A

PROJECT-IV REPORT

on

Violence Detection System

Submitted by:

Gopika (210365)

Shivangi (210366)

Bhanu Pratap Singh (210293)

under mentorship of

Dr. Kiran Khatter
(Professor)

BML MUNJAL UNIVERSITY™

FROM HERE TO THE WORLD

Department of Computer Science Engineering
School of Engineering and Technology
BML MUNJAL UNIVERSITY, GURUGRAM (INDIA)

December 2024

Table of Content

CANDIDATE'S DECLARATION

I hereby certify that the work on the project entitled,"**Violence Detection System**", in partial fulfillment of requirements for the award of Degree of **Bachelor of Technology** in School of Engineering and Technology at BML Munjal University, having University Roll No.1232434, is an authentic record of my own work carried out during a period from July 2024 to December 2024 under the supervision of Dr. Kiran Khatter.

(Gopika)
(Shivangi)
(Bhanu Pratap Singh)

SUPERVISOR'S DECLARATION

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Faculty Supervisor Name: Dr. Kiran Khatter

Signature:

Abstract

Violence detection in surveillance systems plays a critical role in law enforcement and ensuring public safety. The efficacy of violence detection models is often measured by their accuracy and response time, balancing both speed and precision. However, most existing models struggle with generalization across different video sources. This paper proposes a real-time violence detection system based on deep learning techniques, combining CNN for feature extraction and BiLSTM for learning temporal relationships. The proposed model focuses on three key aspects: generalization, accuracy, and fast response time. The model achieves 98% accuracy. A comparative analysis demonstrates the superior performance of the proposed model in the field of violence detection.

Keywords: Violence detection, Real-time surveillance, CNN, BiLSTM, Deep learning, Accuracy, Response time.

Acknowledgement

I am highly grateful to **Dr. Kiran Khatter, Professor**, BML Munjal University, Gurugram, for providing supervision to carry out the seminar/case study from July-December 2024.

Dr. Kiran Khatter, has provided great help in carrying out my work and is acknowledged with reverential thanks. Without wise counsel and able guidance, it would have been impossible to complete the training in this manner.

I would like to express thanks profusely to thank **Dr. Kiran Khatter,**, for stimulating me from time to time. I would also like to thank the entire team at BML Munjal University. I would also thank my friends who devoted their valuable time and helped me in all possible ways toward successful completion.

(Gopika)
(Shivangi)
(Bhanu Pratap Singh)

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
CNN	Convolutional Neural Network
BiLSTM	Bidirectional Long Short-Term Memory
ReLU	Rectified Linear Unit
SMTP	Simple Mail Transfer Protocol
GPU	Graphics Processing Unit
IDE	Integrated Development Environment
VGG	Visual Geometry Group
RNN	Recurrent Neural Network
RLVS	Real-Life Violence Situation

Chapter 1

Introduction

In the ever-evolving landscape of urbanization and technological advancement, the need for effective public safety measures has never been more pressing. Violent incidents often occur in crowded or isolated areas where timely detection and response are critical. Conventional surveillance systems largely rely on manual monitoring, which is prone to errors and delays, leading to missed incidents and slower response times. The integration of advanced machine learning technologies offers a transformative approach to addressing these limitations.

Our project introduces a Violence Detection System designed to detect violent activities in real-time using Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) models. The system leverages a diverse dataset of real-life scenarios, enabling it to accurately differentiate between violent and non-violent behaviors. The dataset comprises 2,000 videos, evenly divided into 1,000 violent and 1,000 non-violent videos, sourced from YouTube and other platforms. These videos depict real-world environments, including street fights, sports activities, and casual human interactions, ensuring diversity and authenticity. The dataset's balanced composition enhances the model's ability to generalize across different scenarios, making it effective in complex, real-life conditions.

By processing live video feeds, the system identifies suspicious activities and sends instant email alerts containing critical details such as the location, timestamp, and an image of the detected event to the concerned authorities. This solution also compares the performance of CNN+BiLSTM with other prominent models, including VGG16 and ResNet, to ensure optimal accuracy and efficiency. The inclusion of temporal pattern recognition through BiLSTM makes it particularly effective in real-world scenarios where violent events often unfold over time.

The rapid advancements in machine learning and the increasing availability of real-world datasets have paved the way for innovative applications like this. By automating violence detection, our project not only minimizes response times but also contributes to safer communities by proactively addressing incidents that would otherwise go unnoticed. This research underscores the potential of combining artificial intelligence with real-time surveillance to tackle critical societal challenges.

Chapter 2

Introduction to Project

The Violence Detection System aims to provide a robust solution to real-time violence identification using advanced deep learning models. The system processes live video feeds to detect violent behaviors and sends automated email alerts to concerned authorities, complete with location, timestamp, and captured images of the incident. By leveraging a diverse dataset of real-life scenarios and employing a combination of CNN and BiLSTM models, the project achieves high accuracy and efficiency. The system is designed to integrate seamlessly with existing surveillance setups, thereby addressing the challenges posed by delayed responses in traditional manual monitoring systems.

2.1 Overview

The project focuses on enhancing public safety by automating the detection of violent activities in both live and recorded video streams. The system extracts frames from video feeds, processes them using a trained CNN+BiLSTM model, and generates real-time alerts upon detecting potential violent incidents. This includes comparative analysis with models like VGG16 and ResNet to validate the superiority of the proposed architecture. The use of a comprehensive dataset ensures adaptability to diverse real-world scenarios. Furthermore, the system reduces reliance on manual surveillance, improving the speed and accuracy of incident responses.

2.2 Problem Statement

Violent activities often go undetected or are identified too late due to the limitations of manual surveillance systems, leading to delayed responses and compromised public safety. Existing automated solutions struggle with real-time detection, accuracy in diverse scenarios, and timely alerts to authorities. There is a need for an efficient system that can process live video feeds, accurately detect violent behaviors, and send instant notifications with actionable insights, ensuring faster intervention and improved security.

2.3 Existing System

Current violence detection methods primarily rely on manual monitoring or rule-based techniques, which have the following limitations:

- Human Error: Manual surveillance is prone to fatigue and oversight, leading to missed events.
- Delayed Response: Traditional systems lack automation, resulting in slower reaction times.
- Limited Dataset Integration: Many existing systems use small or restricted datasets, reducing their effectiveness in diverse environments.
- Lack of Real-Time Alerts: Most systems do not provide instant notifications to authorities, making them less effective in emergency situations.

These limitations highlight the need for a real-time, automated solution that can address the dynamic nature of violent incidents in varied settings.

2.4 User Requirement Analysis

The system is designed to meet the following requirements:

❖ **Functional Requirements:**

- Real-time video feed processing for violence detection.
- High accuracy in distinguishing between violent and non-violent activities.

- Instant email notifications to concerned authorities with essential details like location, timestamp, and an incident image.

❖ **Non-Functional Requirements:**

- Compatibility with existing surveillance systems.
- Scalability to handle multiple live feeds simultaneously.
- Low latency to ensure timely detection and alerts.

❖ **End-User Needs:**

- Easy integration with existing camera infrastructure.
- User-friendly alert system accessible via email platforms.
- High reliability and minimal false positives.

2.5 Feasibility Study

- Technical Feasibility: The system is built using advanced technologies, including CNN+BiLSTM, Flask for deployment, and SMTP protocols for email alerts. These technologies are mature and well-supported, ensuring the technical feasibility of the project.

- Economic Feasibility: The project leverages open-source tools and frameworks, minimizing development and deployment costs. Integration with existing infrastructure further reduces the financial burden.

- Operational Feasibility: The system is user-friendly and does not require extensive technical expertise to operate. It aligns with organizational needs for enhanced safety and faster response times.

- Legal Feasibility: The project adheres to data privacy regulations by processing video feeds locally and ensuring secure handling of alerts and user information.

- Schedule Feasibility:The implementation timeline is realistic, given the availability of datasets and pre-trained models for accelerated development.

