**Independent University** **Bangladesh (IUB)**

**IUB Academic Repository**

Computer Science and Engineering | Undergraduate Thesis

2025-08

# Enhancing Violence Detection: Comparative Analysis of Frame Difference and Conventional Methods Using the UCF Crime Dataset

Islam, Md. Zahidul

IUB

https://ar.iub.edu.bd/handle/11348/1075

*Downloaded from IUB Academic Repository*


# Enhancing Violence Detection: Comparative Analysis of Frame Difference and Conventional Methods Using the UCF Crime Dataset

**Md. Zahidul Islam (2022504)**

**Saniul Islam Sani (2022605)**

Dissertation submitted in partial fulfillment for the Degree of Bachelor of Science in Computer Science and Engineering
Department of Computer Science & Engineering
Independent University, Bangladesh

**August 2025**

# Attestation

I understand the nature of plagiarism, and I am aware of the university's strict policy on the matter. I certify that this is an original work by me. However, others' written work or software used in any part of this project is properly cited following internationally accepted academic guidelines.

**Signature:**

I. Md. Zahidul Islam: _____

II. Saniul Islam Sani: _____

# Evaluation Committee

Supervision Panel

…… ……

Supervisor Co-Supervisor

External Examiners

……

External Examiner 1

Internal Examiners

……..

Internal Examiner 1 Internal Examiner 2

Office Use

……

Program Coordinator Head of the Department

……..……..

Director Graduate Studies, Research and Industrial Relations

..... Library

# Acknowledgment

I would like to express my deepest gratitude to my supervisor, Saadia Binte Alam, PhD, for her expert guidance, insightful feedback, and unwavering support. Without her direction, we would not have been able to complete this thesis.

Additionally, I would like to extend my sincere thanks to MD. Rashedur Rahman, a doctoral Fellow at the University of Hyogo, Japan, and an Associate Professor at Independent University, Bangladesh, for his invaluable contributions and guidance as my co-supervisor.

Lastly, my heartfelt thanks to Mahmudul Haque, former Adjunct Faculty at Independent University, Bangladesh, for his valuable suggestions and contributions during the experiments. His assistance and encouragement were instrumental throughout this research.

I would also like to acknowledge the use of AI-assisted tools for language refinement and grammatical improvement only. All research ideas, analyses, and conclusions presented in this thesis were developed and carried out by the authors.

# Abstract

Violence is one of the biggest concerns in the world. It leads to big problems in both society and the economy. This is an important issue, according to both the World Health Organization and the Institute for Economics and Peace. Video surveillance has become a big part of keeping people safe. However, someone has to view the video in traditional systems for them to be effective. Viewers are limited and tend to miss violent events in time. It becomes increasingly difficult to monitor violent events when there are too many cameras. This proves the importance of automated real-time detection for violent events. The aim of this thesis is to create a deep learning method that makes detecting these events more accurate, fast, and efficient. It suggests a hybrid approach that combines two powerful tools: Convolutional Neural Network (CNNs) and Recurrent Neural Network (RNNs). It uses a pre-trained ResNet50 model for extracting spatial features. It then follows up with a Long Short-Term Memory (LSTM) network for extracting temporal features. This approach works very well for both space and time features. It doesn't need as much processing power as more complicated models like 3D CNNs. It explores two ways in which the video data is preprocessed for the model. The first is through evenly spaced RGB frames. The other is where it explores the differences between frames. This involves calculating the absolute difference between frames. The disparities in frames highlight motion. This is important for the identification of violence. They also remove noise in stationary backgrounds. This makes it simpler to locate things more surely and using less computer power. This makes it practical for real-time applications. The UCF Crime dataset is used for testing. It contains more than 1,900 real-life videos. These videos contain issues such as light intensity levels, diverse angles, and an imbalance between violent and non-violent videos. Both methods were tested under equal conditions. Accuracy and the Area Under the Receiver Operating Characteristic Curve (AUROC) were used for testing the models. Both the frame difference approach and the frame difference method were correct 85% of the time. The frame difference approach also outperformed the other by having a higher AUROC. This is 0.9091 compared to 0.8892 by the other. This is more effective at separating the two. It also improved training time by 2.9% and inference time by 4.2%. This indicates it is more accurate and efficient for CNN-LSTM models in detecting violence. The model is very effective in practical scenarios. It can apply in a range of scenarios concerning surveillance. Through the focus on motion, it is able to separate the violent scenes from the non-violent scenes even in crowded settings. Due to its effectiveness, it is also appropriate for resource-poor devices. The model is able to apply in the current systems used for surveillance. This is even in public centers or public transport. This development is an improvement in the systems for intelligent surveillance. This is a practical approach towards public safety in a complex city setting. This is a practical approach towards public safety. This is a practical approach towards public safety. This is a practical approach towards public safety. This can be implemented in schools, airports, and other busy places where it is necessary to detect violence immediately. This method ensures that people are protected because it solves present problems and allows people to develop.

# List Of Publication

S. Islam Sani, S. B. Alam, Z. Islam, S. T. Bhuiyan, S. Wasi, M. Haque, and S. I. Haque, "Assessment of Accelerating Violence Detection: A Comparative Study of Frame Difference and Traditional Methods on UCF Crime Dataset," in Proceedings of the *International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, Bali, Indonesia, July 12–15, 2025. (Accepted)

# Table of Contents

# List of Figures



# List of Table

# Chapter 1: Introduction

Violence is a worldwide problem. The social and financial costs are high. The World Health Organization and the Institute for Economics Beyond Peace both highlight the scope of the problem. The video surveillance of public space provides better protection, yet traditional systems rely on employee screen watching. Human eyes tire, miss cues, and seldom find violence in a timely manner. These chances are further diminished when a city is covered by thousands of cameras. Such limits illustrate the need for an automatic real-time violence detector.

## 1.1 Motivation

Violence nowadays is one of the big concerns in the world and, annually, it continues affecting millions. Its aftermath is not confined to a few visible injuries or emotional shock at the time of an incident. In many cases, the damage lasts much longer, influencing families, communities, and social stability. Healthcare services are often overwhelmed, and governments are forced to spend large amounts of money addressing the consequences of violent acts. According to the World Health Organization (WHO), more than five million people die each year as a direct result of violence, which shows how serious this issue is from a public health perspective. Apart from the human cost, violence also creates a significant economic burden. The Institute for Economics & Peace reported that in 2019, violence cost the global economy over $14 trillion, reducing funds that could otherwise be used for education, healthcare, and development initiatives [1]. The combined effects of human suffering and economic loss clearly show that better and more effective measures are needed to prevent violence and reduce its impact [1]. As safety has become a growing concern, people around the world are increasingly depending on technology for support. Video surveillance systems are now widely used and have become a common part of urban environments. They are an important element of smart city initiatives, where technology is applied to improve public safety, security, and overall quality of life. This also includes a network of street surveillance cameras, cameras at public areas, and business locations that deter criminal activities, aid law enforcement agencies in investigations, and respond with much speed in areas of emergency. However, these systems generate an extremely large amount of video data, which is difficult to analyze by human operators. This research focuses on combining public safety needs with recent technological developments and proposes intelligent systems that are capable of automatically detecting violent incidents and turning traditional surveillance cameras into more effective public safety tools [2].

Traditional video surveillance systems depend entirely on human monitoring for viewing modified images shot by a number of cameras. The truth is that doing so manually is a relatively difficult task. In fact, it is most likely that future video surveillance systems will consist of HCs literally in the thousands that will be running 24/7. Not to mention the fact that it will generate an unprecedented volume of images that are not really amenable to human monitoring for certain significant events to occur. After prolonged observation, it is likely that observers will tire,

allowing them to get distracted or lose focus on what they are doing. For this reason, manual observation of surveillance systems is very demanding and unreliable. There is also limited usage of this type of surveillance implemented for purposes of surveillance after an incident has occurred, such as crimes or accidents [3]. They are generally not effective at identifying threats in real time. These limitations have led to a significant shift in surveillance research, moving away from systems that rely only on human supervision toward automated and intelligent approaches. The main objective of this transition is to assist or replace human operators with algorithms that are able to analyze video streams continuously and at a much larger scale than any individual could manage. Unlike manual monitoring, automated analysis enables the surveillance system to react proactively, as opposed to acting upon the occurrence of the event. Through the use of artificial intelligence and machine learning algorithms, intelligent surveillance systems can monitor live video streams round the clock and alert the relevant authorities upon the occurrence of suspicious or violence acts [4]. Automation of the analysis of the data acquired by the surveillance cameras is viewed as going beyond mere efficiency; it is necessary if the full potential of video surveillance is to be actualized in early intervention and prevention of crimes [5].

Deep learning has contributed greatly in this paradigm shift in order to achieve intelligent surveillance. This subfield of machine learning has brought a paradigm shift in the way computers analyze pictures and recognize patterns. State-of-the-art automatic violent content detection systems are typically based on deep neural networks capable of detecting violent behaviors directly in video streams. Convolutional Neural Networks (CNNs) are the most important part of these systems. They are made to work with grid-topology input like photographs [6]. There are layers of filters in CNNs that learn how to use these spatial information. The earlier layers learn how to use simple features like edges and textures, and the later layers learn how to use more complicated features like objects and layouts of environments [7]. Network models such as VGG, which involves stacking small filters and Residual Networks (ResNet) that utilizes shortcut connections to train very deep models have become ubiquitous tools for efficient image processing [6]. However, for video processing, there comes into play yet another critical dimension - time. Unlike images, which are spatial entities, a video is an ordered sequence of images or frames that possess temporal relationships, which are critical for their interpretation. To tap this temporal information, researchers have employed models such as RNNs and LSTM [7]. These models are designed to process sequences by maintaining an internal memory of the frames while processing any new frame in a sequence. Thus, by joining CNNs with LSTNs, researchers have discovered novel models that tap not only spatial but also temporal knowledge [8]. In this fairly successful paradigm, CNNs act as perceptual front-ends to extract high-level abstracts from images, while LSTNs interpret the temporal evolution of such spatial image representations to understand actions/events. The CNN-LSTM paradigm, on which this thesis will explore specific models, presents an efficient means to interpret spatiotemporal data such as videos.

