, the same training configuration, the same dataset, and the same evaluation protocol. The only thing that changes between experiments is the form of the video fed into the model.

The pipeline proceeds in two stages. In the first stage, the raw video clips from the RWF-2000 dataset [4] are transformed into one of two preprocessed representations. The proposed method applies frame differencing first, producing a video in which static background regions become dark and only regions of inter-frame pixel change remain visible, and then applies MoveNet [12] pose estimation to each of those motion-residual frames, annotating 17 human skeletal keypoints per detected person. This is referred to throughout as MoveNet on Frame Difference. The ablation comparison reverses the order: MoveNet [12] is applied first to the original RGB frames, and frame differencing is then computed on the resulting skeleton-annotated sequence. This is referred to as Frame Difference on MoveNet [12]. Both preprocessing pipelines feed into an identical classification model.

In the second stage, the preprocessed video is classified by a fixed EfficientNet-B0 [13] and LSTM [5] architecture. EfficientNet-B0 [13] acts as the spatial feature extractor, mapping each sampled frame to a compact feature vector. An LSTM network [5] reads the sequence of those feature vectors across time and produces a single video-level representation. A linear classification layer then maps that representation to a binary prediction: Fight or NonFight. Figure 1 illustrates the complete pipeline of proposed method from raw input through preprocessing and classification to output.

The experiment also includes three additional preprocessing variants for the purposes of broader comparison: raw frame difference, MoveNet skeleton overlay on original RGB frames, and MoveNet skeleton without background. These five conditions, together with the raw video baseline taken from Bhuiyan et al. [8], form the full set of results reported in Chapter 5.

4.2 Preprocessing

Two preprocessing operations are at the core of this study: frame differencing and MoveNet pose estimation[12]. This section describes each operation independently before explaining how they are combined.

4.2.1 Frame Difference

Frame differencing is a technique for isolating motion in a video by removing the parts of the image that did not change between consecutive frames. All frames first extracted from each video. Then, the absolute difference between two consecutive frames is calculated to get the frame difference.

$$\Delta F_t = |F_t - F_{t-1}| \quad (4.1)$$

where $F(t)$ is the current frame and $F(t-1)$ is the previous frame [6]. Pixels where the scene was static produce difference values close to zero and appear dark in the output. Pixels where something moved produce larger values and appear bright. Applied to a surveillance clip, this operation suppresses fixed background elements such as walls, floors, parked vehicles, and static objects. What remains visible in the output is primarily the region occupied by moving bodies. In a violent clip, those moving bodies are the people involved in the altercation. In a non-violent clip with little physical activity, the difference images are sparse and predominantly dark throughout the clip. The downstream classifier therefore receives a representation in which the spatial distribution of motion is the primary signal rather than the appearance of the full scene. The first preprocessing variant used in this study applies this operation to the original RGB video and uses the resulting motion-residual video directly as the classifier input.

4.2.2 MoveNet

MoveNet is a pose estimation model developed by Google and made available through TensorFlow Hub [12]. The multipose Lightning variant was used in this study because it supports detection of multiple people within a single frame, which is necessary for violence scenarios that involve two or more individuals. MoveNet [12] is a bottom-up estimator: rather than detecting each person first and then estimating their joints, it predicts all keypoints across the full image and groups them into individual person instances afterward. Its architecture consists of a feature extractor based on a modified MobileNet backbone and prediction heads whose design is inspired by the CenterNet object detection framework [12].

For each detected person in each frame, MoveNet predicts 17 body keypoints: nose, left and right eye, left and right ear, left and right shoulder, left and right elbow, left and right wrist, left and right hip, left and right knee, and left and right ankle. Each keypoint is returned as a normalized (y, x, confidence) triplet [12]. Skeletal connections between adjacent keypoints are then drawn using a predefined set of 18 limb pairs, producing a stick-figure representation of the person’s posture[12].

For the MoveNet variant, these skeletal representation are drawn on top of the original RGB frame, preserving the background scene. For the MoveNet Without Background variant, the original frame is replaced with a black canvas, so the output contains only the skeleton representation against a dark background.

4.2.3 Proposed Method: MoveNet on Frame Difference

The proposed preprocessing pipeline combines the two operations above in a specific order: frame differencing is applied first, and MoveNet pose estimation [12] is applied second. The frame difference videos produced by Equation 1 serve as the direct input to MoveNet. Rather than estimating human pose from the raw RGB surveillance videos, MoveNet [12] receives a motion-aware video in which the static background has already been removed.

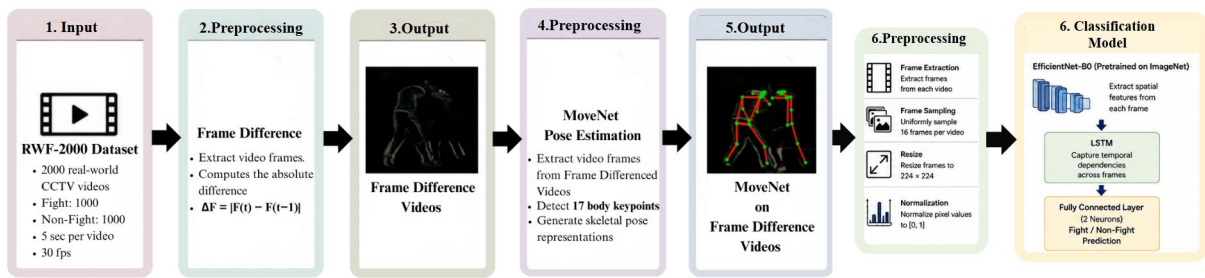


Figure 4.1: Proposed Method Overview

The design reasoning behind this ordering is as follows. When MoveNet [12] processes a full RGB surveillance video, it must locate body parts against a background that may contain texturally complex or visually busy elements: patterned floor tiles, parked vehicles, signage, or groups of bystanders. Frame differencing removes most of this complexity before pose estimation runs. The regions of the video that remain bright after differencing correspond to where bodies were moving, and MoveNet’s task is therefore focused on those regions from the outset. The resulting output is a video in which the skeleton is drawn on a

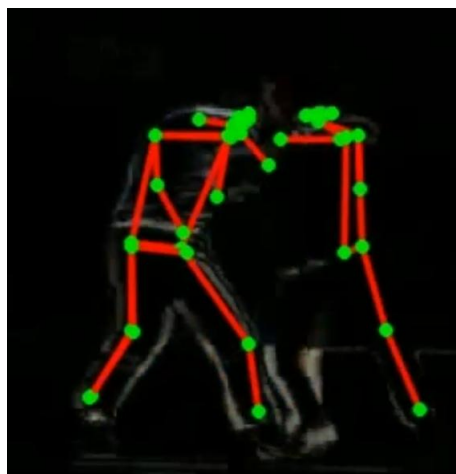


Figure 4.2: Sample Output Frame from MoveNet on Frame Difference Preprocessing

motion-residual background rather than the original scene, combining the motion-isolation property of frame differencing with the structural body-geometry information provided by skeletal keypoint annotation.

4.2.4 Ablation: Frame Difference on MoveNet

To test whether the ordering of these two operations matters, an ablation condition reverses the sequence. MoveNet [12] is applied first to the original RGB frames, producing skeleton-annotated video in which the full background remains visible. Frame differencing is then applied to this annotated sequence, computing the absolute difference between consecutive skeletal representative frames. The output captures where the skeletal representations moved between frames rather than where the original pixels changed.

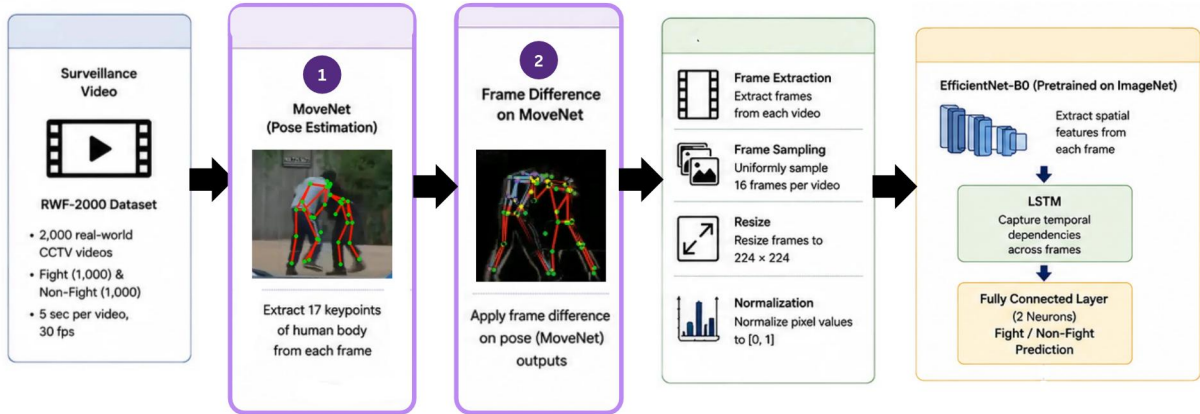


Figure 4.3: Ablation Method Overview

This reversed order has a known structural weakness that the ablation is designed to confirm quantitatively. MoveNet’s [12] keypoint predictions are not perfectly stable frame to frame: even for a person standing still, individual joint detections shift slightly between frames due to the stochastic nature of neural network inference on slightly different pixel inputs. When frame differencing is applied to the resulting skeleton video, these small



Figure 4.4: Sample Output Frame from Frame Difference on MoveNet Preprocessing

fluctuations appear as nonzero values in the output even in the absence of actual body movement. The difference signal therefore contains a mixture of genuine limb motion and detection noise, and the classifier must separate them. Whether this degradation is severe enough to meaningfully hurt performance relative to the proposed ordering is an empirical question that the ablation addresses directly.