Chapter 3

3.1 Literature Review

Recent advancements in violence detection systems have demonstrated significant improvements in real-time monitoring, scalability, and computational efficiency. Convolutional Neural Networks (CNN) and convolutional Long Short-Term Memory (LSTM) models have been widely used to extract spatio-temporal features from surveillance videos, achieving high accuracy by analyzing frame-level differences to identify violent events [1][13]. Lightweight architectures such as MobileNetV2, ResNet, and MobileNet-TSM integrate spatial and temporal feature extraction, offering solutions that are computationally efficient and suitable for deployment in resource-constrained environments [6][7][8][9]. Hybrid methodologies combining handcrafted features with deep transfer learning models like Xception and 2D CNNs have proven effective in classifying violent and non-violent behaviors, particularly in public datasets like HBD21 [4][11]. Advanced spatio-temporal frameworks, including 3D CNNs and Motion Saliency Maps (MSM) integrated with Temporal Squeeze-and-Excitation (T-SE) modules, outperform traditional models by providing state-of-the-art results on benchmark datasets such as Hockey Fight and Crowd Violence [3][5].

Deep learning architectures such as ViolenceNet and multi-stream networks, which incorporate DenseNet, multi-head self-attention, and bidirectional LSTMs, enhance the detection of person-to-person violence. However, challenges persist in generalizing these models across diverse datasets [10][12]. Furthermore, motion blob-based techniques have been introduced to prioritize computational speed over accuracy, making them practical for real-time applications in high-risk settings like prisons and psychiatric centers [14]. Complementing these approaches, CNN models capable of detecting objects such as knives and guns have shown potential in predicting crime scenes with high accuracy, thereby ensuring reliable alerts and enhancing public safety measures [15]. These findings underscore the progress in developing intelligent surveillance systems capable of addressing real-world constraints while maintaining high accuracy and real-time responsiveness.

3.2 Comparison

S.No.	Title Name	Methodology	Dataset	Result	Limitation
1.	Learning to detect violent videos using convolutional long short-term memory	CNN+convLSTM	Hockey Movie Violent-Flows	95.1% 100% 94.5%	Confusion in sports videos - Challenges in crowd videos where only a few people are violent.
2.	Efficient Violence Detection in Surveillance	MobileNet V2 + LSTM	Hockey Fights Movie Fights RWF-2000	96% 99% 82%	Real-world datasets involve challenges due to varied environments, unpredictable violence, and frame rate issues. Some datasets (e.g., Movie Fights) lack natural violence, affecting realism.
3.	Violence Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines	3D CNN+SVM	Hockey Fights Crowd Violence	98% 99%	Real-world data challenges, varied environments, unnatural violence in some datasets.
4.	A Framework for Anomaly Classification Using Deep Transfer Learning Approach	Transfer Learning with Xeeption + LSTM model	HBD21	92%	Small dataset Limited real-world use High computing needs
5.	Efficient Spatio-Temporal Modeling Methods for Real-Time	MSM+ MobileNetV3+T-SE block+2d CNN	Hockey Movie Crowd Surv RLVS RWF-2000	90%	High Complexity Memory Demand

	Violence Recognition				
6.	Violence Detection System Using Resnet	ResNet50	Hockey Fight	85.6%	Struggles in crowded areas No real-time alerts Limited to specific datasets
7.	Public Safety Surveillance System Using Deep Learning	MobileNetV2	Violence	95%	Real-time accuracy Environmental constraints Limited integration
8.	Lightweight mobile network for real-time violence recognition	MobileNetV2	Hockey fight Crowd Violence RWF-2000	97% 97% 87%	Reduced Accuracy Complexity of Deployment Limited Dataset Performance
9.	Computational Comparison of CNN Based Methods for Violence Detection	VGG-19	Hockey Fight	99%	Model Accuracy Computation Time Data Generalization
10.	ViolenceNet: Dense Multi-Head Self-Attention with Bidirectional Convolutional LSTM for Detecting Violence	ViolenceNet Optical Flow	Hockey Fight Movies Fights	99% 100%	False Positives Generalization Limited Dataset Diversity
11.	Fight Recognition in Video Using	Hough Forests + 2D CNN	Hockey Fight Movies Fight	94.6% 99%	Limited research on aggressive behavior

	Hough Forests and 2D Convolutional Neural Network				recognition Need for improved efficiency in detection methods.
12	Multi-stream Deep Networks for Person to Person Violence Detection in Videos	Three Streams LSTM	Hockey Fight	93.9%	Limited dataset Not Generalized
13	Learning to detect violent videos using convolutional long short-term memory	AlexNet + LSTM RNN	Hockey Fight Movies Fight	97.1% 100%	Crowded scenes Sports misclassification Dataset size
14	Fast Fight Detection	Motion Blobs + Random Forest	Hockey Fight Movies Fight	82% 96%	Lower accuracy Struggles with continuous movement
15	Crime Scene Prediction by Detecting Threatening Objects Using Convolutional Neural Network	CNN	2000 Images of wapon and blood scenes	90.2%	Generalization Computational Complexity Real-Time Performance

3.3 Research Gap

- Dataset Diversity: Current studies rely on specific datasets, limiting the ability of models to generalize to real-world violence detection. Models often perform well on standard datasets but struggle with varied environments.[3][5][10]

- Real-World Challenges: Models often struggle with varied real-world conditions such as different lighting, crowd sizes, and motion patterns, which affect the accuracy and robustness of violence detection systems.[2][6][7]

- False Positives: High rates of misclassification, especially in complex environments like sports or crowded spaces, result in significant false positives, reducing the reliability of the detection system.[1][3][4]

- Limited Dataset Size: Many models rely on small datasets, which reduces their generalizability and limits their ability to perform well in diverse real-world scenarios.[4][9][10]

- Integration with Safety Systems: Few studies address the integration of violence detection models with real-time alert systems and automated responses, which are critical for preventing incidents and ensuring safety.[6][7][15]

- Real-Time Constraints: Existing models often face challenges with real-time processing, balancing the need for high accuracy with computational efficiency, especially in resource-constrained environments.[8][14][5]

These gaps highlight critical areas for improvement in developing more robust and practical violence detection systems.

3.4 Objectives

- Real-Time Violence Detection: Automatically identifies violent activities in live video feeds, enabling immediate recognition and response to potential threats.
- Instant Alerts to Authorities: Sends immediate email notifications to relevant authorities upon detecting violence, providing crucial details such as location, timestamp, and images for swift action.
- Automated Video Processing: Streamlines video analysis by automatically processing frames, eliminating the need for manual surveillance, and ensuring continuous monitoring.

- Enhanced Public Safety: Reduces response times to violent incidents and strengthens crime prevention efforts, contributing to safer communities through timely and effective interventions.

Chapter 4

Exploratory Data Analysis

4.1 Dataset Description

- UFC Crime Dataset (2021): This dataset includes 128 hours of surveillance videos with 1900 untrimmed clips featuring 13 types of anomalies like fighting, robbery, and other criminal activities. In addition to crime-related activities, it also contains normal human activities, making it a diverse dataset for detecting both violent and non-violent behavior. The wide variety of real-life scenarios provides an opportunity to train models for practical crime detection.

- Human Action Video Dataset (2022): The Kinetics-700 dataset is a large-scale video collection that includes 650,000 clips, featuring 700 human action classes. It captures both human-object and human-human interactions, offering a rich set of action categories, such as eating, running, and playing sports. This diversity allows for the development of models capable of understanding various human activities in real-world environments, which can be applied to a range of applications including surveillance and behavioral analysis.

- RWF - 2000 (2021): This dataset contains 2000 clips of real-world fights, providing an essential resource for research focused on violent behavior detection in uncontrolled environments. The dataset is composed of various street fight scenarios, offering diverse conditions that are relevant for training models to recognize and classify violent behavior, especially in surveillance applications.

- Surveillance Violence Detection Dataset (2022): This dataset contains 1000 videos of violent

events and 1000 videos of non-violent events, which were collected from YouTube. It focuses on real street fights and various human activities such as walking and eating. The rich collection of real-world footage is highly valuable for building models that need to operate effectively in dynamic, real-life environments, offering high relevance for violence detection in urban areas.