## 1.2 Problem Statement and Research Gap

Detecting violence in video in real time is both challenging and essential. One of the main difficulties is balancing high accuracy with low computational cost. Among the most effective methods, models such as 3D Convolutional Neural Networks (3D CNNs) treat a video as a spatiotemporal volume, using 3D filters to capture motion and appearance simultaneously. These models perform very well on many action recognition benchmarks, but at a high cost. They involve a large number of parameters which require substantial computational power and memory. They are therefore difficult to train and are also too slow for real-time detection [9] [10].

This tradeoff between accuracy and efficiency is what underlines the main research problem in violence detection. There is a great need for techniques which can ensure strong accuracy while the needed computer power is low enough to enable real-time processing. True progress is made by techniques which can ensure strong accuracy while the needed computer power is reduced. This is preferable to techniques which ensure slight improvements in accuracy but need large amounts of computer power. This research is founded on the assumption that true progress is made by improving the tradeoff between accuracy and efficiency. To move violence detection systems from the lab to the real world, we need to solve this problem.

## 1.3 Proposed Approach and Contributions

This thesis fills the research gap by investigating a method to achieve a good tradeoff between accuracy and computational complexity for video-based violence detection. The key point of this research is to explore a method where frame differences can serve as inputs to a ResNet-LSTM network to potentially achieve a more accurate and more computationally efficient real-time violence detection system than traditional RGB frame inputs. The proposed method aims to highlight more motion-related information through pre-processing to allow the network to pay more attention to visual cues of violent behaviors. This method not only enables the network to achieve a better distinction of violent occurrences, measured by AUROC, but also avoids redundant computations of unnecessary background information included in conventional RGB frames. The contributions of this thesis can be summarized as follows: This thesis conducts a comprehensive comparison study of two different inputs for violence detection: conventional RGB frames and frame differences. This comparison study has been carried out by adopting a widely used hybrid model of ResNet and LSTM, ensuring a fair comparison of their performance. The experimental results on the UCF Crime dataset demonstrate that more attention to motion-related information by processing frame differences can achieve more stable AUROC performance for real-world surveillance videos.

Secondly, in this thesis, there is also an analysis of computational complexity regarding training and detection time in the proposed solution. The analysis proves not only does it improve detection accuracy but it is more efficient in terms of computation, which is an important factor in this area as it needs fast detection in real-time systems. Lastly, this research work contributes to the general knowledge of how an efficient and effective video analysis system can be achieved. This work

provides a feasible approach to improve the detection accuracy without requiring more complex.

## 1.4 Thesis Outline

**Chapter 1:** Presents the societal context, the technological background, and the research problem that will lead towards a statement on the thesis' contributions.

**Chapter 2:** This chapter is dedicated to the review of the existing literature in the domain of action recognition. This includes basic concepts of action recognition, architectures of spatio-temporal learning, representation of motion, and evaluation techniques.

**Chapter 3:** Deals with deep learning methods used in violence detection, including key models, methods, and milestones that have made it possible to examine violent behavior in videos.

**Chapter 4:** This chapter introduces the UCF Crime dataset, which would be used in the study. This dataset contains more than 1,900 surveillance videos, which provides diversity in the dataset. This dataset is imbalanced.

**Chapter 5:** In this chapter, the design of the proposed solution/model in terms of architecture will be presented. It is at this point that the reasoning behind the design of the solution and architecture and their integration in order to fulfill the research objectives will be presented.

**Chapter 6:** This chapter explains the training and testing process with emphasis on the preprocessing, hyperparameters, and testing criteria used in determining the effectiveness of the model.

**Chapter 7:** In this chapter, the results that were achieved from the experimental work conducted in the thesis will be shown. The performance of the model in terms of accuracy, robustness,

**Chapter 8:** The findings are placed in the context of previous research and their relevance highlighted before there is a brief discussion about the contribution made to research through this thesis. The limitations of this research are also highlighted and areas where improvements are possible are mentioned.

# Chapter 2: Literature Review

The aim of this chapter is to introduce a summary of existing work which serves as a basis of this thesis work. The analysis of existing work contributes to a definition of challenges and achievements in video-based violence detection as a field of study related to action recognition in videos. Being a field of study related to violence detection, the chapter first introduces a summary of work related to action recognition in videos, including traditional handcrafted solutions and deep learning solutions. Finally, it introduces a summary of work related to techniques of motion representation which is a key point in differentiation of violent actions from non-violent ones.

## 2.1 Foundations of Action Recognition in Video

The identification of human actions in videos is a long-standing problem in the computer vision community. The developments made in the identification of human actions in videos have contributed to the current technology used in the identification of violence [11]. Traditionally, the identification of human actions in videos would involve the design of features by human designers [12]. Currently, the technology has been replaced by deep learning, which uses data to develop features.

### 2.1.1 Early Approaches and Handcrafted Features

The early solutions to the problem of human action recognition were largely reliant on hand-designed features. In this method, it was necessary to develop algorithms to encode the essential information about the video. An interesting line of research was the usage of local spatio-temporal patterns [13]. Among these, dense motion paths were shown to work quite well. This technique involves the usage of optical flow to track points in the video and then constructing features along these trajectories to encode the performed action [14] [15].

### 2.1.2 The Deep Learning Revolution in Vision

The emergence of deep learning techniques and more specifically the success of Convolutional Neural Networks (CNNs) in large-scale image classification tasks introduced a dramatic shift in computer vision and subsequently in video analysis as well. Unlike traditional hand-crafted techniques, deep learning techniques can potentially handle feature extraction by automatically learning task-specific features directly from the training samples. The capability of automatically learning task-specific features has been a key factor in improving performance in a wide variety of vision tasks [16]. Various architectural techniques also played a crucial role in achieving these improvements. The initial deep networks demonstrated that depth could be a key factor in improving performance, but researchers quickly encountered a “vanishing gradient” issue in which the gradients become too small to train the network effectively [17]. The VGGNet network tackled this issue by exploiting a very deep and regular network architecture composed of small 3×3 filters in convolutional layers and demonstrated that such networks could be effectively trained [6]. The biggest milestone was achieved by Residual Networks (ResNet) by exploiting skip connections to facilitate gradient flow through multiple layers of a network and hence improving training stability.

This solved the vanishing gradient issue and made it possible to train extremely deep networks, in some cases with more than a thousand layers [6]. ResNet's ability to learn highly robust and expressive visual features has made it a standard choice in computer vision and a core component in many systems, including the spatial feature extraction part of the video analysis model used in this thesis [7] [18].

## 2.2 Architectures for Spatio-Temporal Feature Learning

Video analysis needs models that can handle both the visual content in each frame and the changes that happen across frames over time. Researchers have tried different architectural strategies to achieve this, each with its own advantages and limitations. Some approaches separate spatial and temporal processing, while others attempt to learn both together in a unified model.

### 2.2.1 Spatial Feature Extraction with 2D CNNs

Pre-trained 2D CNNs are often used to get features from each frame of a video. This strategy takes advantage of transfer learning. Before being utilized for something else, a deep network like ResNet-50 is trained on a huge set of photographs that have been sorted and categorized into thousands of groups, like ImageNet. During this training, the model picks up a lot of broad visual cues that can be exploited in many different ways [6] [9].

The pre-trained network is used again for video tasks like finding violence. Most of the time, the last layer of categorization is removed, and the network processes each frame to produce a high-level feature vector. This vector tells you what is on the frame in a few words. Using a pre-trained network is a smart choice since it saves you from having to build a deep model from the ground up, which would require a lot of labeled video data and a lot of processing power. Researchers can utilize a pre-trained model to build on what the network has already learned and spend more time figuring out the video's time-related portions [15].

### 2.2.2 Modeling Temporal Dynamics with Recurrent Networks

After getting the spatial data from each frame, the next critical step is to figure out how these features change over time. This is where RNNs, or Recurrent Neural Networks, come in. An RNN takes the series of feature vectors that the CNN makes and processes them one at a time to learn how the video changes over time [8]. Simple RNNs can find short-term patterns, but they have trouble with long-term dependencies because of gradients that disappear or explode. This is the same difficulty that very deep feedforward networks have [19]. LSTMs are a kind of RNN that has a more complicated internal structure. They have a memory cell, and three gates called input gate, forget gate, and output gate. These gates control how information is read, stored, or erased in the memory cell [20]. This gating mechanism allows LSTMs to remember important information for long periods. This makes them very good at capturing the long-range patterns often seen in human actions in videos. In practice, the LSTM takes the features from individual frames and transforms them into an understanding of the full, dynamic event, which helps the system tell the difference between a normal interaction and a violent one.

### 2.2.3 End-to-End Spatio-Temporal Learning with 3D CNNs

3D CNNs are designed to learn both spatial and temporal features at the same time. They use 3D filters that move across video frames, which allow them to directly capture motion and dynamic patterns. Models such as the Inflated 3D ConvNet (I3D) build on conventional 2D CNN architectures by inflating their filters into the temporal dimension, allowing the network to jointly model spatial appearance and temporal motion, and they have achieved strong performance on action recognition tasks [9]. Despite their strengths, these models require a lot of computation and memory, and they can easily overfit when there is limited training data. This makes them hard to employ in real life.

### 2.2.4 The CNN-LSTM Hybrid: A Pragmatic Compromise

The CNN-LSTM hybrid combines the strengths of both 2D CNNs, which capture information in space in each image, and LSTMs, which look for patterns in time. This makes it quite efficient and effective for analyzing videos. The hybrid is composed of many components, making it easy to modify one of them. This makes it quite adaptable and ready for any future advancements. The hybrid is quite popular in research work as it allows you to analyze your videos quite effectively without incurring much cost in computation as in 3D CNNs [21].

## 2.3 Motion Representation for Efficient Violence Detection

Violence is all about motion, and it is therefore crucial to ensure that there is accuracy in capturing this motion in order to distinguish violent behavior from non-violent behavior, and at the same time ensure that the processing is fast enough.