4.3 Classification Model

For the video classification stage, an EfficientNet-B0 + LSTM [13] [5] architecture was selected. EfficientNet-B0 [13] was used to extract spatial features from individual video frames, while a Long Short-Term Memory (LSTM) network [5] was used to model the temporal relationships between consecutive frames. This architecture handles the spatiotemporal nature of video classification by processing spatial and temporal information at separate, dedicated stages rather than jointly through a single volumetric operation, which reduces computational demand while preserving the model’s ability to reason across time.

4.3.1 Frame Sampling

Each video in the dataset [4] contains 150 frames. Feeding all 150 frames through the classifier would be computationally expensive and would include a great deal of redundant information, since consecutive frames in a 30fps video are highly similar. Instead, 16 frames are sampled uniformly from each clip by dividing the clip into 16 equal intervals and selecting one frame from each interval. This gives a compact but temporally distributed representation of the clip’s content. For videos shorter than 16 frames, the last valid frame is repeated until the sequence reaches the required length.

Each sampled frame is resized to 224×224 pixels to match EfficientNet-B0’s [13] expected input resolution, converted from BGR to RGB channel ordering, and normalized to the range $[0, 1]$ by dividing pixel values by 255.

4.3.2 EfficientNet-B0: Spatial Feature Extraction

Each of the sixteen sampled frames was passed independently through EfficientNet-B0 [13], a convolutional neural network pretrained on the ImageNet dataset [37]. EfficientNet-B0 [19] was selected because it provides strong spatial feature extraction at a relatively low parameter count. The network’s original classification head was replaced with an identity mapping so that the backbone outputs a 1,280-dimensional feature vector per frame rather than a probability distribution over 1,000 ImageNet [37] classes. No convolutional layers were frozen during training; all backbone parameters were updated through backpropagation together with the LSTM weights and the classification head. This allows the pretrained filters to adapt to the visual characteristics of each preprocessed variant rather than remaining fixed to general natural-image features.

The backbone processes each frame independently, without access to adjacent frames. Temporal information is not captured at this stage. The output is a sequence of sixteen feature vectors, each of dimension 1,280, which together encode the spatial content of the sixteen sampled time points. This sequence is passed to the LSTM [5] for temporal modeling.

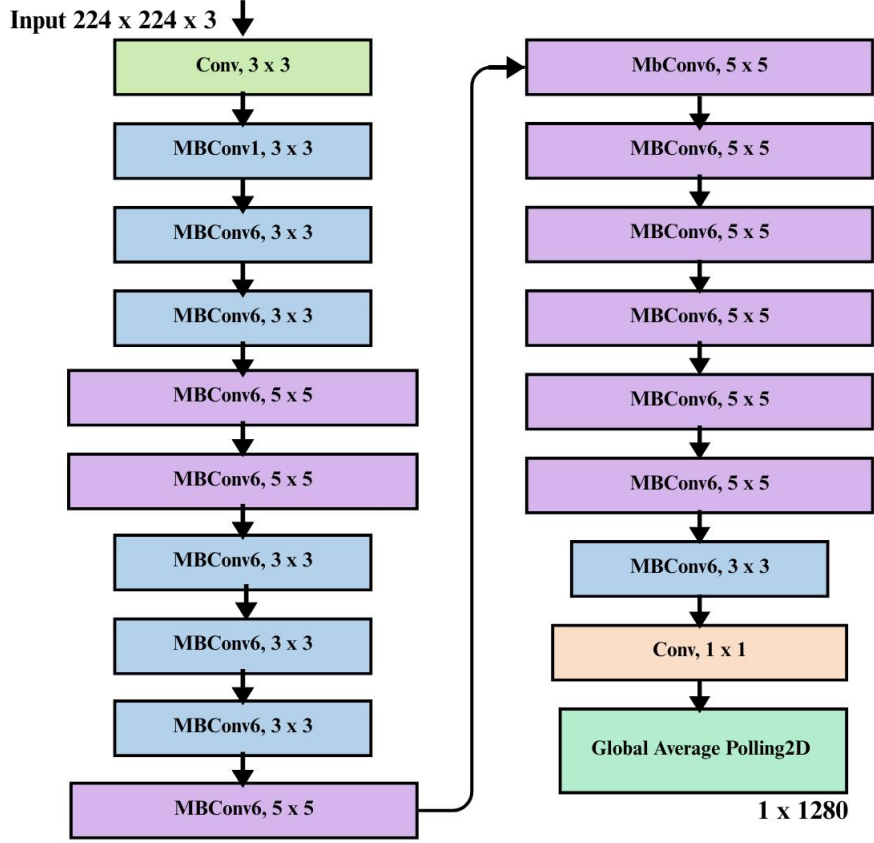


Figure 4.5: EfficientNetB0 Achitecture Diagram

4.3.3 LSTM: Temporal Aggregation

The sequence of sixteen 1,280-dimensional feature vectors is fed into a single-layer Long Short-Term Memory [5] network with 128 hidden units. The LSTM is configured with `batch-first=True` so that the input tensor has dimensions (batch size, 16, 1,280). The network processes one feature vector per time step, updating its internal cell state and hidden state at each step according to the standard LSTM gating equations. The hidden state at the final time step, after the network has processed all sixteen vectors, is retained as the video-level representation. Intermediate hidden states from earlier time steps are discarded. Figure 4.6 [38] shows the Long Short-Term Memory Diagram.

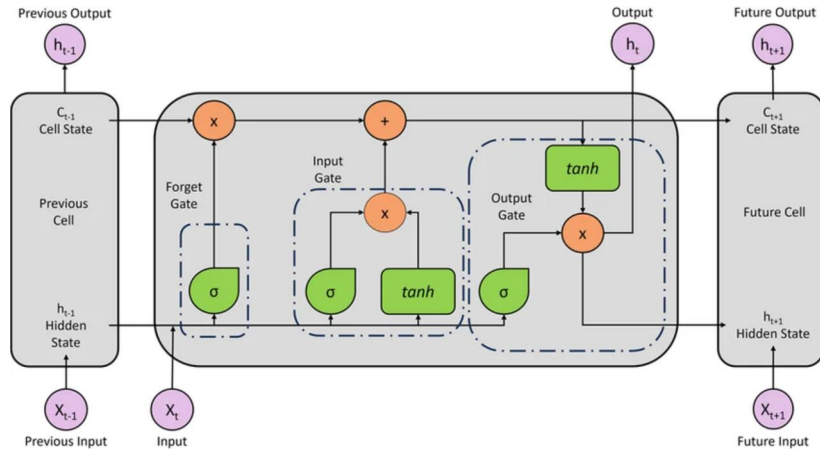


Figure 4.6: Long Short-Term Memory Model Diagram

This design allows the LSTM [5] to accumulate contextual information across the full sampled temporal sequence before producing the summary that drives classification. Unlike average pooling or maximum pooling over the feature sequence, the LSTM’s recurrent connections allow later time steps to be interpreted in light of what was seen at earlier steps, which is relevant for violence detection because a violent event often builds over time rather than appearing instantaneously in a single frame.

4.3.4 Classification Head

The 128-dimensional final hidden state from the LSTM [5] is passed to a fully connected linear layer with two output units. These two units correspond to the NonFight and Fight classes respectively. No activation function is applied at this stage within the model; the two raw logits are passed directly to the cross-entropy loss function during training. At inference time, a softmax function is applied externally to convert the logit pair to a probability distribution over the two classes. The class with the higher probability is taken as the model’s prediction. The Fight class probability is also separately retained for each test clip and used to compute Receiver Operating Characteristic curves as described in Section 4.5.

4.4 Training Configuration

All five preprocessing variants were trained using an identical configuration. The Adam [39] optimizer was used with a fixed learning rate of 0.0001 and no learning rate schedule or decay. Cross-entropy loss served as the training objective throughout all 100 training epochs. The batch size was set to 2, a value determined by GPU memory capacity: processing a batch of videos, each represented as a sequence of sixteen 224×224 frames through EfficientNet-B0 [13], places a substantial memory demand on the device, and a batch size of 2 was the largest that could be accommodated without memory errors.

Table 4.1: Hyperparameters and training settings applied identically to all five preprocessing variants.

Hyperparameter	Value
Frame Sampled per Video	16
Input Frame Resolution	224×224
Batch Size	2
Number of Training Epochs	100
Optimizer	Adam [39]
Learning Rate	0.0001
LSTM Hidden Units	128
Loss Function	Cross-Entropy

At the end of every training epoch, the model was evaluated on the validation set. The epoch that produced the lowest validation loss was identified, and the corresponding model

checkpoint was saved to disk. Final evaluation on the test set used this saved checkpoint rather than the weights from the last training epoch. This selection criterion guards against the scenario where a model achieves its best generalization at an intermediate epoch and then overfits during subsequent training, which would make last-epoch evaluation misleading. Table 4.1 summarizes the complete training configuration applied uniformly across all five variants.