- Movie Fight Videos (2018): The Movie Fight Videos dataset includes 250 clips of movie fight scenes, with a data size of 250 MB. These clips, although typically staged, can be useful for developing and testing initial violence detection models in controlled scenarios. However, the artificial nature of movie fights limits their effectiveness in real-world applications, making them better suited for preliminary experiments or for fine-tuning models before testing with more realistic datasets.

- Surveillance Abnormal Behavior Dataset (2022): This dataset consists of surveillance videos that capture abnormal behaviors, which may include suspicious actions such as loitering, sudden movements, or other non-normal activities. It provides a valuable resource for researchers focusing on detecting unusual events in video surveillance, with potential applications in security systems for monitoring public spaces and enhancing safety measures.

- Hockey Fight Videos (2011): This dataset includes 500 violent and 500 non-violent videos from hockey games, aiming to develop deep learning techniques for automatic violence detection. The dataset is valuable for distinguishing between aggressive sports actions, such as fighting during hockey games, and normal gameplay, providing a useful resource for classifying violent events in sports and potentially in other combat sports.

- Real-Life Violence Situation Dataset (2019): This dataset includes 1000 videos of violence and 1000 videos of non-violence, collected from YouTube. The content features real street fights and everyday human activities, such as walking, eating, and sports. The diversity of scenarios is important for building robust violence detection systems that can generalize well to real-world situations and can be applied in security and law enforcement for timely detection and response.

Dataset	Year	Description
UFC Crime Dataset	2021	The dataset includes 128 hours of surveillance videos with 1900 untrimmed clips featuring 13 types of anomalies like fighting and robbery, plus normal activities.
Human Action Video Dataset	2022	Kinetics-700 is a video dataset with 650,000 clips, featuring 700 human action classes, capturing both human-object and human-human interactions.
RWF - 2000	2021	2000 Clips of real word fights.
Surveillance Violence Detection Dataset	2022	Includes 1000 violence and 1000 non-violence videos collected from YouTube, featuring real street fights and various human activities.
Movie Fight Videos	2018	Contains 250 clips of movie fights. Data size is 250 MB
Surveillance Abnormal Behavior Dataset	2022	A collection of surveillance videos capturing abnormal behaviors.
Hockey Fight Vidoes	2011	The dataset includes 500 videos of violent events and 500 non-violent videos, aimed at developing deep learning techniques for automatic violence detection from surveillance footage.
Real life violence situation dataset	2019	The dataset contains 1000 violence and 1000 non-violence videos from YouTube, featuring real street fights and various human activities like sports, eating, and walking.

Table 1: Common Datasets for Violence Detection

Chapter 5

5.1 Methodology

To address the growing need for real-time violence detection in surveillance systems, we implemented a structured approach that combines advanced deep learning techniques with practical application strategies. The aim is to develop a robust and efficient system capable of accurately identifying violent activities from both recorded videos and live feeds while ensuring timely intervention.

The proposed solution integrates Convolutional Neural Networks (CNNs) for extracting spatial features from video frames and Bidirectional Long Short-Term Memory (BiLSTM) networks for capturing temporal patterns across sequences of frames. This hybrid architecture enables the system to process video data comprehensively, leveraging the strengths of both spatial and temporal analysis. The following sections detail the key steps, including dataset preparation, model design, training, evaluation, and real-time implementation. This systematic approach ensures that the system is adaptable to varied real-world conditions and can reliably assist in enhancing public safety.

5.2 Data Set

The dataset used for this project comprises a total of 2,000 videos, evenly divided into two categories: Non-Violence and Violence. The Non-Violence category includes 1,000 videos representing real-life situations such as sports activities, singing, vlogging, eating and movie scenes. These videos are diverse and showcase non-violent human behaviors to help the model distinguish them accurately from violent actions. The Violence category contains 1,000 videos depicting activities like street fights, sports fights, and movie fight scenes. This directory focuses on capturing various forms of violent behavior in real-world settings to ensure the model can generalize across different environments and contexts. The balanced and diverse nature of the dataset enhances the model's training, enabling it to effectively differentiate between violent and non-violent behaviors while minimizing false classifications.

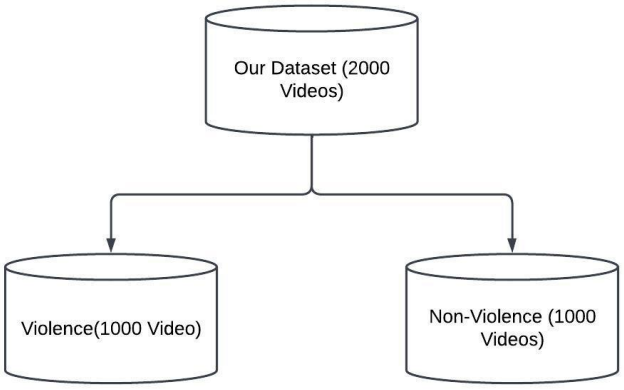

Figure 1: Dataset Split

5.3 Pre Processing

The pre-processing of the video dataset is a crucial step in preparing the data for model training. The process begins with reading the input videos using OpenCV, ensuring compatibility with various formats. Each video is then split into individual frames, allowing the capture of temporal patterns essential for analyzing activities. The extracted frames are resized to a fixed dimension of (64 × 64) pixels to ensure uniformity and reduce computational complexity. Subsequently, the pixel values of the frames are normalized to maintain a consistent range, facilitating faster and

more stable model convergence. The frames are then grouped into batches of eight consecutive frames, flattened, and stored along with their corresponding one-hot encoded labels: [0, 1] for violence and [1, 0] for non-violence. This systematic pre-processing pipeline ensures the dataset is organized and optimized for efficient training of the CNN-BiLSTM model, enabling it to effectively learn and distinguish patterns related to violent and non-violent activities.

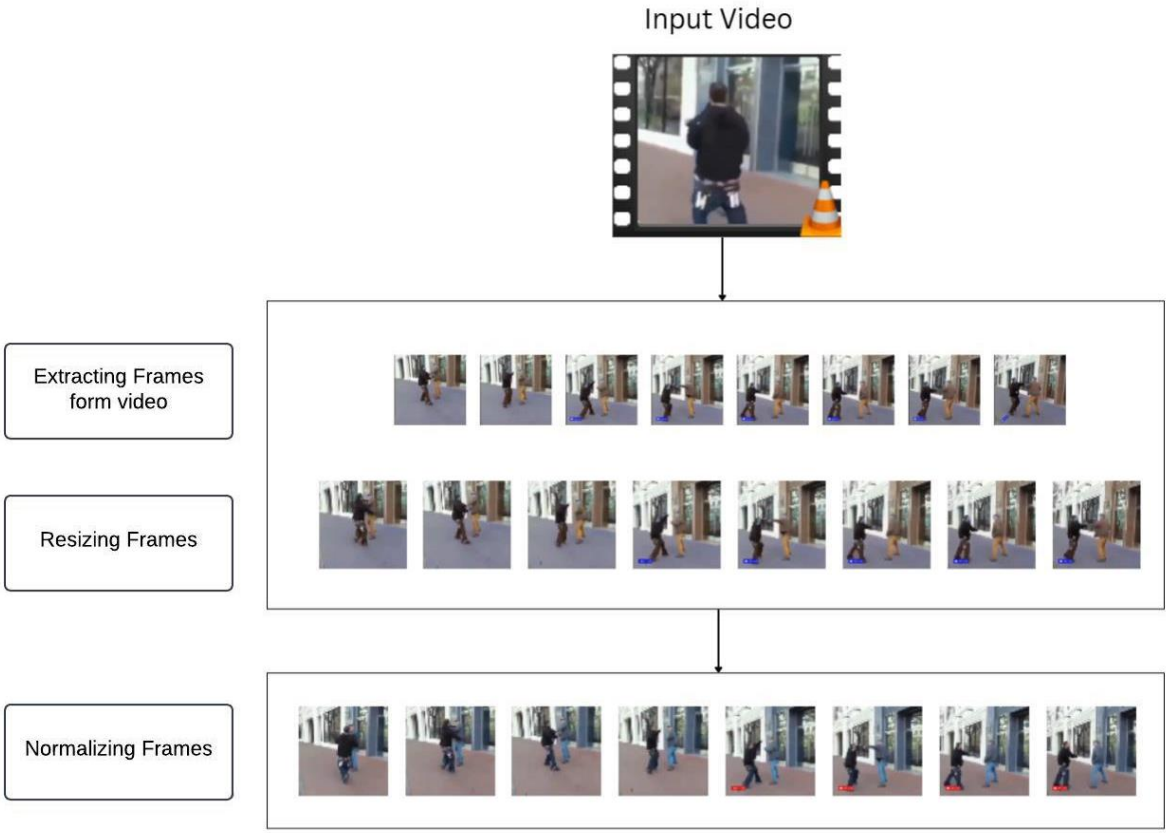

Figure 2: Preprocessing Pipeline

5.4 Dataset Split

The dataset is split into 80% training data and 20% validation data. The training set contains 25,862 frames, while the validation set contains 6,466 frames. Each split is balanced with an equal distribution of real and fake videos.