### 2.3.1 Optical Flow as a Rich Motion Descriptor

Optical flow is a widely used method for representing dense motion. It calculates motion vectors for each pixel between consecutive frames, providing detailed information about movement [22]. Many high-performing systems, such as two-stream networks, rely on optical flow. In these networks, one stream processes RGB frames to capture appearance, while the other uses optical flow to focus on motion. The disadvantage of this algorithm is that it is computationally expensive in calculating the dense optical flow. This makes it quite difficult to implement in real-time applications [23].

### 2.3.2 Frame Differencing: A Lightweight Alternative

Frame differencing is a very efficient and straightforward method for analyzing the motion present in videos. The process involves subtracting one frame from another. This helps in pointing out the regions where there is motion while ignoring the stationary background. This strategy eliminates redundant information while concentrating on the patterns that are most significant for analyzing violent behaviors. Violent behaviors tend to involve very rapid motions [9]. The frame differencing mechanism can be considered as a sort of attention mechanism. It is able to focus on the regions where there is motion without needing any weights [24].

# Chapter 3: Background

There are four basic types of deep learning algorithms employed in detecting violence in images, which include supervised learning algorithms, semi-supervised algorithms, unsupervised algorithms, and a combination of algorithms. The four types of algorithms employ a different approach to learning and a different model of a neural network in image processing and analysis. The differences in approach and model of a neural network result in some algorithms performing better than others in certain circumstances.

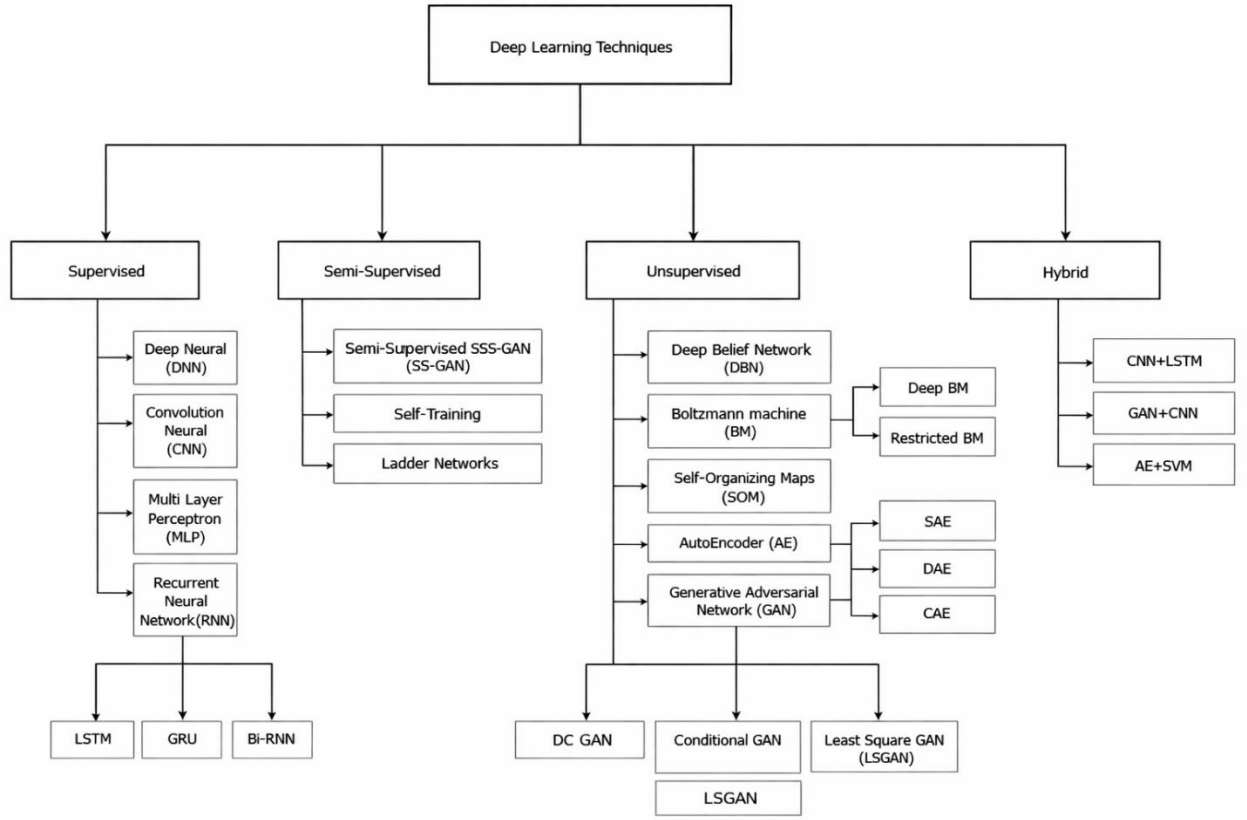


Figure 3.0. Deep learning techniques

## 3.1 Supervised Learning

One area where supervised learning is used in deep learning is violence identification. In this process of supervised learning, the machine is trained using the data labeled with their corresponding outputs. This assists the machine in comprehending the patterns in the images and their association with certain categories such as aggressive and non-aggressive images. There are various tasks in Supervised Learning that can be done by Deep Neural Networks (DNNs). DNNs can learn difficult patterns in large amounts of data. As a result of the presence of multiple layers, DNNs are capable of creating hierarchical representations of the data. This is quite useful in tasks such as violence detection, where it is required to extract small visual cues [25].

The Convolutional Neural Networks (CNNs) are very effective at dealing with spatial data from images. It is because of this that 2D CNNs are commonly used to detect signals of violence from individual frames of video or images. One of the main advantages of the use of CNNs is that the important features are learned automatically by the network, and thus one does not have to design them by hand [18].

Scientists were able to develop 3D Convolutional Neural Networks (CNNs) based on 2D CNNs in order to capture dynamics in video files effectively. The model is able to work with spatial and temporal information at once, which makes it possible to track variations in violent behavior in each pair of frames and is essential for correct detection in real life [26].

Region-based CNNs (R-CNNs), such as Fast R-CNN and Faster R-CNN, do not process the entire frame equally. Rather, these models pay attention to particular areas of the image. Such models can improve the accuracy of detection by pointing to areas of interest. Such models are very useful when dealing with crowded scenes where it might be more difficult to distinguish violence [26].

Recurrent Neural Networks (RNNs) are designed to handle data which is of the sequence type, and hence they are very suitable for the analysis of videos. Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU), are two of the variations of RNNs which perform very well at the job of finding patterns in data, and hence are very suitable for determining what people do when they are violent and analyzing sequences of acts [27].

In the Bidirectional RNNs' case, the method is further extended because the RNNs process the data in the forward as well as the backward passes. This gives the model more knowledge about the time aspect, which might make it more precise in identifying violence [22]. Finally, the feed-forward neural network or Multi-Layer Perceptrons (MLPs) is made up of layers of neurons which are interconnected. These networks are also utilized by humans in tasks such as classification or regression. In the violence-detecting models, MLPs are applied to evaluate or classify the vectors of features that are extracted from the images or picture frames. This assists in determining the violence content depending on the representations of the data [26] [28].

## 3.2 Semi-Supervised Learning

Semi-supervised learning is in the middle between supervised and unsupervised learning since it uses both labeled and unlabeled data to train. This method is very useful for finding violence, because it takes a lot of time, money, and often personal opinion to collect and categorize huge video datasets. Semi-supervised approaches can increase model performance and generalization by using a lot of unlabeled data and a smaller quantity of labeled examples. This makes less use of manual labeling. One common method used in Semi-supervised learning is self-training. This involves training a machine using data that is labeled. After which the machine assigns labels to some data that is not already labeled. This not-labeled data is added to the training pool, thereby improving the machine. In the world of violent action detection, self-training aids the machine in identifying a range of body motions from a large pool of data [29].

The Semi-supervised Generative Adversarial Networks (SS-GANs) are an extension of the traditional GAN framework, in which the discriminator is able to make use of both labeled and unlabeled data. By doing this, it essentially undertakes two tasks simultaneously – classification and differentiating between real and fake data. This allows it to work towards discovering more meaningful features and making use of large amounts of unlabeled data as well. The potential of the former in violence detection has proved quite significant when there is limited labeled data available. Initial works indicate that such models have the potential to enhance the accuracy of classification and provide more stable performance with limited amounts of data compared to fully supervised methods [30].

Ladder Networks combine both supervised and unsupervised tasks by optimizing both the classification and reconstruction losses on multiple layers. Such architecture facilitates stable and noise-resistant representation learning. Ladder Networks have been able to extract vital spatial and temporal information and generally generalize properly on various video signals for violence detection tasks [31].

## 3.3 Unsupervised Learning

Unsupervised learning is about drawing insight from data with no pre-assigned labels. As such, it does not rely on labeled examples but instead queries the data itself to find natural patterns, clusters, and underlining structures [28]. That capability makes it of special value for knowledge discovery, such as identifying unusual or deviant behavior, and automatic extraction of key features that are relevant for violent tasks.

Self-Organizing Maps (SOMs) are a kind of artificial neural network that learns how to make a low-dimensional (usually 2D) picture of the input space of the training samples all by itself. This is why they are so good at finding hidden groups or patterns in data with a lot of dimensions [16]. Stacked autoencoders are a composition of several autoencoders in which each subsequent autoencoder takes the output of the preceding autoencoder. This allows the model to build up increasingly abstract data representations, and it is very good at minimizing redundant dimensionality while emphasizing what it considers to be the most important features [32].

Denoising Autoencoders are designed to unravel clean data from noisily input ones. Through the training of the network to reverse the noise injected into the data, the network is able to focus on the intrinsic characteristics of the data. This helps the Denoising Autoencoders develop strong models which remain intact even with dirty data present in reality [33].

Contractive Autoencoders (CAEs), on the other hand, develop stable feature representations through the inclusion of a new regularization term. This particular term helps in adding a penalty to the model to ensure that the hidden layer does not respond vigorously to minute variations in the input. This particular outcome results in a feature space which remains relatively stable when a minute variation in the input is observed. This makes CAEs effective in identifying aggressiveness, where the differentiation between normal and abnormal patterns is important [34].

Generative Adversarial Networks (GANs): The Deep Convolutional GAN (DCGAN) integrates convolutional networks into the GAN structure, which helps the network produce better images. You can use the concept of the DCGAN to augment the dataset or create realistic samples for training [28].