4.5 Evaluation

Test set performance was measured using five standard binary classification metrics. Throughout, the Fight class is treated as the positive class. True positives (TP) are Fight clips correctly predicted as Fight; false positives (FP) are NonFight clips incorrectly predicted as Fight; true negatives (TN) are NonFight clips correctly predicted as NonFight; false negatives (FN) are Fight clips incorrectly predicted as NonFight.

Accuracy is the proportion of all 400 test clips classified correctly:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.2)$$

Precision reports what fraction of clips predicted as Fight were genuinely violent, penalizing false alarms:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.3)$$

Recall measures what fraction of all actual Fight clips the model identified correctly, penalizing missed detections:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.4)$$

The F1-score is the harmonic mean of precision and recall. It provides a single summary metric that balances both concerns and is particularly informative when the consequences of false positives and false negatives are not identical:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.5)$$

A confusion matrix was recorded for each of the five variants, reporting the raw TP, FP, TN, and FN counts. These matrices allow a finer-grained reading of model behaviour than accuracy alone: they reveal whether a model’s errors were concentrated in one direction, systematically missing violent clips, or systematically flagging non-violent clips as violent or distributed more evenly. This distinction matters for real-world deployment, where the costs of a missed violent event and a false alarm are typically asymmetric.

The Fight class probability produced by the softmax output was retained for every test clip and used to compute Receiver Operating Characteristic curves. The area under each

ROC curve (AUC) quantifies a model’s ability to rank violent clips above non-violent ones across all possible decision thresholds, without committing to a specific operating point. It is calculated as follows:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}) d(\text{FPR})$$

where TPR denotes the True Positive Rate and FPR denotes the False Positive Rate.

Training and validation loss and accuracy were recorded at each epoch for all five variants to allow examination of convergence behavior and to identify any overfitting patterns that were specific to a particular preprocessing condition.

Taken together, the proposed MoveNet-on-Frame-Difference approach is the central focus of the experimental methodology presented in this chapter. By applying frame differencing before pose estimation, the method aims to emphasize motion-related information while reducing the influence of static background content during pose extraction. Its effectiveness was examined alongside four other preprocessing variants and a reversed-order ablation using the same EfficientNet-B0 and LSTM classifier, training configuration, dataset partitioning, and evaluation protocol. This controlled setup allows the impact of the preprocessing strategy and the order of operations to be evaluated independently. Model performance was assessed on the held-out test set using accuracy, precision, recall, F1-score, confusion matrices, and AUC. The results of these experiments are presented and compared in the following chapter.

Chapter-5: Results and Analysis

5.1 Overview

This chapter reports the test-set performance of the EfficientNet-B0 + LSTM [13] [5] classifier trained on each of the five preprocessed input representations derived from the RWF-2000 dataset [4], along with the raw video baseline taken from Bhuiyan et al. [8]. All results are measured on the same held-out test partition of 400 clips (200 Fight, 200 NonFight), using the best-performing model checkpoint selected by minimum validation loss during training. Five metrics are reported throughout: accuracy, precision, recall, F1-score, and AUC. The chapter is organized into three sections. Section 5.2 presents the main comparative results across all five preprocessing variants and the external baseline. Section 5.3 presents the ablation study comparing the two sequential combination orderings. Section 5.4 summarizes the key findings.

5.2 Main Results: Preprocessing Variants Compared

Table 5.1 reports test-set performance for the preprocessing variants and the external raw video baseline. The variants are listed in ascending order of accuracy, from the weakest to the best-performing configuration.

Table 5.1: Comparative Results Across Preprocessing Methods

Variant	Accuracy	Precision	Recall	F1-Score	AUC
Raw Video [8]	68.50%	65.81%	77.0%	0.7090	–
MoveNet	72.50%	68.44%	83.5%	0.7523	0.7846
Frame Difference	76.50%	80.81%	69.5%	0.7473	0.8619
MoveNet W/o Background	78.00%	80.77%	73.5%	0.7696	0.8285
MoveNet on Frame Difference	81.50%	77.63%	88.5%	0.8271	0.8661

The proposed method, MoveNet on Frame Difference, achieves the highest performance across accuracy (81.50%), recall (88.5%), F1-score (0.8271), and AUC (0.8661). It is the only variant to exceed 80 percent accuracy on the test set.

The accuracy comparison in Figure 10 presents the classification performance obtained using the preprocessing methods. The accuracy values vary across the different approaches,

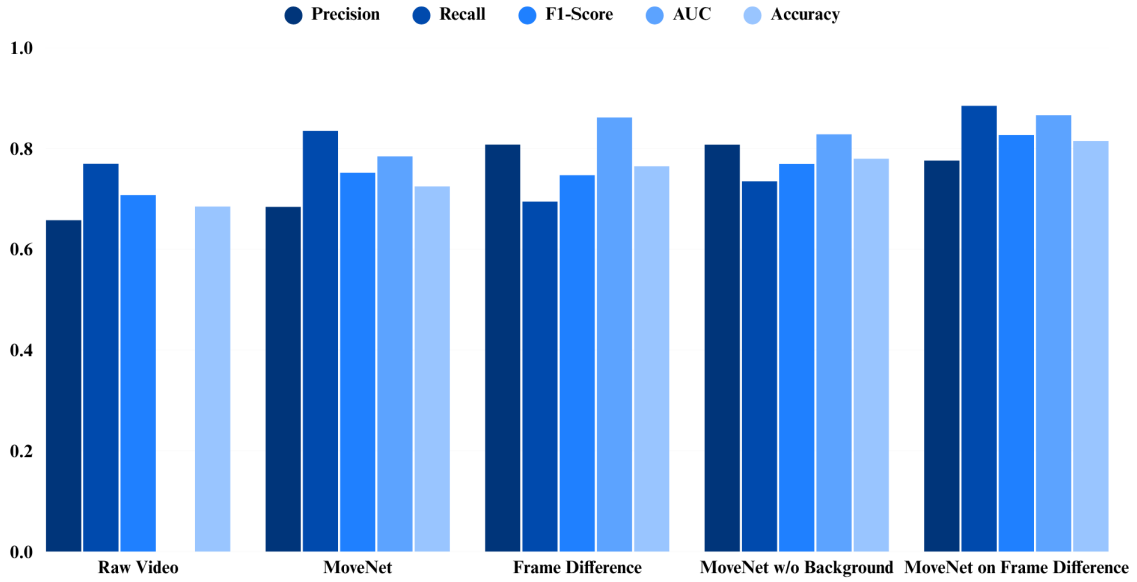


Figure 5.1: Performance comparison of different preprocessing variants on the RWF-2000 dataset.

ranging from **72.50% to 81.50%**. Frame Difference achieved an accuracy of **76.50%**, while MoveNet and MoveNet Without Background obtained **72.50%** and **78.00%**, respectively. The combination of MoveNet on Frame Difference resulted in **81.50%**, whereas Frame Difference on MoveNet achieved **74.00%**. These results show the differences in classification accuracy across the preprocessing configurations.

5.2.1 Raw Video Baseline

The raw video baseline, taken from Bhuiyan et al. [8] who used the same dataset with a ResNet50V2 + LSTM classifier on unprocessed RGB footage, achieved 68.50 percent accuracy with a precision of 65.81 percent, recall of 77.0 percent, and F1-score of 0.7090. AUC is not available for this condition as it was not reported in the source study. This baseline sets the lower bound for comparison: every preprocessing variant tested in this study outperforms it on accuracy.

The relatively high recall (77.0%) against the low precision (65.81%) in the baseline condition indicates that the raw RGB classifier was biased toward predicting Fight. It caught a reasonable proportion of actual violent clips but also flagged a large number of non-violent ones as violent. This pattern is consistent with a model that has not learned to reliably distinguish violent from non-violent motion patterns, and instead leans toward the positive class under uncertainty.

5.2.2 MoveNet

The MoveNet variant, which overlays skeletal annotations on the original RGB frames, achieved 72.50 percent accuracy. Among the five variants, it produced the second highest recall (83.5%), meaning it was the second most sensitive detector of genuinely violent clips after MoveNet on Frame Difference. However, its precision was the second lowest of any variant (68.44%), and its AUC was also the lowest (0.7846). The combination of high recall and low precision reflects the same false-alarm tendency seen in the raw baseline, but more pronounced: the model correctly identified the majority of violent clips but also flagged non-violent clips as violent at a high rate.

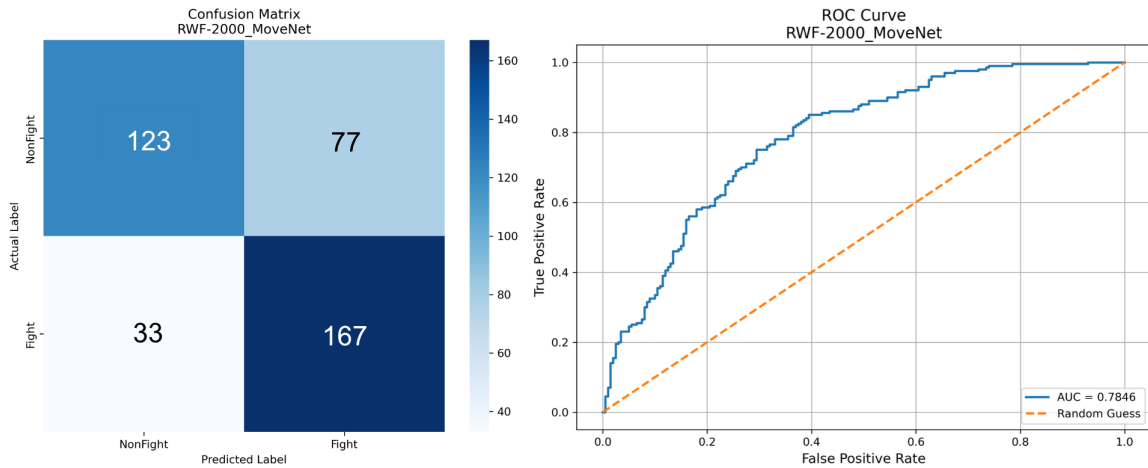


Figure 5.2: Confusion matrix and ROC curve for the MoveNet variant.