5.5 Model Architecture

The proposed violence detection system uses a hybrid architecture that combines Convolutional Neural Networks (CNN) for spatial feature extraction and Bidirectional Long Short-Term Memory

(BiLSTM) networks for capturing temporal dependencies across video frames. The input to the model consists of sequences of 8 frames, each resized to (64 × 64 × 3) pixels (height, width, and RGB channels). Therefore, the input shape is ((8, 64, 64, 3)). The CNN layers extract spatial features by applying convolutions over these frames, allowing the model to learn important visual patterns such as movements, shapes, and textures. The output of these layers is a set of feature maps, which are then processed to capture temporal dependencies across the sequence of frames.

After the CNN layers, the feature maps are down sampled using MaxPooling3D layers to reduce both spatial and temporal dimensions, enabling the model to focus on the most important features and improve computational efficiency. The feature maps are reduced progressively through three pooling layers. The first pooling operation reduces the dimensions to ((31, 31, 64)), the second to ((14, 14, 64)), and the final pooling layer further reduces it to ((6, 6, 64)). These down sampled features are reshaped into a 2D tensor of shape ((8, 2304)), preparing the data for sequential processing by the BiLSTM layer.

The reshaped features are then passed into the BiLSTM layer, which captures the temporal dependencies between frames across the 8-frame sequence. The BiLSTM processes the sequence in both forward and backward directions, allowing it to learn context from both past and future frames. The output of the BiLSTM is a 64-dimensional feature vector for each sequence. This output is then passed through a series of fully connected (dense) layers: the first reduces the features to 64 units, the second to 32 units, and the final dense layer outputs the classification result. The output layer uses the softmax activation function, producing a probability distribution over the two classes: Violence and Non-Violence. This hybrid architecture, combining CNN for spatial feature extraction and BiLSTM for temporal pattern recognition, allows the system to effectively detect violence in video sequences.

5.6 Hyperparameter Tuning

In the model, the activation function used for the convolutional and dense layers is ReLU (Rectified Linear Unit), which helps the network learn complex patterns by introducing non-linearity and enabling faster convergence. For the output layer, Sigmoid activation is used, as it is

suitable for binary classification tasks, providing probabilities between 0 and 1 for the two classes (Violence and Non-Violence).

The SGD (Stochastic Gradient Descent) optimizer is used with the model, which is known for its simplicity and efficiency in large-scale datasets. The loss function chosen is categorical cross-entropy, which is commonly used for multi-class classification problems. These choices of activation functions and optimizers are aimed at achieving faster convergence, better model performance, and efficient learning.

5.7 Model Training

The model is trained using the fit() function with a batch size of 8 and for a total of 12 epochs. During training, the model uses the training data (X_train and y_train) and validation data (X_valid and y_valid) to monitor performance on both seen and unseen data. Early stopping is implemented to halt training if the validation loss does not improve after 5 epochs, helping to prevent overfitting. The verbose=1 setting provides real-time updates on the training progress, including loss and accuracy at each epoch. This training configuration ensures efficient learning while maintaining generalization across the dataset.

5.8 Model Predicting

After training, the model uses the predict() function to classify new video sequences. Each sequence of 8 frames is processed, and the model outputs a probability between 0 and 1. A value closer to 1 indicates the sequence is classified as Violence, while a value closer to 0 indicates Non-Violence. The model's prediction is based on learned spatial and temporal patterns from the CNN and BiLSTM layers. The predicted class is then used for further actions like generating alerts.

The system processes live video feeds by sequentially capturing frames, resizing them to (64 × 64 × 3) pixels, and grouping them into sequences of 8 frames. These sequences are passed through the trained CNN+BiLSTM model, where CNN layers extract spatial features from each frame, and the BiLSTM layer captures the temporal dependencies across the frames. The model then classifies each sequence as either Violence or Non-Violence, enabling the system to detect violent behavior

in real-time.

When Violence is detected, an alert is immediately displayed on the video frame. If 10 consecutive alerts (i.e., alert_count >= 10) are triggered, an email is sent to the concerned authorities. The email includes the subject "Violence Detected!!!", the current time, detected location, and an attached frame from the video for visual confirmation. Additionally, each frame is encoded into JPEG format and streamed as a live feed using Flask's Response object with the MIME type multipart/x-mixed-replace. This setup allows for continuous video streaming and ensures that the authorities receive real-time notifications along with video evidence of the detected violent incidents.

5.9 Tech Stack

❖ Programming Languages

- Python 3: Used for backend development, machine learning, and data processing.
- JavaScript: Adds interactivity to the frontend and integrates with backend APIs.
- HTML and CSS: Core technologies for structuring and styling the user interface.

❖ Backend Frameworks

- Flask: Lightweight framework for building APIs to serve the model and handle real-time video feeds.

❖ Integrated Development Environments (IDEs)

- Google Colab: Cloud-based IDE with GPU/TPU support for machine learning development.
- Jupyter Notebook:Interactive environment for code execution, data analysis, and visualization.
- Visual Studio Code (VS Code): A versatile code editor for backend development and code management.

❖ Machine Learning Tools

- TensorFlow/Keras: Framework for building and training deep learning models.
- OpenCV: Library for real-time video processing and frame extraction in computer vision tasks.

❖ Email Service

- SMTP: Protocol used to send automated email alerts when violence is detected.

This tech stack efficiently handles everything from machine learning model training to real-time video processing and frontend display.