The conditional GANs are a modification of the traditional GAN architecture that enable the use of extra information like class labels. This means that the network can create more specific data than before instead of generating it randomly. This is very useful in research on violence detection since it allows you to simulate kinds of violence at any time you want [35].

Another variant is called Least Squares GAN (LSGAN), and it replaces the loss function used in the training of the discriminator with a least-squares loss function. This is beneficial because it prevents vanishing gradients, which are associated with slow training, during training and thus tends to train much better than previous methods [36].

## 3.4 Hybrid Models

Hybrid models are a very complex combination of techniques used in supervised and unsupervised learning paradigms. They may employ other techniques that utilize the strengths of supervised and unsupervised learning. The objective of such models is to enhance their performance and ability to generalize in complex tasks such as violence detection, where large amounts of supervised and unsupervised data are accessible. However, there are challenges associated with data types.

**GAN + CNN:** The combination of GAN and CNN makes an effective tool for data and feature enhancement. The role of the GAN is to produce artificial and realistic training data, which is then employed to train the CNN, thereby increasing the level of accuracy in violence identification [37].

**AE + SVM (Support Vector Machine):** The autoencoder (AE) is employed in the reduction of dimensions and the identification of features, which transform the high-dimensional data to more meaningful data. The features are then used in the classification process by the Support Vector Machine (SVM), which is quite effective in classifying data where there is an obvious boundary between the classes [38].

**R-CNN + Graph Models:** The Region-based Convolutional Neural Networks (R-CNNs) identify features from specific regions of interest that are analyzed using graph models incorporating relational information among the regions. This is appropriate for violence that is implicitly indicated by interaction information rather than visual information [39].

## 3.5 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are employed in violence detection tasks owing to their robust capability to detect spatial features in video frames. Based on their discriminative ability to learn visual features, CNNs are capable of detecting features such as aggressive human postures, forceful human movement, and unusual human interactions that are normally evident in violent acts. Based on optical flow features and CNNs, it is possible to detect signs of violence such as

aggressive human postures that are normally characteristic of human fights and violence [40] [41] [42].

The CNNs learn hierarchical feature representations, beginning with basic visual features like edges, corners, and textures and advancing to more complex features like gestures, interactions of humans, and object movement. This learning ability at different hierarchical levels helps the CNN-based models perform well under challenging situations involving cluttered environments, partial occlusion, cluttered backgrounds, or complex camera views [43]. This type of ability is very essential in practical surveillance applications, as lighting is usually uncontrolled.

CNNs such as ResNet, DenseNet, and EfficientNet provide a superior form of representation as well as generalization capabilities with high computational efficiency through architectural advancements such as residual learning, dense connections, and compound scaling of the neural network [42] [44] [45]. This tradeoff between accuracy and speed of the algorithms makes CNN-based solutions very suitable for large-scale implementations and real-time surveillance applications. These architectures also demonstrate the ability to deal with practical challenges of the application domain, such as lighting, camera motion, and occlusion, which can be encountered in violence detection from videos [46].

## 3.6 Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, are very good at detecting such patterns in videos because of their capability of simulating such relationships. Usually, violent activities are observed to occur in a sequence of acts and not as a result of a single act. RNNs have the capability of learning such sequences of activities and are hence able to distinguish between acts of aggression and fast, unintentional, and harmless movements [42] [45]. RNNs have the capability of storing information from previous images in their memory.

When applied together in a hybrid framework, RNNs and Convolutional Neural Networks (CNNs) enable the description of the space and time in videos. The CNNs extract high-level spatial features from videos, such as body position, motion, and interaction, from individual frames. The RNNs, on the other hand, describe changes in these features over time to describe action dynamics [44].

Another sophisticated extension is Convolutional LSTM (ConvLSTM). The model improves results by introducing convolutional layers to the LSTM architecture itself. Thus, ConvLSTM is effective for video-based surveillance systems because this model maintains spatial correlations and also captures temporal connections [47]. Hybrid CNN-RNN models also proved effective for real-world surveillance applications where there are possibilities of moving cameras, obstacles, viewpoints, and congested situations [43].

# Chapter 4: Dataset Description

For the development and evaluation of efficient violence detection models, it is necessary to understand the nature of the datasets employed for the purpose of training and evaluation. The size and nature of the dataset, as well as its ability to resemble real-world scenarios, play very important roles in determining the effectiveness of the developed model. This chapter stresses the need for the use of real-world datasets. The dataset employed in the research is discussed below and is the most common dataset employed in violence detection research.

## 4.1 The Need for Real-World Surveillance Datasets in Violence Detection

Real-world surveillance datasets are crucial to furthering systems of violence detection. Many current datasets have their origins in sports footage or films. Nonetheless, they do cover a wide range of actions, contexts, and camera angles that help models pick up useful patterns. At the same time, there is an increasing focus on creating datasets that more closely approximate real surveillance environments capturing motion from cameras and complex interactions of people in crowds so models can work when deployed for real monitoring.

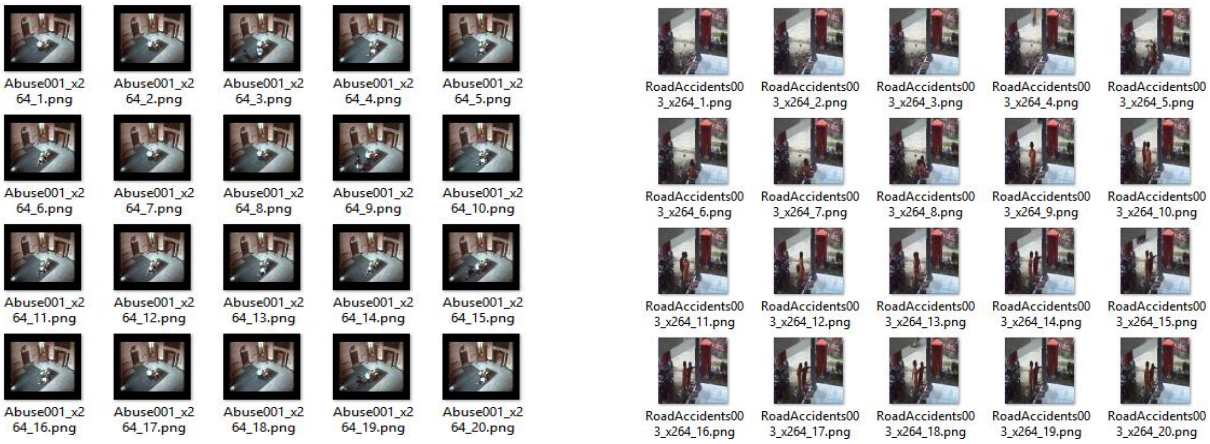


Figure 4.1. UCF Crime Dataset

## 4.2 The UCF Crime Dataset

The UCF Crime dataset is one of the biggest standard datasets used in the literature in the domain of violence detection because of its complexity [48]. This dataset is specifically designed to assist in the development of automated surveillance systems that could identify abnormal or aggressive actions in reality. This dataset contains 14 different actions, some normal and some violent. These categories hold a number of violent actions such as fighting, robbery, burglary, explosion, and assault, and some actions which are neither violent nor normal actions such as strolling, standing, and loitering [44]. The current dataset holds more than 1,900 video clips and these are all from

real-world surveillance systems. The diversity and complexity of the UCF Crime dataset are two of its greatest strengths. The videos were taken under varying lighting, from varying angles, and at varying resolutions to show the complexity of real-world surveillance [44]. This makes it extremely difficult for automated detection systems to handle because they have to handle issues like occlusions, background clutter, blur, and camera shake. There are both lengthy and shorter videos in the dataset, which allow researchers to test the ability of their algorithm to handle time in actuality, like the ability to spot unusual or violent behaviors in the past. The dataset is well-structured and provides segregation between normal and deviant behaviors. This makes it possible to train and test the performance of the model on actual criminal detection tasks. The UCF Crime dataset is an ideal tool to test new approaches to violence detection because it is rich, difficult, and informative. It provides the best of the lab environment and the complexity of actual surveillance.

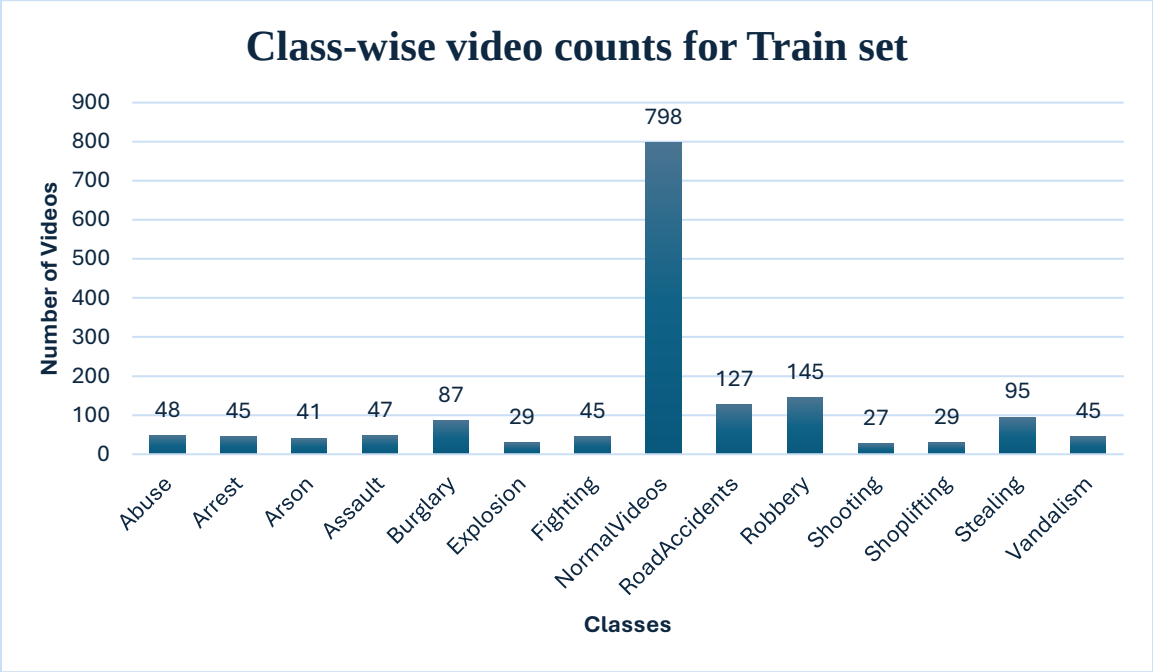


Figure 4.2. Class-wise video counts for Train set

Figure 4.2: This is a bar graph showing the number of training videos for each class. This graph shows a huge imbalance in class sizes. Look at how many videos there are in the "Normal" class compared to the other classes. The "Abuse" class has only 48 videos, the Explosion class has only 29 videos, and there are only 27 videos for the shooting class. The large number of people who fall into the "Normal" class shows that it is not evenly distributed, which can have implications for

model training. To counter this issue, you can choose to apply techniques to oversample or change class weights.