One likely explanation is that skeletal annotations drawn over a busy background scene still leave the classifier with substantial irrelevant visual information to process. Crowd scenes, complex environments, and non-violent physical activity such as people running or gesturing can produce skeleton configurations that superficially resemble aggressive postures when the background context is still present. Despite higher accuracy than the raw video baseline by four percentage points, MoveNet [12] on raw RGB is the weakest of the five variants this study evaluated.

5.2.3 Frame Difference

The Frame Difference variant achieved 76.50 percent accuracy. Its precision was 80.81 percent, the highest of any variant, indicating that when the frame-difference model predicted Fight, it was correct more than four times in five. However, its recall was the lowest of all variants at 69.5 percent, meaning it missed nearly one in three genuinely violent clips.

This precision-recall profile is the inverse of the MoveNet [12] pattern. Frame differencing produces a representation in which static background content is suppressed and only motion-

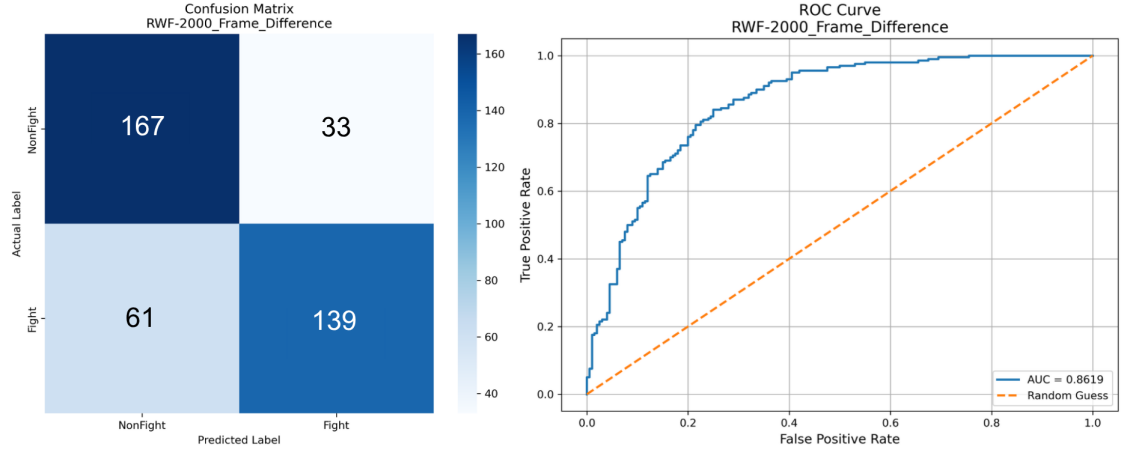


Figure 5.3: Confusion matrix and ROC curve for the Frame Difference variant.

active regions remain, but it carries no information about the structural configuration of the human body. A slow or subtle violent interaction, one where the object move gradually rather than rapidly, produces a sparse frame-difference signal that looks similar to non-violent low-motion footage. The classifier therefore tends toward caution: when it does predict Fight, it is usually right, but it misses a meaningful fraction of actual fights. The AUC of 0.8619 is the second highest of all variants, suggesting that the model’s underlying ranking of clips by violence probability is fairly good even when the threshold-based recall is low.

5.2.4 MoveNet without Background

Removing the background before rendering skeletal annotations improved performance over the MoveNet-with-background variant by 5.5 percentage points in accuracy (78.00% vs 72.50%). Precision rose from 68.44 to 80.77 percent and AUC improved from 0.7846 to 0.8285. Recall, however, fell from 83.5 to 73.5 percent. This trade-off indicates that eliminating the scene-level background reduced false alarms substantially: the model was less likely to misclassify a non-violent clip as violent when the background was absent. The cost was a reduction in sensitivity, with more actual violent clips going undetected relative to the MoveNet [12] variant.

This result supports the interpretation that background scene content in the MoveNet-with-background variant was introducing misleading visual information. When the classifier received full-color frames annotated with skeletons, it could associate incidental background features; crowd density, low lighting, specific location types - with the Fight class rather than learning from body-posture geometry alone. Removing the background forced the model to rely on the skeleton representation exclusively, which reduced both false positives and true positives relative to the condition where background context was available. MoveNet Without Background achieved the second highest accuracy of any variant and

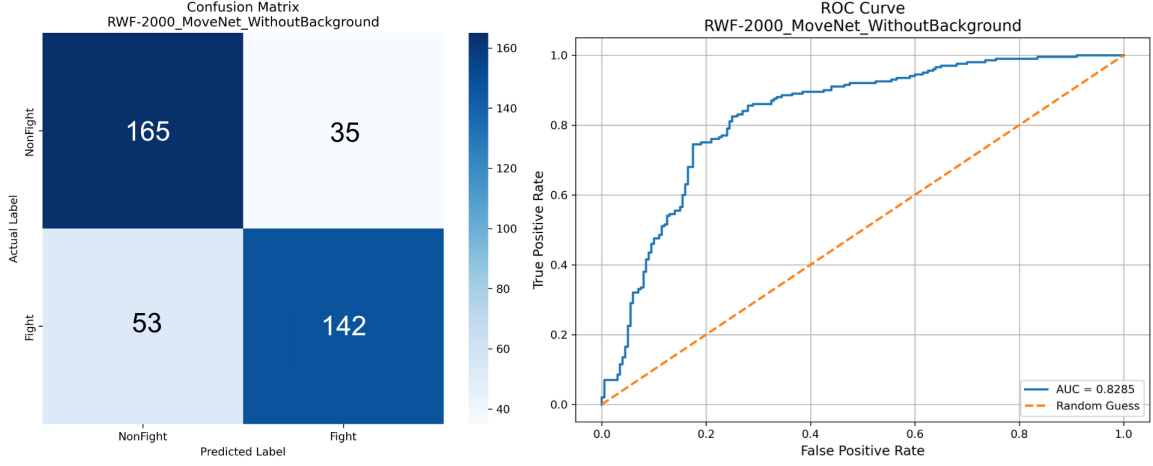


Figure 5.4: Confusion matrix and ROC curve for the MoveNet without Background variant.

its F1-score of 0.7696 also ranked second overall.

5.2.5 MoveNet on Frame Difference

The proposed method, MoveNet on Frame Difference, outperformed all other variants on accuracy, recall, F1-score, and AUC. At 81.50 percent accuracy, it represents a 13-percentage-point improvement over the raw video baseline, a nine-point improvement over the MoveNet variant, a five-point improvement over Frame Difference alone, and a 3.5-point improvement over MoveNet Without Background.

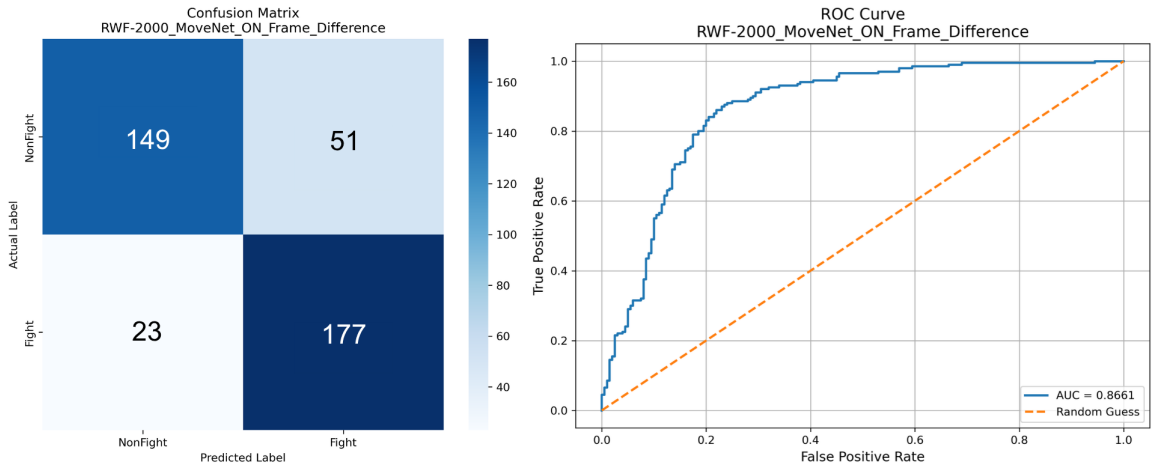


Figure 5.5: Confusion matrix and ROC curve for the MoveNet on Frame Difference variant.

The recall of 88.5 percent is the highest of any variant. Out of 200 genuine Fight clips in the test set, the model correctly identified 177. Only 23 violent clips were missed. This is the most operationally important metric in a surveillance context, where failing to detect a real violent event carries greater consequence than raising a false alarm.