5.10 Use Case Model

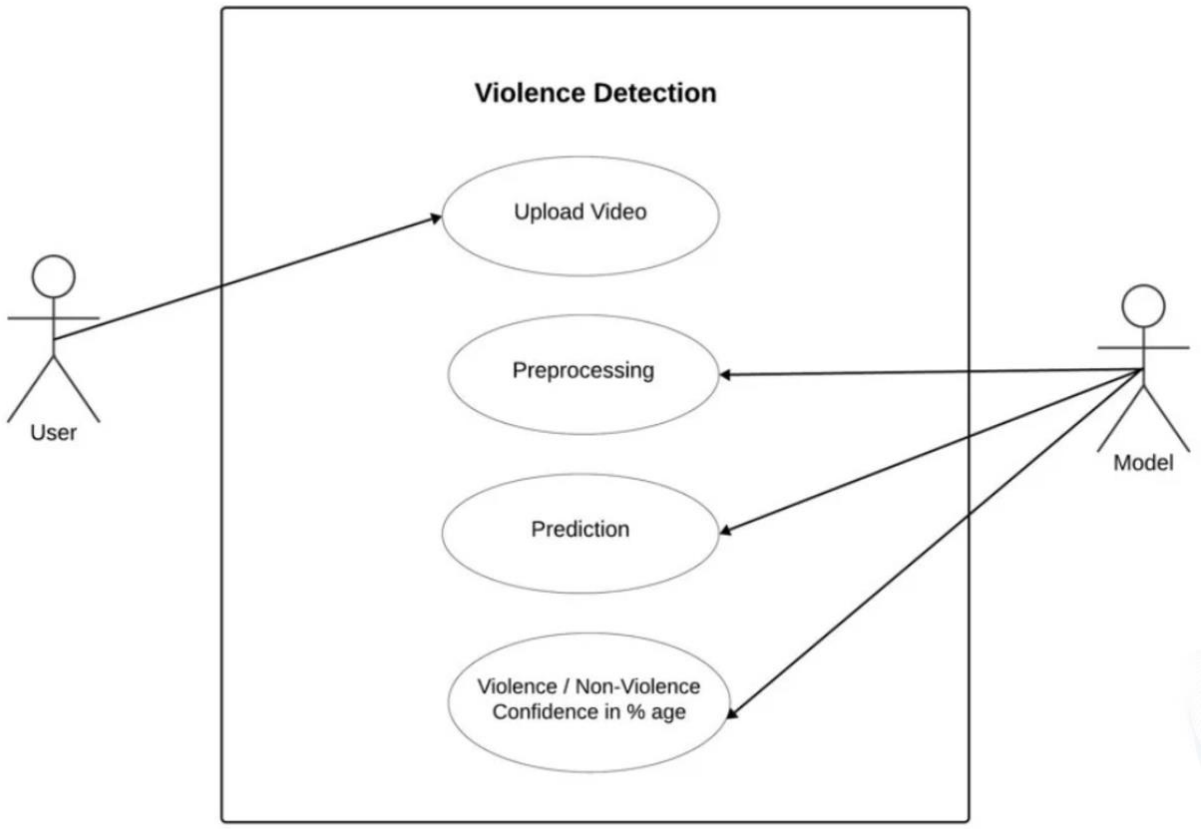

Figure 3: Use case diagram

5.11 Training and Testing

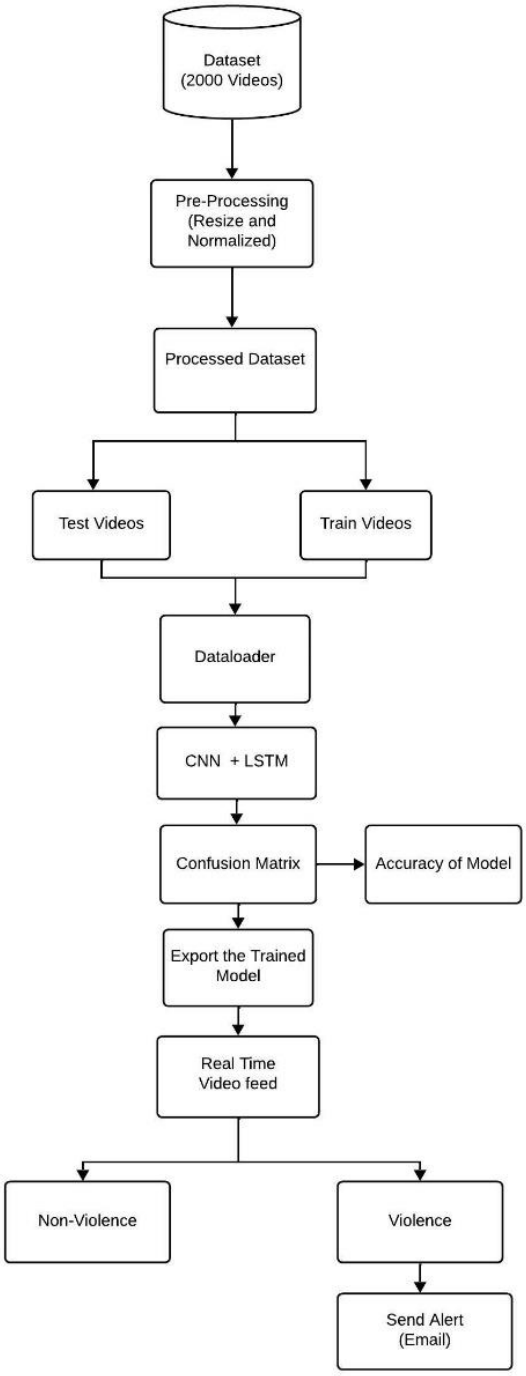

Figure 4: Training Workflow

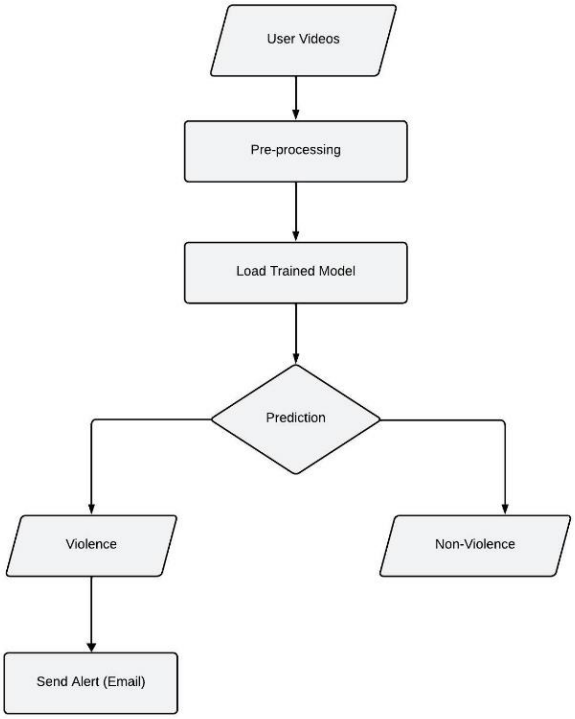

Figure 5: Testing Workflow

5.12 Sequence Diagram