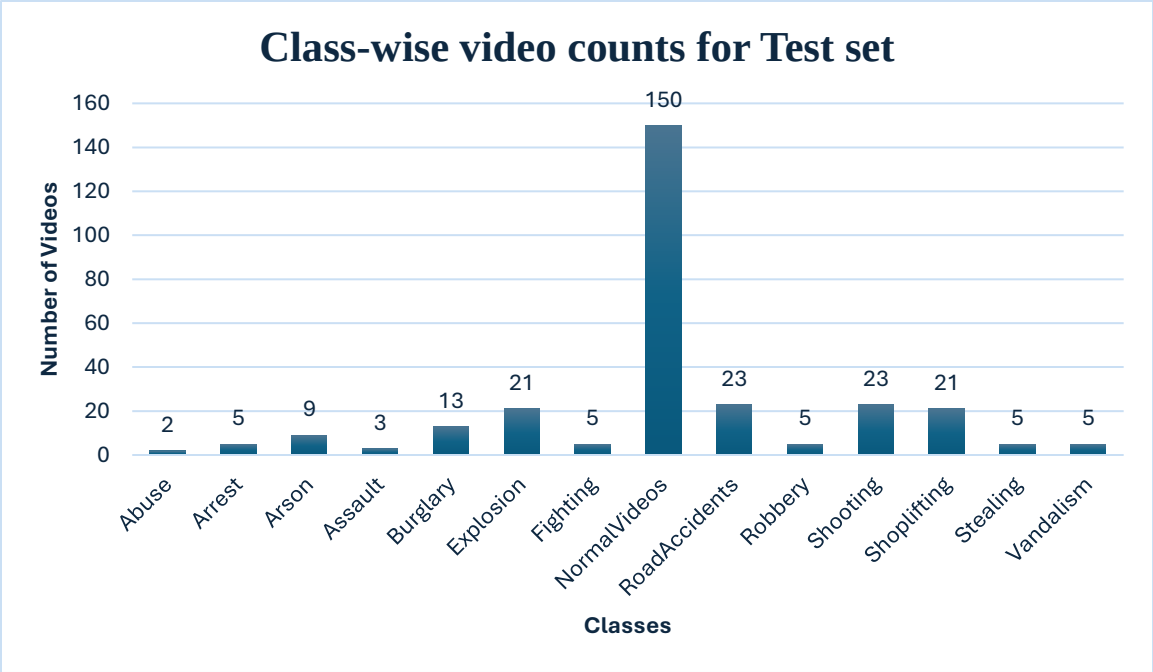


Figure 4.3. Class-wise video counts for Test set

A bar chart in Figure 4.3 illustrates the distribution of classes in the test dataset. From the graph, it is evident that there are many instances of "Normal Videos" since there are a total of 150 videos. The other classes are quite rare since their counts range between 2 and 23 videos. This is an indication that the test dataset is composed mostly of normal events. This may impact the evaluation of the model since it will be quite easy to identify the normal events.

# Chapter 5: Method

Two different methods for automated violence detection in video frames have been explored in this research. The first technique involves using ResNet50-LSTM on regular frames from videos. The second technique involves using the difference between frames from videos. The primary aim of this research is to compare the two methods to establish if the frame difference technique can improve processing speed without compromising the Area Under the Receiver Operating Characteristic Curve (AUROC) or classification accuracy. The comparison of results is important for establishing if the frame difference technique can provide the best of both worlds, which is an important consideration for real-time systems.

The overall method will be explained in this chapter, starting from data pretreatment up to model design. Every design consideration will be discussed in relation to the overall objectives of this study, which are to improve the accuracy and computational efficiency of violence detection for the UCF Crime dataset.

## 5.1 Data Preprocessing

Data preprocessing is a very important part of any deep learning pipeline because raw video data often has a lot of extra information and changes a lot [49]. The goal of preprocessing in this work is to filter the input data so that only the most important motion patterns that show violent actions are kept and the least important background information is removed.

This research is based on the UCF Crime dataset, which is made up of real-world surveillance videos of different lengths and resolutions [44]. To ensure consistency in model training, it is necessary to extract a fixed sequence length from each video. A sequence length of 20 frames (or frame differences) was chosen for this study. This value was selected based on a compromise among three factors: findings from previous research, the computational resources at hand, and the imperative to obtain adequate temporal context for accurate identification of violent activities.

Common sequence lengths for video-based recognition tasks are between 8 and 64 frames [32]. Sequences that are too short (e.g., $\leq 10$ frames) often don't show how an action changes over time. Violent events are often marked by a buildup of tension, such as yelling or pushing, that leads to a peak action, such as hitting or fighting. A short sequence may miss these important cues here, which could lead to more false negatives [48]. On the other hand, very long sequences (like 30 frames or more) put a lot of strain on computers. Adding more frames makes it harder for the ResNet50 feature extractor and the LSTM network to do their jobs. Initial tests showed that using 30 frames added about 40% more time to training and only improved AUROC by less than 0.01. Based on research on similar action recognition datasets [50], a sequence length of 20 frames was chosen as a good balance between being long enough to include most violent events and short enough to be useful for training and inference.

The value of the step size used to sample the frames from a video with a total number of T frames is given by step = T // 20. Then, frames are sampled uniformly (at positions: 0 step, 1 step, 2 step, and so on, up to 19 step) from the initial frame of the video. To ensure that the input size of ResNet50 is satisfied, each frame or each computed frame difference is resized to 224×224 pixels. Finally, dividing each value by 255 ensures that the intensity values of the pixels lie in the unit interval [0,1]. This ensures that the 8-bit intensity values of the pixels (ranging from 0 to 255) lie in a unit interval, which makes the learning process less sensitive to illumination changes.

In this study, two different methodologies of data preprocessing have been used:

:

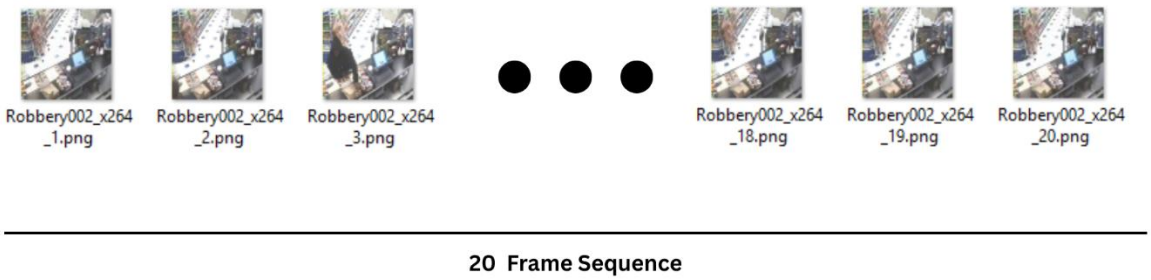


Figure 5.1. Frame Sequence

- **Normal Frames:** This approach involves the direct extraction of 20 color frames from the original video. This approach provides the system with a complete understanding of space and time, including both static and dynamic aspects of the scene. The major advantage of this approach is that it maintains all appearance features (such as objects, scenes, and human poses) that could be useful in classification tasks [49]. However, one of the major drawbacks of this approach is that there is a significant amount of static information that is not useful in motion-related detection tasks.

- **Frame Differences:** The second approach takes differences between frames as inputs in order to give more weight to the motion information. The approach is very effective in locating violence as violence usually involves rapid and violent motions [19]. To form a set of 20 different frames, a set of 21 frames is first extracted from the video. The absolute difference between frames is then computed for every channel:

$$\Delta F = F(t) \sim F(\mathrm{t}_1) \quad \text{----------(1)}$$

In this case, (F(t)) is the frame that is currently used, while (F(t1)) is the frame that existed earlier. The variation in each frame shows how the video varies from one frame to another. This shows change and/or movement in the video. This is useful in removing unnecessary information presented in the video and focuses on the dynamic aspect of the movement, which might be abnormal and therefore show violence [20].

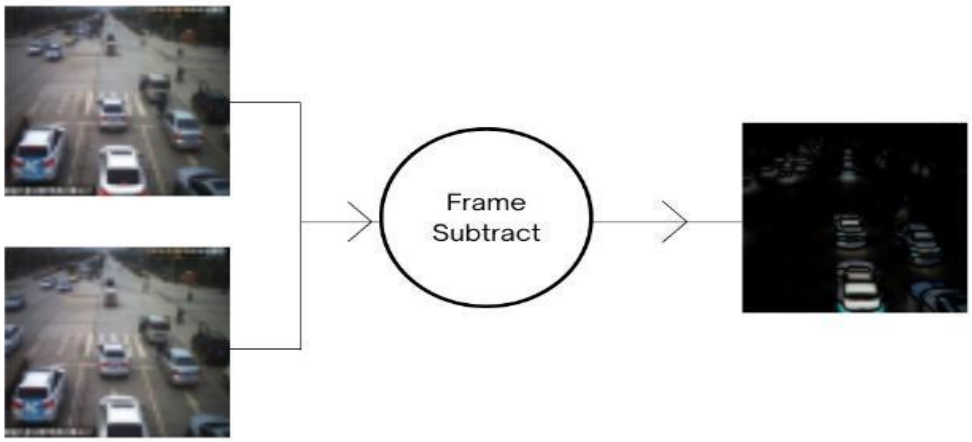


Figure 5.2. Consecutive frame


Figure 5.3. Frame Differences

## 5.2 Dataset Splits

The official splits for training and testing come with the UCF Crime dataset. The official test set was kept totally separate for this analysis and used only at the end to test this new model on it to ensure that this testing was done fairly on unseen data. The official training set was further split, with 80% of the data used for training this model while 20% of it was set aside for validation. This type of split is very common in machine learning models [51]. The idea here is to allow this machine learning model to learn excellent feature representations from 80% of this very diverse set of data while still having this 20% set aside for validation [26].