The precision of 77.63 percent, while not the highest of any individual variant (both Frame Difference and MoveNet Without Background achieved approximately 80.8 percent), is notably higher than the raw video baseline and the MoveNet variant.

The F1-score of 0.8271 is the highest of any condition, indicating that the improvement in recall was not achieved by simply shifting the decision boundary to predict Fight more aggressively at the cost of precision. The AUC of 0.8661 is also the highest, meaning the model’s underlying probability scores rank violent clips above non-violent ones more reliably than any other variant across the full range of decision thresholds.

The performance gains of this variant relative to Frame Difference alone suggest that adding skeletal structure on top of the motion-residual images provides genuine discriminating information beyond what motion isolation alone captures. The gains relative to MoveNet Without Background suggest that having frame differencing as the substrate for pose estimation is more informative than having a black background, likely because the motion-residual image actively directs MoveNet’s attention toward the regions of the frame where people are moving.

5.3 Ablation Study: Order of Fusion

The ablation study of this research is the Frame Difference on MoveNet preprocessing variant. Table 5.2 compares the two sequential combination strategies across all five evaluation metrics, including accuracy, precision, recall, F1-Score, and AUC. The first is the proposed method, MoveNet on Frame Difference, and the second is the ablation, Frame Difference on MoveNet.

Table 5.2: Ablation Study: Order of Sequential Fusion

Variant	Accuracy	Precision	Recall	F1-Score	AUC
MoveNet on Frame Difference	81.50%	77.63%	88.5%	0.8271	0.8661
Frame Difference on MoveNet	74%	72.22%	78.00%	0.7500	0.8163

The two variants differ by 7.5 percentage points in accuracy, 5.41 points in precision, 10.5 points in recall, 0.0771 in F1-score, and 0.0498 in AUC. On every metric, MoveNet on Frame Difference outperforms Frame Difference on MoveNet. The differences are large enough across all five dimensions to conclude that the order in which these two operations are applied has a meaningful and consistent effect on classification performance.

The performance gap is largest in recall: MoveNet on Frame Difference correctly identified 88.5 percent of Fight clips, while Frame Difference on MoveNet identified only 78.0 percent. Ten and a half percentage points more violent clips were missed by the reversed-order pipeline. This metric most directly reflects the operational impact of the ordering decision.

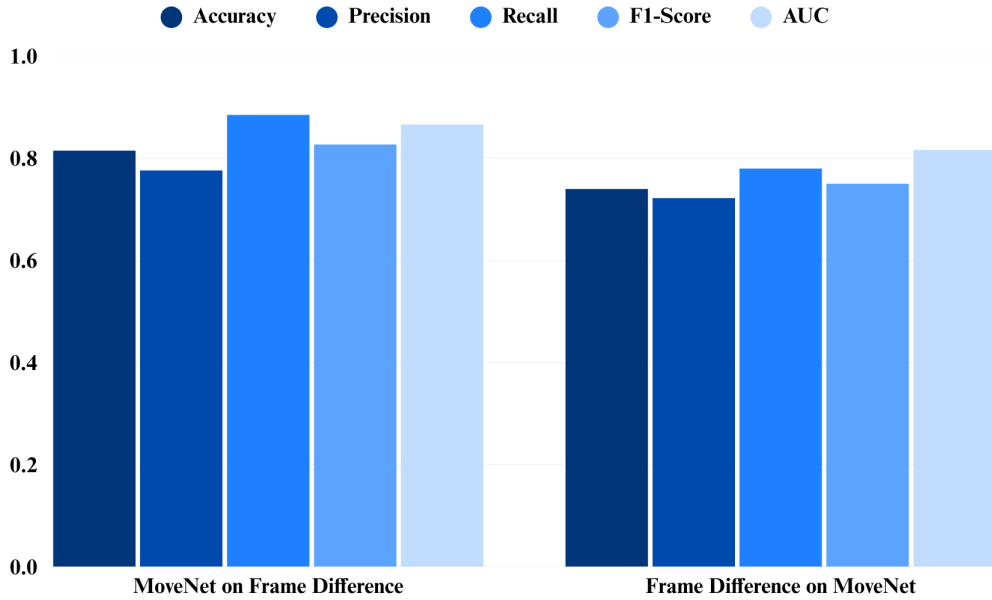


Figure 5.6: Performance comparison of preprocessing order in the ablation study.

The structural explanation for this gap relates to what each operation receives as input. In MoveNet on Frame Difference, MoveNet processes a frame-difference image in which the static background has already been suppressed. The input to pose estimation is a representation where the most visually prominent regions are the moving bodies. In Frame Difference on MoveNet, MoveNet [12] processes the original full-color RGB frame, which contains background complexity that may complicate keypoint detection. The frame differencing then applied to the MoveNet [12] output subtracts consecutive skeleton frames from each other. Because MoveNet’s [12] keypoint predictions are not perfectly stable frame-to-frame, individual joint positions shift slightly even for a stationary person due to inference-level noise; this differencing operation amplifies detection noise rather than capturing genuine body movement. The resulting signal is a mixture of real limb motion and MoveNet’s own estimation jitter, which provides a less reliable basis for classification than the motion-first approach.

The AUC of 0.8163 for Frame Difference on MoveNet, while lower than the proposed method’s 0.8661, is still above 0.8, indicating that the reversed-order pipeline has meaningful discriminative ability. It also outperforms the raw video baseline (68.50% accuracy) and the standalone MoveNet variant (72.50% accuracy) on accuracy, although its recall of 78.0 percent is lower than the standalone MoveNet’s 83.5 percent. The ablation therefore shows that combining the two operations in either order is better than the raw video baseline, but that the specific order is consequential, and applying frame differencing before pose estimation is clearly the more effective choice.

5.4 Summary of Key Findings

The results in all six conditions lead to four conclusions. Preprocessing improves over raw video. Each preprocessing variant tested in this study outperformed the raw video baseline with a precision of 68.50 percent. The gains range from four percentage points (MoveNet)

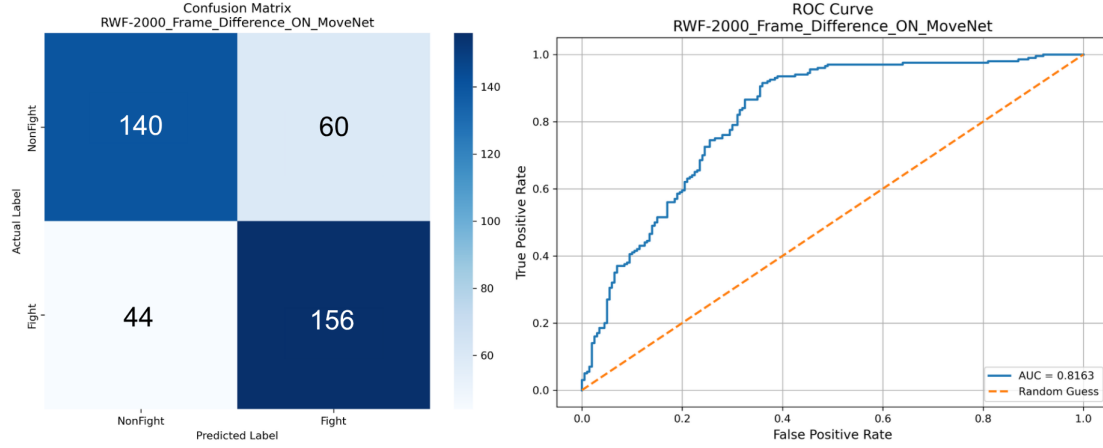


Figure 5.7: Confusion matrix and ROC curve for the Frame Difference on MoveNet variant.

to thirteen percentage points (MoveNet on Frame Difference). This confirms the core premise of the study: transforming the input before classification reduces the learning difficulty for the downstream model. The proposed method, MoveNet on Frame Difference, is the best-performing variant. It achieved the highest score on accuracy (81.50%), recall (88.5%), F1-score (0.8271), and AUC (0.8661). It is the only variant to exceed 80 percent accuracy and the only one to exceed 88 percent recall on the test set. Removing background content from pose-estimated video improves performance. MoveNet Without Background outperformed MoveNet by 5.5 percentage points in accuracy and substantially improved precision. This suggests that background scene content in annotated RGB frames acts as noise for the classifier rather than helpful context. The order of sequential fusion matters. MoveNet on Frame Difference outperformed Frame Difference on MoveNet by 7.5 percentage points in accuracy and 10.5 points in recall. Applying frame differencing before pose estimation produces a more informative and less noise-contaminated input than the reverse order.

Chapter-6: Discussion

6.1 Overview

This study was aimed to achieve three objectives, discussed in Section 1.3: propose a preprocessing method that combines motion and pose information in a specific order, compare it against four alternative preprocessing strategies and a raw-video baseline, and confirm the design choice through an ablation that reverses the order of combination. This chapter discusses the results in that same order. Sections 6.2 and 6.3 address the primary objective directly: why MoveNet on Frame Difference, the proposed method, performed best, and why Frame Difference on MoveNet, its reversed-order ablation, performed worse. Section 6.4 brings the two findings together into a single account of what worked and what did not work in all five variants. Section 6.5 turns to the secondary objective: situating all five preprocessing methods and the raw RWF-2000 [4] baseline reported by Bhuiyan et al. [8], against one another. Section 6.6 compares the best result obtained here with previous work on the same data set, and Section 6.7 considers what the findings mean for a deployed system.