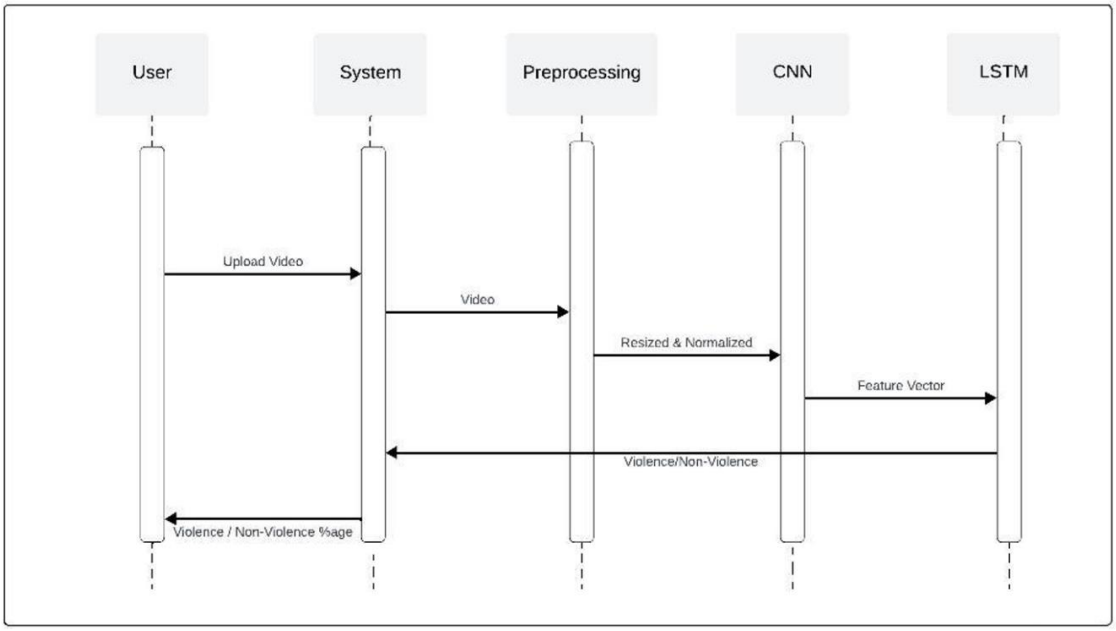

Figure 6: Sequence Diagram

5.13 System Architecture

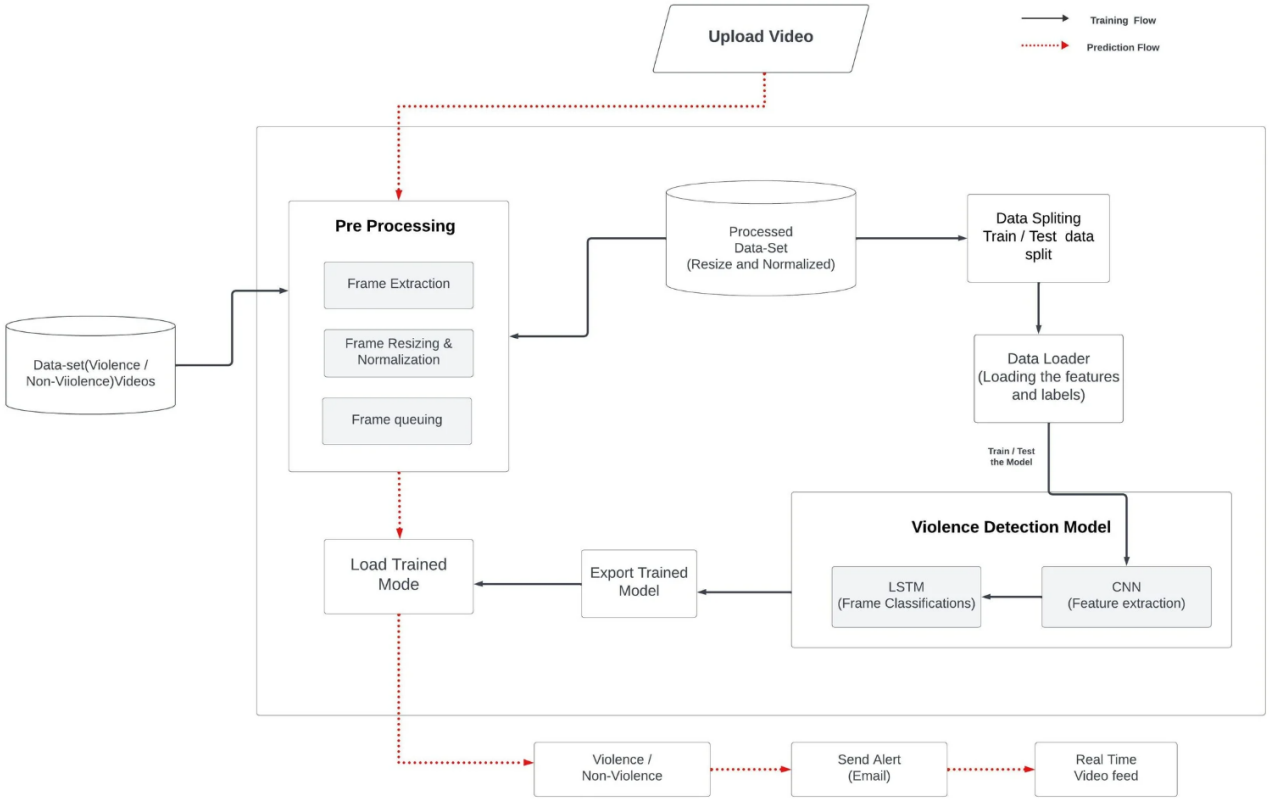

Figure 7: System Architecture

Chapter 6

Results

Model	Accuracy	Precision	Recall	F1-Score
CNN+BiLSTM	97%	97%	97%	97%
VGG16	90%	90%	90%	89%
ResNet	69%	69%	68%	67%