## 5.3 Model Architecture

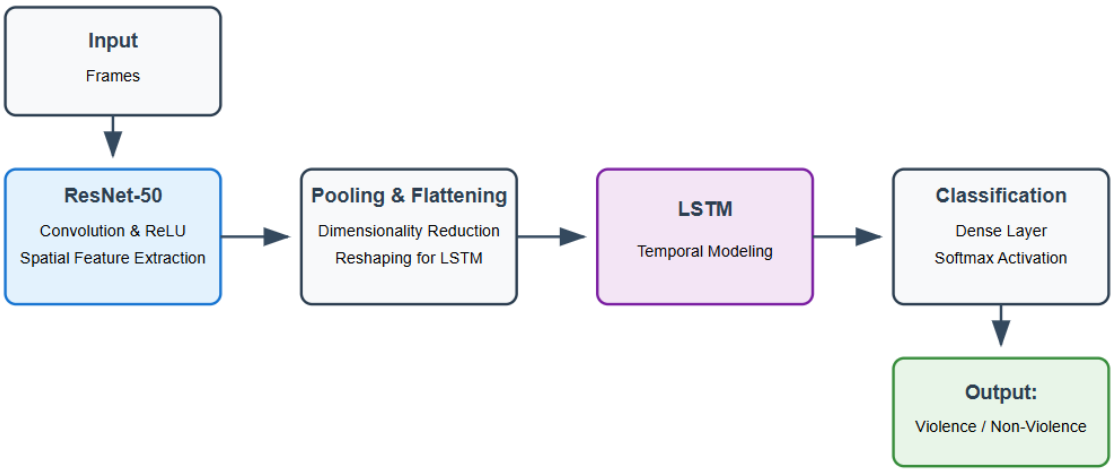


Figure 5.4. ResNet50-LSTM Model

The architecture is referred to as a hybrid architecture, where an LSTM architecture is used for handling the temporal nature of data, while a ResNet50 architecture is used for handling the spatial nature of data. This architecture is able to utilize Recurrent Neural Networks for handling temporal information from video frames and Convolutional Neural Networks for handling spatial information from video frames [44].

Figure 5.4 represents the overall architectural design of the proposed model. Here, the input sequence of data of the form (20,224,224,3), which represents the sequence of either 20 normal frames or frame difference frames, is provided to the network, and the spatial features are passed through the LSTM network for classification, finally producing a decision through the softmax-activated dense network to determine whether the given sequence of frames is violent or non-violent, respectively. The model is implemented using the Anaconda Jupyter Notebook environment.

### 5.3.1. Spatial Feature Extraction (ResNet50)

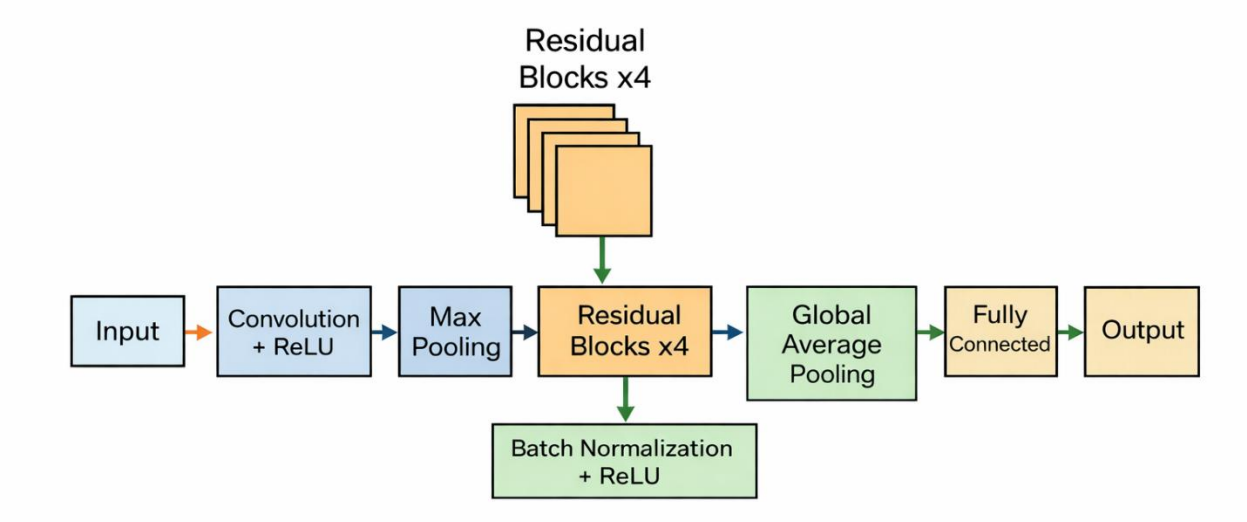


Figure 5.5. ResNet-50 Architecture

In order to extract useful information from these images, we resort to the ResNet50 model, which has been trained on ImageNet [49]. This model has residual connections to counter vanishing gradient values, making it capable of training deep models effectively. The early convolutional layers of these models detect low-level information like edges and textures. The later layers detect more semantic information like objects and shapes [32]. ReLU activation functions add less linearity to these models, and batch normalization ensures faster convergence. With transfer learning, you freeze the first few layers of your ResNet50 model to retain its feature detectors. The last layers of these models are then adjusted to detect violence.

### 5.3.2. Avg-Pooling and Flattening

After the ResNet50 has processed each of the frames, it appears as a high-dimensional feature map, such as 7x7x2048. We apply the Global Average Pooling (GAP) layer to each of the feature maps to transform them into a 1D vector, which is comprised of 2048 components. This is beneficial in the sense that it decreases the parameters and assists in the avoidance of overfitting, as well as enhancing the translational invariance [33]. This produces a set of 20 vectors and each vector is 2048 bytes long. This data is what the temporal model takes as input.

### 5.3.3. Temporal Modeling (LSTM Layer)

A flow of feature vectors of size 2048 is passed to an LSTM layer. LSTMs are suitable for modeling long dependencies in a sequence as their gate control mechanism (input gate, forget gate, and output gate) allows greater control over the flow of information [20]. This is a very crucial aspect of modeling how a potentially violent act might evolve in a temporal fashion. It was observed after experimentation that a lone LSTM layer with a unit size of 128 was optimal for performing this task. 5.3.4. Classification (Dense Layer with SoftMax The final hidden state from

the LSTM layer is fed into a fully connected (Dense) layer with two units. A softmax activation function is applied to the output to get the chances for every class. A dropout layer (rate=0.5) is added before the classification layer to prevent overfitting further. During training, it drops neurons [35].

.

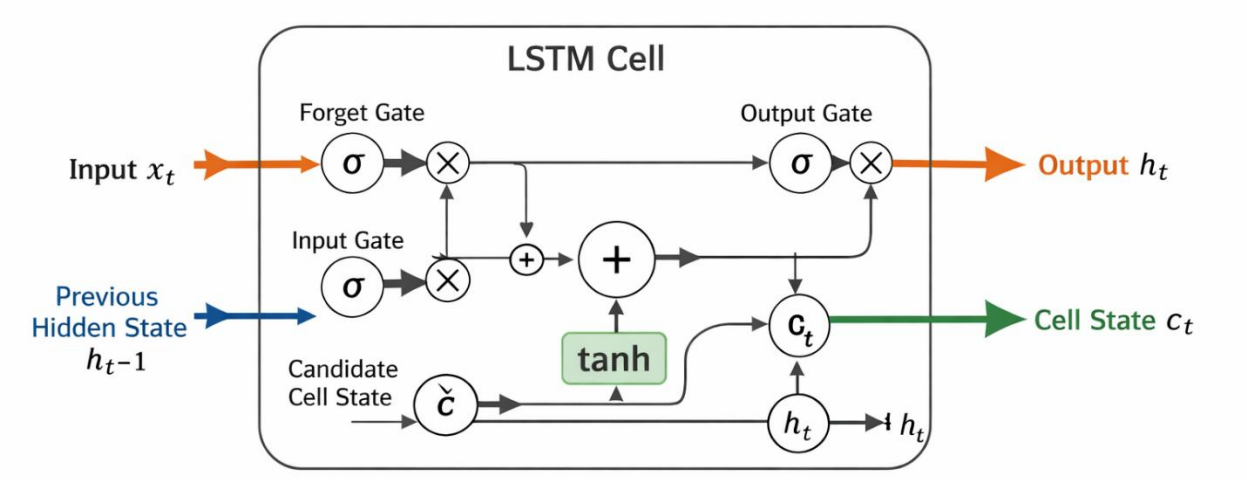


Figure 5.6. LSTM Architecture

### 5.3.4. Classification (Dense Layer with SoftMax Activation)

The last hidden state from the LSTM layer goes to a fully linked (Dense) layer with two units. A softmax activation function turns the outputs into the chances of each class. A dropout layer (rate=0.5) is used before the classification layer to help stop overfitting even more. During training, this randomly drops neurons [35].

# Chapter 6: Training and Evaluation

The training & testing phase is supposed to improve the ResNet50-LSTM network for violence detection and test it. The chapter describes to you what to do to train your computer, what training entails, what parameters were used, and what results were achieved. To compare them fairly, two models were trained under exactly the same parameters: one took images, while the other took differences between images. The aim here is to test to what extent they can sort things out fast enough, which is important for practical usage.

## 6.1 Fundamental Classification Outcomes

There are four key results of a confusion matrix to evaluate how effectively a model performs:

1. True Positive (TP): A violent clip is labeled as violent.
2. True Negative (TN): A non-violent video is correctly labeled as non-violent.
3. False Positive (FP): A video that is not violent but classified as violent mistakenly.
4. False Negative (FN): A violent video that is missed by the detector.