Across the six conditions tested (the raw baseline and five preprocessing variants), accuracy ranged from 68.50 to 81.50 percent, Fight recall from 69.5 to 88.5 percent, and AUC from 0.7846 to 0.8661, with the classifier held identical throughout. Preprocessing alone, in other words, was enough to move test accuracy by thirteen points. The sections that follow explain why.

6.2 Why MoveNet on Frame Difference Performed Well

The proposed method of this study, MoveNet [12] on Frame Difference, was the best performer in four evaluation metrics in. It achieved 81.50 percent accuracy, 88.5 percent recall, an F1-score of 0.8271, and an AUC of 0.8661. It correctly identified 177 of the 200 Fight clips in the test set, missing only 23, the fewest missed violent clips of any variant.

The explanation can be followed backwards starting with the inputs to each stage of the pipeline. First, frame differencing will be activated and any object that didn't move in the clip will be set to near zero; floors, walls, furniture, background signage etc. will all be near zero, and what will be visible is the concentration on areas where a body actually moved. MoveNet [12] subsequently infers pose in this already simplified image, not the RGB frame. The background noise has been reduced in most cases, which means that

MoveNet [12] doesn't have to compete with irrelevant visual information when making its predictions, and the keypoints it detects are drawn to the place where there is actual movement, instead of being distracted by background noise or similar moving objects or people who are standing still.

The result is a representation that carries two aligned signals at once: where in the frame motion is occurring, and what shape the body is making while it moves there. Neither signal alone gets close to this performance. Frame Difference by itself reached 76.50 percent accuracy and 69.5 percent recall. It knows where motion happened but has no information about body structure, so a slow or subtle violent interaction can look like ordinary low-motion footage. MoveNet [12] Without Background reached 78.00 percent accuracy and 73.5 percent recall. It knows what the body looks like but has no explicit motion signal to draw on. Combining the two, in this order, closed both gaps at once: the fusion is not redundant, it supplies information neither input carried on its own.

6.3 Why Frame Difference on MoveNet Performed Poorly

Frame Difference on MoveNet [12], the ablation that reverses the fusion order, reached 74.00 percent accuracy, 7.5 points below the proposed method, and its recall of 78.0 percent left 10.5 more points of violent clips undetected. Its AUC of 0.8163 was likewise the second-lowest of the five preprocessing variants. Since both pipelines use exactly the same two operations, frame differencing and MoveNet [12] pose estimation, the gap is attributable entirely to the order in which they run.

In this ordering, MoveNet[12] runs first, directly on the raw, unprocessed video. Its keypoint predictions are not perfectly stable from one frame to the next: the estimated joint positions drift slightly even when a person is holding still, and MoveNet [12] occasionally misplaces a keypoint altogether. Frame differencing is then applied to the resulting skeleton-overlay frames, subtracting one annotated frame from the next. But a rendered skeleton is a thin arrangement of lines and dots rather than a solid, extended shape like a real body, so two consecutive skeleton frames overlap very little even when the underlying person barely moved. What the differencing step measures, then, is mostly MoveNet's own frame-to-frame jitter, estimation noise, rather than genuine motion in the scene.

This does not mean the ablation failed outright. Frame Difference on MoveNet still beat the raw video baseline by 5.5 points and the standalone MoveNet [12] variant by 1.5 points on accuracy, so combining motion and pose information in this order still carries some benefit over either signal alone. The comparison with the proposed method shows that the benefit is capped by where the noise enters the pipeline: applying frame differencing before pose estimation gives MoveNet [12] a cleaner input to work from, while applying it after pose estimation amplifies MoveNet's own instability into the final signal. The fusion concept itself is sound in both directions; only one direction avoids compounding the noise.

6.4 What Works and What Does not

The five preprocessing variants point to a consistent pattern. Frame Difference works, in part. It isolates motion effectively and produced the joint-highest precision of any variant (80.81%), but it carries no structural information about the body, so it misses subtle or slow violent interactions. Its recall, at 69.5 percent, was the lowest of the five. MoveNet [12], on its own, does not work well. It captures body pose but retains the full original background, and that retained scene content diluted the classifier’s signal rather than adding useful context. It was the weakest of the five variants on accuracy (72.50%), precision (68.44%), and AUC (0.7846). Removing the background works. MoveNet Without Background outperformed the background-retaining version by 5.5 points of accuracy using the identical pose-estimation step, which indicates that scene content was actively hurting classification rather than helping it. MoveNet on Frame Difference works, and works best. Combining motion isolation with pose estimation, in that order, produced the strongest result on every metric reported, accuracy, recall, F1-score, and AUC, by supplying the classifier with both signals at once instead of either alone. Frame Difference on MoveNet does not work as well. Combining the same two operations in the reverse order let MoveNet’s own frame-to-frame jitter get amplified by the differencing step, producing a noisier signal and a real drop in performance, though it still improved on the raw baseline.

The pattern across all five is the same one: get the sequence right, and motion and pose information reinforce each other; get it wrong, and one signal actively degrades the other rather than adding to it.

6.5 Comparing All Five Preprocessing Methods

The secondary objective of this study was to compare all five preprocessing pipelines against one another and against a raw-video baseline, in order to place the proposed method’s performance in context. The raw video baseline, taken from Bhuiyan et al. [8] who used the same dataset [4] with a ResNet50V2 + LSTM classifier on unprocessed RGB footage, reached 68.50 percent accuracy, 65.81 percent precision, and 77.0 percent recall. Every preprocessing variant tested in this study outperformed it on accuracy, and its relatively high recall against low precision suggests a classifier biased toward predicting Fight under uncertainty rather than one that had learned to reliably separate violent from non-violent motion.

The MoveNet variant, which overlays skeletal annotations on the original RGB frame, improved on the raw baseline by only four accuracy points (72.50%) and was the weakest of the five preprocessing conditions. Its recall was the second-highest of the study (83.5%), but its precision was the lowest (68.44%), and it produced the most false positives of any variant: 77 NonFight clips were misclassified as Fight. Crowd scenes, complex environments, and ordinary physical activity such as running or gesturing can produce skeleton configurations that superficially resemble aggressive postures once background context is left in the frame, and that appears to be what happened here.

Stripping that background out, in MoveNet Without Background, corrected much of this.

Accuracy rose to 78.00 percent, false positives fell from 77 to 35, and NonFight recall rose from 61.5 to 82.5 percent. This is consistent with the pattern reported in [8], where the same comparison between RGB input and pose input found pose-based representations outperforming raw RGB, while also noting that inaccurate pose estimation can undercut the classification result. Both how much pose information is used, and how it is presented to the classifier, appear to matter.

Frame Difference alone reached 76.50 percent accuracy with the second-highest AUC of the study (0.8619) and the joint-highest precision (80.81%), but its recall of 69.5 percent was the lowest of the five variants: 61 genuinely violent clips were classified as NonFight. This is a known limitation of motion-based representations generally: they highlight frame-to-frame change effectively, but a violent interaction with limited or gradual movement produces a comparatively weak motion signal, one that a purely motion-based classifier is prone to under-detect.

MoveNet on Frame Difference, discussed in Section 6.2, sits above all four of these at 81.50 percent accuracy, and Frame Difference on MoveNet, discussed in Section 6.3, sits between the single-signal variants and the proposed method at 74.00 percent. Taken as a whole, the six-way comparison shows that no single preprocessing choice is free of trade-offs. Frame Difference favors precision over recall, MoveNet favors recall over precision, and only the correctly-ordered fusion of the two improves on both simultaneously.

6.6 Comparison with Previous Studies

The best result of this study was 81.50% accuracy on RWF-2000. It is higher than the accuracy of the CNN-LSTM approach described in [9], which reached 68.00 percent accuracy in the studies reviewed in Chapter 2. It is also close to the 81.98 percent reported in [20] with a lightweight network, and the 82.00 percent reported in [21] with an Efficient Violence Detection approach.

Meanwhile, some studies obtained better accuracy than the proposed method. For example, a prior work [18] presented a VGG16 + ConvLSTM model, with the frame-difference-based VGG16 achieving 92.40 percent accuracy, and another [25] presented a staged frame-difference approach reaching 88.50 percent accuracy. Pose- and motion-based fusion methods also achieved higher results, including 90.25 percent in [10], 95.05 percent in [11], and 93.28 percent in [29]. The accuracy reported in [23] was 96.21 percent on RWF-2000 [4] using an attention-based method.

This comparison shows that the proposed approach does not offer the best absolute accuracy among the studies reviewed, but the architectures being compared are not equivalent. Many of the higher-scoring studies use two- or multi-branch architectures, where different sources of information are processed through separate networks before being combined. In [11], for instance, pose and motion are processed and moved separately before being merged; in [29], RGB features are fused with pose information through additional weighting mechanisms. The present study, by contrast, was designed to hold the classifier fixed across every condition and isolate what a single sequential preprocessing choice contributes on its own. The comparison here is more useful for understanding how preprocessing

affects the outcome than for ranking overall architectures by peak accuracy.

Against that backdrop, the 68.00 percent lower bound from [9] and the 81.50 percent achieved here (a 13.5-point gap) confirms that this study’s preprocessing choice does substantive work on its own. The remaining gap to the higher-scoring fusion and attention-based architectures suggests that pairing MoveNet on Frame Difference with a more elaborate classification architecture is a reasonable next step, rather than evidence that the preprocessing choice itself was ineffective.