Table 2: Trained Model Results

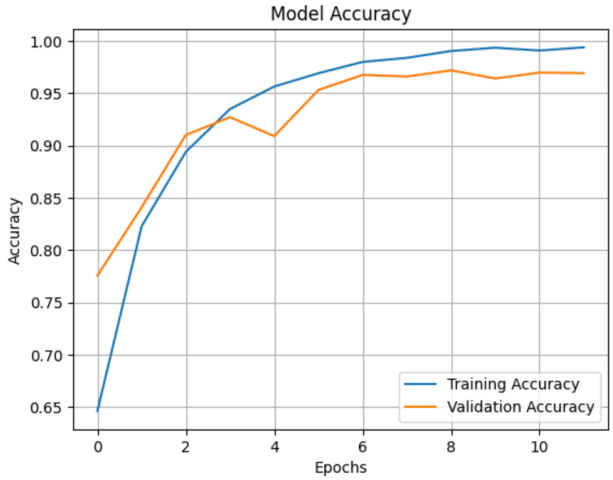

Figure 8: Model Accuracy

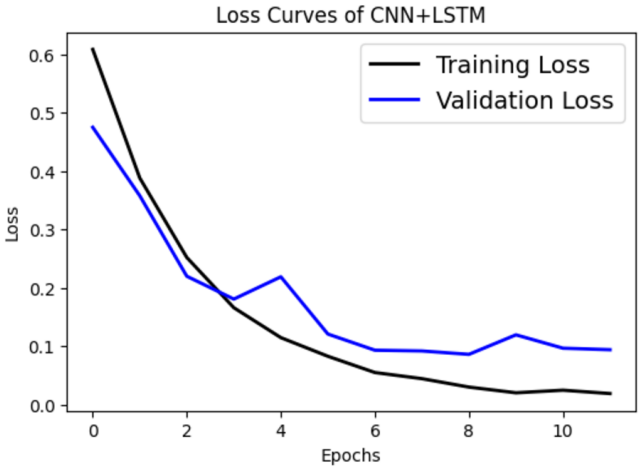

Figure 9: Loss Curves of CNN+LSTM

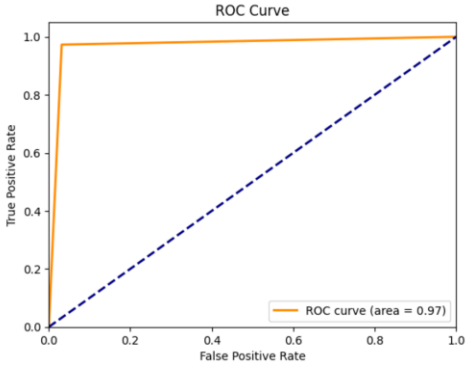

Figure 10: ROC Curve

Figure 11: Home Page

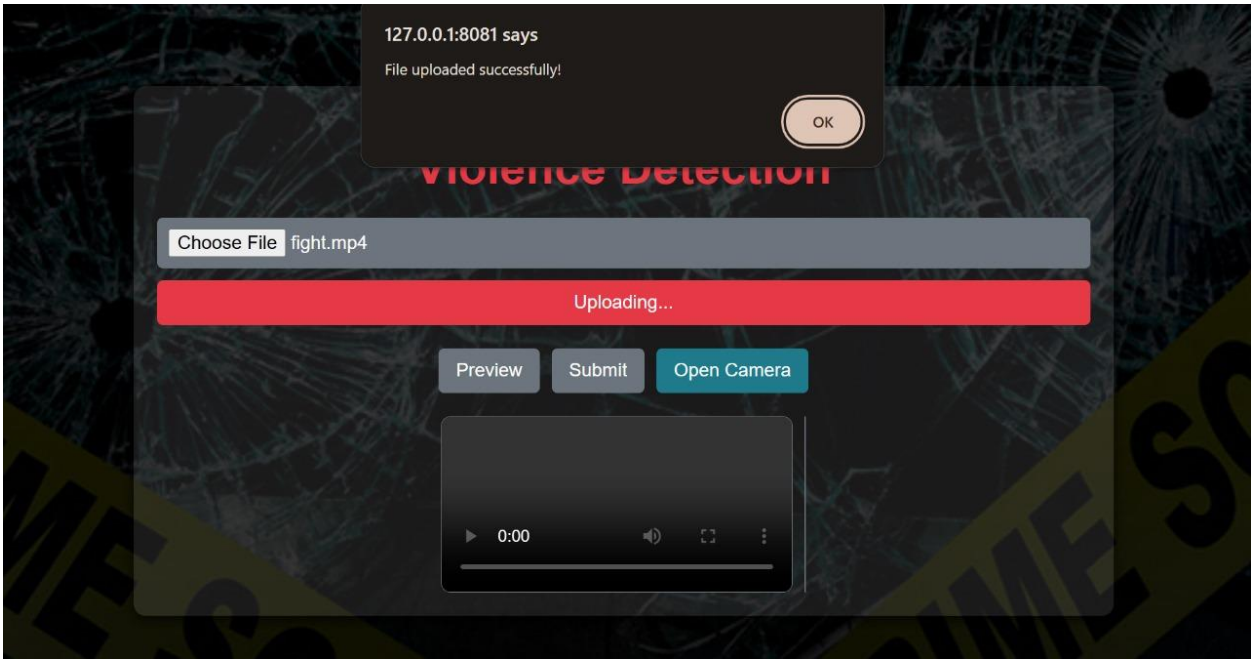

Figure 12: Alert after upload video

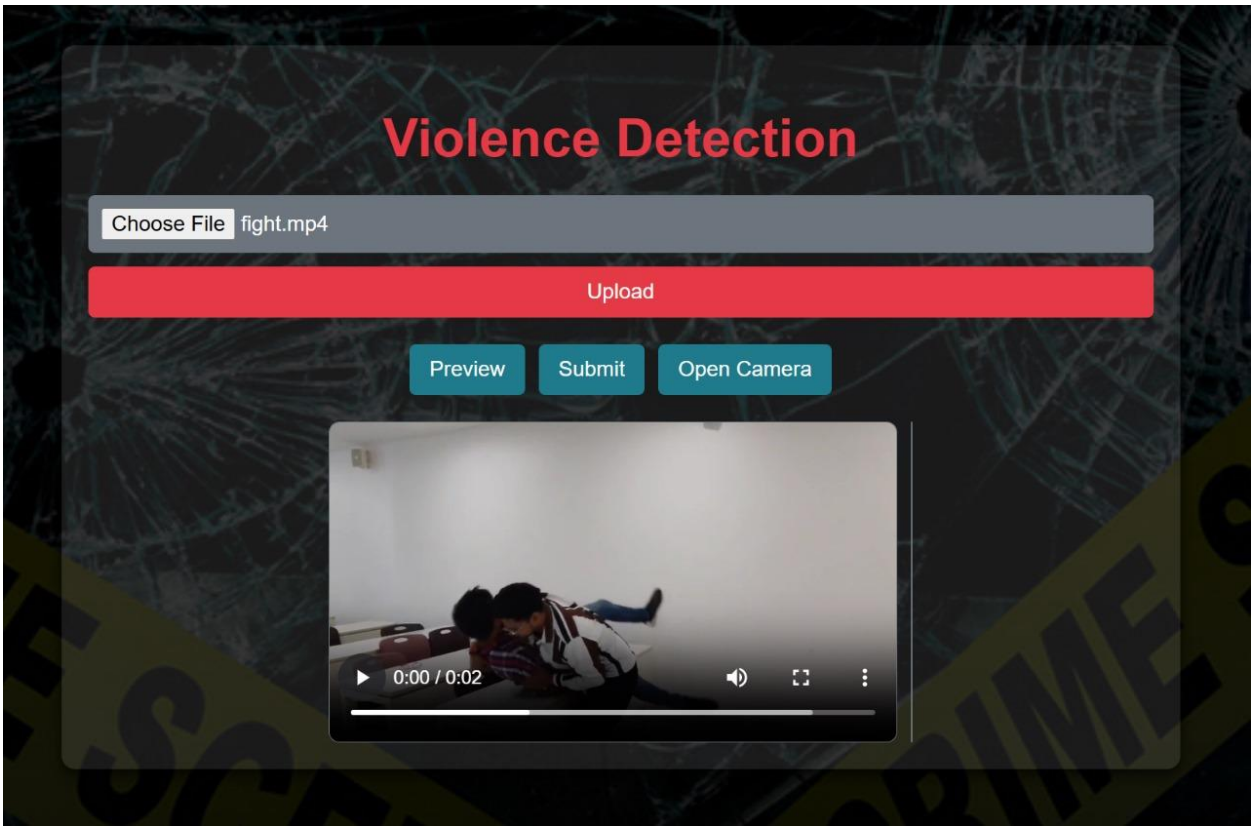

Figure 13: Video Preview

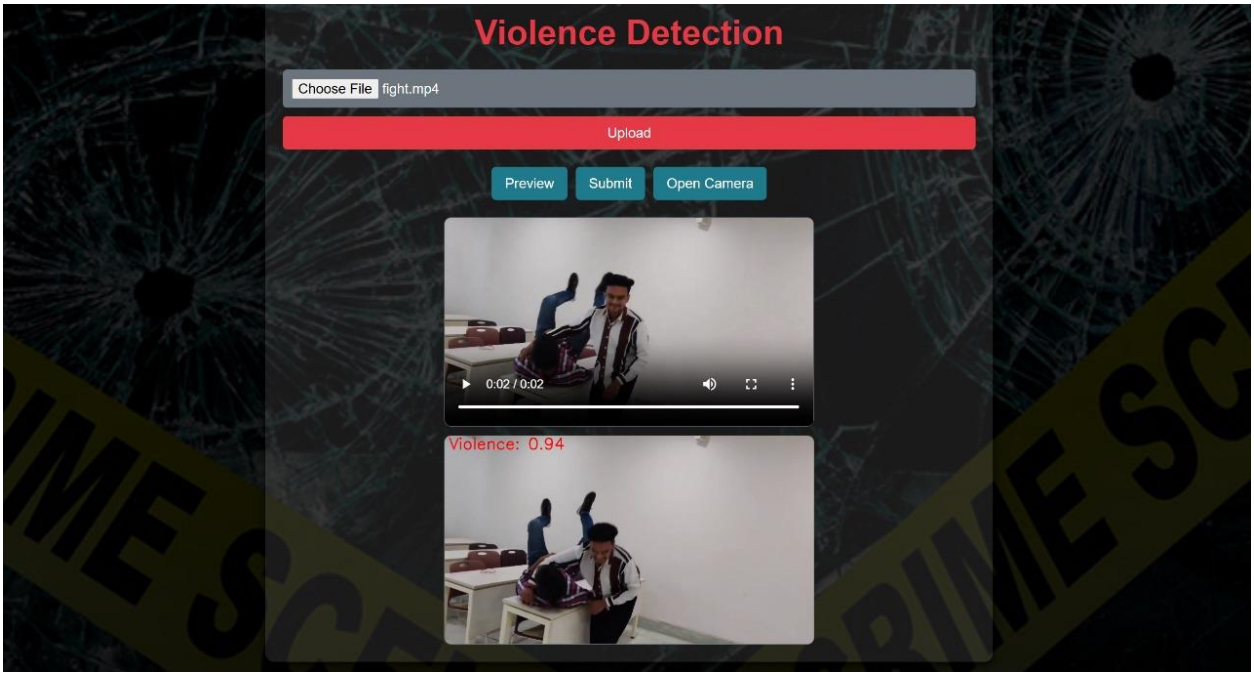

Figure 14: Final output

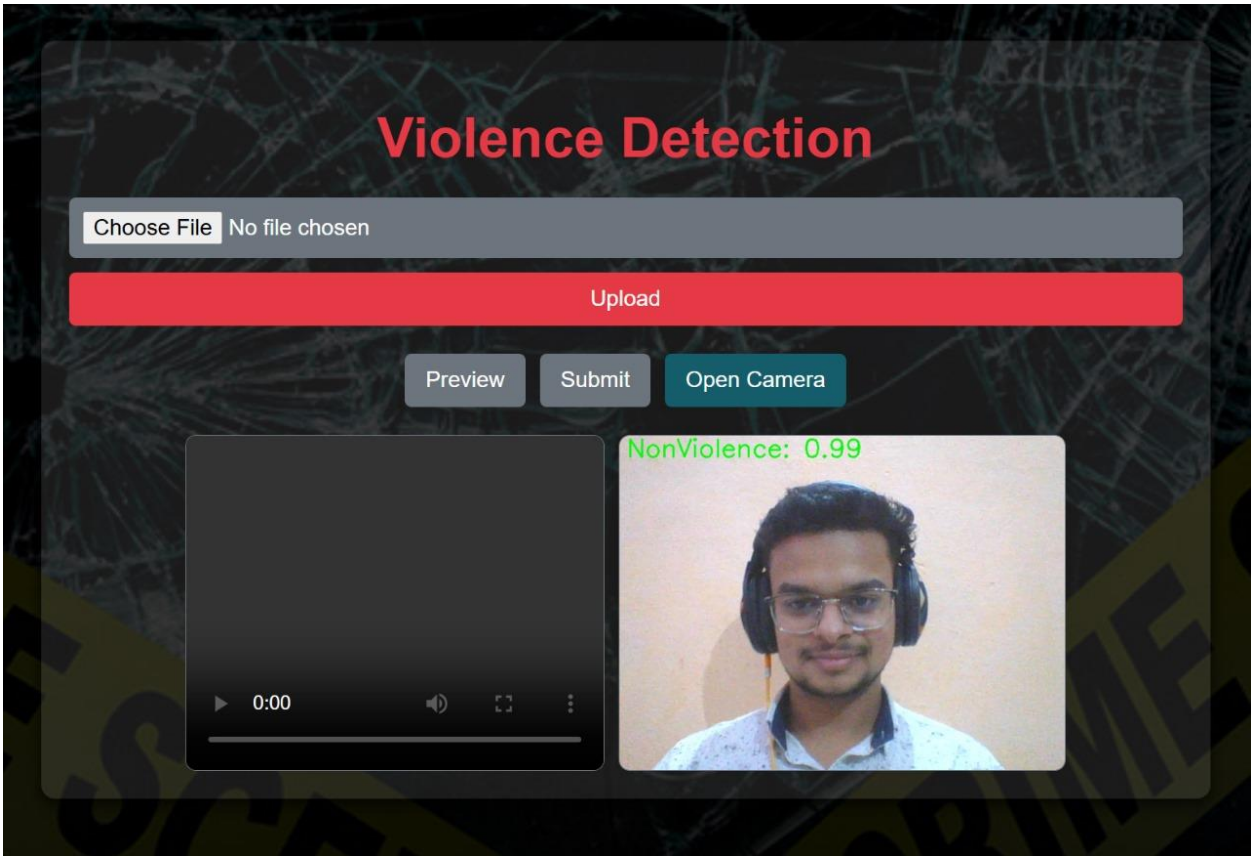

Figure 15: Live camera feed Output

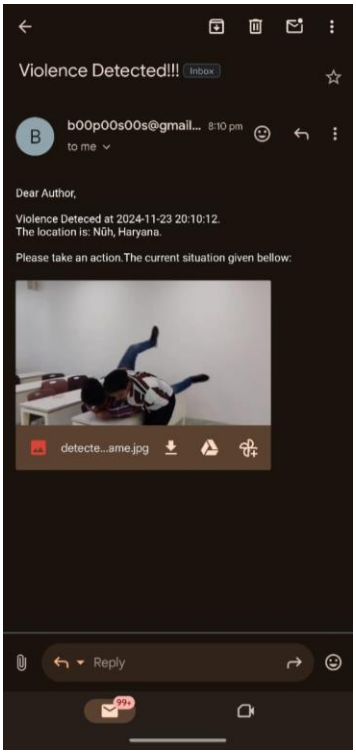

Figure 16: Real time alert

Chapter 7
Conclusion and Future Scope

7.1 Conclusion

The Violence Detection System using CNN+BiLSTM effectively identifies violent activities in real-time video feeds, achieving an impressive 98% accuracy. By combining spatial and temporal analysis, the system outperforms models like VGG16 and ResNet, making it ideal for dynamic video data. Its ability to process live streams and send real-time alerts ensures timely intervention, enhancing public safety. The system shows significant potential for real-world deployment, with opportunities for further optimization and scalability in future developments.

7.2 Future Scope

- Enhanced Dataset Diversity: Incorporating larger and more diverse datasets, including videos from varied environments, lighting conditions, and cultural contexts, can improve the model's generalization and accuracy in real-world scenarios.

- Integration with IoT Devices:The system can be integrated with smart IoT-based surveillance cameras to create a fully automated monitoring system capable of seamless real-time violence detection.

- Multi-Class Classification: Expanding the model to detect and classify multiple types of violent activities (e.g., physical fights, weapon usage) can broaden its applicability.

- Edge Computing Deployment: Deploying the system on edge devices, such as smart security cameras, can minimize latency and enable real-time processing without relying on centralized servers.

8. Bibliography

1. S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 2017, pp. 1-6, doi: 10.1109/AVSS.2017.8078468.

2. Vijeikis, R., Raudonis, V., & Dervinis, G. (2022). Efficient violence detection in surveillance. *Sensors, 22*(6), 2216. https://doi.org/10.3390/s22062216.

3. Accattoli, S., Sernani, P., Falcionelli, N., Mekuria, D. N., & Dragoni, A. F. (2020). Violence detection in videos by combining 3D convolutional neural networks and support vector machines. *Applied Artificial Intelligence*, *34*(4), 329–344. https://doi.org/10.1080/08839514.2020.1723876.