These results are the foundation of all other performance metrics that follow [36].

| | **Predicted Violent** | **Predicted Non-Violent** |
|---|---|---|
| Actual Violent | True Positive (TP) | False Negative (FN) |
| Actual Non-Violent | False Positive (FP) | True Negative (TN) |

Table 6.1 Confusion matrix table

Table 6.1 shows the confusion matrix of the proposed model on the test set. It illustrates the distribution of TP, TN, FP, and FN predictions. This analysis provides insight into misclassification patterns between violent and non-violent videos.

## 6.2 Training Process and Hyperparameters

To provide a fair point of comparison, the conventional frames and frame difference models are trained in the same environment. Both models are optimized for their ability to perform the violence detection task using the same set of hyperparameters.

**1. Batch Size and Epochs**

- **Batch Size**: The model processes 16 samples for each training iteration because the batch size is 16. A smaller batch size would allow the model to discover more complex patterns and would require slightly more time for training [37].

- **Epochs**: The number of epochs for both models is 50. The number of epochs means that the models will get to process the entire training data set multiple times.

**2. Learning Rate**

Learning rate was fixed to 0.0001. This value makes it simpler to introduce slight modifications to the parameters of the models. This ensures models converge to a perfect solution. A higher value may lead to models failing to identify optimal points. On the other hand, a lower value ensures models make steady progress [49].

**3. Optimizer: Adam**

We employ the Adam optimizer to adjust the learning rate in our model. We find Adam to be suitable for this purpose because it adjusts the learning rate based on the first moment (mean) and second moment (variance) of the gradient [19]. This property ensures that the training process remains stable and also improves convergence speed, even when working with complex models and large datasets. Adam optimizer has proven to be very effective since it has combined the advantages of both momentum and adaptive learning techniques [19].

## 6.3 Evaluation Metrics

We use the following metrics to compare the performance of both models:

- **Accuracy**: Calculates the proportion of correct predictions on the video samples. This is an easy method of gauging the model's overall performance according to the formula,

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **ROC (Receiver Operating Characteristic Curve)**: It is a graphical representation showing the goodness of the model based on the true positive rate (TPR) versus the false positive rate (FPR) values [36].
- **AUROC (Area Under the ROC Curve)**: This number shows how well the model can tell the difference between violent and non-violent classes. A higher AUROC score means better discrimination, and a score of 1.0 means that the two classes are perfectly separated [20].

With the usage of accuracy and AUROC, we can evaluate the effectiveness of the model in classification as well as its ability to recognize the details that are required for the detection of violence.

## 6.4. Computational Time Measurement

We analyzed the efficiency of the computer by measuring the time it took for training and testing. The environment used for testing is the same for all experiments. The environment consists of an

NVIDIA GeForce RTX 2060 graphics card (6 GB GDDR6 VRAM), an Intel Core i5-10400F processor, and 16 GB RAM. This allows us to provide a fair comparison between the processing speeds of the input modes.

# Chapter 7: Results And Discussion

The result of our research is presented in this chapter, along with the implication of the result to the violence detection systems. In an attempt to make the differences between the two approaches to identify the violent and non-violent scenes clearer, we compare the frame difference method to the traditional frame approach in terms of performance and computing efficiency.

## 7.1 Performance Evaluation

We employed two very important measures: the use of Receiver Operating Characteristic curves and the Area Under the Curve of the Receiver Operating Characteristic. We employed frame differentials as well as frames in determining the effectiveness of our approach.

| Class | Normal Frame AUROC | Frame Difference AUROC |
|---|---|---|
| Violence | 0.8892 | 0.9091 |
| Non-Violence | 0.8988 | 0.9091 |

Table 7.1 Comparison of AUROC for Normal Frame and Frame Difference Approaches

The AUROC values for the violence and non-violence classes are shown in the Table 7.1 below. The performance of the AUROC of the frame difference approach is better compared to the traditional frame approach. The AUROC values for the violence class are 0.8892 and 0.9091, respectively. Therefore, the AUROC values for the non-violence class are 0.8988 and 0.9091, respectively. This is indicative of the effectiveness of the approach in identifying violence and non-violence sequences, which is probably due to its ability to handle motion and scene changes, which is very necessary for the identification of violence. The reason why the frame difference method is more effective is that it has the capacity to obtain more information about the scene. This information includes scenes where there is either motion or environmental change. This information is necessary for the detection of violence. Using the frame difference method, it is possible to distinguish a violent scene from a non-violent scene. The micro-average ROC curves for standard frames and frame differences are shown in Figure 7.2. The superiority of the frame difference approach is also indicated by the high micro-average AUROC values of 0.89 and 0.91,

respectively. Apart from AUROC, we have also measured the accuracy of our method. Accuracy has been shown to remain unchanged for normal images as well as for frame difference images, which has a value of 85%. The confidence level of the frame difference method to attain this accuracy can be measured by its AUROC values.

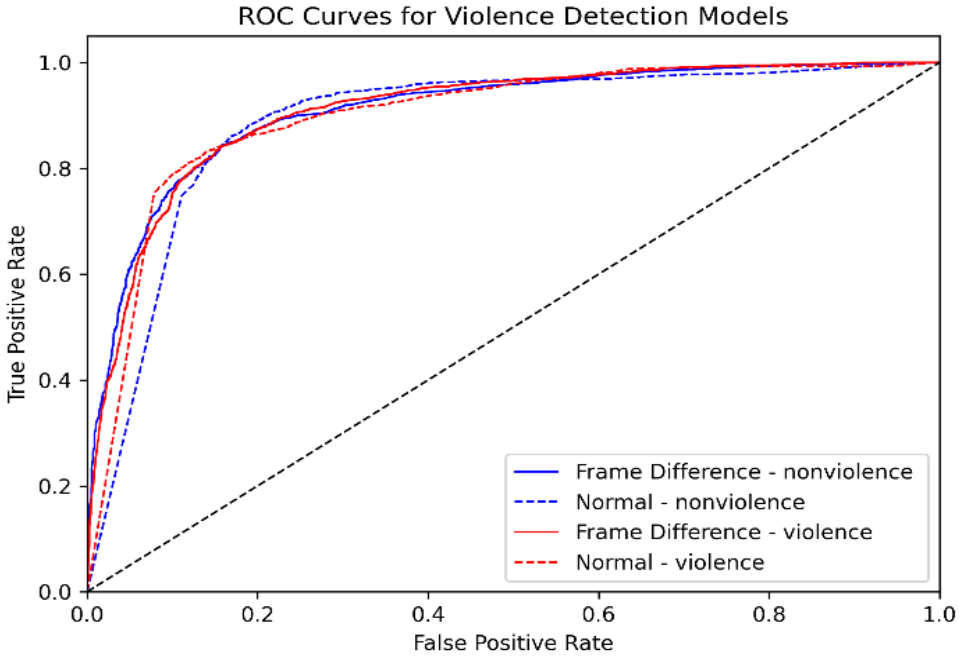


Figure 7.1 Class Wise ROC Curves

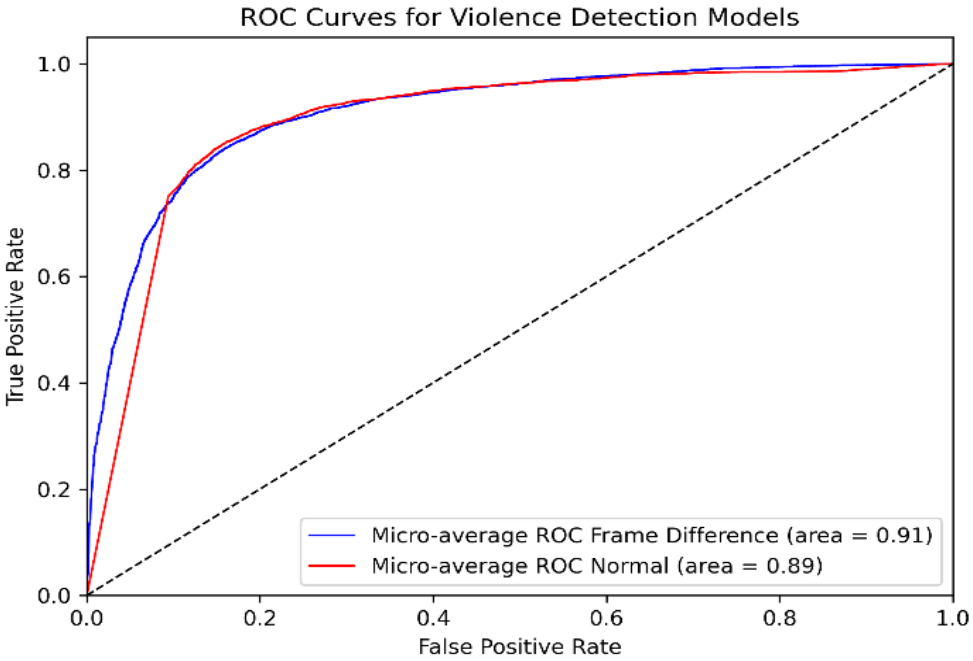


Figure 7.2 Micro-average ROC Curves

## 7.2 Computational Efficiency

In addition to testing the model performances, we are also testing the computational efficiencies of our models. The information is presented in the order in which it occurred. The results indicate that the frame difference method is slightly better in terms of time-saving. There is a difference of about 2.9% in training time and 4.2% in testing time. Although it is not a remarkably large variation, it still makes the frame difference technique more appropriate for use in real-time systems. Small variations in time can make a large variation in the time it takes for the system to respond.

| Approaches | Training Time/epoch (s) | Inference Time on Testset (s) |
|---|---|---|
| Frame Difference | 361.4 | 23 |
| Normal Frame | 372.2 | 24 |

Table 7.2. Time Computation Table

The reason why it has been proved that using the frame difference technique is an efficient way of increasing computing speed is due to its capability of reducing the dimensionality of the feature space. The technique takes less processing power and results in greater processing speed. The reason for this is that it concentrates on extracting more relevant features from the frames of the video.