6.7 Practical Significance

These results carry implications for real-world surveillance-based violence detection. The main strength of MoveNet on Frame Difference is how much of the genuinely violent footage it catches: a Fight recall of 88.5 percent means it correctly flagged 177 of the 200 Fight clips in the test set, the highest of any variant tested.

This kind of representation is best suited to a system whose job is to flag a potential incident for a person to review, not to act as a final decision-maker on its own. Its output would be more appropriately used to direct a human reviewer’s attention to the clips most worth a second look, rather than to trigger an automated response.

The broader implication is that preprocessing is a comparatively simple lever for improving an existing system’s input representation without redesigning the classifier itself. A system does not necessarily need a larger or more complex model to improve; it can also gain by checking whether the representation reaching that model is well suited to the task.

Overall, the findings show that preprocessing plays an important role in violence detection performance. The proposed MoveNet on Frame Difference method provided the best balance of motion and pose information, while the ablation confirmed that their order of application affects performance. These results demonstrate that an appropriate input representation can improve detection without changing the underlying classifier, providing a practical direction for developing more effective surveillance-based violence detection systems.

Chapter-7: Conclusion

7.1 Summary

This thesis asked whether the order in which motion and pose information are combined during preprocessing changes how well a fixed classifier can detect violence in real surveillance footage. Three things were done to answer that, matching the objectives set out in Section 1.3: A novel sequential preprocessing pipeline was proposed. MoveNet [12] applied to Frame Difference first isolates motion in a clip through frame differencing, then estimates human pose from that motion-isolated signal, rather than from the original RGB frame.

The proposed method was compared against four alternative pipelines and a raw-video baseline. Frame Difference, MoveNet Skeleton Overlay, MoveNet Without Background, and Frame Difference on MoveNet were each run through the identical EfficientNet-B0 + LSTM [13] [5] classifier, trained under the same hyperparameters, alongside the raw-video figures reported by Bhuiyan et al. [8]. MoveNet on Frame Difference came out ahead on every metric reported: 81.50 percent accuracy, 88.5 percent recall, an F1-score of 0.8271, and an AUC of 0.8661.

An ablation reversed the fusion order. Frame Difference on MoveNet, built from the same two operations in the opposite sequence, reached only 74.00 percent accuracy and 78.0 percent recall, confirming that the order in which motion and pose are combined has a real, measurable effect on detection performance rather than being an arbitrary implementation detail.

Nothing about the classifier changed between any of these five runs: the same architecture, the same hyperparameters, the same training procedure. Test accuracy still moved by thirteen points and AUC by close to a tenth, purely as a function of how the video was represented before the classifier saw it. MoveNet on Frame Difference came out on top overall; the raw MoveNet overlay, which keeps the full original background sitting behind the skeleton, came out at the bottom. Stripping background out helped in both places it was tried, and running motion isolation before pose estimation clearly beat doing it the other way around.

7.2 Research Questions Answered

Three research questions (RQs) were formulated in Chapter 1 to guide the investigation.

RQ1: Does applying frame differencing before MoveNet pose estimation, as a sequential preprocessing pipeline, produce a more effective input representation for violence detection than either technique applied independently? The proposed MoveNet on Frame Difference pipeline produced the strongest overall performance among the preprocessing representations evaluated in this study. It outperformed the individual preprocessing approaches in terms of accuracy, recall, F1-score, and AUC. The results indicate that applying frame differencing first can reduce irrelevant static background content while retaining regions of motion, allowing MoveNet multipose estimation to extract human skeletal information from a more focused representation. The proposed pipeline is also implementable using publicly available tools, including OpenCV for frame differencing and the TensorFlow Hub MoveNet Multipose Lightning model.

RQ2: Across five distinct preprocessing strategies evaluated under a fixed classifier and identical training conditions on the RWF-2000 dataset, which input representation produces the strongest violence detection performance, and what does the ranking reveal about what information matters most for this task? MoveNet on Frame Difference achieved the strongest overall performance among the five preprocessing strategies. All five variants also outperformed the external raw video baseline [?]. The results show that background removal can improve pose-based representations, while frame differencing alone provides strong precision but comparatively lower recall. Combining motion information with human pose structure produced better overall performance, particularly when the two operations were applied in the proposed order. This suggests that both motion and body-structure information contribute to effective violence detection, while reducing irrelevant background content can further improve the usefulness of the representation.

RQ3: Does the order in which frame differencing and MoveNet pose estimation are applied have a measurable and consistent effect on violence detection performance, and if so, why? The ablation study demonstrated a clear and consistent effect of preprocessing order. MoveNet on Frame Difference outperformed Frame Difference on MoveNet across all evaluated metrics. This indicates that the performance gain is not simply due to combining the two operations, but is also influenced by the order in which they are applied. Applying frame differencing first provides MoveNet with a more focused representation of regions containing motion, whereas applying frame differencing after pose estimation can introduce or retain estimation-related noise from the skeleton overlays. Therefore, the proposed ordering is more effective for the violence detection task evaluated in this study.

Overall, the three research questions are answered by the experimental evidence: the proposed preprocessing pipeline provides the strongest representation, the comparative evaluation demonstrates the value of combining motion and pose information, and the ablation study confirms that the order of these operations has a meaningful effect on performance.

7.3 Analysis of the Social Effects

A working version of the system studied here would be aimed at cutting the time between a violent incident happening and someone actually responding to it, and at taking some of the load off human operators watching camera feeds all day. In places like schools, hospitals, transit stations, and crowded public venues, that faster response time is a real benefit: something that might otherwise go unnoticed until it's already escalated gets flagged while there's still a chance to step in. Chapter 1 already noted that watching hours of footage manually is exhausting work, and people miss things when they're tired. A system that narrows the footage down to the clips most worth a second look can make that job more sustainable rather than eliminating the person doing it.

That same capability has a darker side worth naming rather than glossing over. Rolling out more automated monitoring, even for a genuinely protective purpose, extends how far surveillance infrastructure reaches, and it can quietly shift what people expect in public and semi-public spaces, an effect on ordinary behavior that shows up whether or not anyone's actually doing anything wrong. A system that still gets it wrong sometimes, and the best variant in this study produced 51 false positives out of 200 non-violent test clips, will flag people incorrectly in the real world. Given that, a system like this belongs in the hands of a human decision-maker as a tool that directs attention, not as something left to act on its own.

7.4 Environmental Impact Analysis and Sustainability

Training five models, each run for 100 epochs across 1,280 clips passed frame by frame through a convolutional network, isn't free in terms of electricity, and that electricity comes with a carbon cost that depends entirely on how it's generated. The choices made in this study push in the direction of lowering that cost compared with the alternatives. EfficientNet-B0 was chosen specifically because it delivers solid spatial feature extraction without the parameter count of something like ResNet-50 or VGG-16, and a number of the higher-scoring architectures reviewed in Chapter 2 lean on much heavier machinery, three-dimensional convolutions, transformer-based pose models, multi-head attention, any of which would have driven up both training time and energy use had this study used them instead.

The preprocessing side is lighter still. Frame differencing barely registers as a computational cost, and MoveNet Lightning was built specifically for fast, near-real-time inference rather than heavy batch processing. Because both the preprocessing and the classifier here are relatively cheap to run, this approach is better suited to sustained use across many camera feeds than architectures built around large 3D-convolutional or transformer backbones, which would draw considerably more power if run continuously at that scale. None of this replaces an actual energy audit, though; a real deployment across a city or a large campus would need one before anyone could state its footprint with any confidence. What this study can say is that the choices made here cut that footprint relative to heavier options, not that they remove it.

7.5 Ethical Issues

RWF-2000 [4] was put together by its original authors from publicly posted YouTube footage, not gathered specifically for this thesis, and this study worked entirely from that existing benchmark rather than collecting or publishing any new footage of identifiable people. That’s a meaningful distinction, but it doesn’t make the underlying task ethically neutral. Any system trained this way and later pointed at live cameras would be making automated calls about real people’s behavior, and getting one of those calls wrong isn’t a minor error. Being incorrectly flagged as violent carries a cost for the person involved even if a human reviewer clears the clip minutes later, a cost that a precision or recall figure doesn’t capture.

Fairness is a second issue this study can point to without being able to settle it. RWF-2000’s [4] authors pulled footage using search terms across several languages specifically to broaden geographic and cultural coverage, which is a sensible step against regional bias, but this study never checked whether the trained models perform evenly across different lighting, clothing, skin tones, or camera quality, and an uneven error rate across any of those is entirely plausible without ever having been measured here. There’s also the question of dual use: a model trained to spot violence is, underneath, a general tool for reading human activity from video, and nothing about the model itself stops it from being repurposed for wider behavioral monitoring beyond what it was built for. Preventing that kind of scope creep is more a governance problem than a technical one, and it’s not something this thesis can resolve, but it’s worth stating outright rather than leaving unsaid: a system like this needs clear rules about what it monitors and what it doesn’t, with a person accountable for whatever decisions it feeds into.

7.6 Limitations and Future Work

This study has several limitations related to its research scope. First, evaluation was limited to the RWF-2000 dataset [4]. The findings therefore show how the proposed preprocessing approaches perform on this particular benchmark. Their performance on other surveillance datasets has not been established. Second, only the EfficientNet-B0 + LSTM [13] [5] architecture was evaluated. The study was intentionally designed around one fixed classifier so that the effect of preprocessing could be compared fairly. As a result, it is not known whether the same preprocessing ranking would remain unchanged when other architectures are used. Third, the system performs binary Fight and NonFight classification only. This means that it identifies whether a video belongs to the Fight or NonFight category but does not distinguish between different forms of violence or other types of abnormal activity.