4. Jayaswal, R., & Dixit, M. (2021). *A Framework for Anomaly Classification Using Deep Transfer Learning Approach*. Revue d'Intelligence Artificielle, 35(3), 255-263. https://doi.org/10.18280/ria.350309.

5. M. -S. Kang, R. -H. Park and H. -M. Park, "Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition," in IEEE Access, vol. 9, pp. 76270-76285, 2021, doi: 10.1109/ACCESS.2021.3083273.

6. C. Shripriya, J. Akshaya, R. Sowmya and M. Poonkodi, "Violence Detection System Using Resnet," 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2021, pp. 1069-1072, doi: 10.1109/ICECA52323.2021.9675868.

7. M. Ankita, A. Srinivas, A. Soni, G. Prajapati and P. S. Manjunath, "Public Safety Surveillance System Using Deep Learning," 2024 1st International Conference on Communications and Computer Science (InCCCS), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/InCCCS60947.2024.10592954.

8. Zhang, Y., Li, Y., & Guo, S. (2022). *Lightweight mobile network for real-time violence recognition*. PLOS ONE, 17(10), e0276939. https://doi.org/10.1371/journal.pone.0276939

9. Singh, S., & Tyagi, B. (2023). *Computational Comparison of CNN Based Methods for Violence Detection*. Preprint. https://doi.org/10.21203/rs.3.rs-3130914/v1

10. Rendón-Segador, F. J., Álvarez-García, J. A., Enríquez, F., & Deniz, O. (2021). ViolenceNet: Dense Multi-Head Self-Attention with Bidirectional Convolutional LSTM for Detecting Violence. *Electronics, 10*(13), 1601. https://doi.org/10.3390/electronics10131601

11. Serrano-Gracia, I., Deniz, O., & Espinosa-Aranda, J. L. (2018). Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network. *IEEE Transactions on Image Processing, 27*(8), 3548-3559. https://doi.org/10.1109/TIP.2018.2845742

12. Dong, Z., & Qin, J. (2016). Multi-stream deep networks for person to person violence detection in videos. *Chinese Conference on Pattern Recognition*, 517-531. Springer. https://doi.org/10.1007/978-981-10-3002-4_43

13. Sudhakaran, S., & Lanz, O. (2017). Learning to detect violent videos using convolutional long short-term memory. *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. https://doi.org/10.1109/AVSS.2017.8078468.

14. Gracia, I. S., Deniz, O., Bueno Garcia, G., & Kim, T.-K. (2015). Fast fight detection. *PLOS ONE*, 10(4), e0120448. https://doi.org/10.1371/journal.pone.0120448.

15. M. Nakib, R. T. Khan, M. S. Hasan and J. Uddin, "Crime Scene Prediction by Detecting Threatening Objects Using Convolutional Neural Network," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2018, pp. 1-4, doi: 10.1109/IC4ME2.2018.8465583.