## 7.3 Discussion

The frame difference method is efficient in identifying violence in videos, as evidenced by our tests. The frame difference method and traditional frame method work well in sorting out issues, although the frame difference method is slightly faster and requires fewer processing powers. The above results are important in crime detection systems as they show that the frame difference method is able to get a correct result with less effort.

This algorithm is very helpful for real-time systems that have limited speed and computational power because it works very well. The frame difference algorithm is more helpful for applications that require rapid response time, such as surveillance cameras and real-time video monitoring applications, because it is able to detect objects quickly without losing precision. This algorithm can be employed in place of traditional frame-based algorithms because it maintains efficiency while reducing the workload that has to be processed by the computer. Traditional frame-based algorithms consume more computer power and are not suitable for applications involving time constraints.

# Chapter 8: Conclusion and Future Directions

This final chapter concludes the overall pathway of the study in summarizing its key findings and how it has put forward advancements in automated surveillance technology as a field of study. Among other factors, it also explains the limitations of its own study as well as how to follow along in its pathway in future studies.

## 8.1 Conclusion

This work was informed by one key, global issue-the high social and economic impact of violence-and ways of improving automated video surveillance systems to help detect these situations. The goal wasn't just to record more but to develop smart systems capable of real-time detection of violent behavior for even quicker responses.

The investigation started by tracing the changes of video analysis from the traditional methods, where security officers were watching hours of tapes, slowly and with mistakes, to the new method using AI. But a new breed of learning algorithms, specifically CNNs that interpret images, or LSTMs that interpret a sequence of actions in time, has revolutionized this area of research.

A major challenge in this field is balancing accuracy and speed. A system that is perfectly accurate but too slow to process video in real-time is not helpful for live monitoring. This research aims to tackle this exact problem. The study proposed and tested a specific method to improve efficiency: using frame differences rather than full-color (RGB) video frames.

The frame difference method focuses only on areas where change is observed in a video, thereby completely ignoring the static parts. This makes the task less complicated for the AI model. This study applied a very powerful hybrid model which combined ResNet50 (for understanding the features) with LSTM (for understanding how those features change). This model was trained and tested on the realistic UCF Crime dataset.

The analysis of experimental results emphasizes the performance of the frame difference model. The model using frame differences performed slightly better at distinguishing violence from non-violence (achieving an AUROC score of 0.91 vs. 0.89) while matching the accuracy (85%) of the model using full RGB frames. Most importantly, it did this faster, reducing the time needed for both training and testing by 2.9% and 4.2% respectively. This shows that the system can operate more effectively by focusing on the most important motion signals and can even improve at detecting violence.

In conclusion, from this research, it has been found that the use of frame difference for focusing on the motion part of the video is an effective approach for automatic violence detection. It is one of the solutions that offer good accuracy and fast speed.

## 8.2 Limitations

In the research assessing frame-based and traditional methods for identifying violence in data from UCF Crime videos, it's found that there are a number of major issues with both approaches. On one side, it's been considered as a possible limitation that perhaps the data does not contain all of the different scenarios that may come up in real life. This may make it more difficult for it to apply to other situations. The architecture is fairly robust, but for finding intricate patterns in video, it may not work well. However, it is likely that using the frame difference method could have resulted in faster processing times and possibly used data with even lower accuracy levels. It is also difficult for one to understand the effectiveness of the system in challenging scenarios, such as when the light is poor or when there are too many people around. This further leads us to question the effectiveness in real-world applications. There are a lot of flaws in Optical Character Recognition, making it difficult for one to replicate the tests or compare the findings effectively.

## 8.3 Future Work

The plan is to improve our model and make it way more reliable when used in actual practice. To achieve this, we will train it with data taken under different circumstances, such as in low lighting, different viewpoints of the camera, and different cultural settings. We would also like to experiment with this model in more challenging scenarios such as dim lighting, crowded areas, or situations involving partially visible persons. The datasets to be generated for such testing purposes will aid in better understanding of this model in practical scenarios. We can make our system more accurate by using transformer models or incorporating attention mechanisms instead of LSTM. This will help us find long-lasting violent events better.

The frame-difference method is faster, but it loses some image details. We will combine this with optical flow to get a clearer picture of motion while keeping important static information. Datasets created for these test conditions will help to shed more light on the performance of the model in the real world.

In addition, transfer learning will be used to make the development process more efficient. By building on models already trained on large-scale video datasets, we expect to reduce training time and improve overall accuracy. Together, these steps move us closer to a dependable real-time system for detecting violent activity in practical settings.

# References


[1] World Health Organization, Injuries and Violence: The Facts 2014. Geneva, Switzerland: WHO Press, 2014.

[2] Y. Myagmar-Ochir and W. Kim, "A survey of video surveillance systems in smart city," Electronics, vol. 12, no. 17, Art. no. 3567, Sep. 2023.

[3] M. Gadelkarim, M. Khodier, and W. Gomaa, "Violence detection and recognition from diverse video sources," in Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN), Padova, Italy, Jul. 2022.

[4] M. M. Abid et al., "Computationally intelligent real-time security surveillance system in the education sector using deep learning," PLoS ONE, vol. 19, no. 7, Art. no. e0301908, Jul. 2024.

[5] V. Thamilarasi, R. Hema, A. N. M. Juliet, A. Sheeba, G. Ghule, and A. Raja, "Real-time runway safety: UAV-based video analysis with ICSO enhanced deep learning," Int. J. Comput. Exp. Sci. Eng. (IJCESEN), vol. 11, no. 1, pp. 1314–1329, 2025.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

[7] S. Vosta and K.-C. Yow, "KianNet: A violence detection model using an attention-based CNN–LSTM structure," IEEE Access, to be published, 2023.

[8] S. Sharma, B. Sudharsan, S. Naraharisetti, V. Trehan, and K. Jayavel, "A fully integrated violence detection system using CNN and LSTM," Int. J. Electr. Comput. Eng. (IJECE), vol. 11, no. 4, pp. 3374–3380, Aug. 2021.

[9] I. Mugunga et al., "A frame-based feature model for violence detection from surveillance cameras using ConvLSTM network," in Proc. 6th Int. Conf. Signal Process. Commun. Syst. (ICSPCS), Sydney, Australia, 2021, pp. 1–7.

[10] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Venice, Italy, Oct. 2017, pp. 3154–3160.

[11] S. Saif, S. Tehseen, and S. Kausar, "A survey of the techniques for the identification and classification of human actions from visual data," Sensors, 2018.

[12] Y. Li et al., "A comprehensive survey of vision-based human action recognition methods," Sensors, 2019.

[13] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," Int. J. Comput. Vis., vol. 103, no. 1, pp. 60–79, 2013.

[14] H. Wang and C. Schmid, "Action recognition with improved trajectories," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Sydney, Australia, Dec. 2013, pp. 3169–3176.

[15] M. Patel, "Real-time violence detection using CNN–LSTM," arXiv:2107.07578, 2021.

[16] J. Amin et al., "Detection of anomaly in surveillance videos using quantum convolutional neural networks," Image Vis. Comput., vol. 135, Art. no. 104710, 2022.

[17] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Trans. Neural Netw., vol. 5, no. 2, pp. 157–166, 1994.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv:1512.03385, 2015.

[19] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1, learning rate, batch size, momentum, and weight decay," Naval Research Laboratory, Washington, DC, USA, Tech. Rep., 2018.

[20] J. Li, “Area under the ROC Curve has the most consistent evaluation for binary classification,” PLoS ONE, vol. 19, no. 12, p. e0316019, 2024

[21] D. Tran et al., “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015.

[22] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2004.

[23] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the Kinetics dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 6299–6308.

[24] X. Zhou, D. Wang, and X. Zheng, “Object detection with deep learning: A review,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 7, pp. 3218–3230, Jul. 2019.

[25] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[26] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in Neural Networks: Tricks of the Trade, 2nd ed. Springer, 2012, pp. 437–478.

[27] K. Greff et al., “LSTM: A search space odyssey,” IEEE Trans. Neural Netw. Learn. Syst., 2017.

[28] J. An and S. Cho, “Variational autoencoder based anomaly detection using reconstruction probability,” Special Lecture on IE, vol. 2, no. 1, pp. 1–18, 2015.

[29] Y. Lee, S. Park, and J. Kim, “Self-training approaches for video-based action recognition,” Pattern Recognit. Lett., vol. 131, pp. 16–23, 2020.

[30] T. Salimans et al., “Improved techniques for training GANs,” in Adv. Neural Inf. Process. Syst. (NeurIPS), 2016.

[31] A. Rasmus et al., “Semi-supervised learning with ladder networks,” in Adv. Neural Inf. Process. Syst. (NeurIPS), 2015.

[32] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2014, pp. 818–833.

[33] M. Lin, Q. Chen, and S. Yan, "Network in network," in Proc. Int. Conf. Learn. Represent. (ICLR), 2014.

[34] S. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 6479–6488.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

[36] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006.

[37] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," arXiv preprint arXiv:1804.07612, 2018.

[38] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, 2006.

[39] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” arXiv:1609.02907, 2016.

[40] K. Lee et al., “Detection of real-world fights in surveillance videos,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Brighton, UK, May 2019.

[41] E. Martinez et al., “Violence detection in videos by combining 3D convolutional neural networks and support vector machines,” Appl. Artif. Intell., vol. 34, no. 4, 2020.

[42] X. Wang et al., “A novel violent video detection scheme based on modified 3D convolutional neural networks,” IEEE Access, vol. 7, pp. 129827–129835, 2019.

[43] S. Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” in Proc. IEEE Int. Conf. Adv. Video Signal-Based Surveillance (AVSS), 2017.

[44] J. Donahue et al., “Long-term recurrent convolutional networks for visual recognition and description,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015.

[45] Q. Johnson et al., “Physical violence detection in videos using keyframing,” in Proc. Int. Conf. Intell. Syst. Commun. IoT Secur., 2023.

[46] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[47] X. Shi et al., “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” in Adv. Neural Inf. Process. Syst. (NeurIPS), 2015.

[48] V. Kantorov and I. Laptev, "Efficient feature extraction, encoding and classification for action recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2014, pp. 2593–2600.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[50] X. Mao et al., “Least squares generative adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2017, pp. 2794–2802.

[51] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," arXiv preprint arXiv:1811.12808, 2018.