These limitations lead to several possible directions for future research. Evaluation on larger and more diverse violence-detection datasets. The five preprocessing pipelines can be tested on additional datasets to determine whether the findings observed on RWF-2000 [4] also apply to other surveillance environments. Evaluation with different classification architectures: Future studies can investigate CNN-LSTM, ConvLSTM, 3D CNN, transformer-based, or other lightweight architectures to determine whether the

benefit of MoveNet on Frame Difference is consistent across different models. Multi-class violence detection: Instead of only distinguishing Fight and NonFight, the framework could be extended to identify different types of violent activities, such as physical assault, group fighting, or other categories depending on the dataset. Anomaly detection: The framework could also be extended beyond predefined violence classes to identify unusual or abnormal events in surveillance videos.

7.7 Closing Thoughts

The findings of this thesis demonstrate that preprocessing design plays an important role in real-world violence detection. The main contribution of this work is the proposal of **MoveNet on Frame Difference**, which combines motion and pose information by applying frame differencing before pose estimation. This sequence was designed to emphasize motion-related regions before extracting human body structure, reducing the influence of static background information while retaining meaningful pose information. The experimental results support the effectiveness of this approach, as it achieved the highest performance among the evaluated preprocessing strategies, with 81.50% accuracy, 88.50% recall, 0.8271 F1-score, and 0.8661 AUC.

The ablation results further show that the order of preprocessing operations is an important part of the proposed design. Applying frame differencing after pose estimation resulted in lower performance across the evaluation metrics, indicating that simply combining the two techniques does not provide the same benefit when their order is reversed. Since all preprocessing variants were evaluated using the same EfficientNet-B0 and LSTM classifier and experimental settings, the observed differences highlight the contribution of the proposed input representation rather than changes to the classification architecture. These findings suggest that carefully designed preprocessing can provide a practical way to improve violence detection performance without requiring a more complex model. More broadly, the proposed approach demonstrates the potential of combining motion and pose information in a deliberate sequence for developing more effective surveillance-based violence detection systems.

Bibliography

- [1] W. H. Organization *et al.*, *Injuries and violence: the facts 2014*. World Health Organization, 2014.
- [2] Institute for Economics & Peace, “Global Peace Index 2026: Measuring Peace in a Complex World,” Institute for Economics & Peace, Sydney, Tech. Rep., Jun. 2026. [Online]. Available: <https://gpi.economicsandpeace.org/>
- [3] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, “A review on state-of-the-art violence detection techniques,” *IEEE Access*, vol. 7, pp. 107 560–107 575, 2019.
- [4] M. Cheng, K. Cai, and M. Li, “Rwf-2000: An open large scale video database for violence detection,” in *2020 25th International conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 4183–4190.
- [5] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [6] S. I. Sani, R. Rahman, Z. Islam, S. T. Bhuiyan, M. Haque, S. Wasi, S. I. Haque, and S. B. Alam, “Accelerating violence detection: A comparative study of frame difference and traditional methods on ucf crime dataset,” in *2025 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*. IEEE, 2025, pp. 78–83.
- [7] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 6479–6488.
- [8] S. T. Bhuiyan, M. Haque, H. Khatun, R. Rahman, S. I. Haque, A. Islam, and S. B. Alam, “Evaluating pose estimation’s contribution to violence detection using deep learning,” in *2025 International Conference on Machine Learning and Cybernetics (ICMLC)*. IEEE, 2025, pp. 56–61.
- [9] M. Barapatre, A. Padia, M. Mokhere, V. Jaiswal, and P. Sonsare, “Eye on violence: Real-time violence detection in surveillance video using a cnn-lstm spatiotemporal framework,” in *2026 IEEE International Conference for Convergence in Computing Technology (I3CTCON)*. IEEE, 2026, pp. 1–5.
- [10] G. Garcia-Cobo and J. C. SanMiguel, “Human skeletons and change detection for efficient violence detection in surveillance videos,” *Computer Vision and Image Understanding*, vol. 233, p. 103739, 2023.

- [11] S. T. Y. Aung, W. Kusakunniran, and Y. K. Hooi, "Leveraging fusion methods of human pose and motion dynamics for accurate violence detection in video surveillance," *IEEE Access*, 2025.
- [12] TensorFlow, "MoveNet: Ultra Fast and Accurate Pose Detection Model," <https://www.tensorflow.org/hub/tutorials/movenet>, 2024, accessed: Aug. 20, 2026.
- [13] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PmLR, 2019, pp. 6105–6114.
- [14] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *International conference on Computer analysis of images and patterns*. Springer, 2011, pp. 332–339.
- [15] S. Mohammadi, H. Kiani, A. Perina, and V. Murino, "Violence detection in crowded scenes using substantial derivative," in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2015, pp. 1–6.
- [16] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violence detection in surveillance video using low-level features," *PLoS one*, vol. 13, no. 10, p. e0203668, 2018.
- [17] T. Z. Ehsan and S. M. Mohtavipour, "Vi-net: A deep violent flow network for violence detection in video sequences," in *2020 11th International Conference on Information and Knowledge Technology (IKT)*. IEEE, 2020, pp. 88–92.
- [18] I. Mugunga, J. Dong, E. Rigall, S. Guo, A. H. Madessa, and H. S. Nawaz, "A frame-based feature model for violence detection from surveillance cameras using convlstm network," in *2021 6th international conference on image, vision and computing (ICIVC)*. IEEE, 2021, pp. 55–60.
- [19] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir, and M. Farazi, "Efficient two-stream network for violence detection using separable convolutional lstm," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.
- [20] W. Wang, S. Dong, K. Zou, and W. Li, "A lightweight network for violence detection," in *Proceedings of the 2022 5th International Conference on Image and Graphics Processing*, 2022, pp. 15–21.
- [21] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient violence detection in surveillance," *Sensors*, vol. 22, no. 6, p. 2216, 2022.
- [22] D. Choqueluque-Roman and G. Camara-Chavez, "Weakly supervised violence detection in surveillance video," *Sensors*, vol. 22, no. 12, p. 4502, 2022.
- [23] S. Vosta and K.-C. Yow, "Kiannet: A violence detection model using an attention-based cnn-lstm structure," *IEEE Access*, vol. 12, pp. 2198–2209, 2023.
- [24] V. D. Huszar, V. K. Adhikarla, I. Negyesi, and C. Krasznay, "Toward fast and accurate violence detection for automated video surveillance applications," *IEEE Access*, vol. 11, pp. 18 772–18 793, 2023.
- [25] H. A. Huillcen Baca, F. d. L. Palomino Valdivia, and J. C. Gutierrez Caceres, "Efficient human violence recognition for surveillance in real time," *Sensors*, vol. 24, no. 2, p. 668, 2024.

- [26] S. Santhosh, A. Shenoy, K. Prabhu, K. Patgar, S. A. Ahmed, and S. Shravan, "Real-time violence detection using lightweight deep learning," in *2026 International Conference on Artificial Intelligence and Data Engineering (AIDE)*. IEEE, 2026, pp. 134–139.
- [27] H. Khan, X. Yuan, L. Qingge, and K. Roy, "Violence detection from industrial surveillance videos using deep learning," *IEEE Access*, vol. 13, pp. 15 363–15 375, 2025.
- [28] L. Zhou, "End-to-end video violence detection with transformer," in *2022 5th International conference on pattern recognition and artificial intelligence (PRAI)*. IEEE, 2022, pp. 880–884.
- [29] P. Zhang, W. Lei, X. Zhao, L. Dong, and Z. Lin, "Rtvd-net: An real-time violence detection method based on pre-training of human skeleton images," in *Proceedings of the 2023 12th International Conference on Networks, Communication and Computing*, 2023, pp. 260–266.
- [30] S. Blunsden and B. Fisher, "The behave video dataset: ground truthed video for multi-person behavior classification," *Annals of the BMVA*, vol. 2010, no. 4, pp. 1–11, 2010.
- [31] P. Rota, N. Conci, N. Sebe, and J. M. Rehg, "Real-life violent social interaction detection," in *2015 IEEE international conference on image processing (ICIP)*. IEEE, 2015, pp. 3456–3460.
- [32] C.-H. Demarty, C. Penet, M. Soleymani, and G. Gravier, "Vsd, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation," *Multimedia Tools and Applications*, vol. 74, no. 17, pp. 7379–7404, 2015.
- [33] M. Perez, A. C. Kot, and A. Rocha, "Detection of real-world fights in surveillance videos," in *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019, pp. 2662–2666.
- [34] E. B. Nievas, O. D. Suarez, G. B. Garcia, and R. Sukthankar, "Hockey fight detection dataset," in *Computer Analysis of Images and Patterns*. Springer Heidelberg, 2011, pp. 332–339.
- [35] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 2012, pp. 1–6.
- [36] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 2012, pp. 28–35.
- [37] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [38] G. Nair, H.-Y. Li, J. Schwenk, K. Martinez, C. Manore, and C. Xu, "Increasing mosquito abundance under global warming," *Earth's Future*, vol. 13, no. 6, p. e2024EF005629, 2025.

- [39] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.